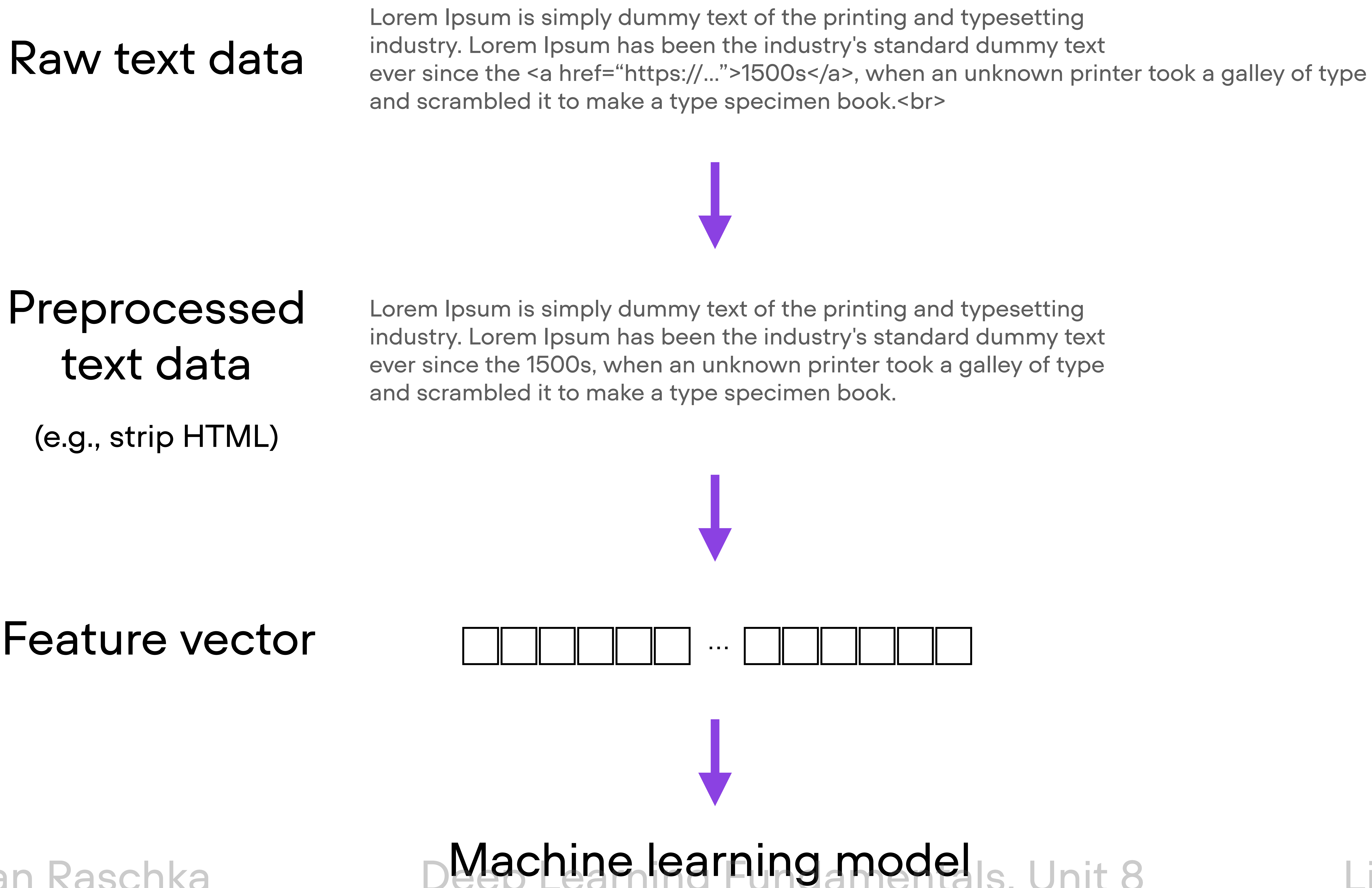
8.1

Working With Text Data

Part 1: Text Modeling Considerations

Sebastian Raschka and the Lightning AI Team



Raw text data

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the <a href=“https://…”>1500s</a>, when an unknown printer took a galley of type and scrambled it to make a type specimen book.<br>

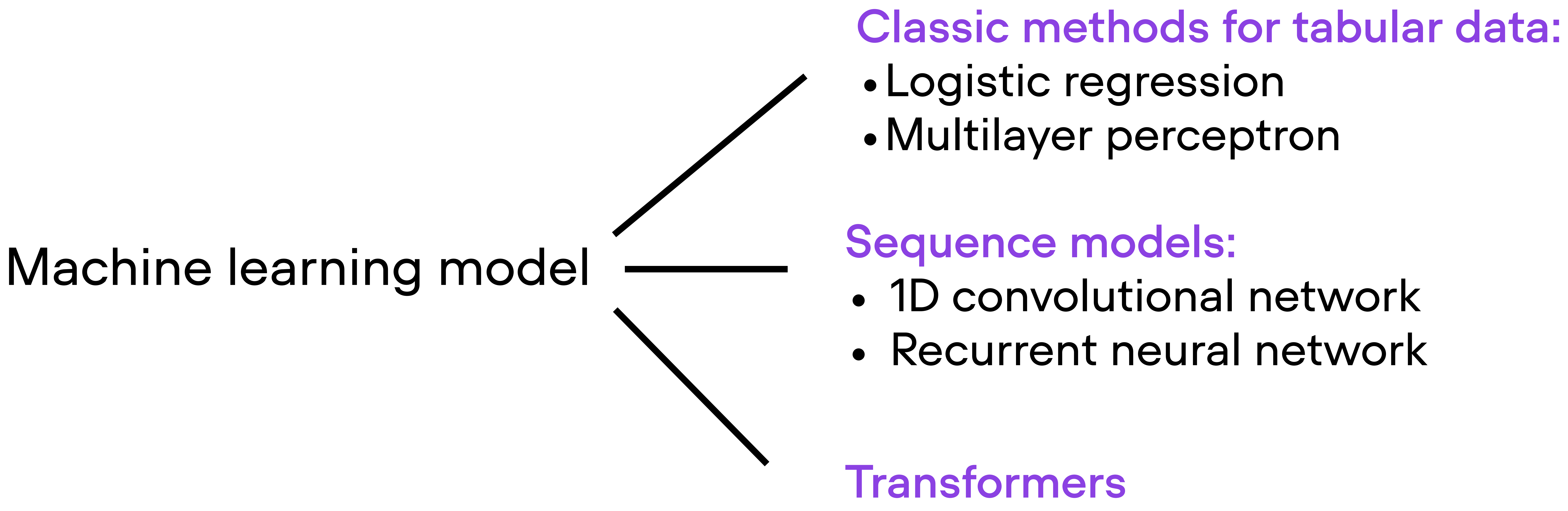
Preprocessed text data

(e.g., strip HTML)

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book.

Feature vector

Machine learning model



Machine learning model

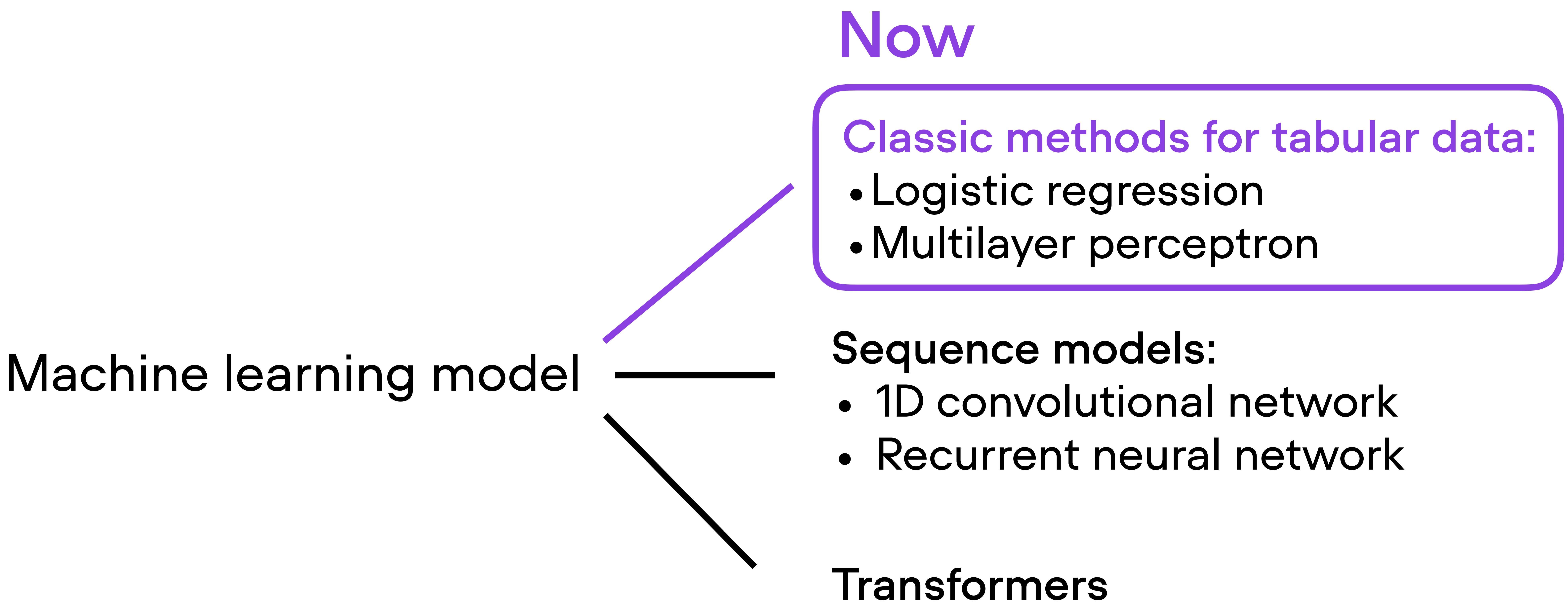
Classic methods for tabular data:

- Logistic regression
- Multilayer perceptron

Sequence models:

- 1D convolutional network
- Recurrent neural network

Transformers



Now

Machine learning model

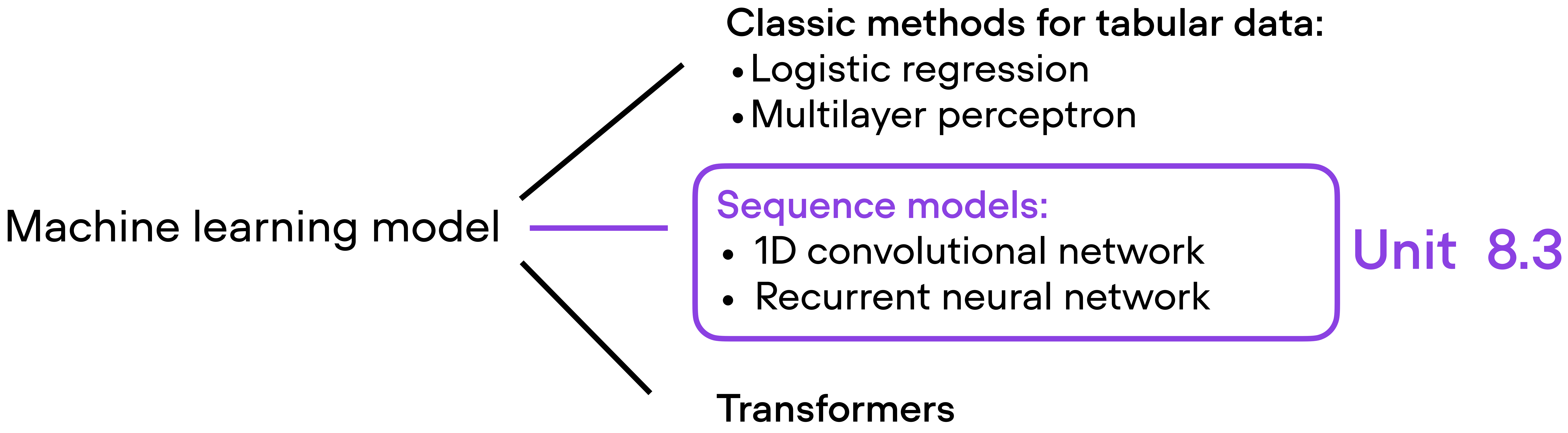
Classic methods for tabular data:

- Logistic regression
- Multilayer perceptron

Sequence models:

- 1D convolutional network
- Recurrent neural network

Transformers



Machine learning model

Classic methods for tabular data:

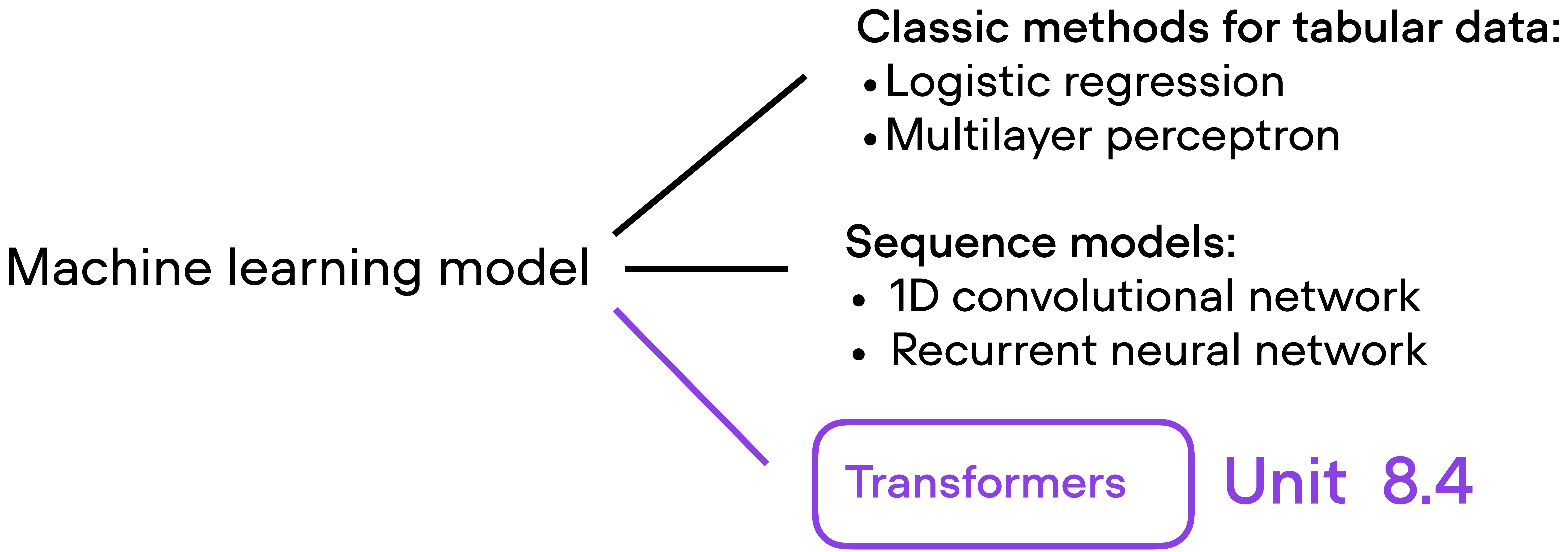
- Logistic regression
- Multilayer perceptron

Sequence models:

- 1D convolutional network
- Recurrent neural network

Unit 8.3

Transformers



Machine learning model

Classic methods for tabular data:

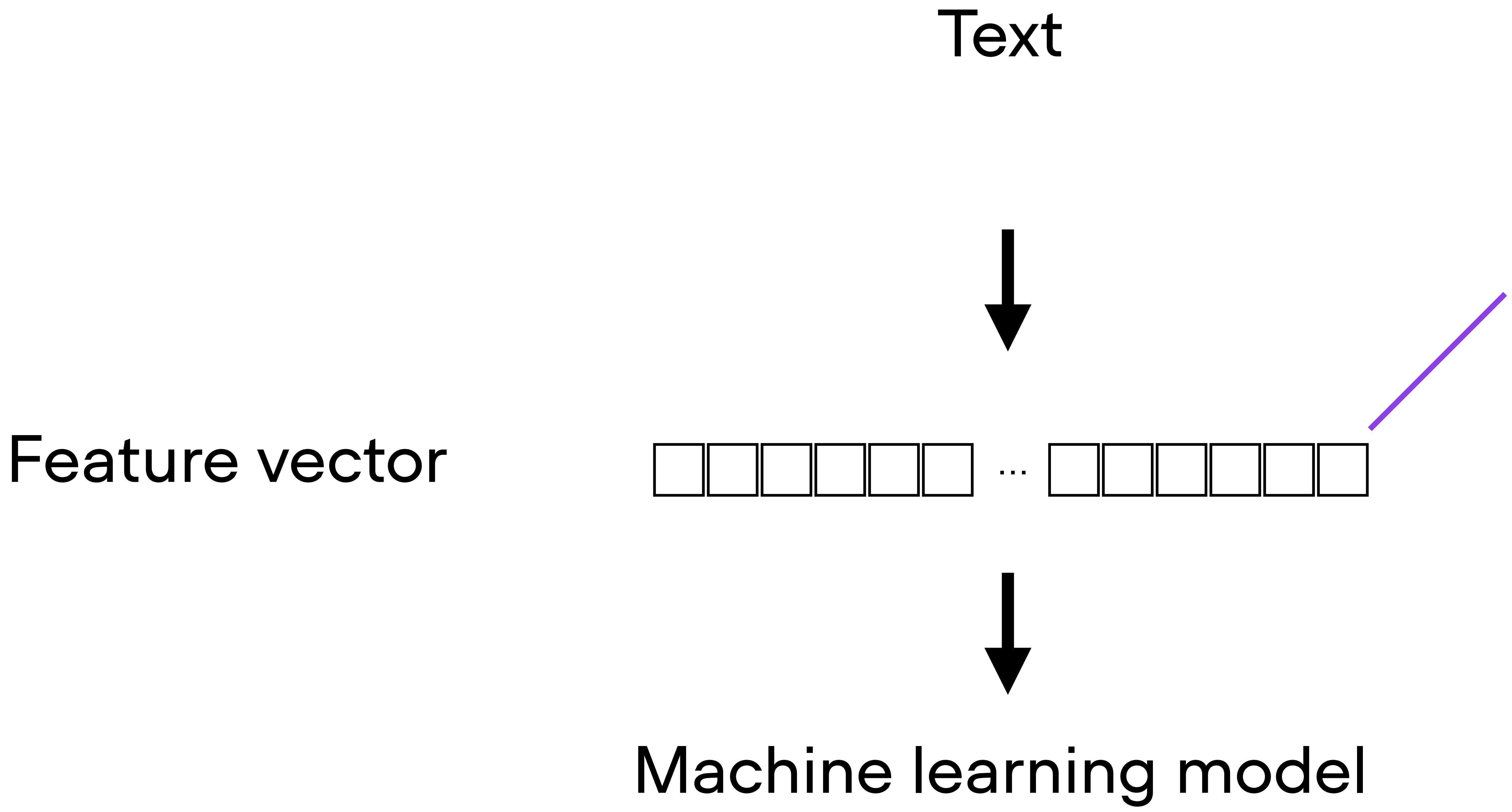
- Logistic regression
- Multilayer perceptron

Sequence models:

- 1D convolutional network
- Recurrent neural network

Transformers

Unit 8.4



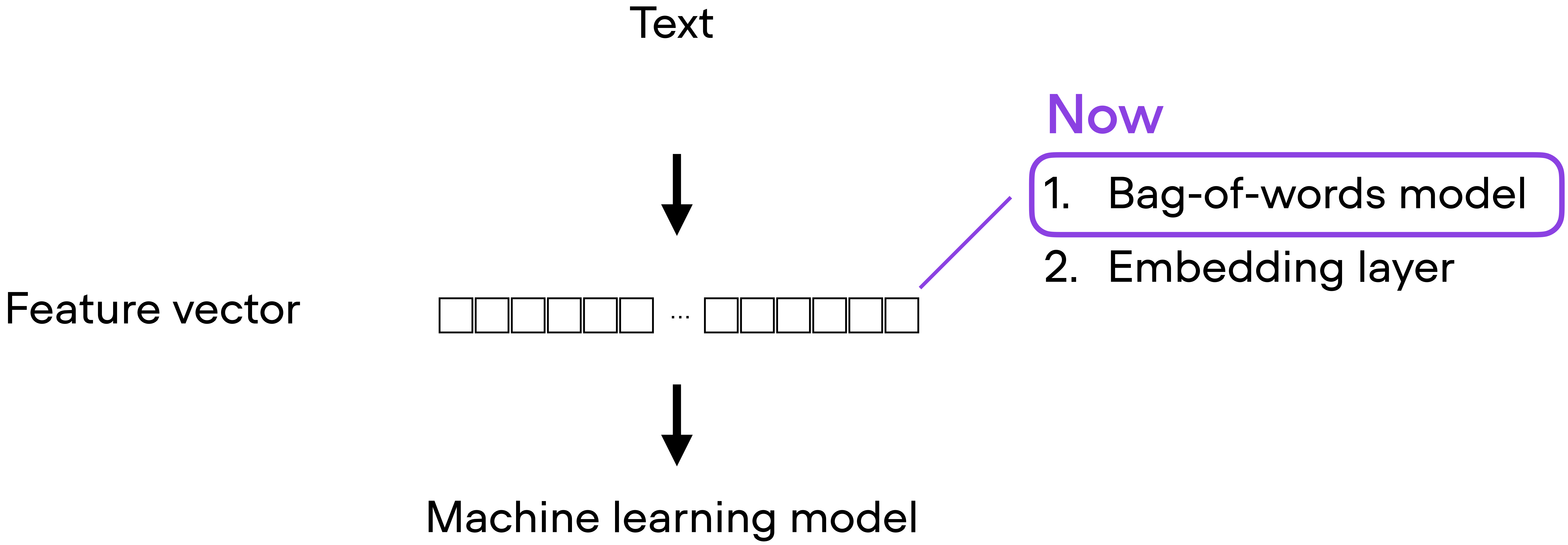
Text

Feature vector

Machine learning model

2 popular options

1. Bag-of-words model
2. Embedding layer



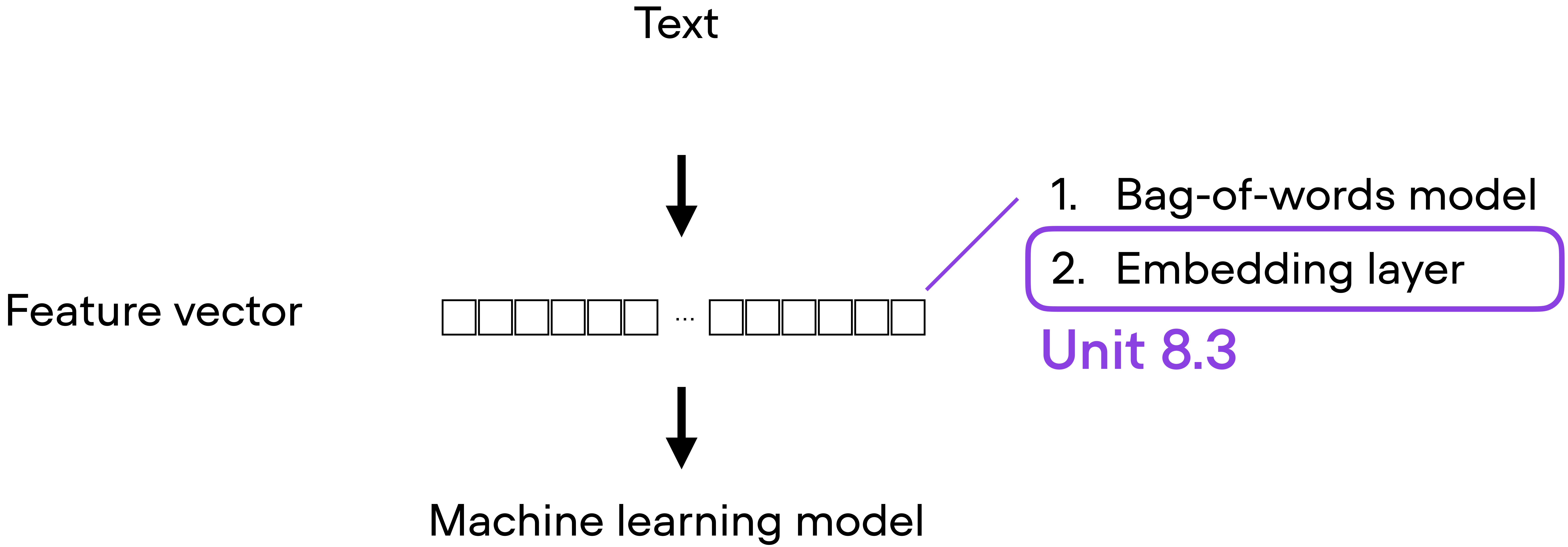
Text

Feature vector

Machine learning model

Now

1. Bag-of-words model
2. Embedding layer



Text

Feature vector

1. Bag-of-words model
2. Embedding layer

Unit 8.3

Machine learning model

Next: The bag-of-words model