

# Language data

## Representing sentences

- Word vectors can be concatenated to represent phrases/sentences
- These concatenated vectors encode the phrase by preserving the context of individual words and the temporal correlation of words
- Extends to other types of sequence data (eg. DNA/RNA, amino acid sequences)

$$\begin{bmatrix} 100 \\ 2 \\ \vdots \\ 240 \end{bmatrix} \quad \begin{bmatrix} 20 \\ 804 \\ \vdots \\ 102 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 2 \\ \vdots \\ 12 \end{bmatrix} \quad \begin{bmatrix} 490 \\ 29 \\ \vdots \\ 300 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

My name is David.

$$\begin{bmatrix} 100 & 20 & 1 & 490 & 0 \\ 2 & 804 & 2 & 29 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 240 & 102 & 12 & 300 & 0 \end{bmatrix}$$

# **Case Study: The Protein Folding Problem and the AlphaFold System**