Sequences and Time-Series

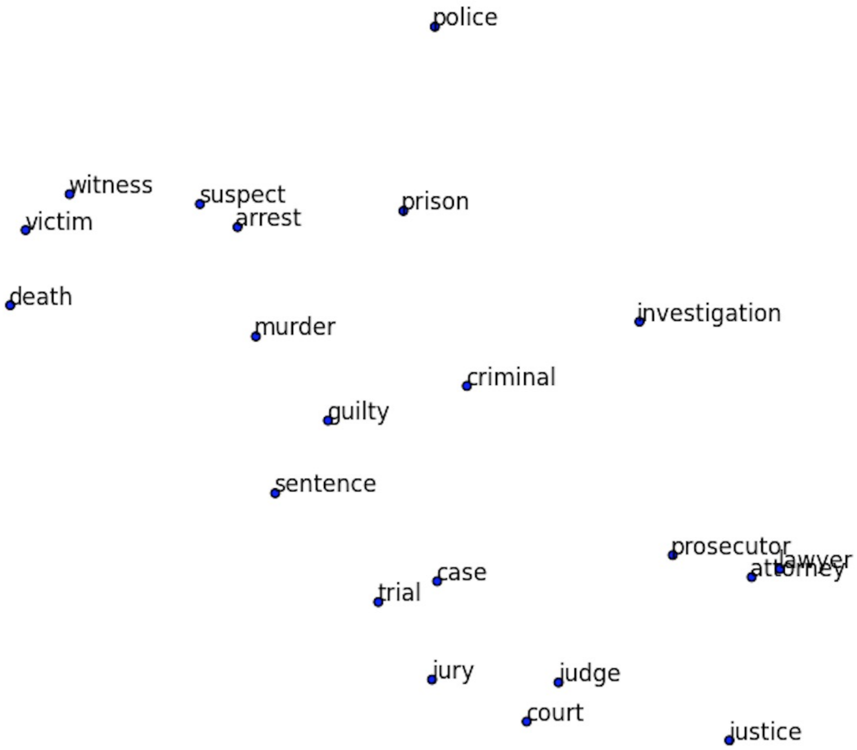
Matthew Engelhard

Recall: Word embeddings allow us to quantify word meaning

If we zoom in on a small region of our word map, it’s all related words.

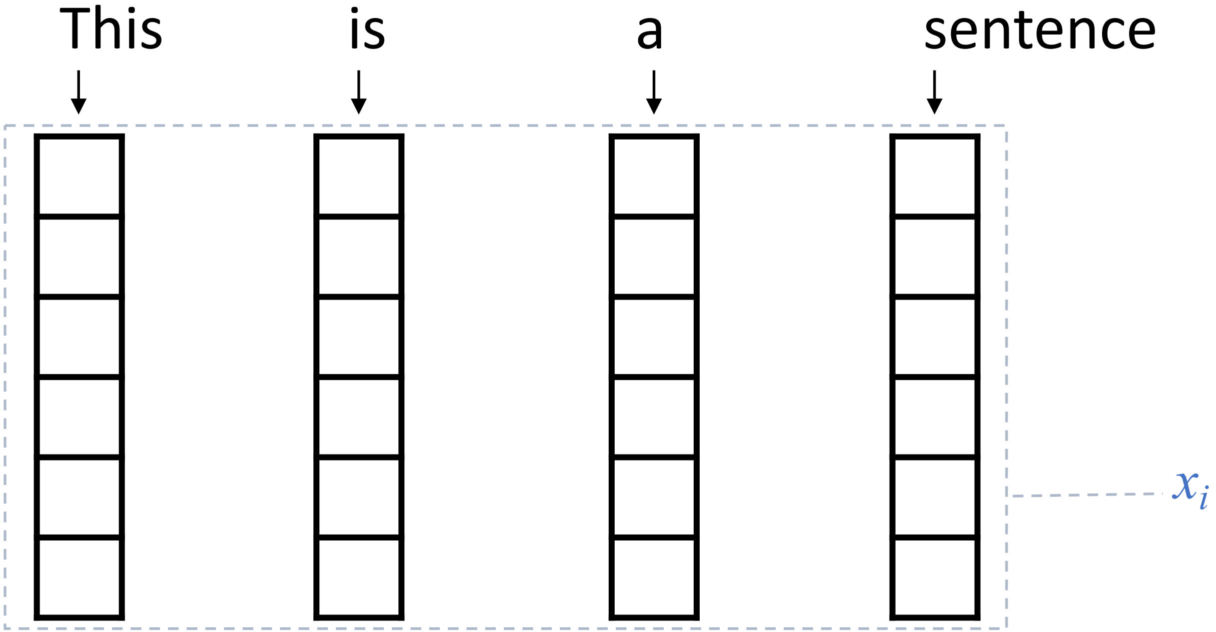
Note the similarity of all the words as a whole, but also of the individual neighbors.

“Lawyer” and “attorney” are nearly identical in space!

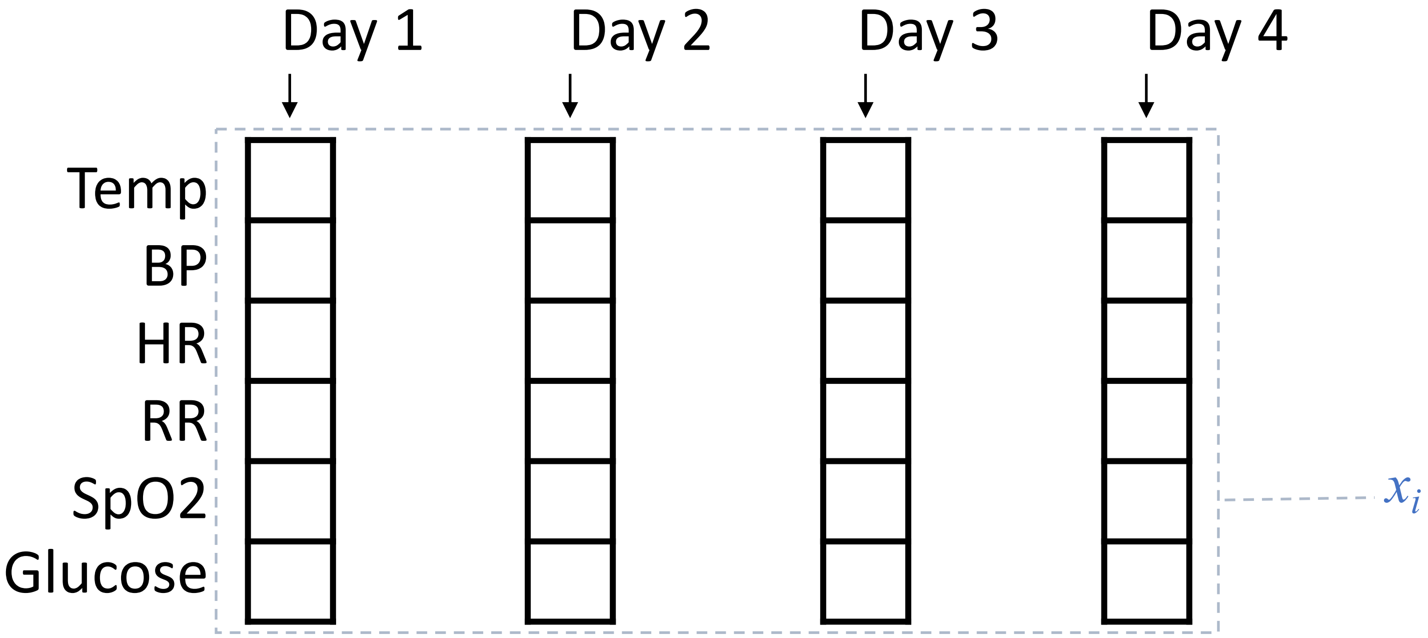


Applying Word Embeddings to a Sentence

- Look up words individually to obtain their vectors
- Construct a sequence of vectors



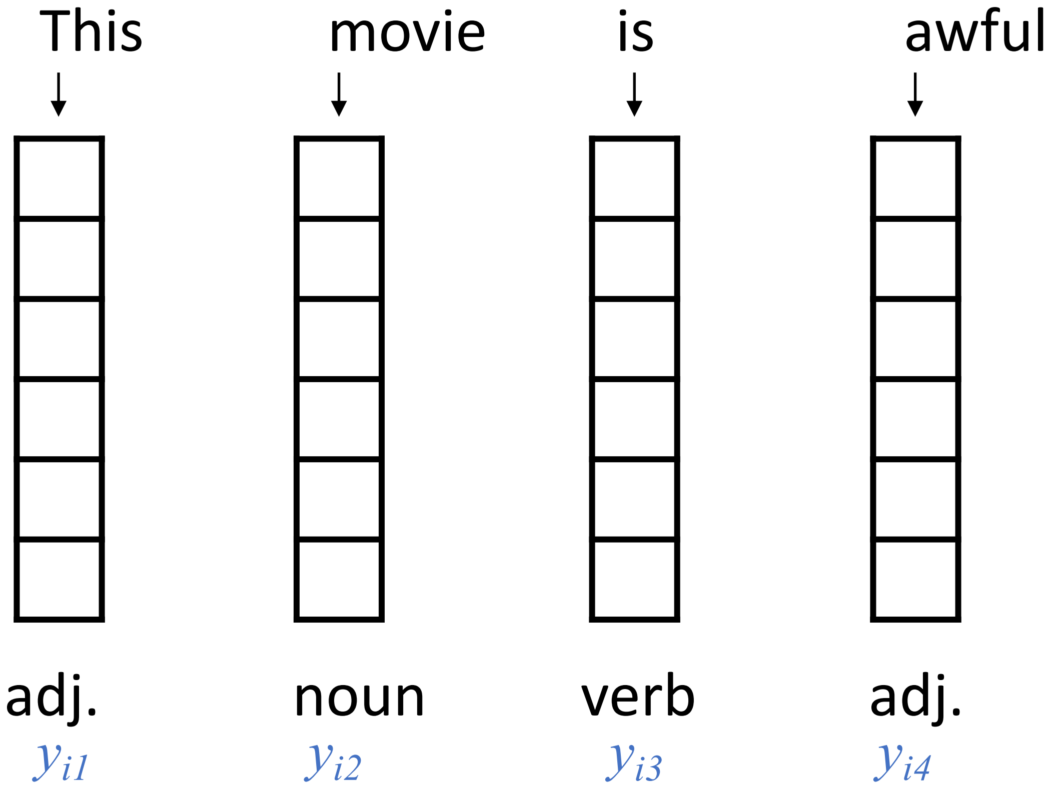
Sequences of measurements: *same structure*



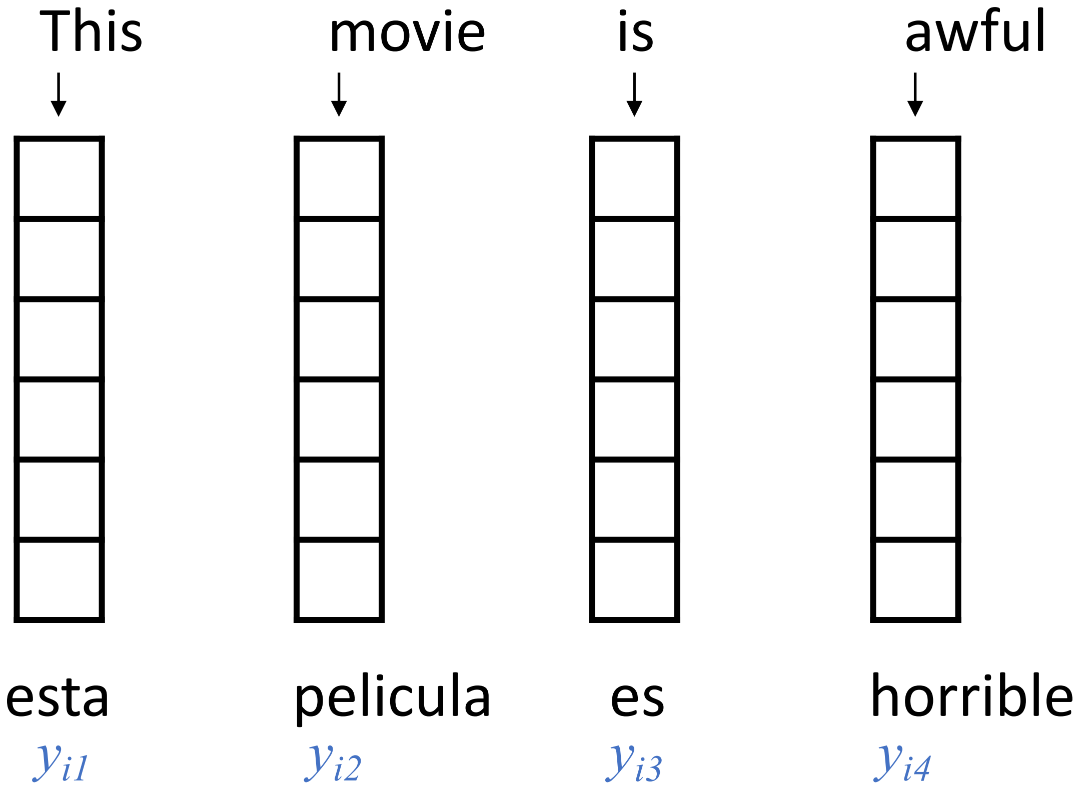
We can make predictions for each:

- Word
- Document (e.g. clinical note)
- Collection of documents (e.g. notes for all patients)

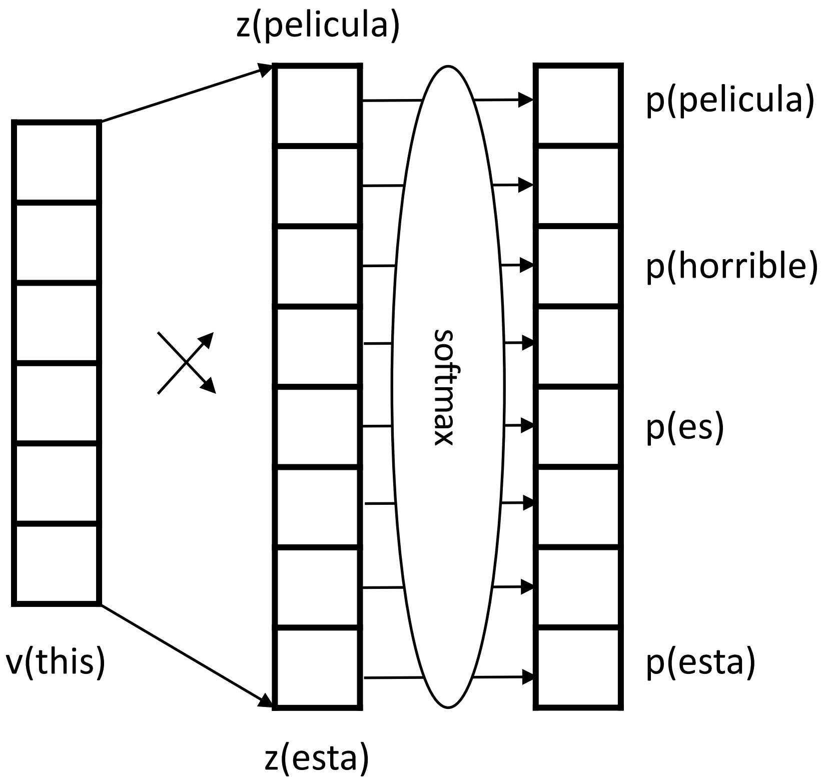
Task 1: Predict a label associated with each word



Task 1: Predict a **label** (?) associated with each word



Multi-Class Logistic Regression (many classes)



Deidentification of Patient Notes

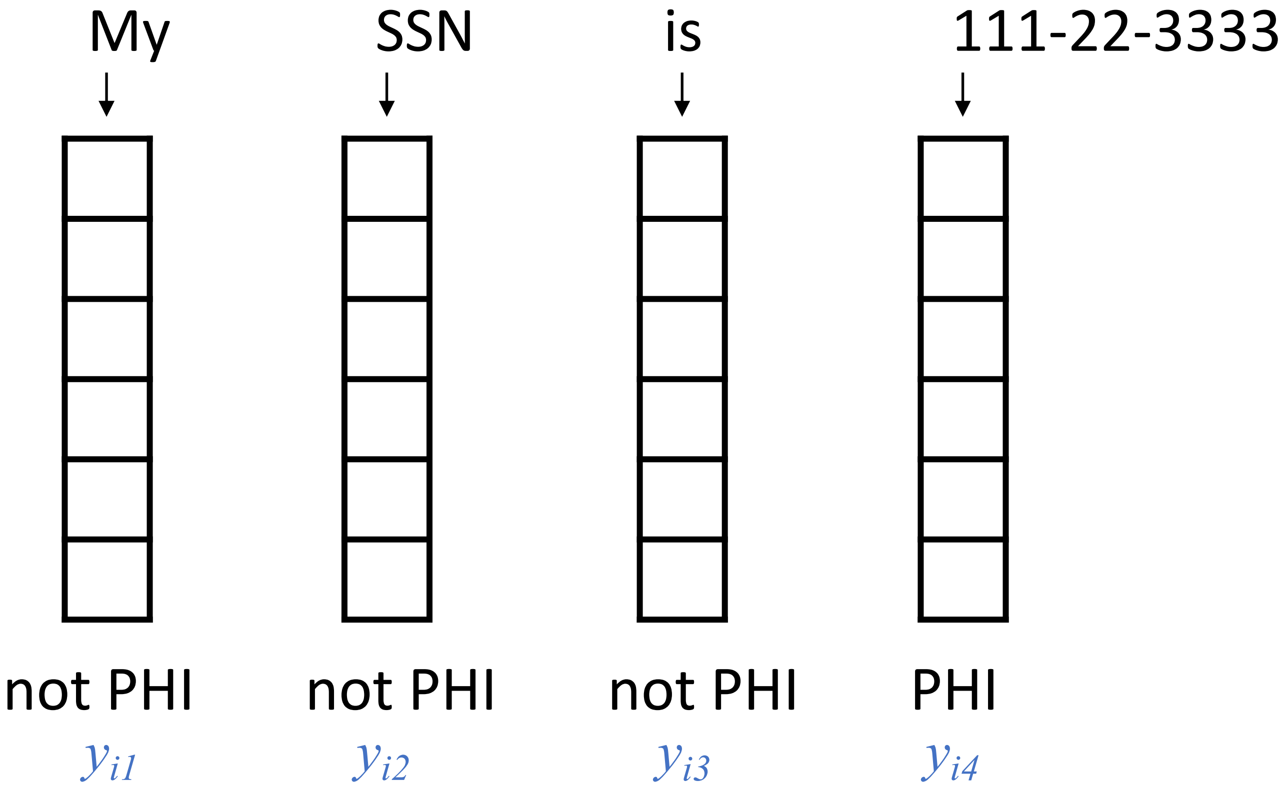
Table 5. Examples of correctly detected PHI instances (in bold) by the ANN

PHI category	ANN
AGE	Father had a stroke at **80** and died of?another stroke at age
	Personal data and overall health: Now **63**, despite his
	FH: Father: Died @ **52** from EtOH abuse (unclear exact etiology)
	Tobacco: smoked from age 7 to **15**, has not smoked since 15.
	History of Present Illness **86F** reports worsening b/l leg pain.
CONTACT	by phone, Dr. Ivan Guy. Call w/ questions **86383**. Keith Gilbert,
	H/O paroxysmal afib VNA **171-311-7974** ======= Medications
DATE	During his **May** hospitalization he had dysphagia
	Social history: divorced, quit smoking in **08**, sober x 10 yrs,
	She is to see him on the **29th** of this month at 1:00 p.m.
	He did have a renal biopsy in teh late **60s** adn thus will look for results,
	Results**02/20/2087** NA 135, K 3.2 (L), CL 96 (L), CO2 30.6, BUN 1
	Jose Church, M.D. /ray DD: 01/18/20 DT: **01/19/:0** DV: 01/18/20

De-identification of patient notes with recurrent neural networks
Dernoncourt F, Lee JY, Uzuner O, Szolovits P
JAMIA 24(3), 2017, 596–606

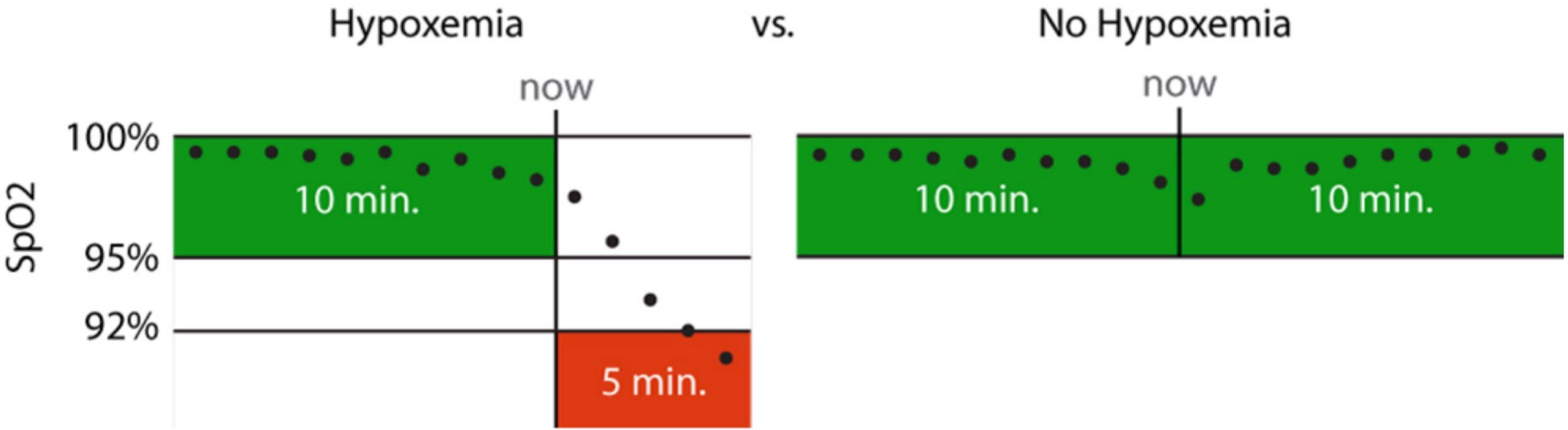
- A bidirectional RNN is used to identify PHI (18 HIPAA fields)
- *i2b2*: 889 discharge summaries, >28k PHI tokens
- *MIMIC*: 1635 discharge summaries, >60k PHI tokens
- State of the art sensitivity and F1 metric on both datasets

Task 1: Predict a label associated with each word

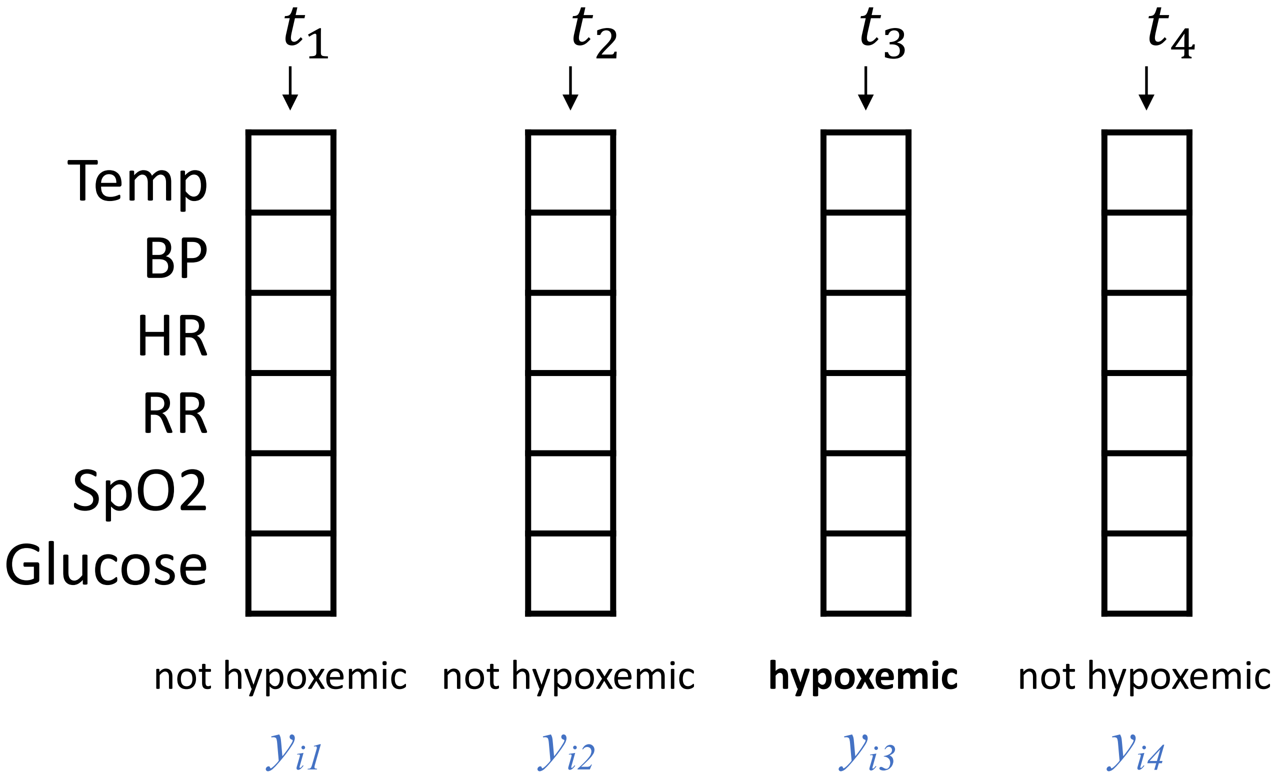


Hypoxemia Prediction during Surgery

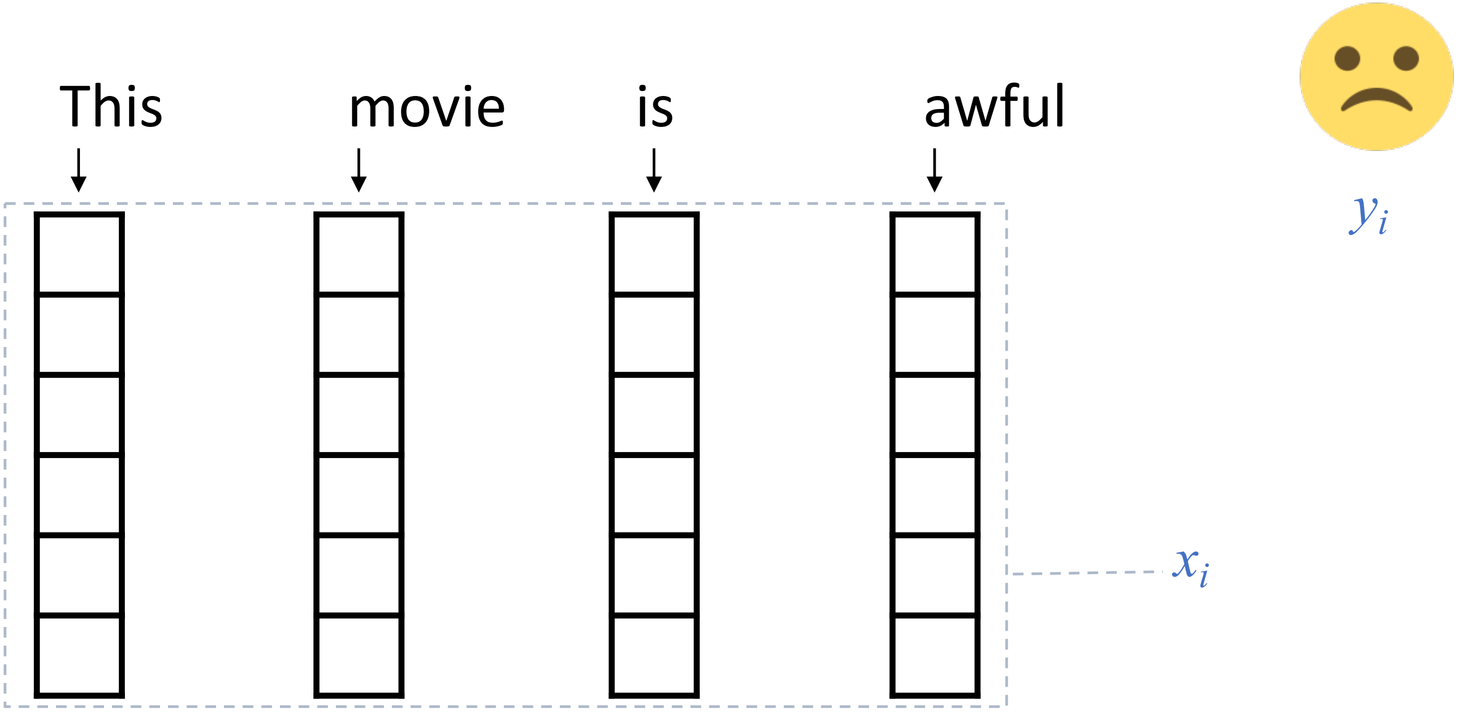
Real-time Prediction Task:
- hypoxemia (yes/no) <u>in the next 5 minutes</u>
- based on data from the Anesthesia Information Management System
- static features + real-time features collected up to that time point



Task 1: Predict label assoc. with each time point

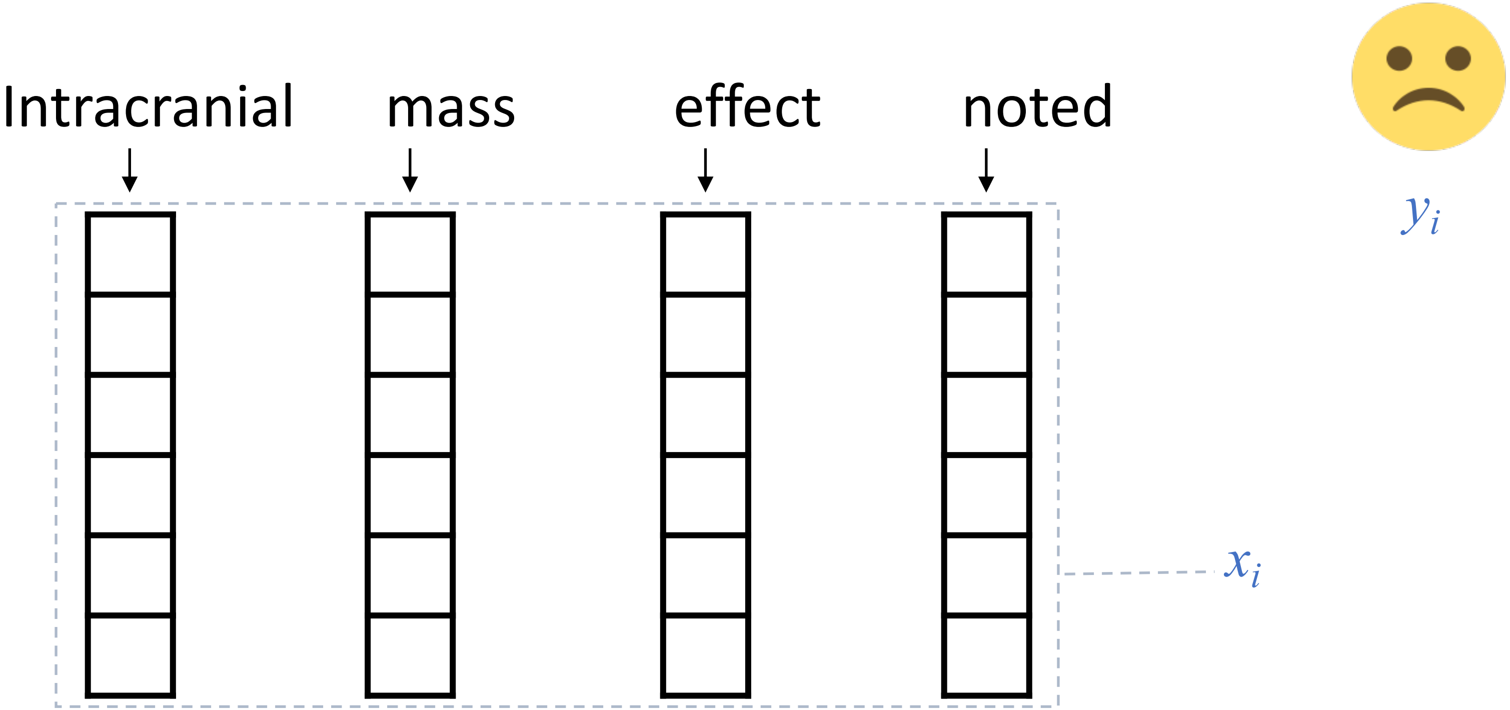


Task 2: Predict a label associated with the document

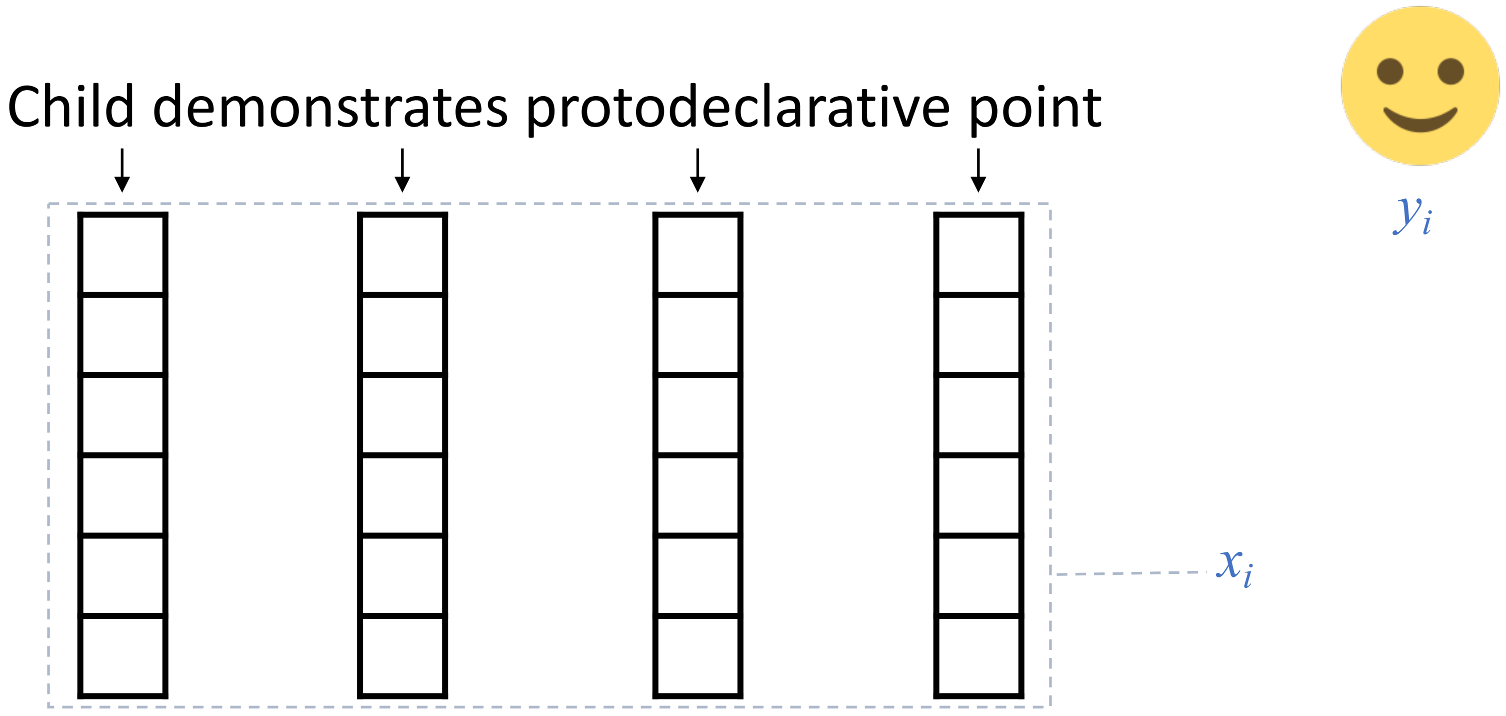


Classification of radiology reports using neural attention models, *IJCNN 2017*

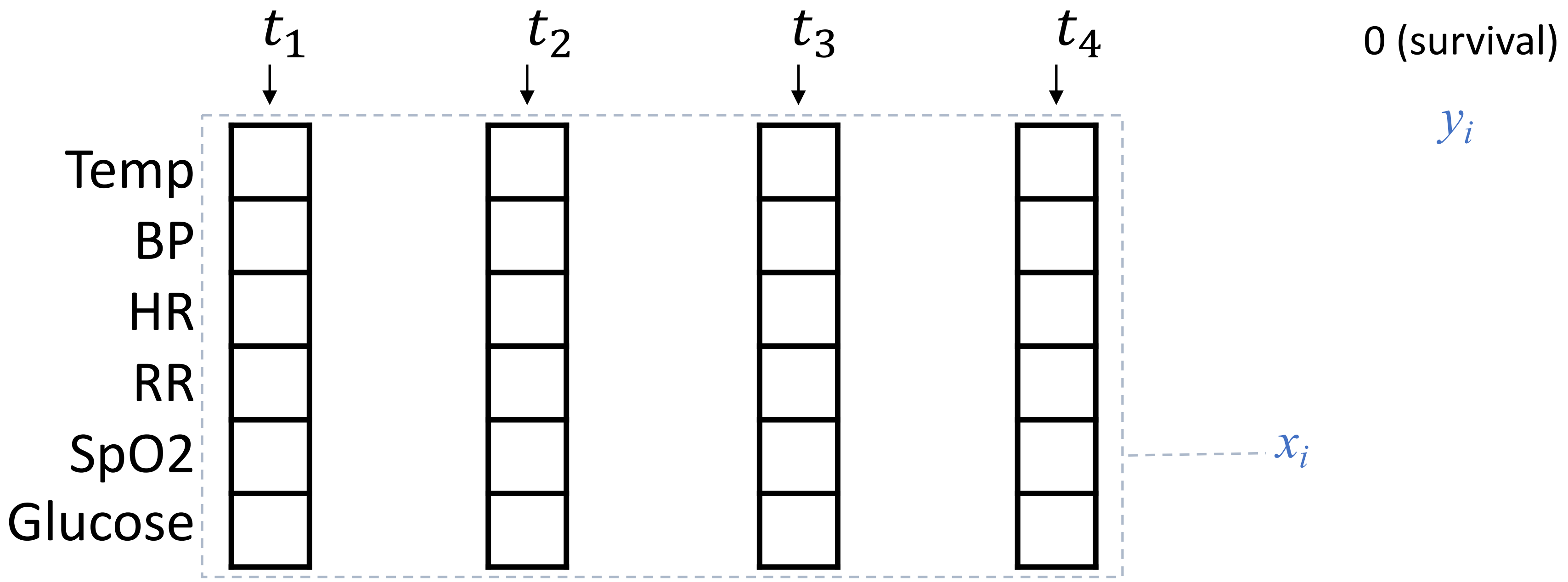
Task 2: Predict a label associated with the report



Task 2: Predict a label associated with the note



Task 2: Predict label assoc. with all measurements



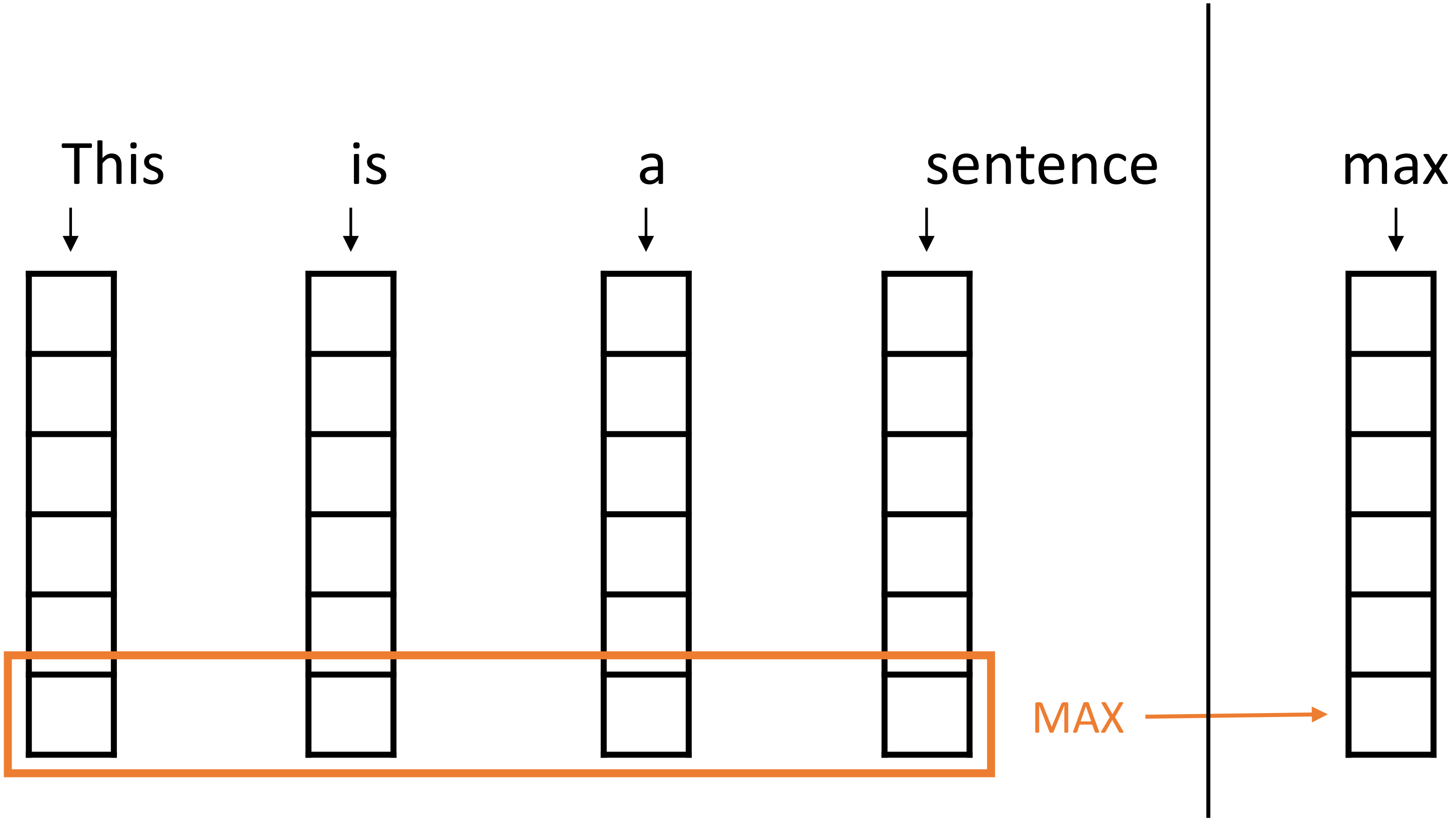
First Challenge: Sequences Vary in Length

- Sentences/text have different # words
- Time-series have different # measurement times

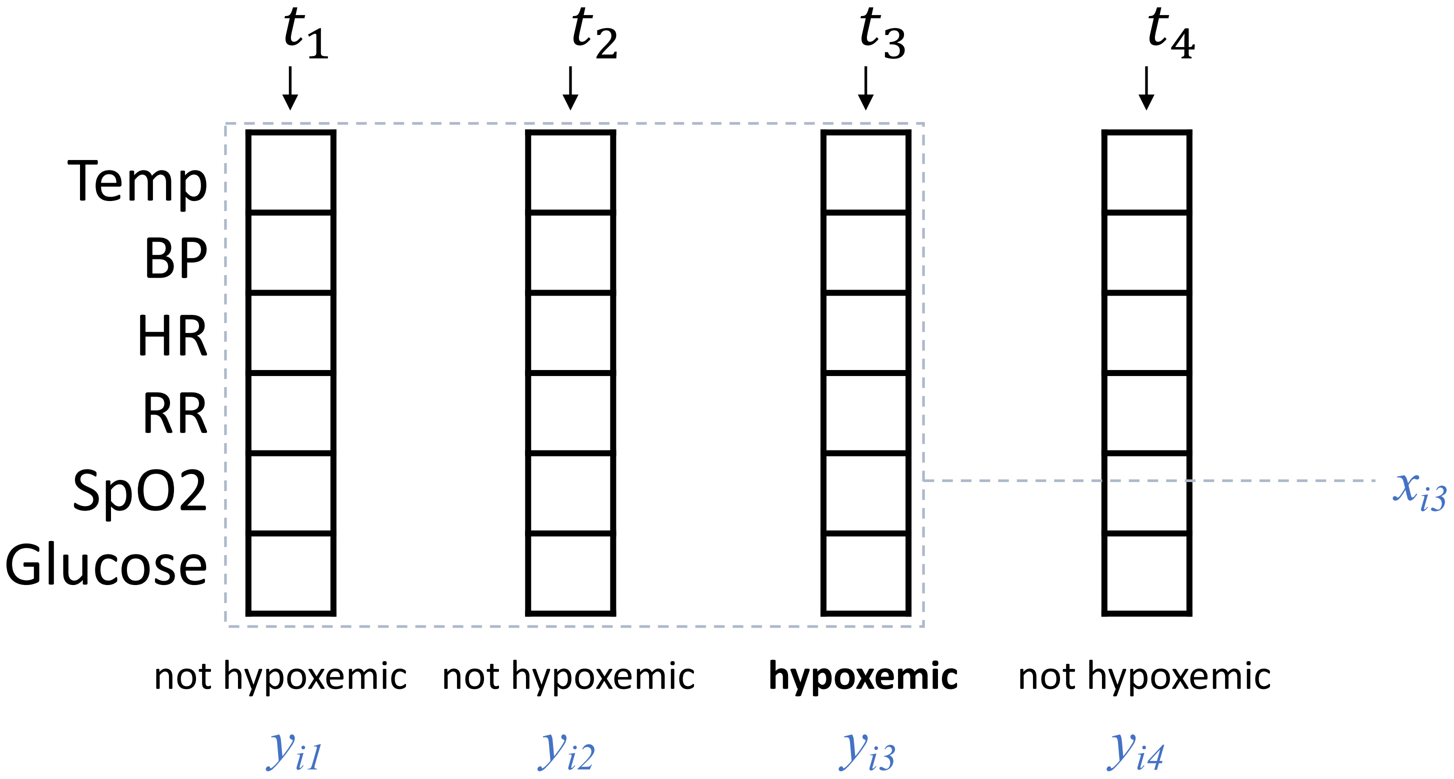
- More generally, even for models where we’re making predictions for each word or time point, we have to deal with the whole history of previous words / measurements

- <u>Easy solution:</u> aggregate over words/time points

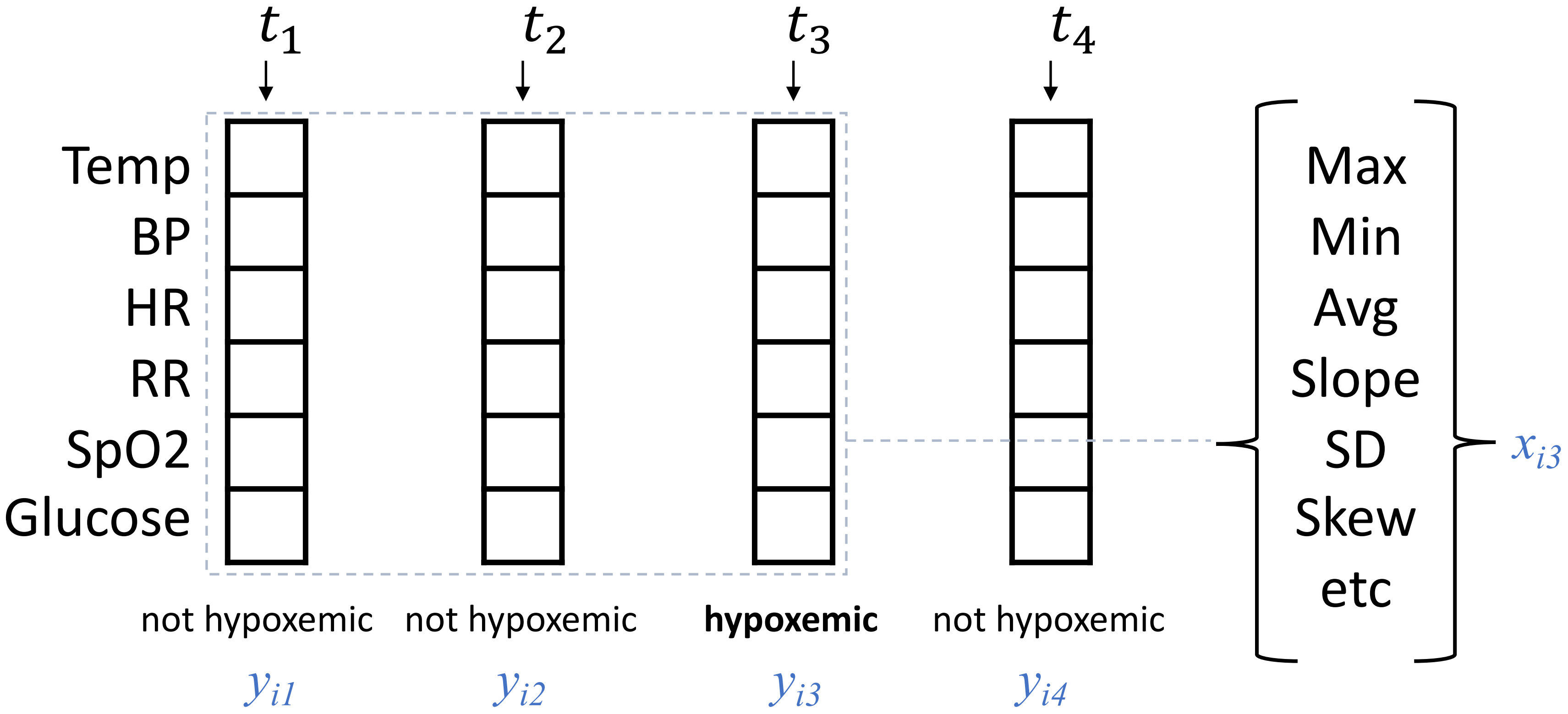
VSWEM allows us to convert a variable-length sentence to a fixed-length feature vector



Similarly, we can aggregate measurements in a time-series



Similarly, we can aggregate measurements in a time-series

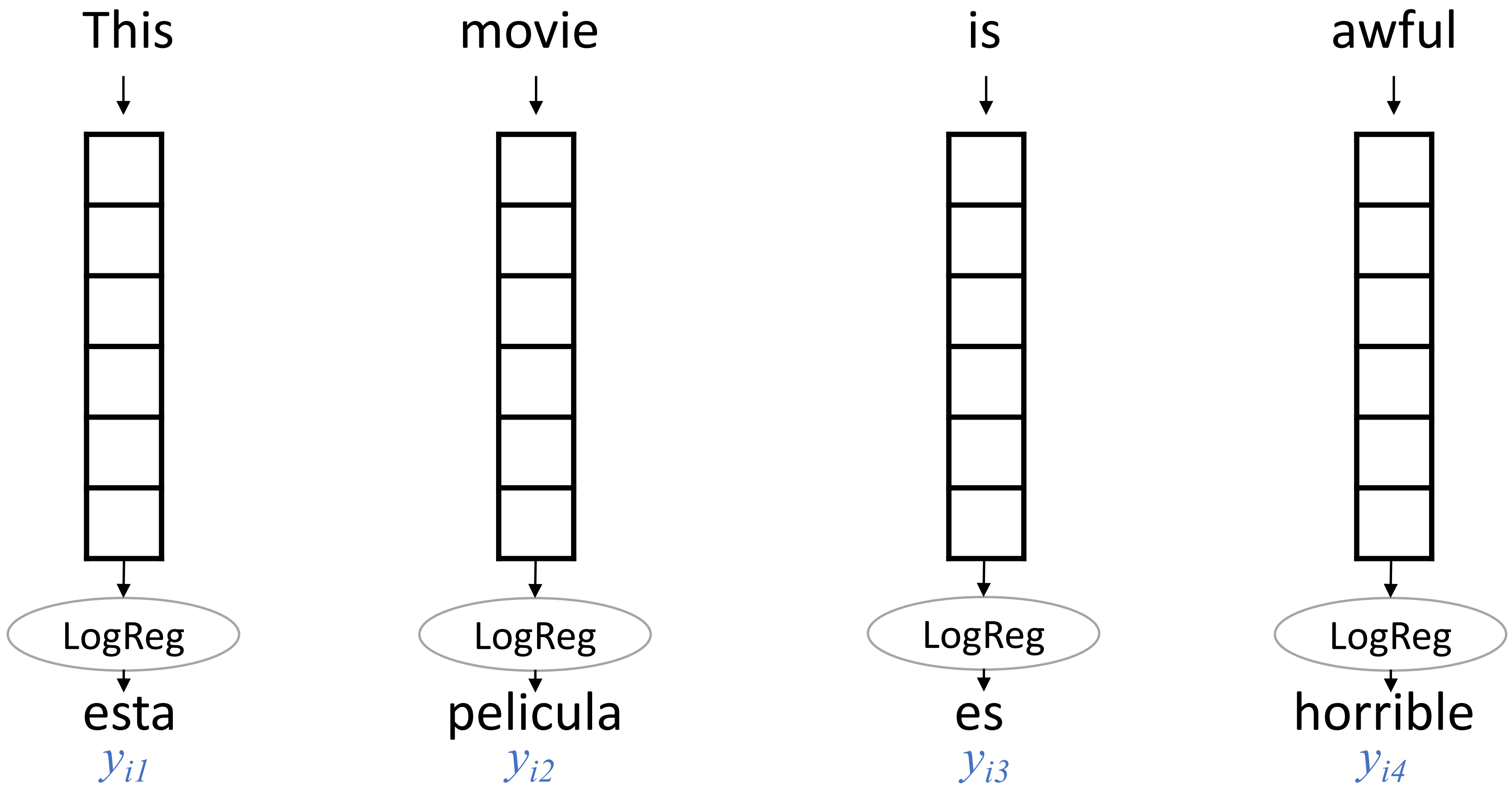


Second Challenge: Is there a better way to aggregate?

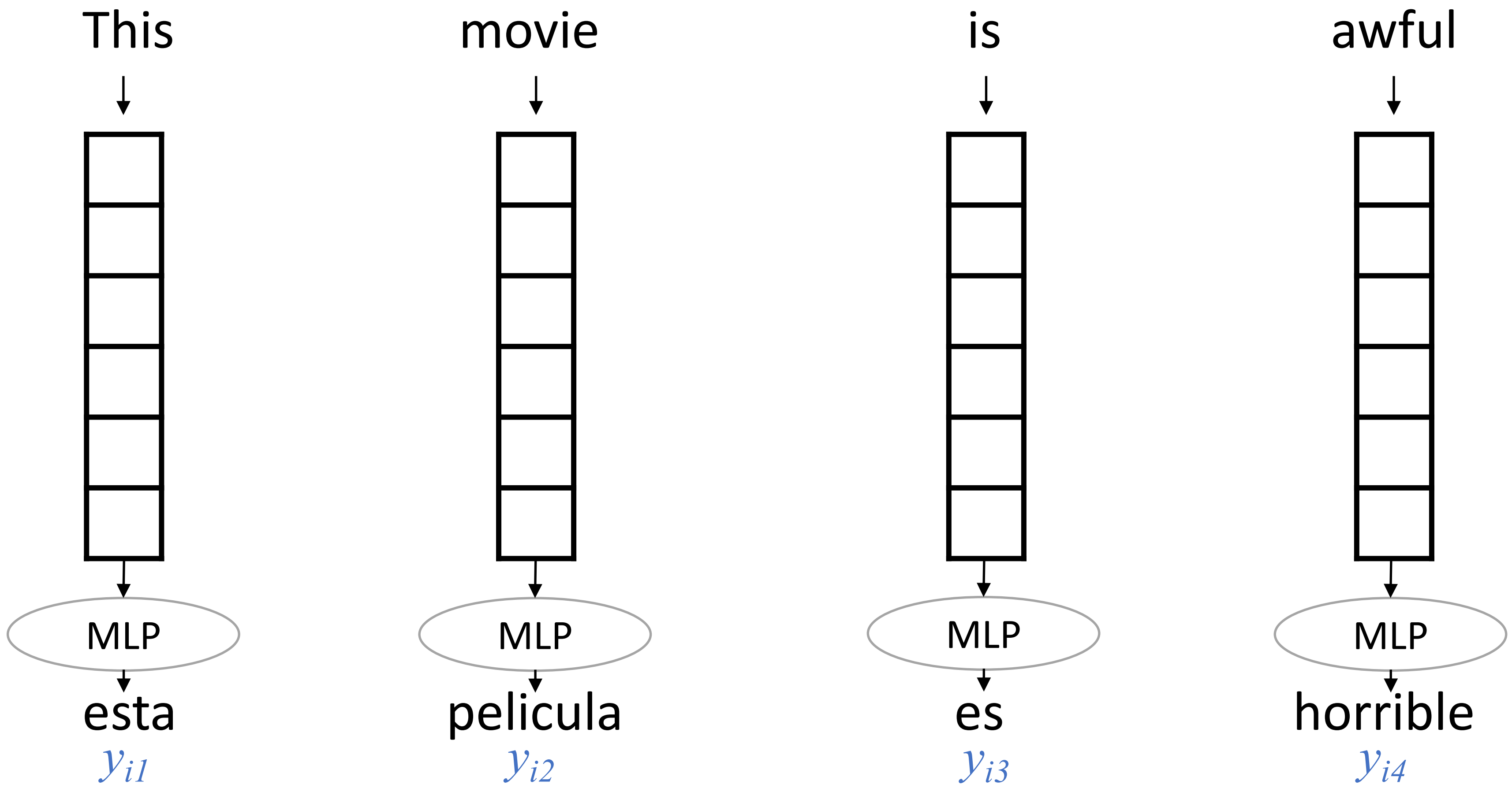
- A sentence is more than the average (or max) of its words
- A time-series is more than the average / min / max / SD of individual measurements
- We'd like to interpret words or measurements *in context*
- <u>Deep learning</u>: we *learn* what's important about the sequence rather than choosing features or summary stats

Recurrent Neural Networks

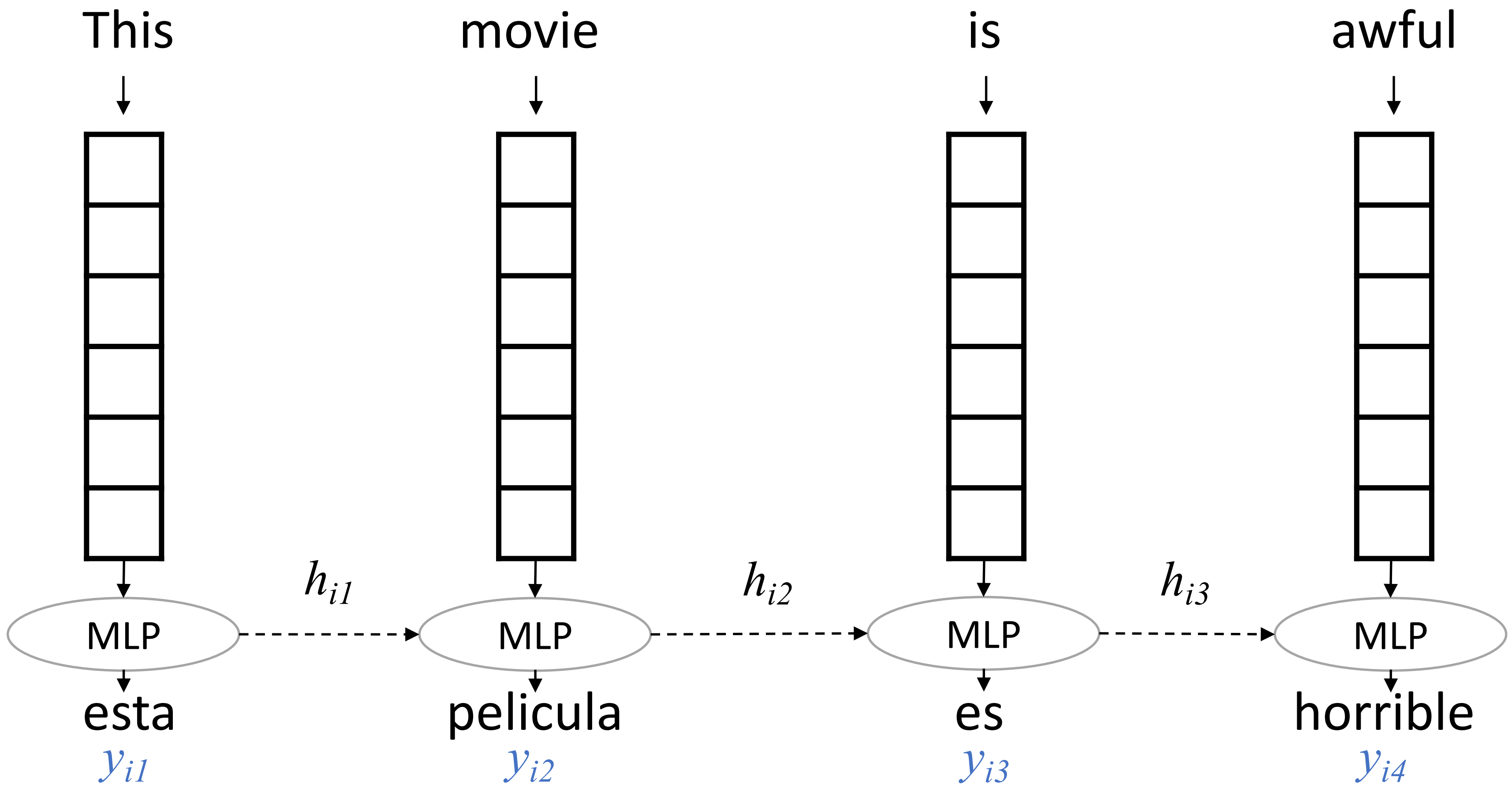
Predict a label associated with each word



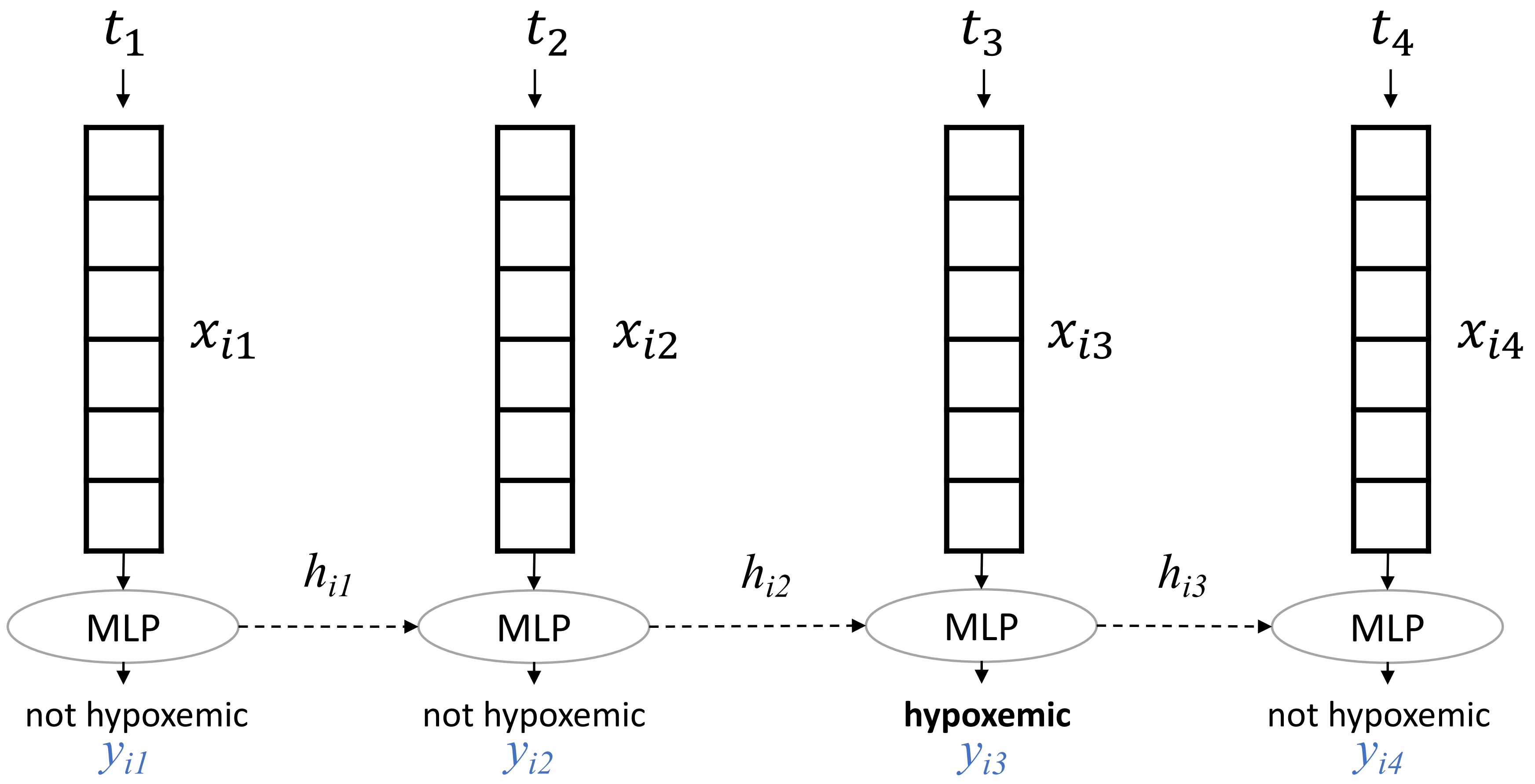
Predict a label associated with each word



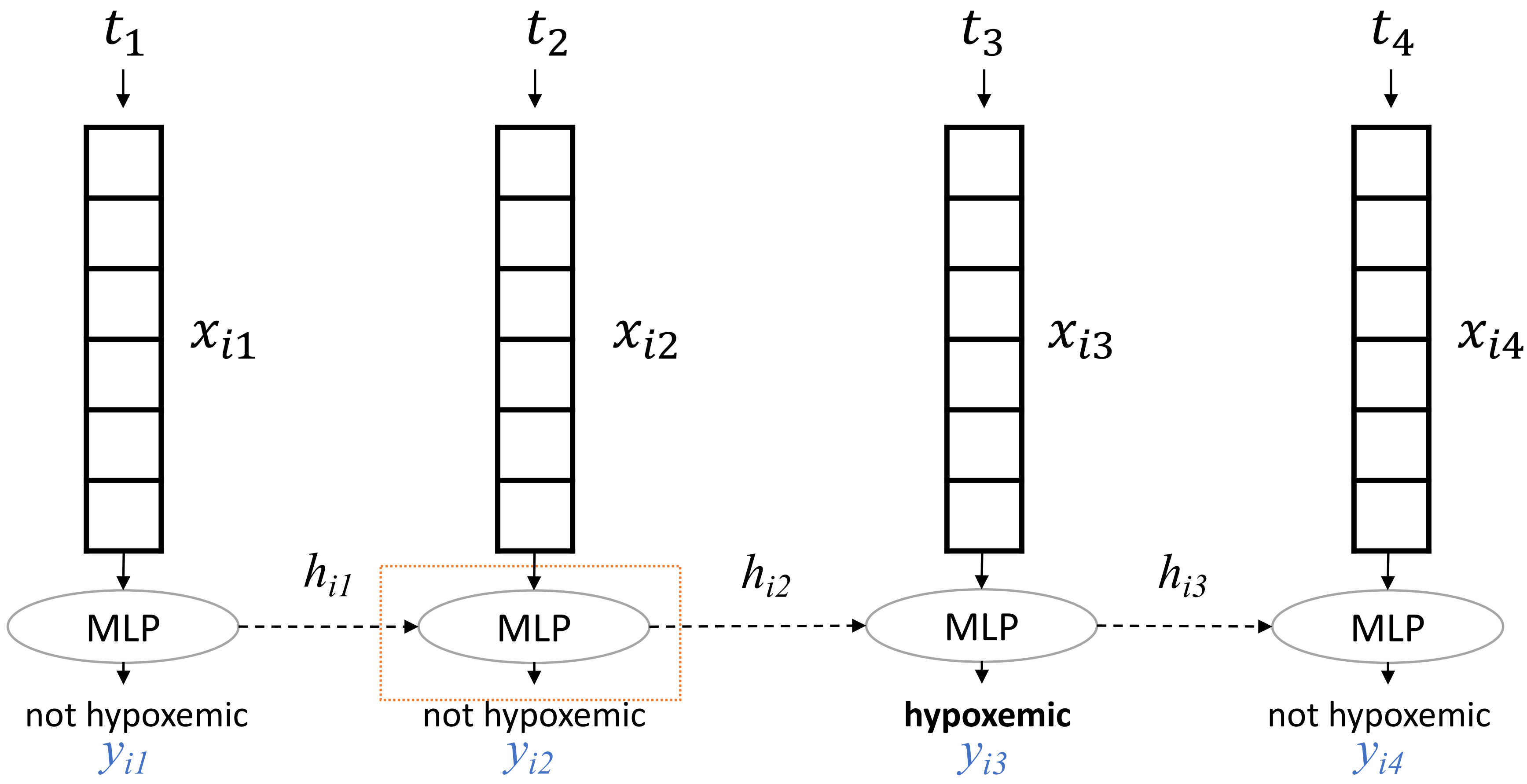
Transfer *relevant* information about earlier words



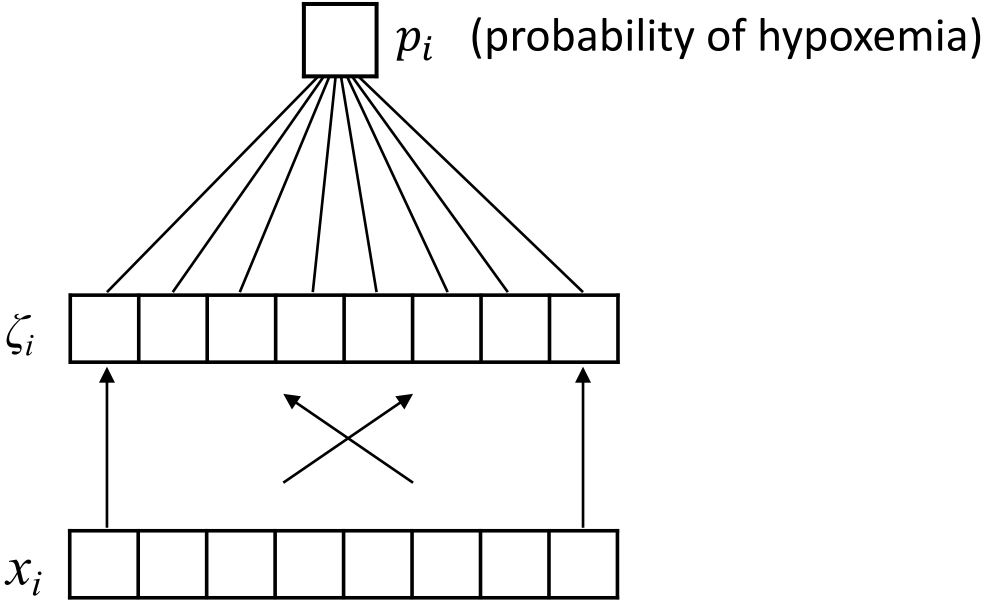
Transfer *relevant* information about earlier values



Transfer *relevant* information about earlier values



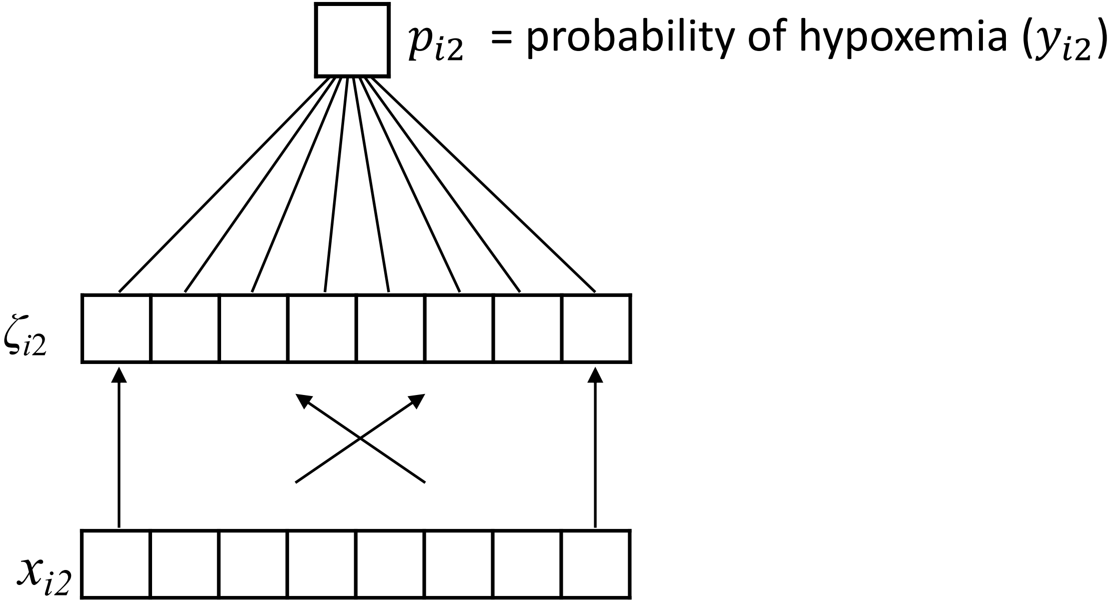
Back to Lectures 2-3…



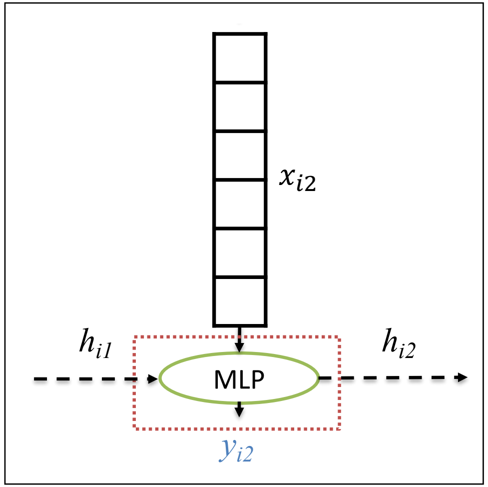
Since they are neither an input nor an output, the features ζ are said to be a “hidden” layer

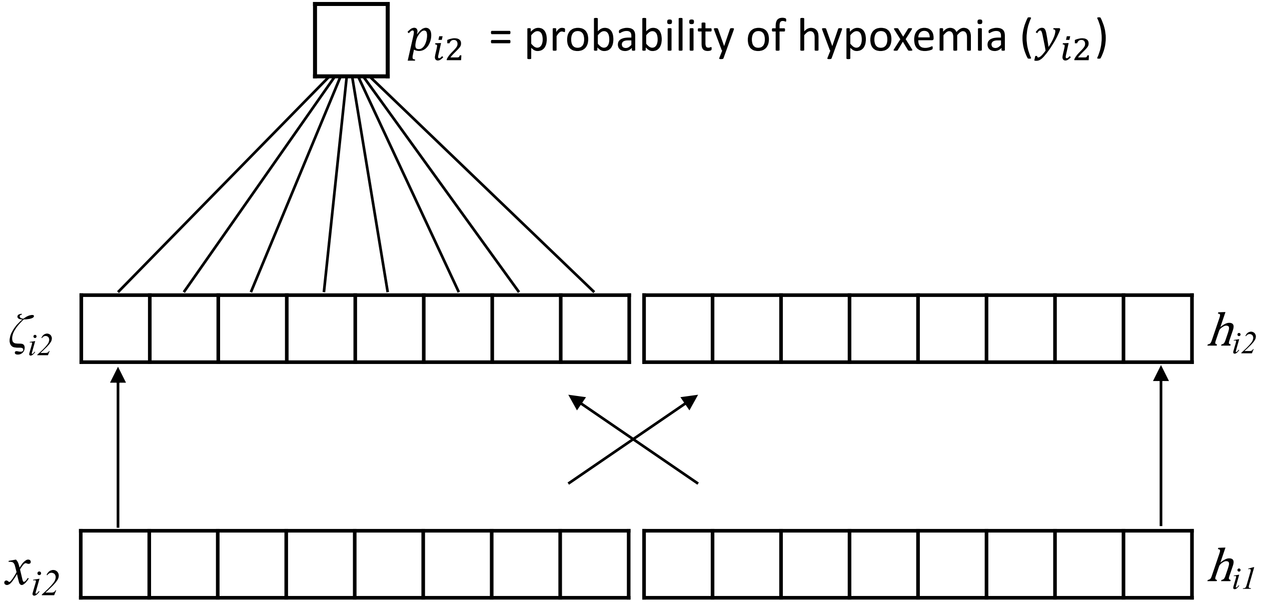
- Instead of predicting p_i directly from our feature vector x, introduce a vector of **“latent” features** ζ (zeta) that we will use to predict p_i

- Think of ζ as a learned representation that is useful for predicting *p*

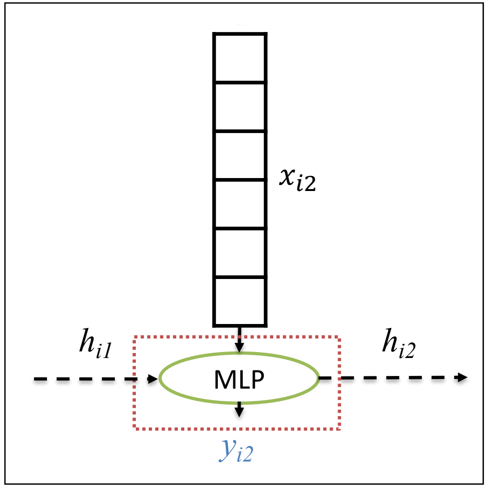


Since they are neither an input nor an output, the features ζ are said to be a "hidden" layer

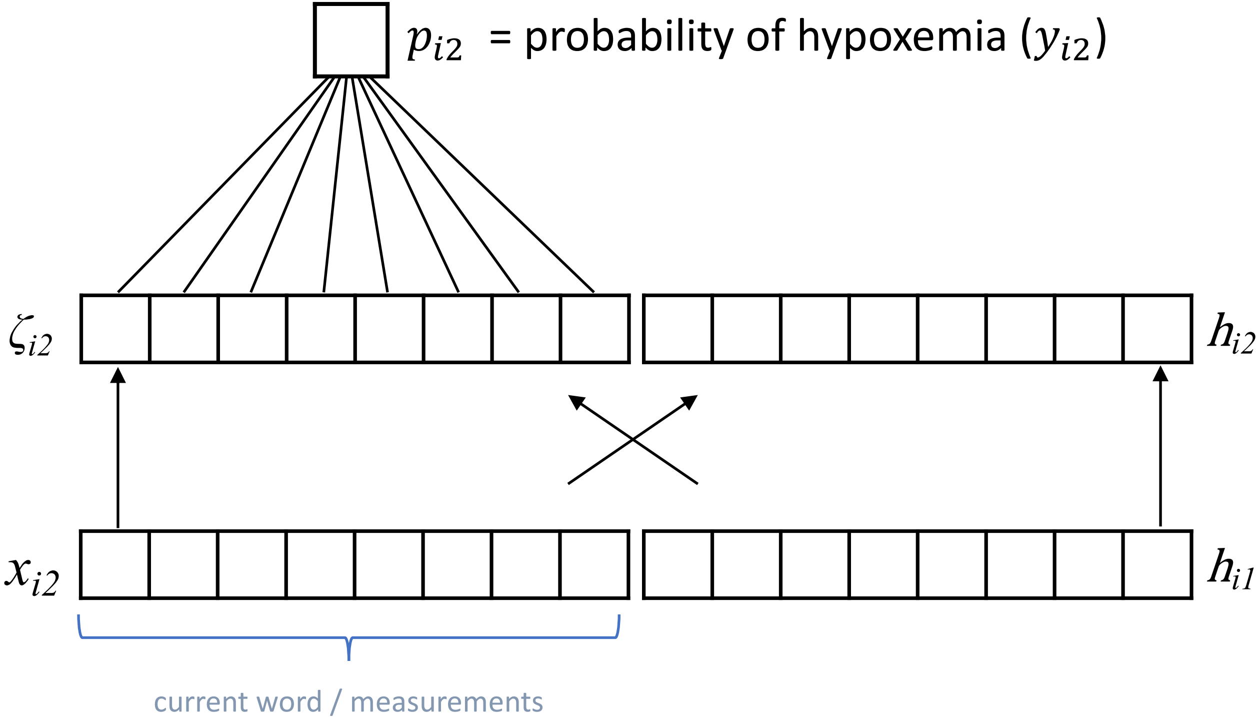




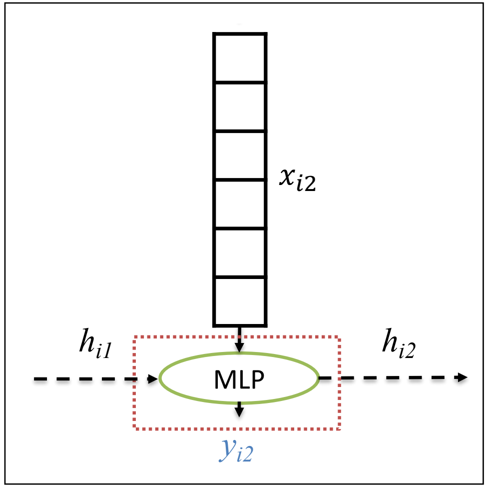
p_{i2} = probability of hypoxemia (y_{i2})
ζ_{i2}
h_{i2}
x_{i2}
h_{i1}



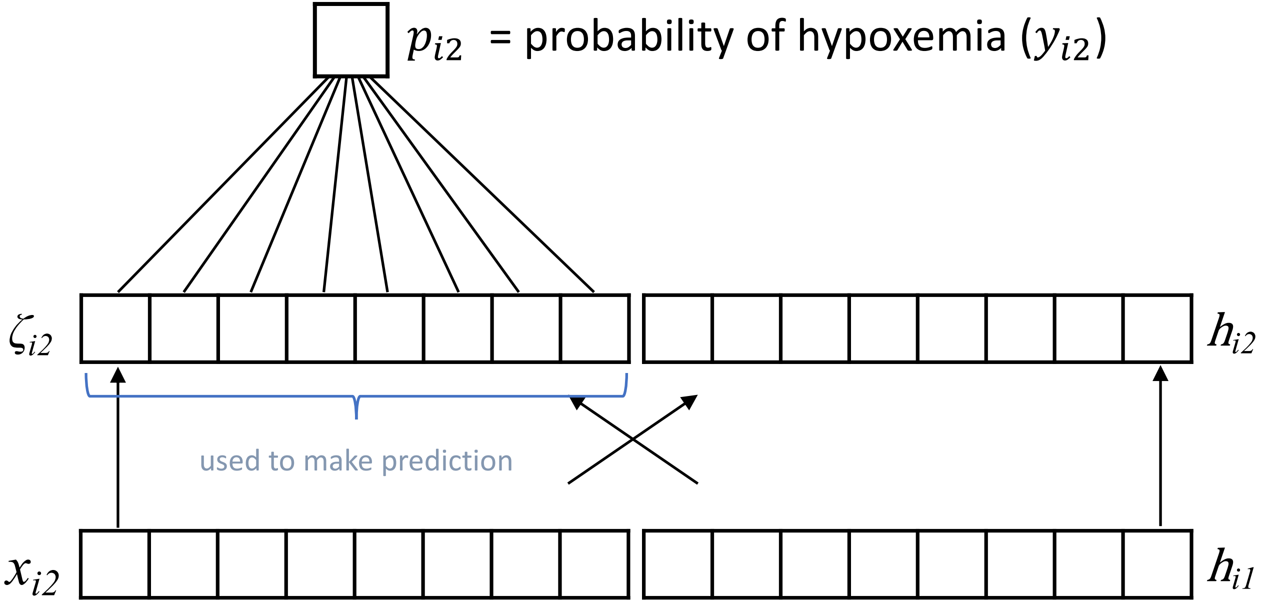
x_{i2}
h_{i1}
h_{i2}
MLP
y_{i2}



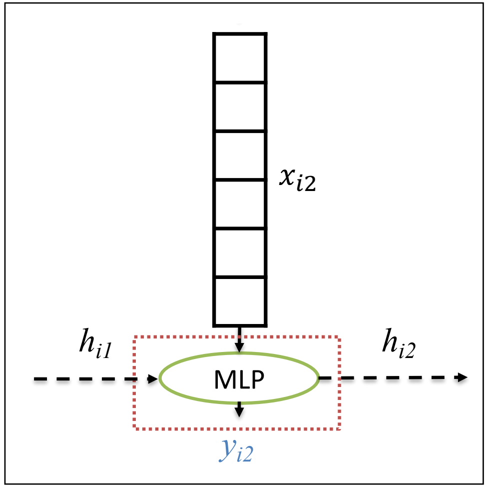
p_{i2} = probability of hypoxemia (y_{i2})
ζ_{i2}
h_{i2}
x_{i2}
h_{i1}
current word / measurements



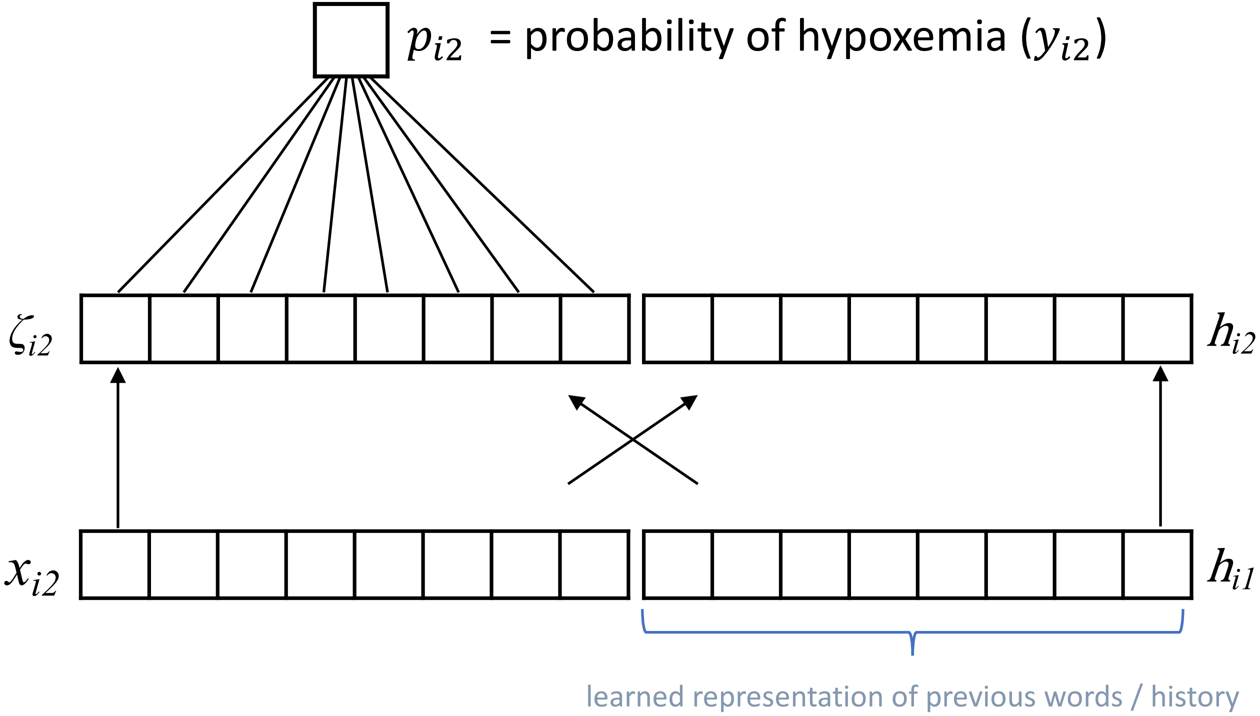
x_{i2}
h_{i1}
h_{i2}
MLP
y_{i2}

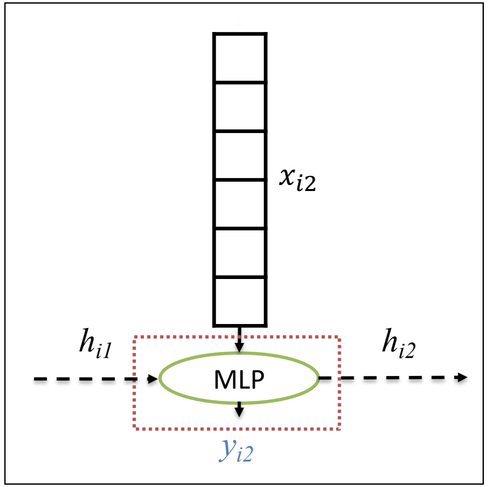


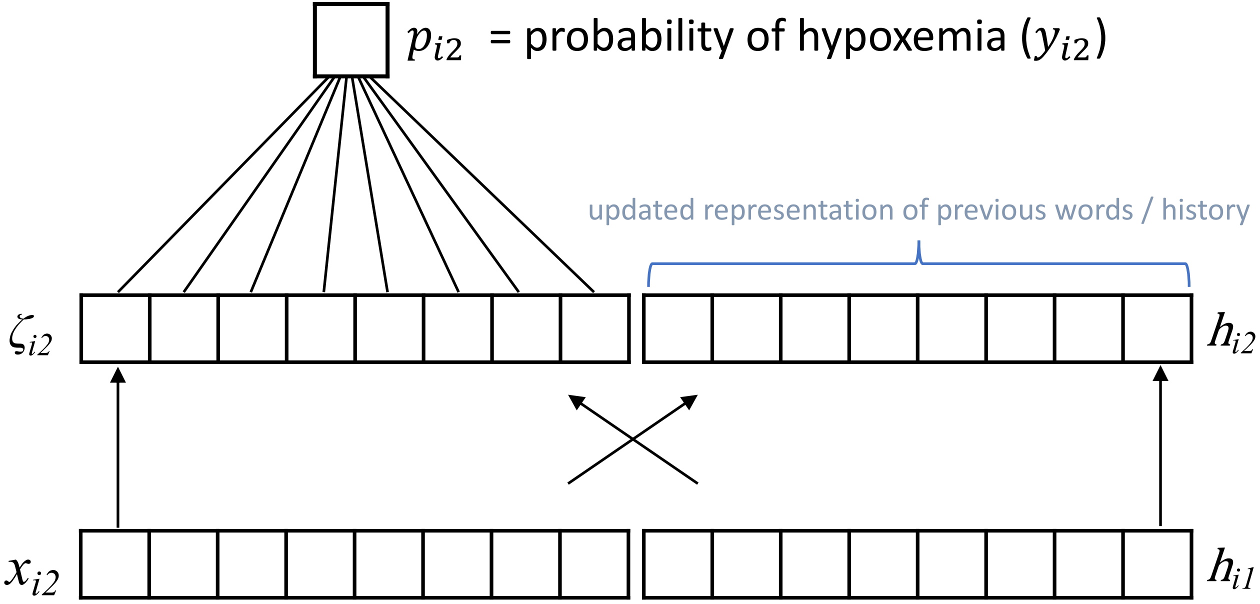
p_{i2} = probability of hypoxemia (y_{i2})
ζ_{i2}
h_{i2}
used to make prediction
x_{i2}
h_{i1}

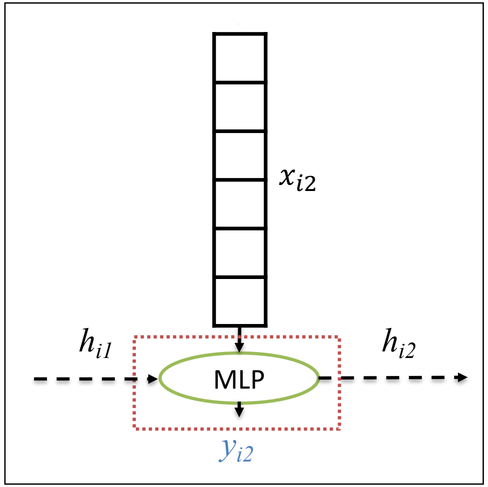


x_{i2}
h_{i1}
h_{i2}
MLP
y_{i2}

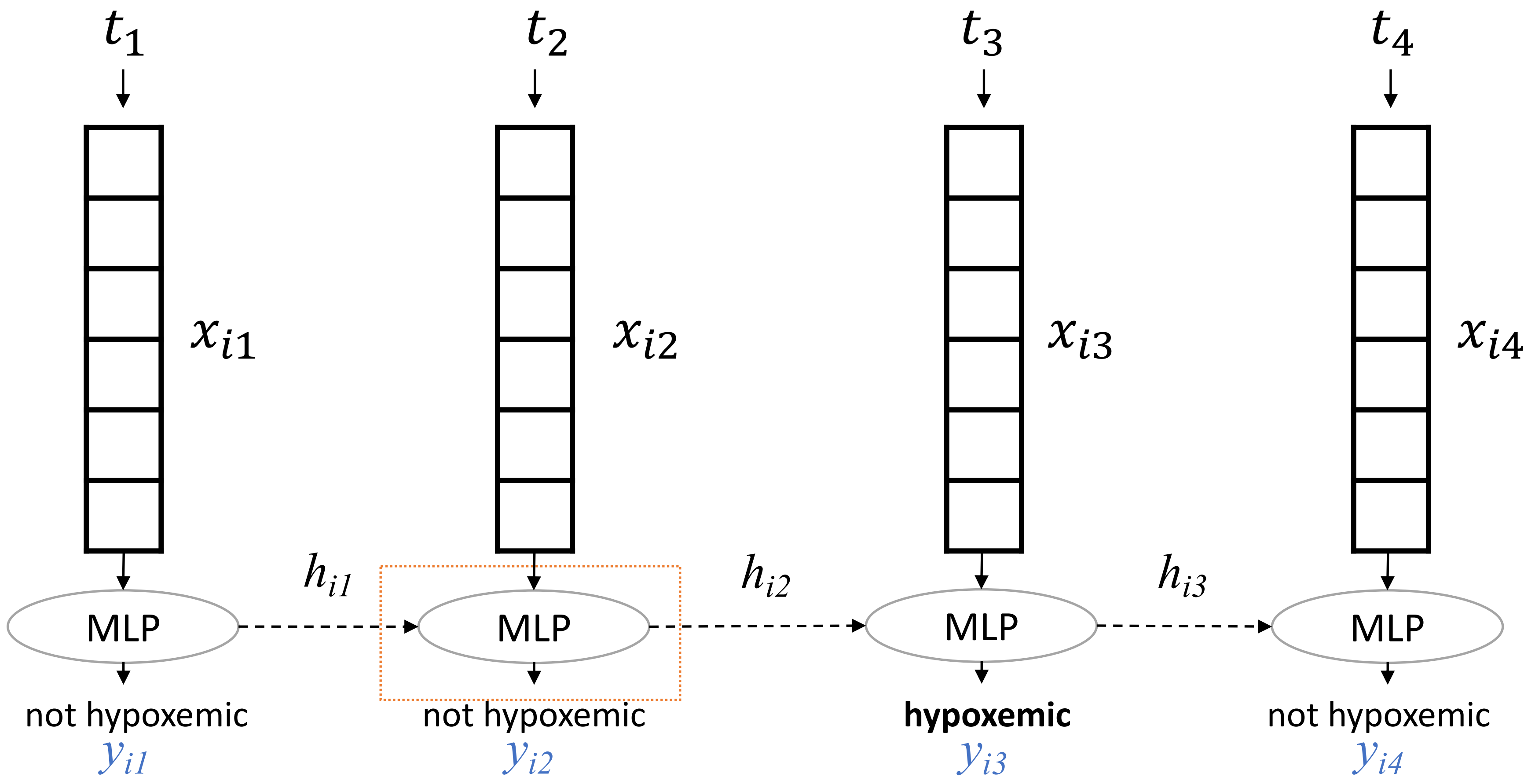

p_{i2} = probability of hypoxemia (y_{i2})
ζ_{i2}
h_{i2}
x_{i2}
h_{i1}
learned representation of previous words / history


x_{i2}
h_{i1}
h_{i2}
MLP
y_{i2}

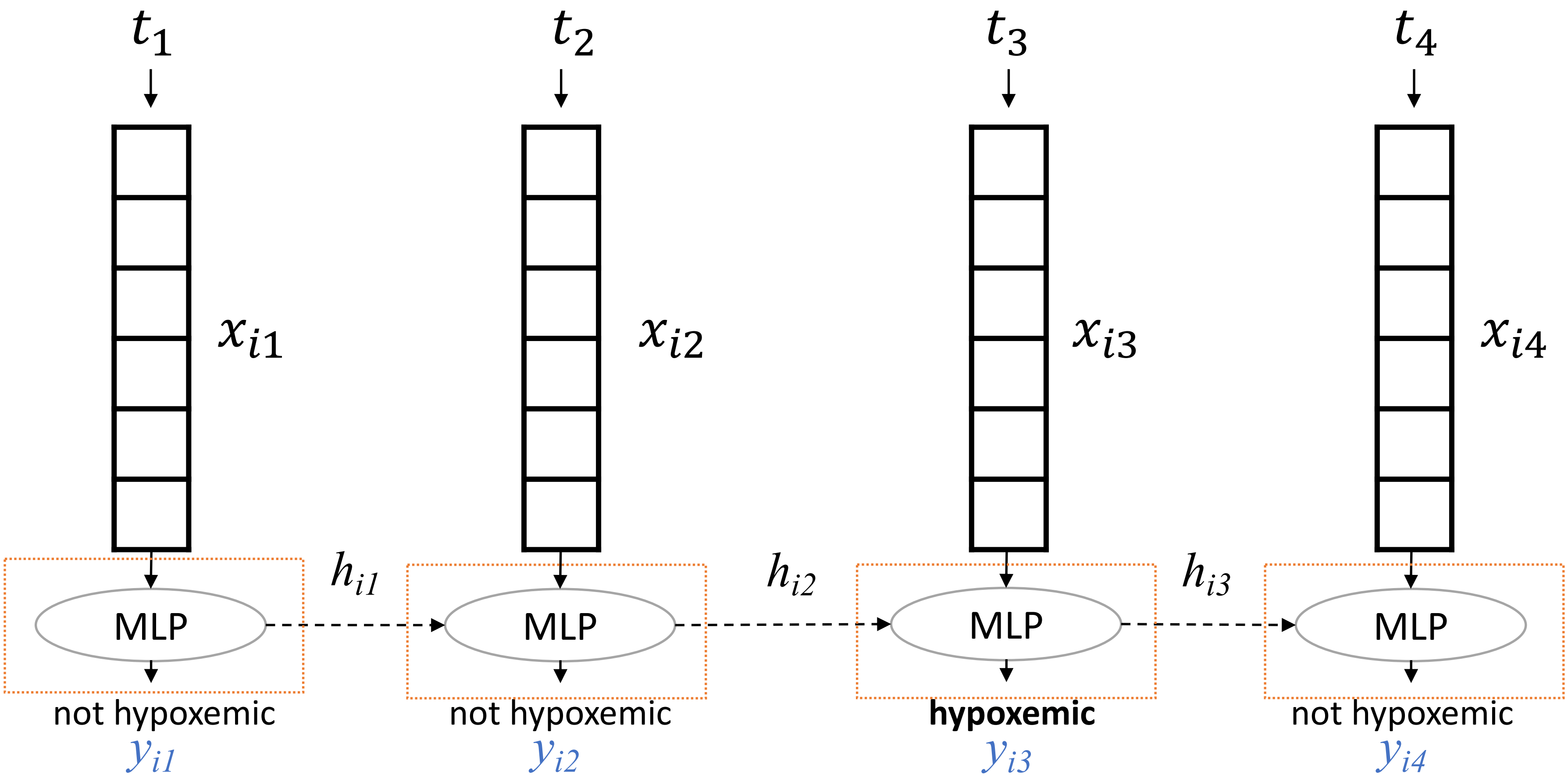

p_{i2} = probability of hypoxemia (y_{i2})
updated representation of previous words / history
ζ_{i2}
h_{i2}
x_{i2}
h_{i1}


x_{i2}
h_{i1}
h_{i2}
MLP
y_{i2}

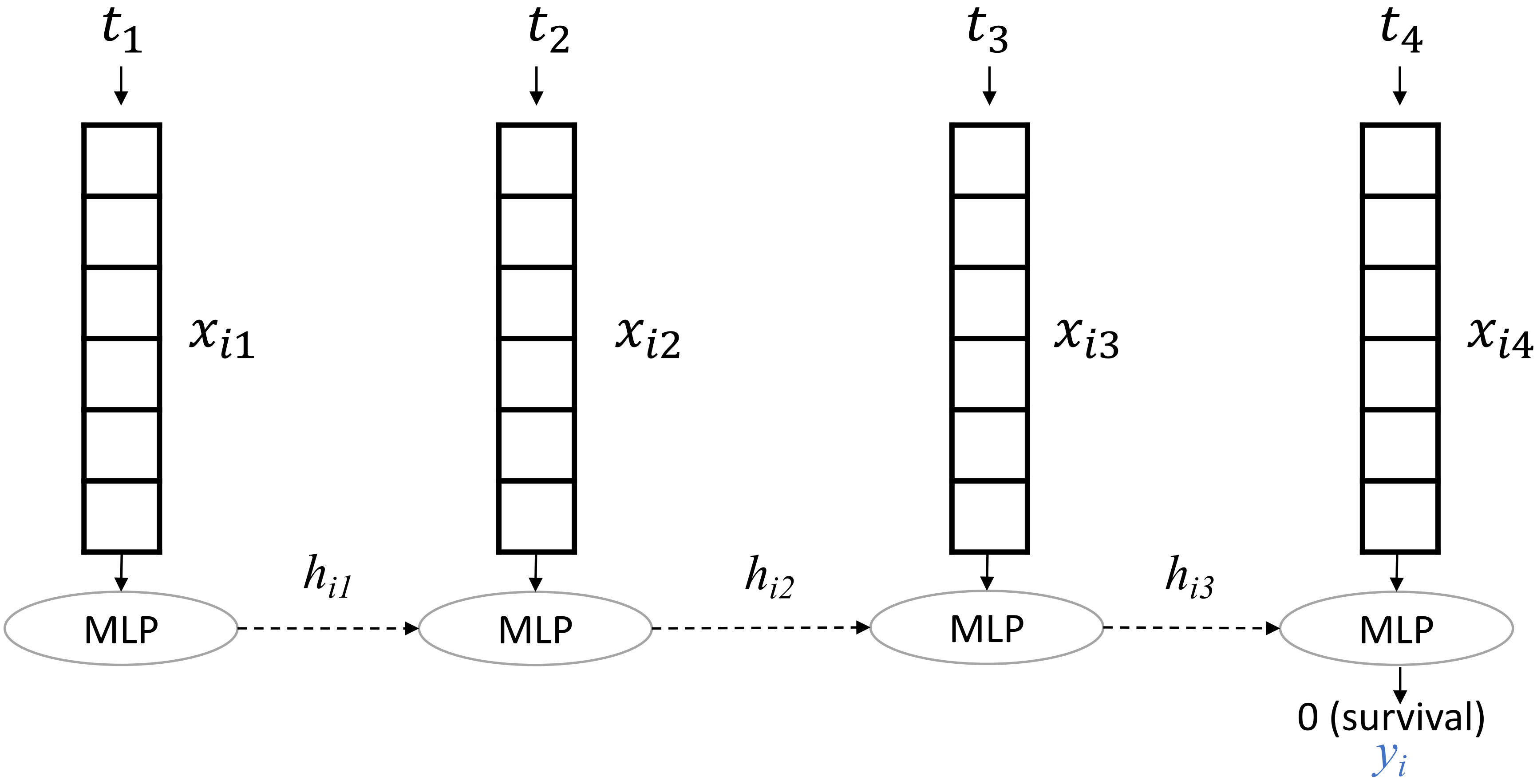
We *learn* what's important about previous values



Recurrent MLP (NN): these are all the same / have same weights



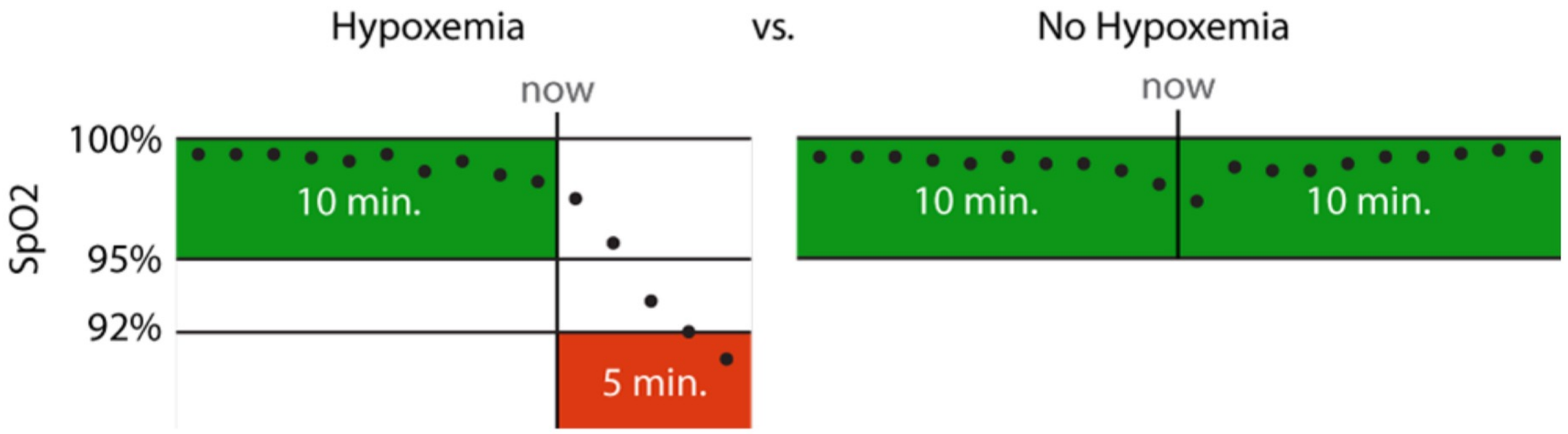
Task 1: Predict a label associated with the sequence



Hypoxemia Prediction: Use learned representation of previous measurements

Real-time Prediction Task:

- hypoxemia (yes/no) in the next 5 minutes
- based on data from the Anesthesia Information Management System
- static features + real-time features collected up to that time point



Common RNN Variants

- Gated Recurrent Unit (GRU)
- Long Short Term Memory (LSTM)
- Bidirectional RNNs
  - Look at previous words and upcoming words
  - Usually not appropriate for time-series

Deidentification of Patient Notes

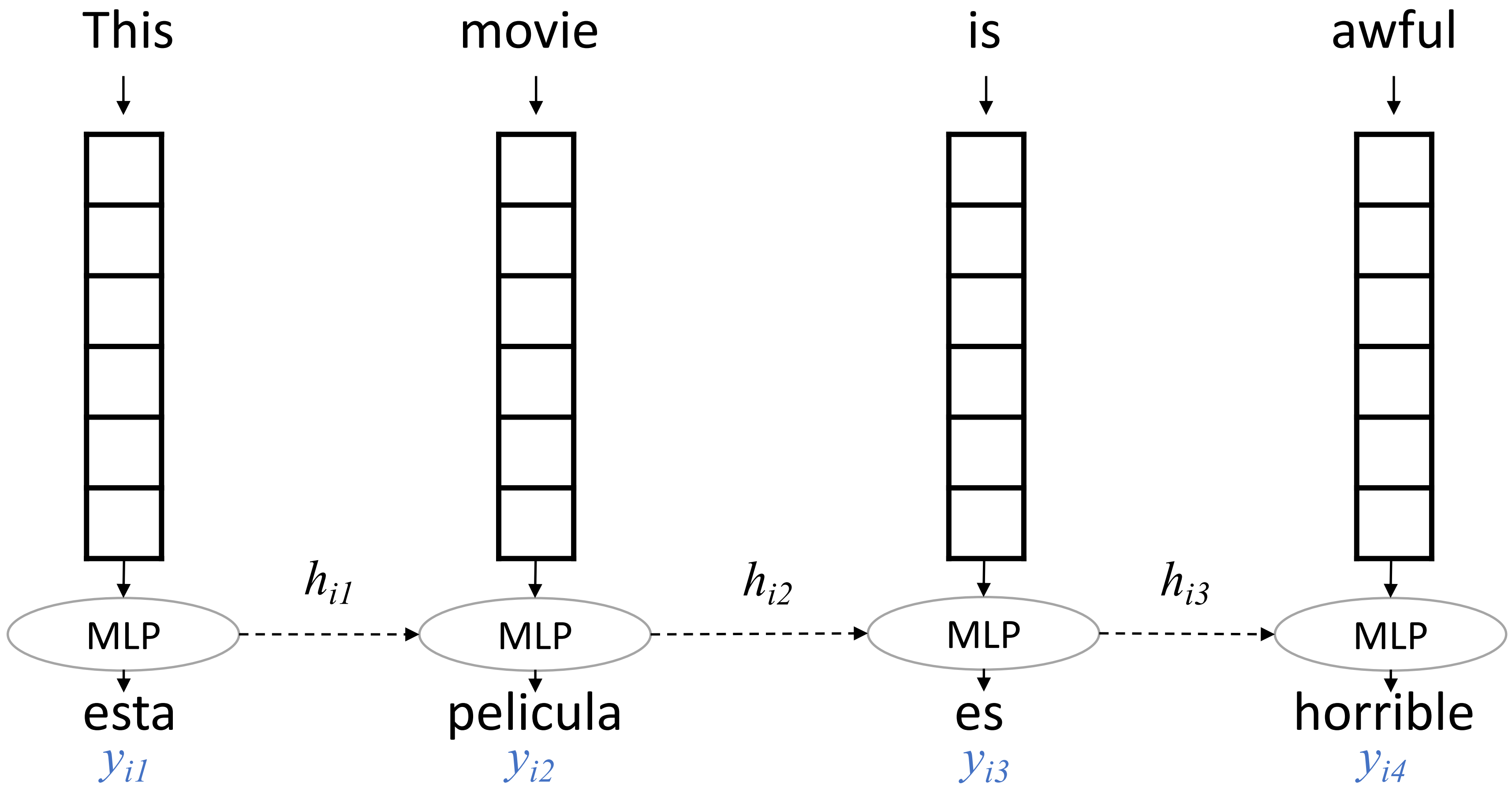
Table 5. Examples of correctly detected PHI instances (in bold) by the ANN

PHI category	ANN
AGE	Father had a stroke at **80** and died of?another stroke at age Personal data and overall health: Now **63**, despite his FH: Father: Died @ **52** from EtOH abuse (unclear exact etiology) Tobacco: smoked from age 7 to **15**, has not smoked since 15.
CONTACT	History of Present Illness **86F** reports worsening b/l leg pain. by phone, Dr. Ivan Guy. Call w/ questions **86383**. Keith Gilbert, H/O paroxysmal afib VNA **171-311-7974** ======= Medications
DATE	During his **May** hospitalization he had dysphagia Social history: divorced, quit smoking in **08**, sober x 10 yrs, She is to see him on the **29th** of this month at 1:00 p.m. He did have a renal biopsy in teh late **60s** adn thus will look for results, Results**02/20/2087** NA 135, K 3.2 (L), CL 96 (L), CO2 30.6, BUN 1 Jose Church, M.D. /ray DD: 01/18/20 DT: **01/19/:0** DV: 01/18/20

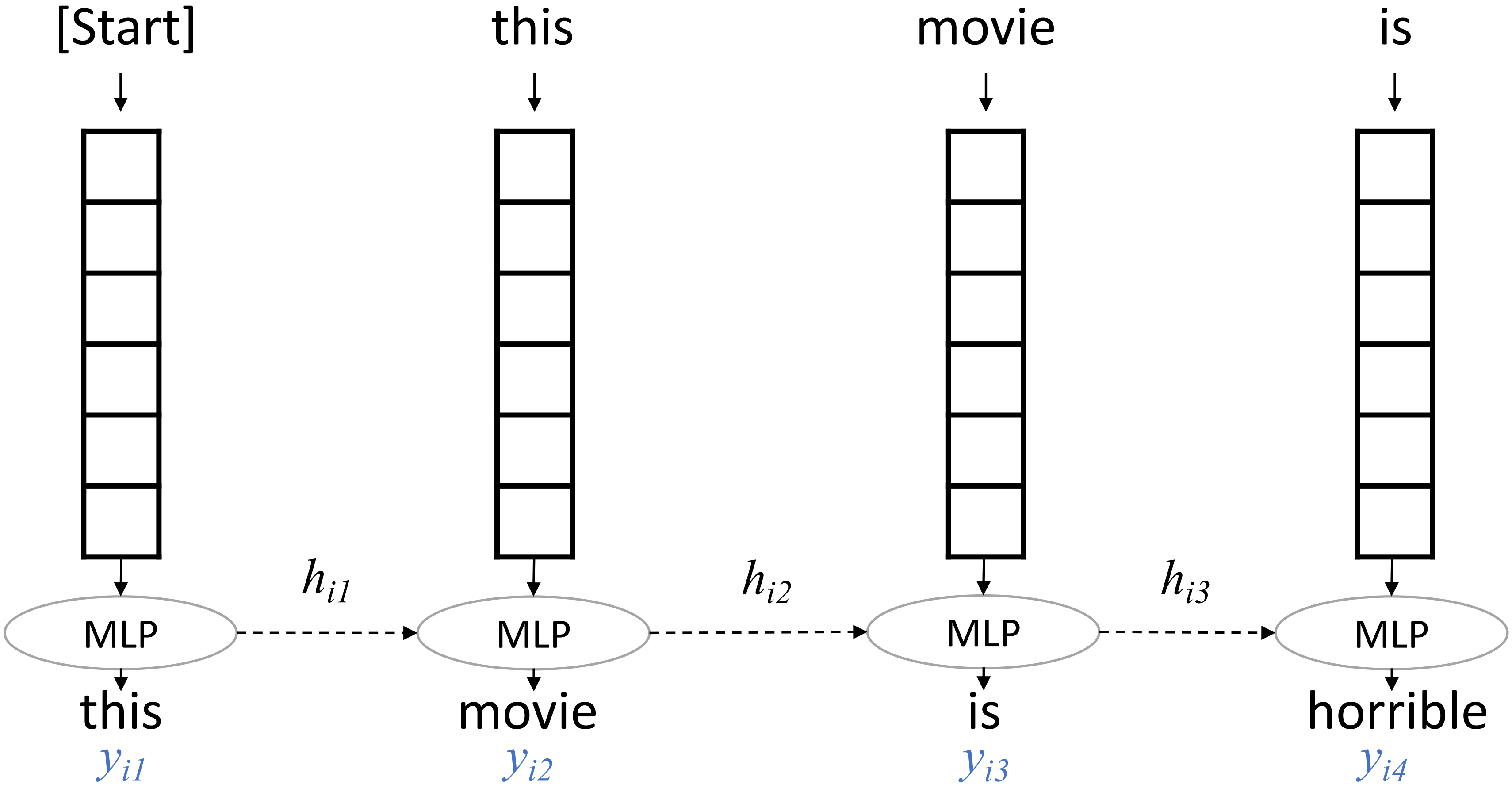
De-identification of patient notes with recurrent neural networks
Dernoncourt F, Lee JY, Uzuner O, Szolovits P
JAMIA 24(3), 2017, 596–606

- A bidirectional RNN is used to identify PHI (18 HIPAA fields)
- *i2b2*: 889 discharge summaries, >28k PHI tokens
- *MIMIC*: 1635 discharge summaries, >60k PHI tokens
- State of the art sensitivity and F1 metric on both datasets

Note: we can also *generate* text this way.



Note: we can also *generate* text this way.



Summary

- For sequential data, a key challenge is how to represent the history of previous measurements or words
- The simplest approach is to choose summary statistics
- Instead, the recurrent neural network *learns* how to summarize earlier information such that prediction performance is maximized
- Very recently, the RNN has been superseded by *transformer* networks, but the principles are largely the same: we use a deep neural network to refine word representations based on context