대한의료인공지능학회
Korean Society of Artificial Intelligence in Medicine

Hands-on Session 3:
Electronic Health Records and Medical Text

RNN & Hands-on
ICU Mortality prediction

KoSAIM 2022

Summer School

22.08.13 (3:30pm-5:30pm)

KAIST AI대학원 문종학

Most slides from KAIST GSAI prof Choi Yoonjae Lectures

Material & Issues

Today’s lecture material and Hands on session code:

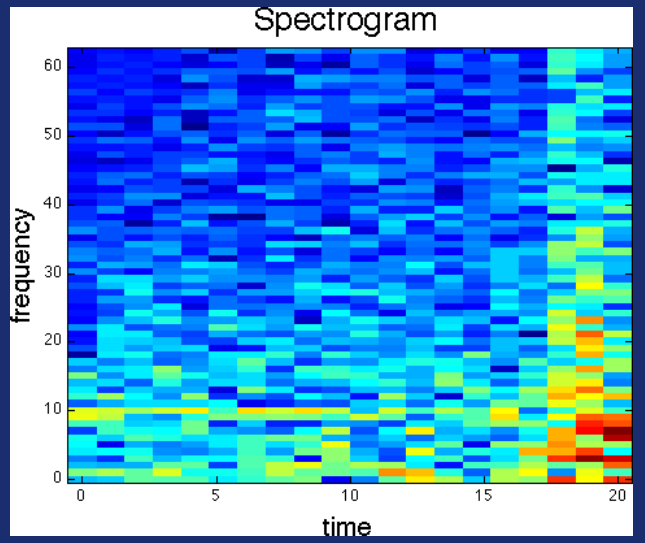
- Github. https://github.com/SuperSupermoon/KoSAIM2022

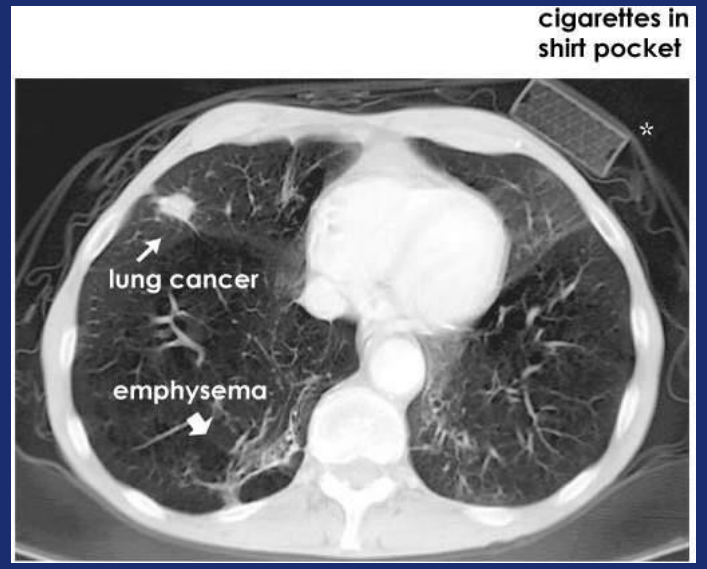
AGENDA

- 세션 소개
- Electronic Health Recoreds (EHR)
- Recurrent Neural Network (RNN)
- Time Series EHR & RNN
- Practice (Hands on session)

EHR

- EHR consists of
  - Structured codes
  - Lab measures
  - Spectrograms (e.g. EEG)
  - Images (CT, MRI, X-ray)
  - Free text
  - Demographics
  - Billing Information
  - Genetic Information





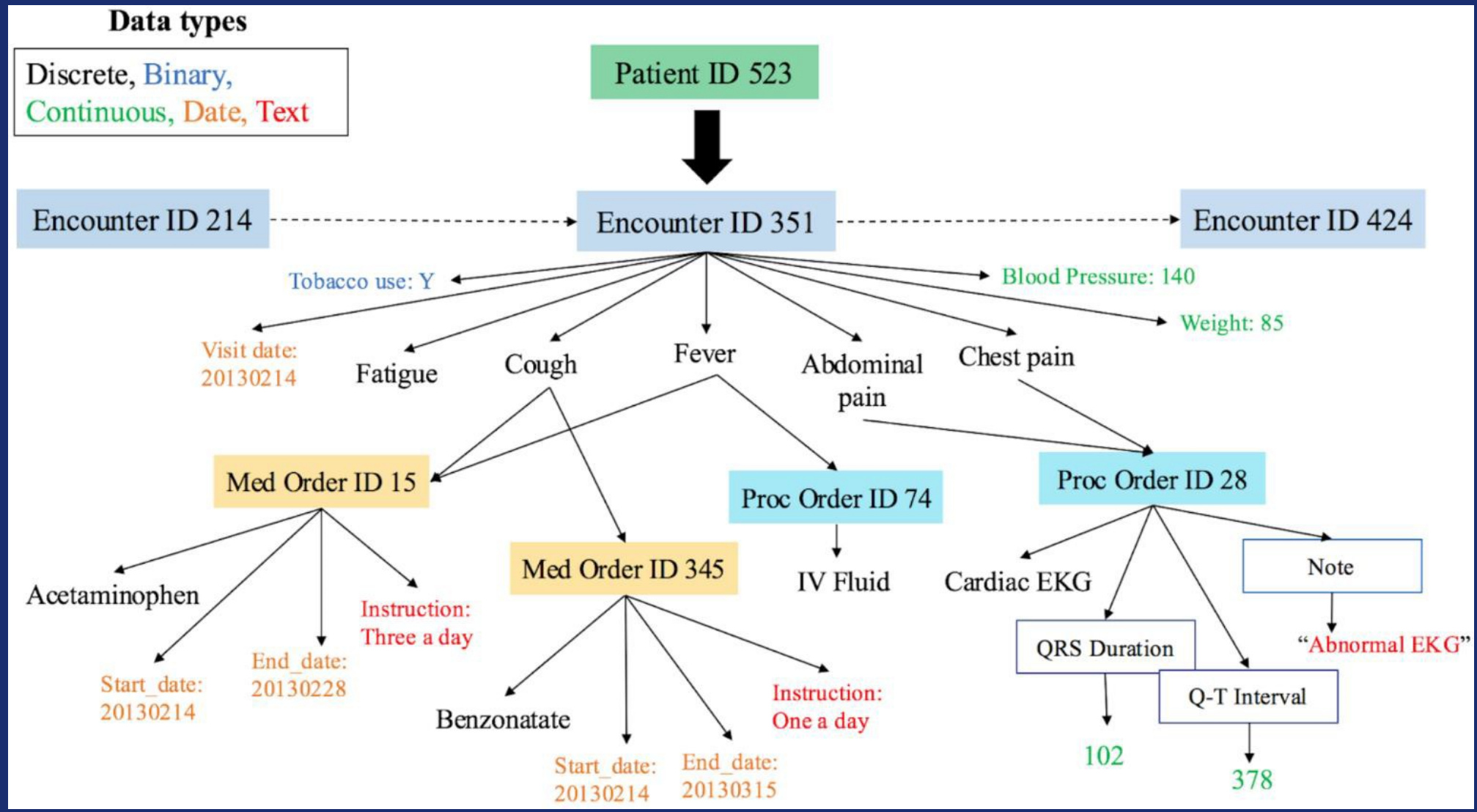
Nurses' Notes

NANDA Dx	Date & Time	Documentation
2 3 1	3/21/11 0800	***(2)*** Mrs. GH alert, awake, and oriented to person and situation but is **confused as to time and place.** She is able to state her name and that she is in the hospital but states that it is afternoon and that she is in the long-term care facility. Was reoriented to time and place. ***(3)*** Skin, warm, dry, pale but without pallor or cyanosis. Bilateral arms have purpura but skin remains intact and without skin tears. No noted decubitus ulcers on coccyx, hips, or heels. Respirations regular and non-labored. ***(1)*** Lung sounds clear except for **crackles noted in left lower lobe** but improved when compared to earlier assessment. Encouraged to cough and deep breathe; crackles lessened after use of incentive spirometer, coughing, and deep breathing. Pulse ox on right index finger showing saturation of 96% on 2 liters O_2 by nasal cannula. Ears and nares checked and are clear of irritation from cannula. Heart rate regular. S_1 and S_2 apical heart sounds clearly heard. Peripheral pulses are +2 at radius and +1 at dorsalis pedis pulses. Equal hand grips; left pedal push is weaker but unchanged since admission. Per graphic flow sheet, voided clear amber urine at 0715. C/O abdominal pain of 7 on 0–10 pain scale. **Abdomen firm, distended, and tender to slight touch. Bowel sounds hyperactive in RUQ and absent in remaining quadrants. States she does not know when she last had a bowel movement. No indication of BM on graphic flow sheet since admission. Refuses breakfast stating that she is nauseous. VS 148/92, 100.6 F, 114, 24. Charge RN notified of nausea, abdominal pain, and distention.** ----------F. Darwin, LVN

Note to student: For instructional purposes, the abnormal findings are put in bold here. Notice that any abnormal finding must include nursing intervention(s) that indicate what nursing actions address that problem/concern.

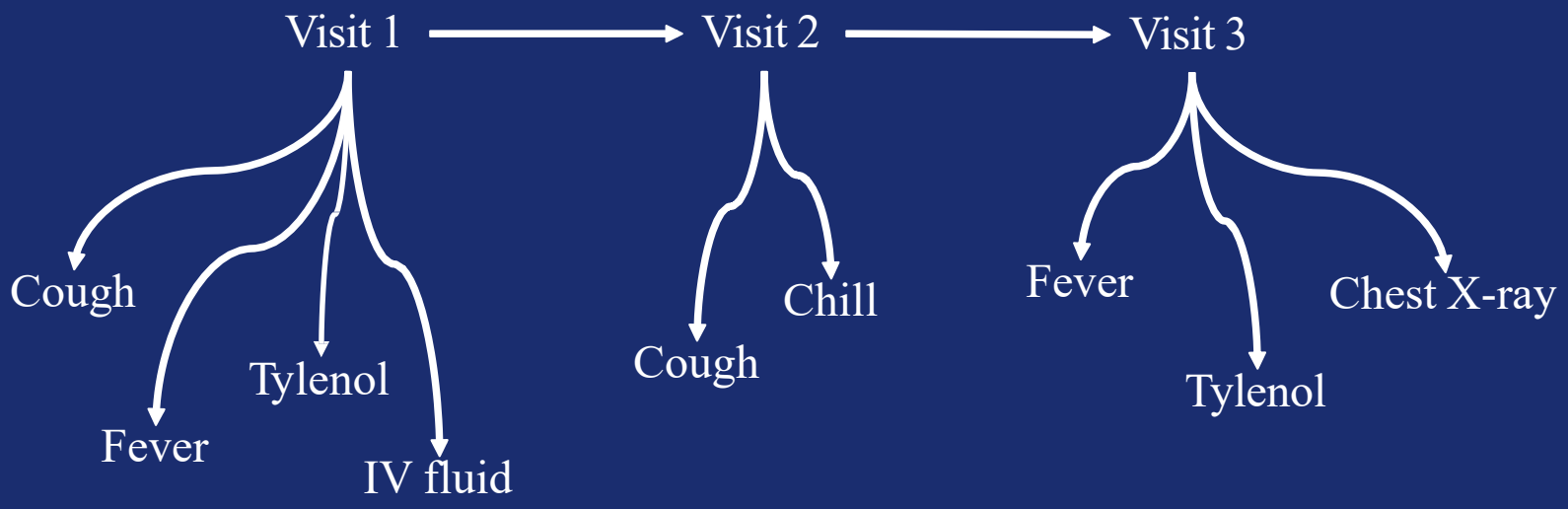
NANDA Diagnoses are listed in the NANDA column and, for our example, within the nurses' notes.

EHR - Structure of Electronic Health Records



EHR

– Structured codes over time



EHR – MIMIC-III

MIMIC –III (Medical Information Mart for Intensive Care)

- Records from Intensive Care Unit (ICU)
- Over 40,000 patients
- Between 2001-2012
- Contains
  - Vital signs
  - Lab test results
  - Medications
  - Caregiver notes
  - Imaging reports (not images themselves)
  - Mortality

Deidentification

- EHR deidentification: procedures to make it impossible to identify an individual given some data
  - Remove all names, location, very rare cases (over 90yo, rare disease, etc)
  - Shift dates
    - Ex: Date of birth 1964.06.12 → 2156.08.24
    - Preserve internal consistency
    - Preserve time of day, day of week, seasonality

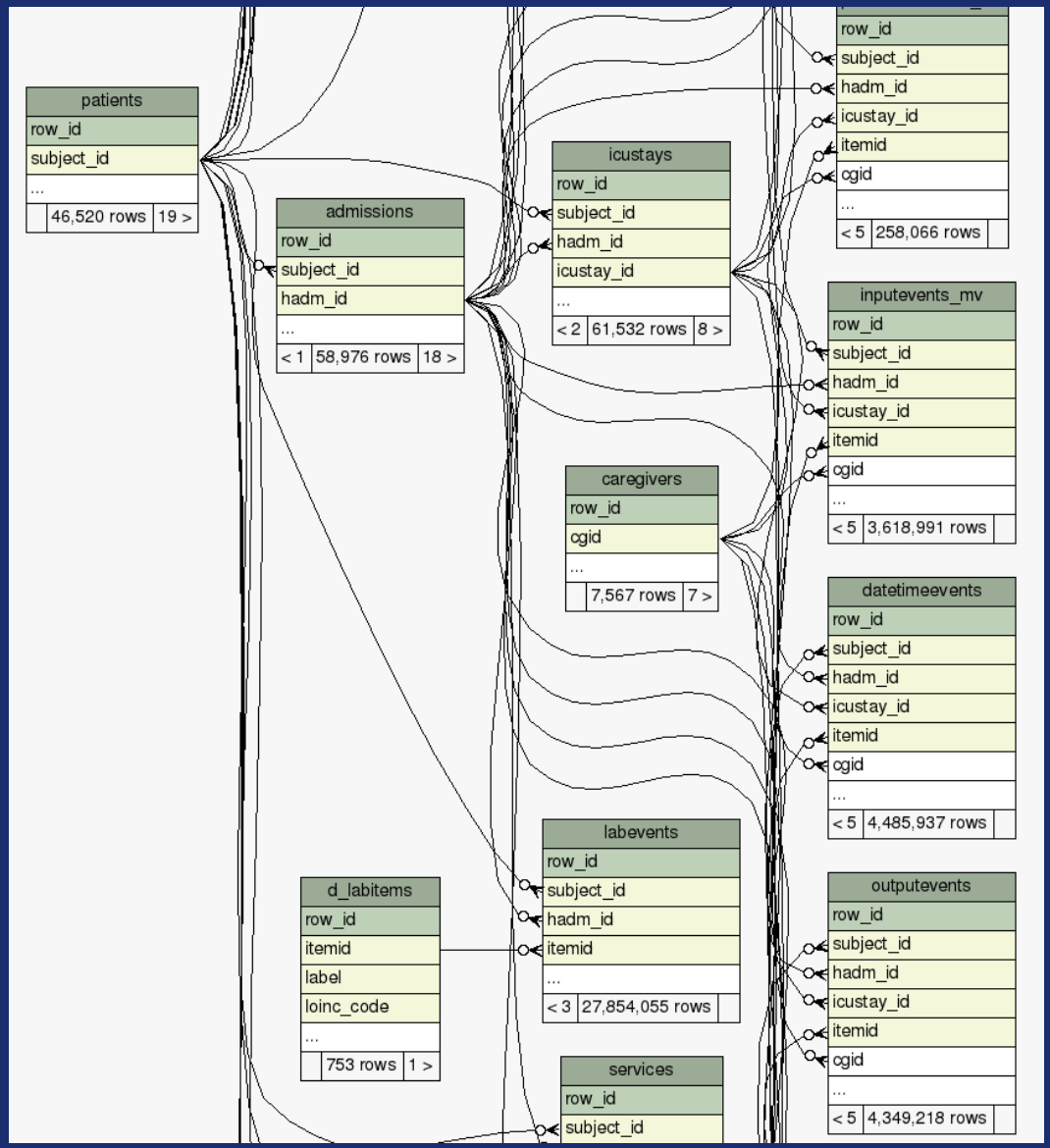
Accessing MIMIC-III

- https://mimic.physionet.org/gettingstarted/access/
- Need to go through CITI training
  - https://www.citiprogram.org/index.cfm?pageID=154&icat=0&ac=0
  - Takes several hours, if done rigorously.

Tables of MIMIC-III

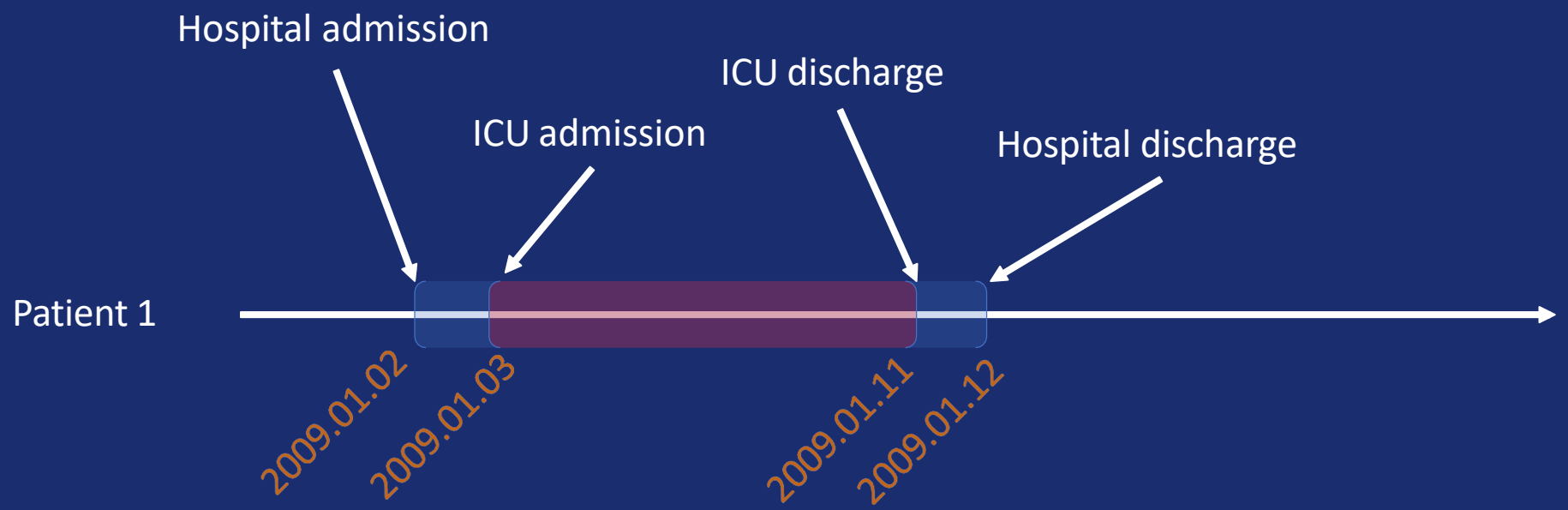
- ADMISSIONS
- CALLOUT
- CPTEVENTS
- DIAGNOSES_ICD
- DRGCODES
- ICUSTAYS
- LABEVENTS
- MICROBIOLOGYEVENTS
- PATIENTS
- PRESCRIPTIONS
- PROCEDURES_ICD
- SERVICES
- TRANSFERS

- CHARTEVENTS
- DATETIMEEVENTS
- INPUTEVENTS_CV
- INPUTEVENTS_MV
- NOTEEVENTS
- OUTPUTEVENTS
- PROCEDUREEVENTS_MV
- …
- …

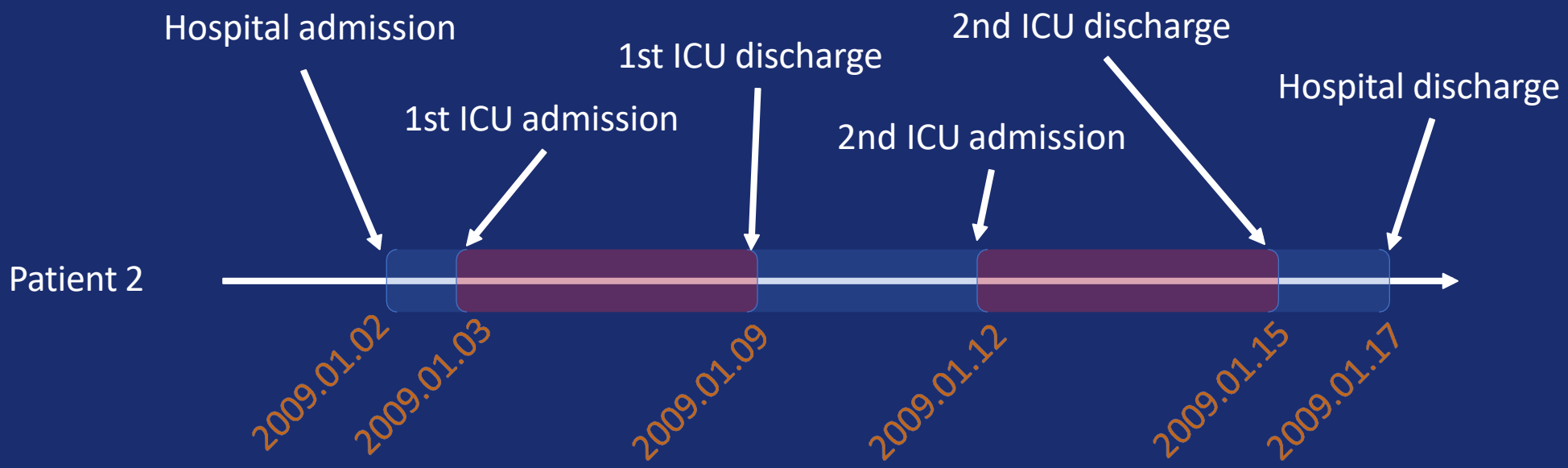


- https://mit-lcp.github.io/mimic-schema-spy/
  - Easy to identify primary keys, and links between tables

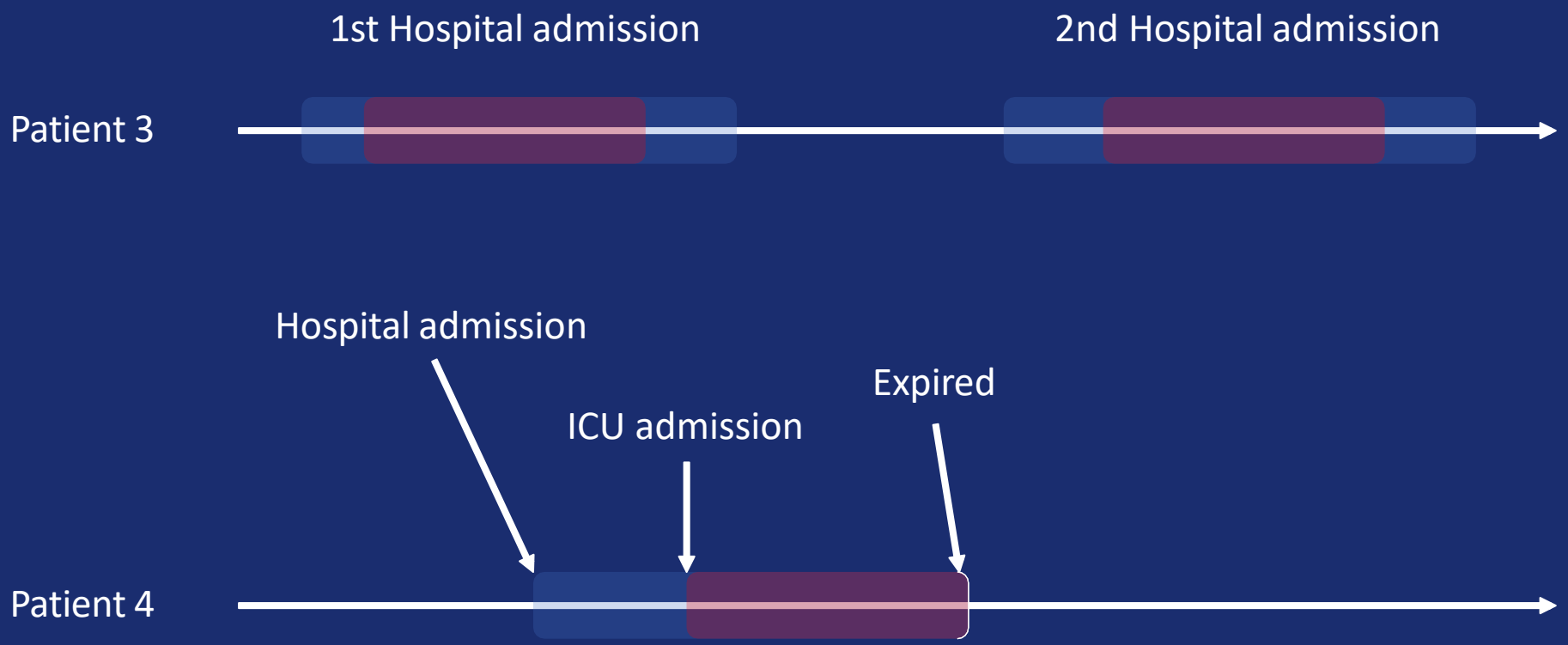
EHR – Addmision patterns



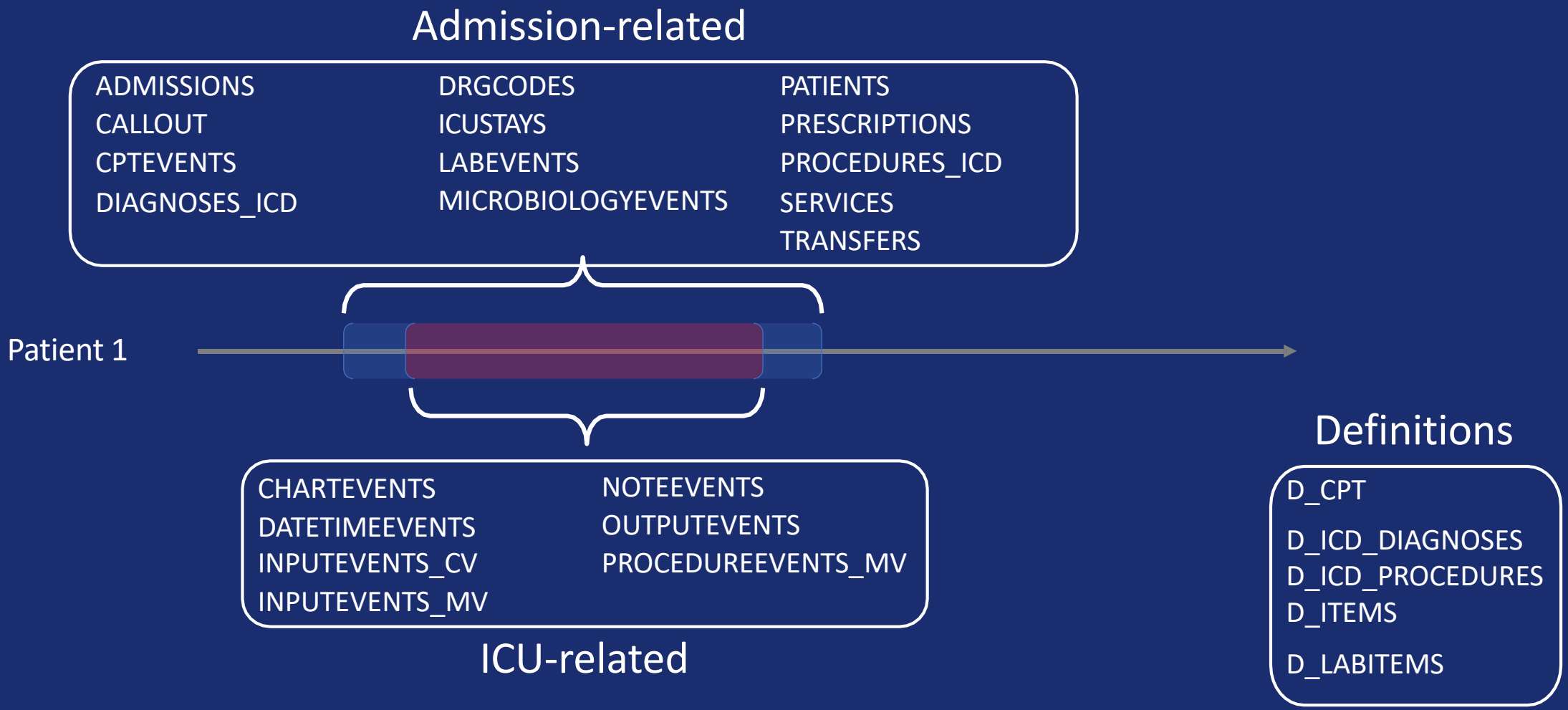
EHR – Addmision patterns



EHR – Addmision patterns



EHR – Table soruces



AGENDA

- 세션 소개
- Electronic Health Recoreds (EHR)
- Recurrent Neural Network (RNN)
- Time Series EHR & RNN
- Practice (Hands on session)

Bag-of-Words

- Classical way to handle variable length sentences/documents
- I gave the ball to John, who gave it to Mary
  - I:1, gave:2, the:1, ball:1, to:2, John:1, who:1, it:1, Mary:1

RNN - Bag-of-Words

- I gave the ball to John, who gave it to Mary
  - I:1, gave:2, the:1, ball:1, to:2, John:1, who:1, it:1, Mary:1

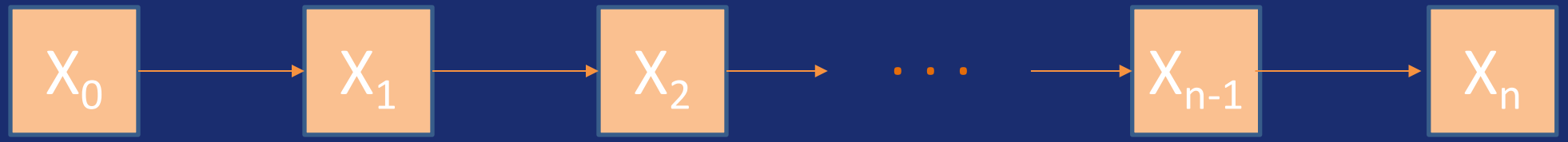
All vocab	Word count
a	0
I	0
ab	0
…	…
…	…
ball	2
…	…
gave	2
…	…

Bag-of-Words

- I gave the ball to John, who gave it to Mary
  - I:1, gave:2, the:1, ball:1, to:2, John:1, who:1, it:1, Mary:1
- I gave the ball to Mary, who gave it to John
  - I:1, gave:2, the:1, ball:1, to:2, John:1, who:1, it:1, Mary:1
- Different meaning, same representation!

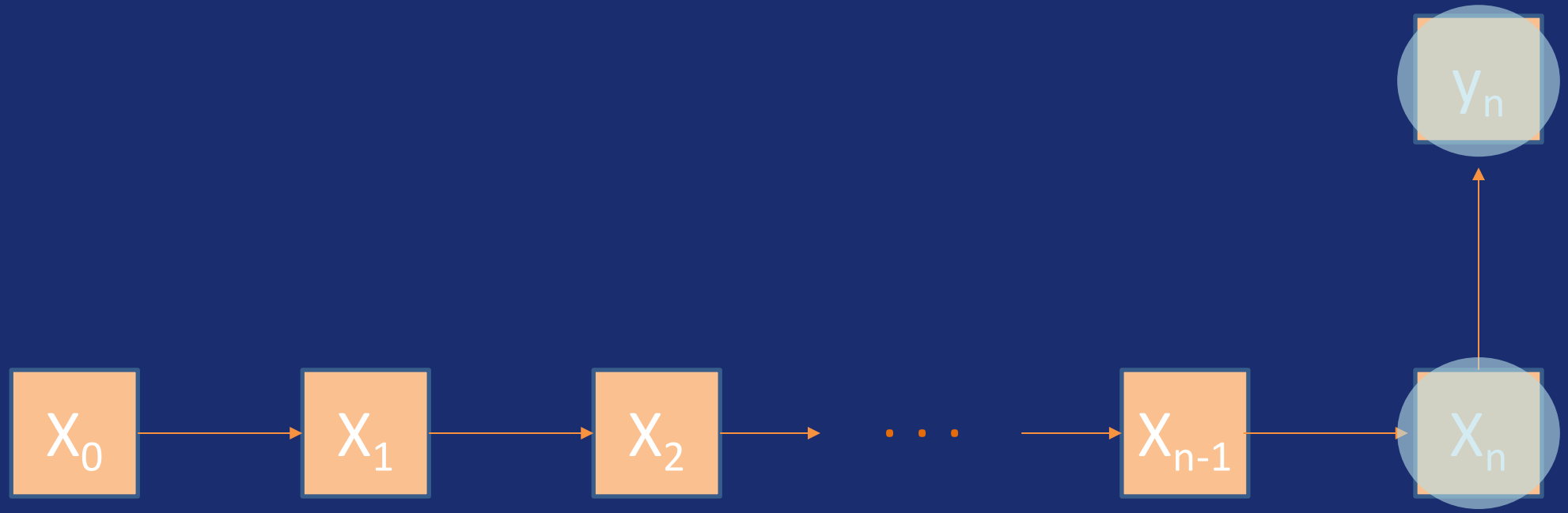
RNN

RNN



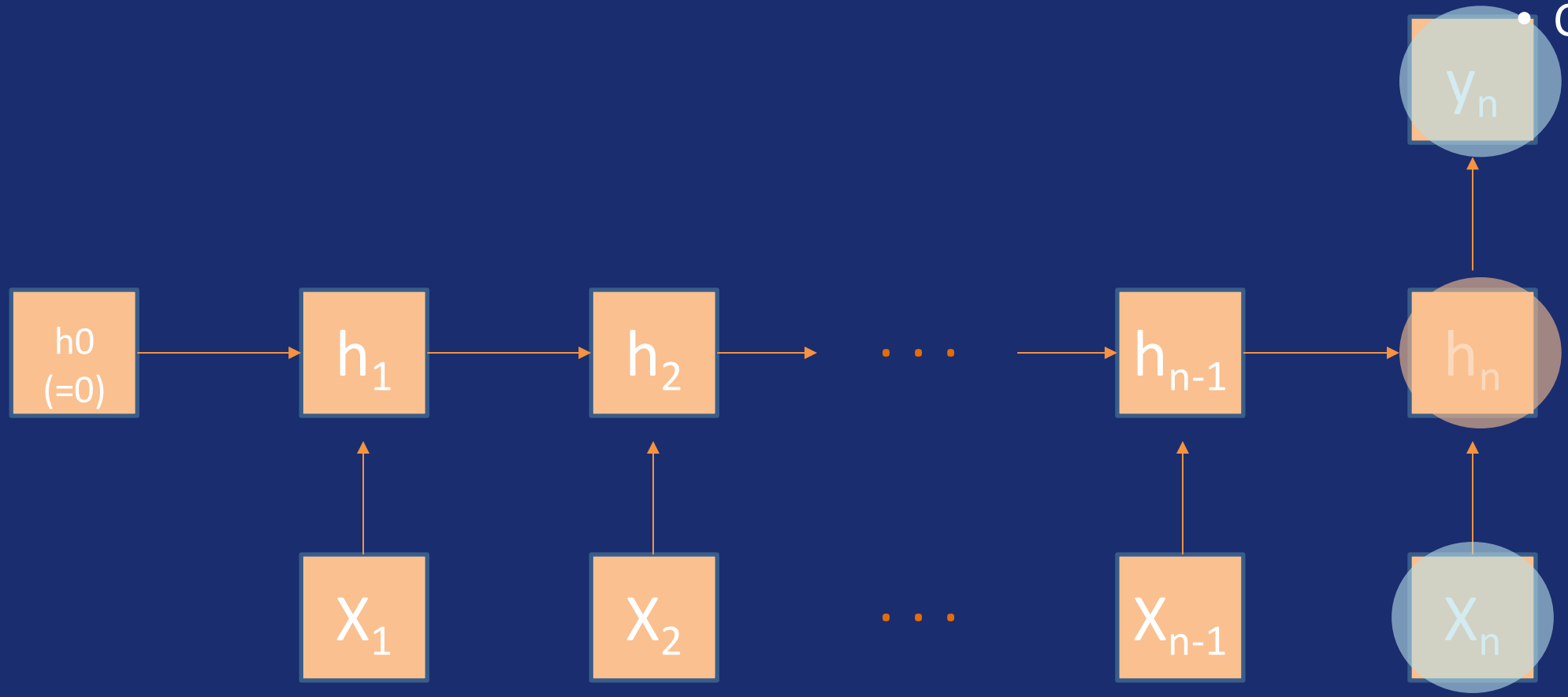
RNN

- Classification

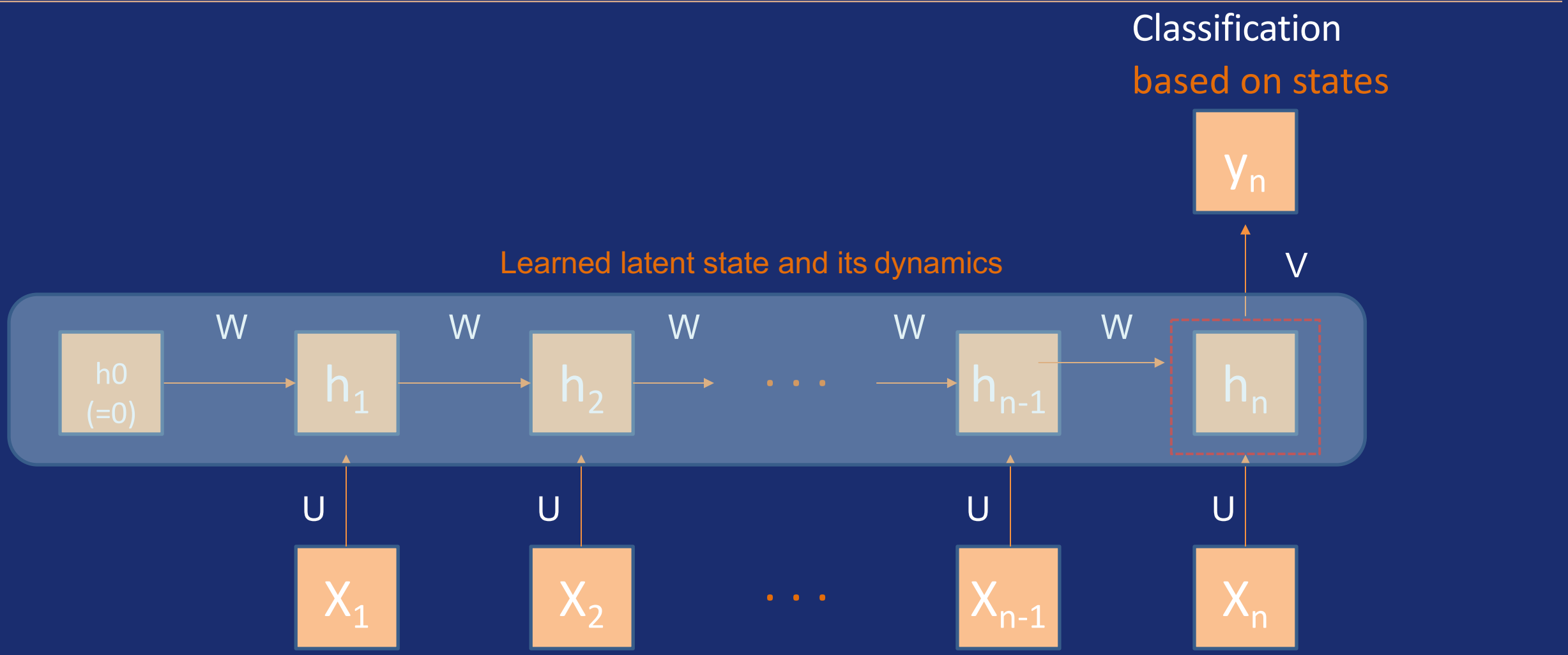


RNN

- Classification

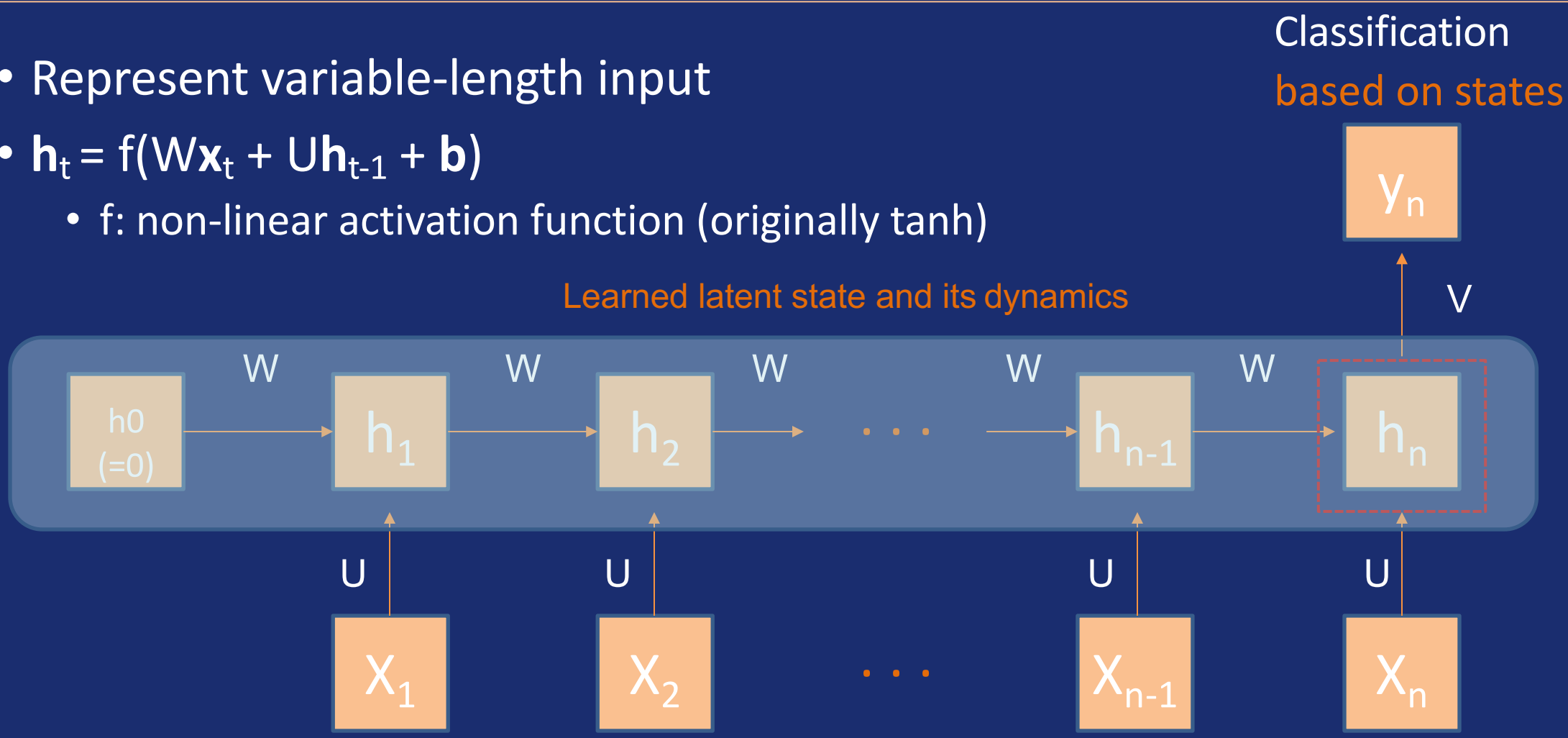


RNN


Classification
based on states
y_n
V
Learned latent state and its dynamics
W
W
W
W
W
h0
(=0)
h_1
h_2
h_{n-1}
h_n
U
U
U
U
X_1
X_2
X_{n-1}
X_n

RNN

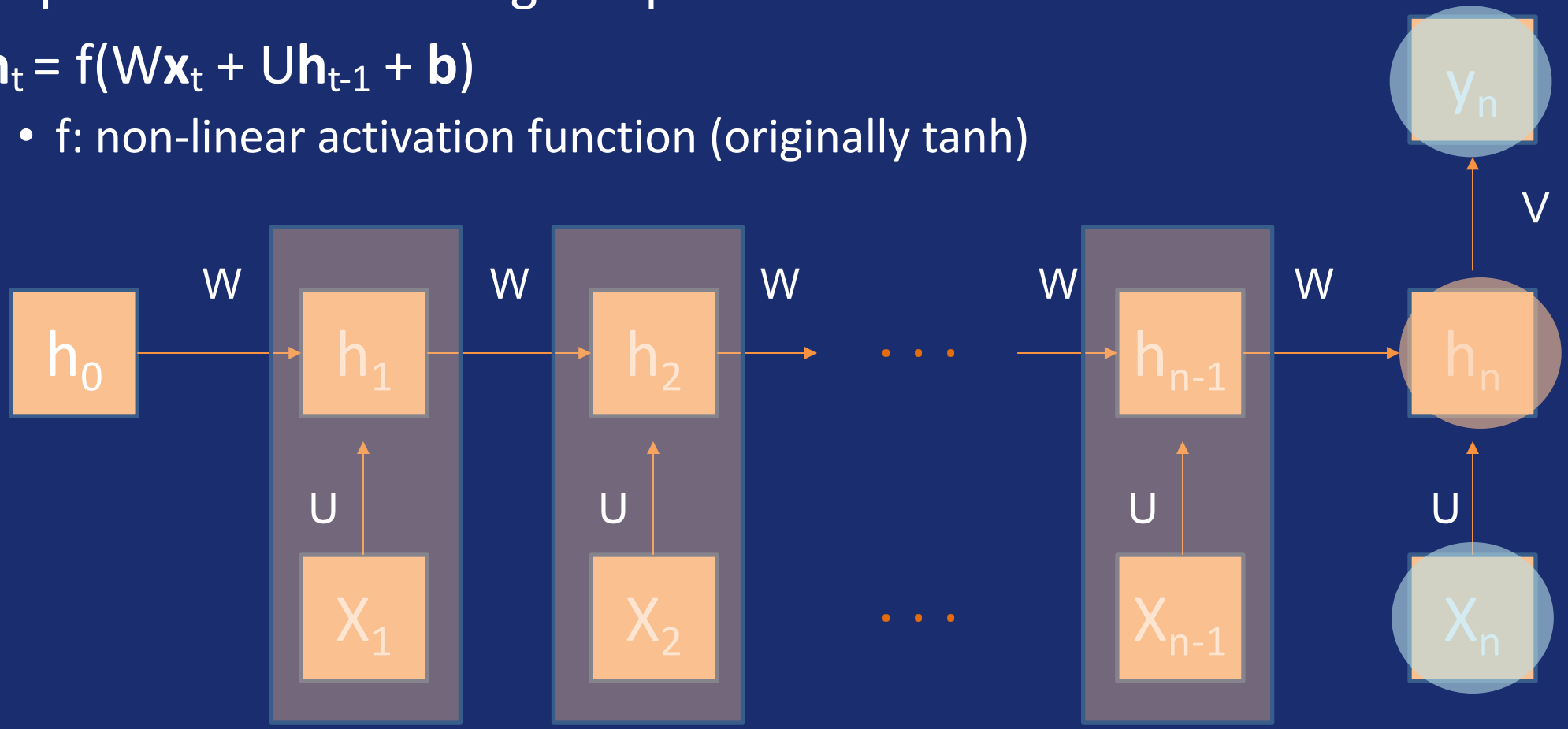
- Represent variable-length input
- $\mathbf{h}_t = f(W\mathbf{x}_t + U\mathbf{h}_{t-1} + \mathbf{b})$
  - f: non-linear activation function (originally tanh)



RNN

- Represent variable-length input
- $\mathbf{h}_t = f(W\mathbf{x}_t + U\mathbf{h}_{t-1} + \mathbf{b})$
  - f: non-linear activation function (originally tanh)

- Classification

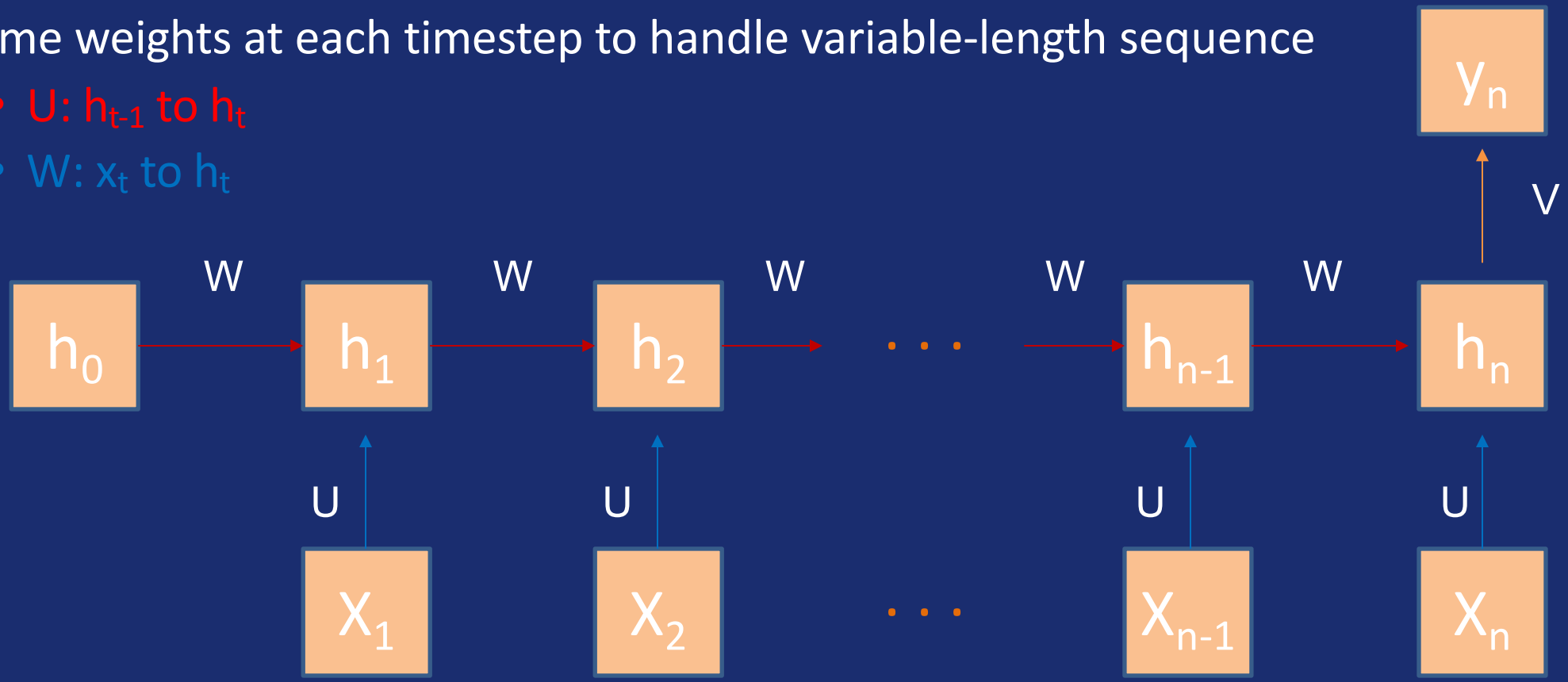


RNN

$\mathbf{h}_t = f(W\mathbf{x}_t + U\mathbf{h}_{t-1} + \mathbf{b})$

- Same weights at each timestep to handle variable-length sequence
  - U: h_{t-1} to h_t
  - W: x_t to h_t

- Classification



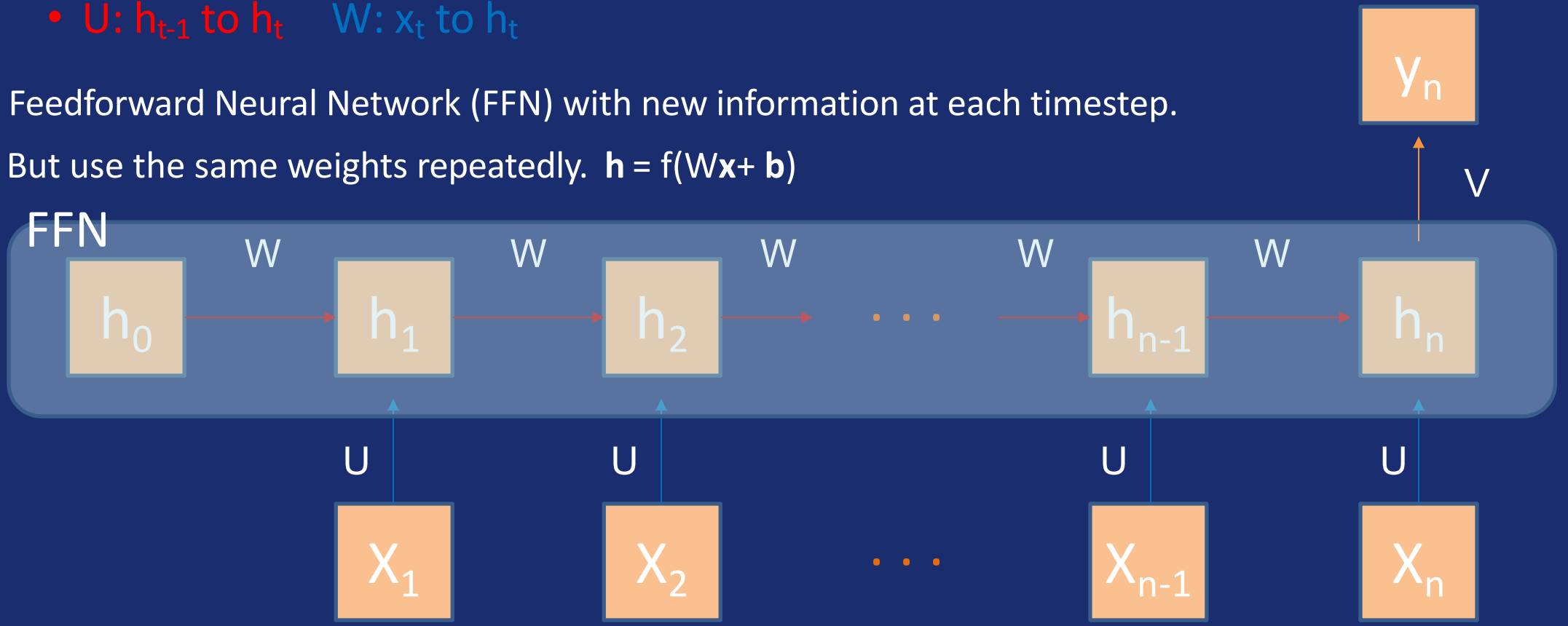
RNN

$\mathbf{h}_t = f(W\mathbf{x}_t + U\mathbf{h}_{t-1} + \mathbf{b})$

- U: h_{t-1} to h_t W: x_t to h_t

- Feedforward Neural Network (FFN) with new information at each timestep.
- But use the same weights repeatedly. $\mathbf{h} = f(W\mathbf{x} + \mathbf{b})$

- Classification



Application

- Sequence-level classification/regression
  - Sentiment classification
  - Topic classification
- Classification/regression at each step.
  - Language modeling
  - Part-of-speech tagging
- Sequence-to-sequence
  - Translation
  - Question answering

RNN - Sequence prediction with RNN

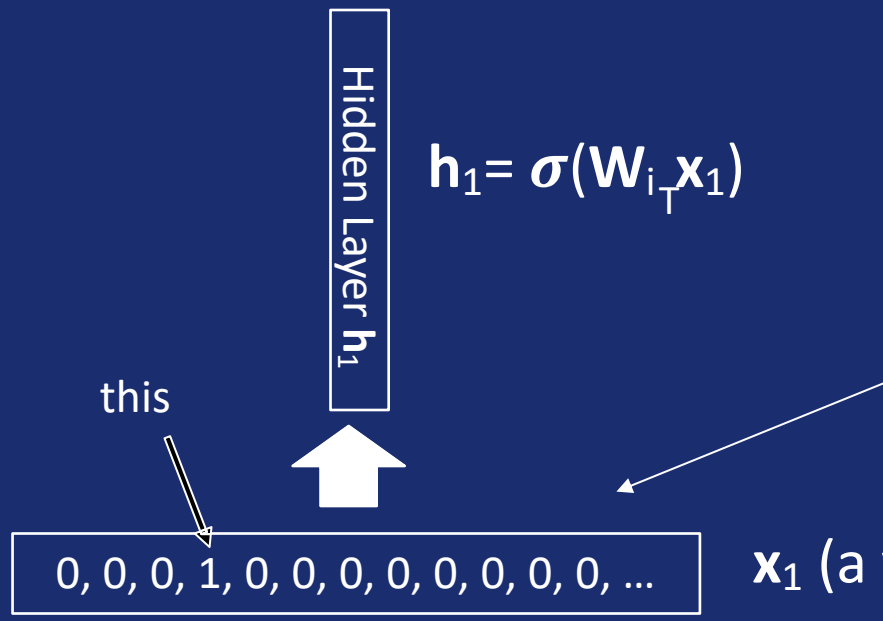
- Sentiment classification: Positive or Negative?
  - “This movie is as impressive as a preschool Christmas play”

→ 1 2 3 4 5 6 7 8 9 10

RNN - Sequence prediction with RNN

- Sentiment classification: Positive or Negative?
  - "This movie is as impressive as a preschool Christmas play"

→ 1 2 3 4 5 6 7 8 9 10



Vocabulary

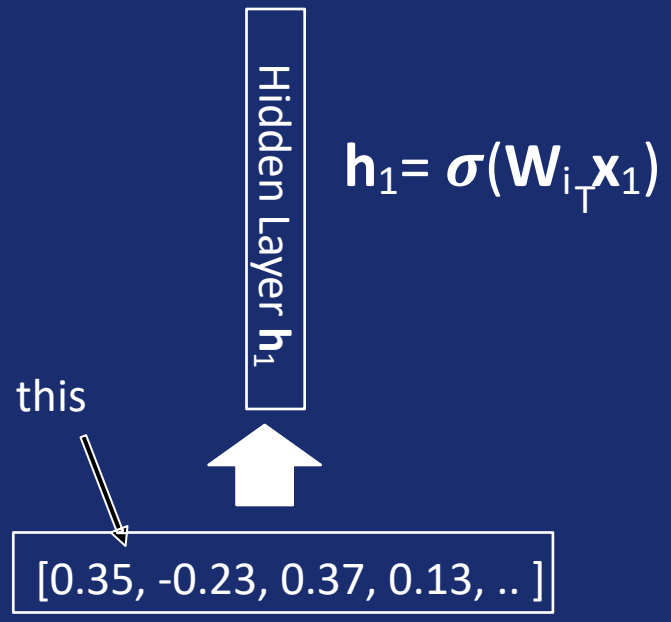
a	1
I	2
Christmas	3
This	4
…	…
impressive	1000
…	…
preschool	10000
…	…

$\mathbf{x}_1$ (a vector with 1M elements. Only "this" is 1.)

RNN - Sequence prediction with RNN

- Sentiment classification: Positive or Negative?
  - “This movie is as impressive as a preschool Christmas play”

→ 1 2 3 4 5 6 7 8 9 10



Vocabulary

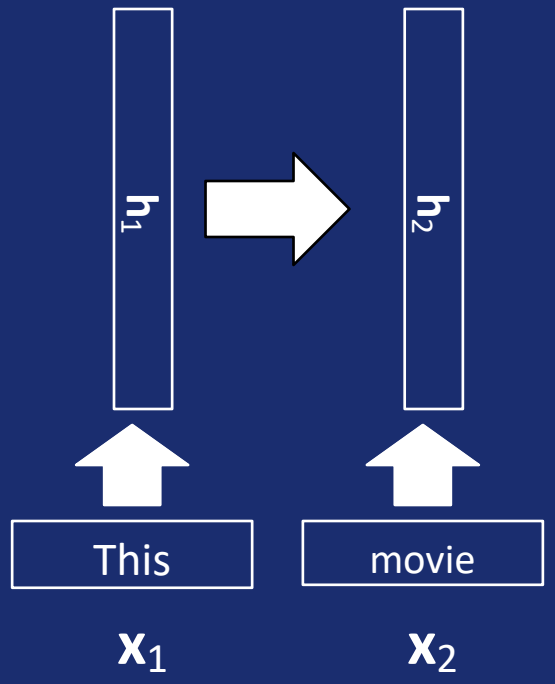
a	1
I	2
Christmas	3
This	4
…	…
impressive	1000
…	…
preschool	10000
…	…

Look up table
embedding vector

[1.38, -3.23, 0.6, 0.8, ..]
[0.52, -3.23, 4.1, 0.59, ..]
…
[0.35, -0.23, 0.37, 0.13, ..]
…
…
…
…
…

RNN - Sequence prediction with RNN

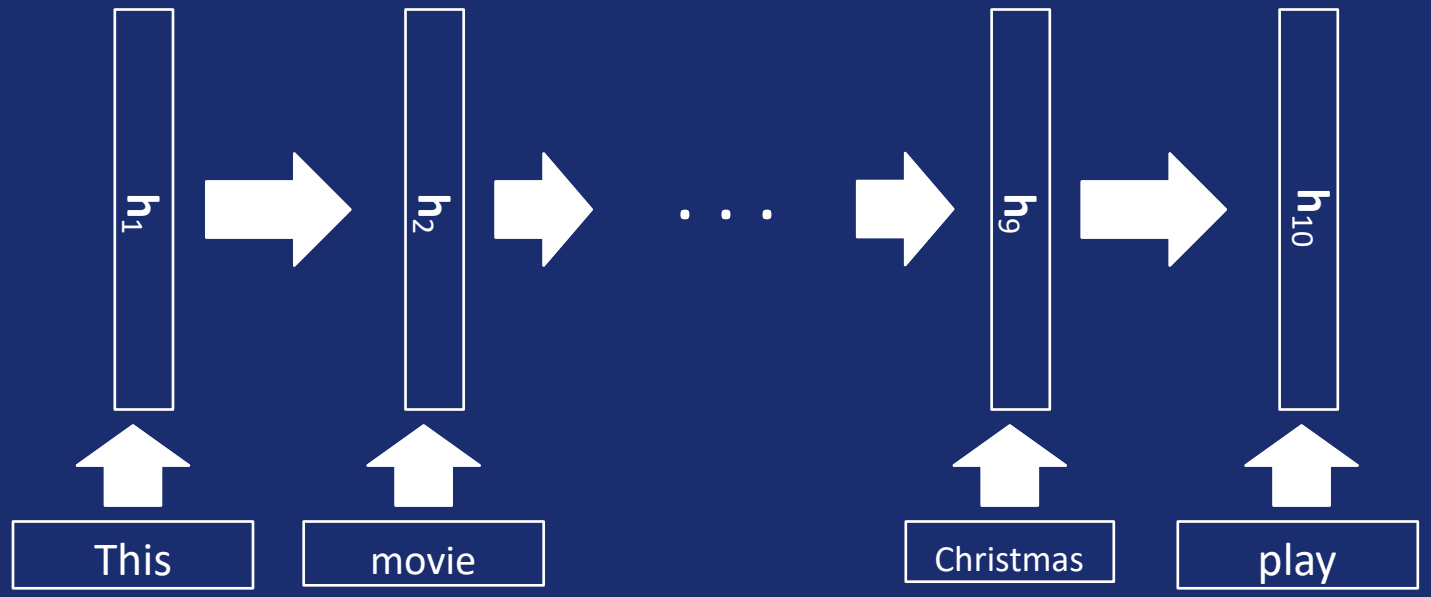
- Sentiment classification: Positive or Negative?
  - "This movie is as impressive as a preschool Christmas play"



$$h_2 = f(W_{hT} h_1 + W_{iT} x_2)$$

RNN

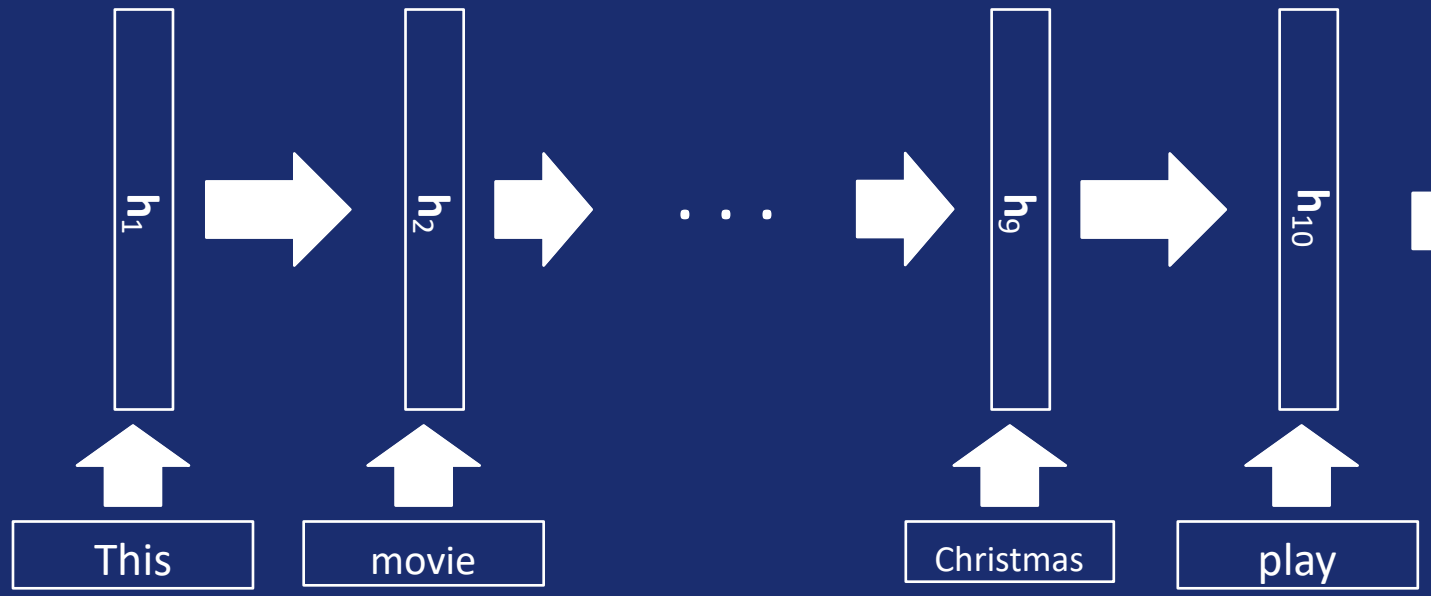
- Sentiment classification: Positive or Negative?
  - “This movie is as impressive as a preschool Christmas play”



$$h_{10} = \sigma(W_{h_T} h_9 + W_{i_T} x_{10})$$

RNN

- Sentiment classification: Positive or Negative?
  - “This movie is as impressive as a preschool Christmas play”



$$h_{10} = \sigma(W_{h_T} h_9 + W_{i_T} x_{10})$$

Output

$$y = \sigma(w_o^T h_{10})$$ O

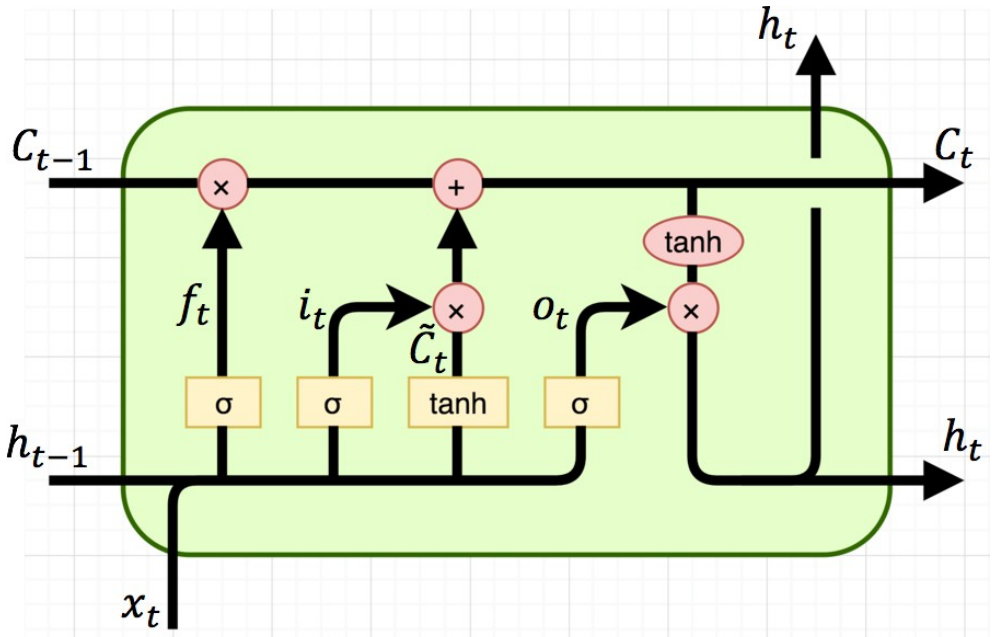
utcome 0.0 ~ 1.0

RNN

- Limitations:
  - Vanishing Gradient
    - Shared W, U으로 인해 점점 기울기 소실.
  - Long-term dependency (장기 의존성 문제)
    - Input sequence length가 길어짐에 따라 초기 time step input의 영향력이 점점 감소.

Long Short-Term memory (LSTM)

- Consists of a memory cell and a set of gating units
  - Memory cell is the context that carries over
  - Forget gate controls erase operation
  - Input gate controls write operation
  - Output gate controls the read operation



$$i_t = \sigma\left(x_t U^i + h_{t-1} W^i\right)$$
$$f_t = \sigma\left(x_t U^f + h_{t-1} W^f\right)$$
$$o_t = \sigma\left(x_t U^o + h_{t-1} W^o\right)$$
$$\tilde{C}_t = \tanh\left(x_t U^g + h_{t-1} W^g\right)$$
$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right)$$
$$h_t = \tanh(C_t) * o_t$$

AGENDA

- 세션 소개
- Electronic Health Recoreds (EHR)
- Recurrent Neural Network (RNN)
- Time Series EHR & RNN
- Practice (Hands on session)

Time Series EHR & RNN

Patient X

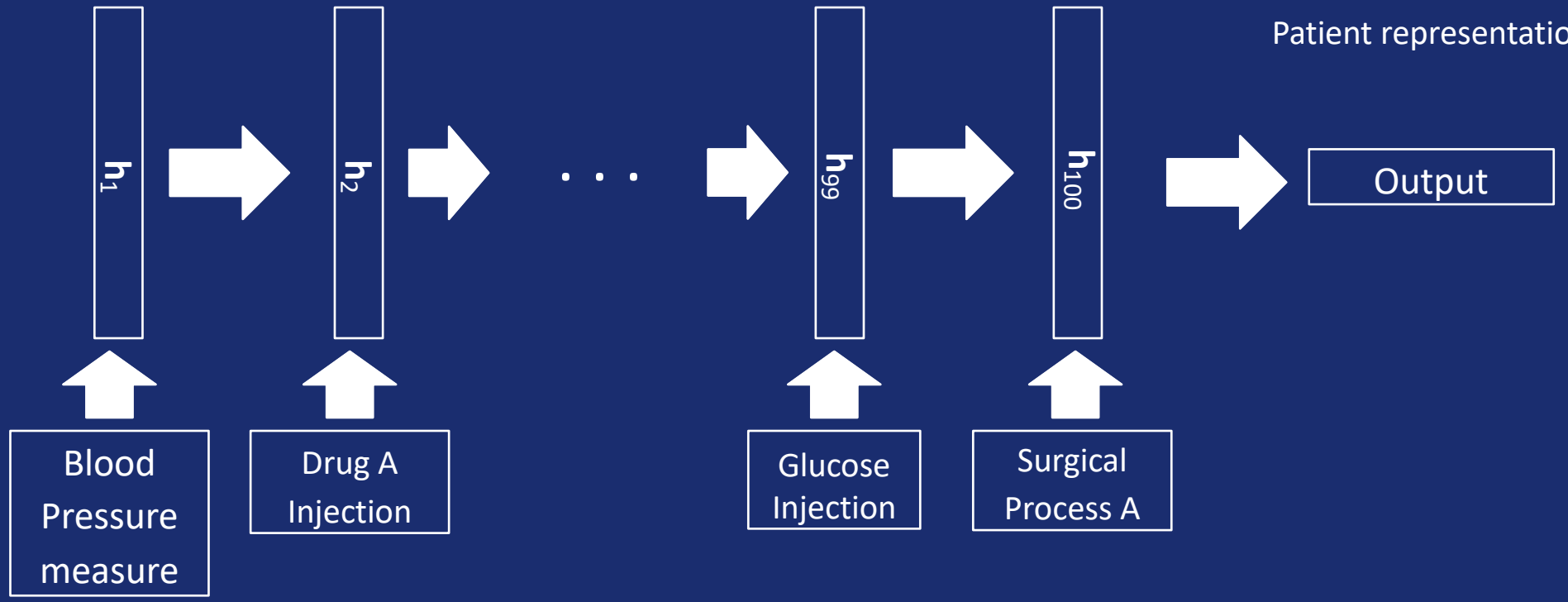
Time

[Lab 4, Med2 ,Inf3, Lab2, Lab5, Inf1, Dx3, …]

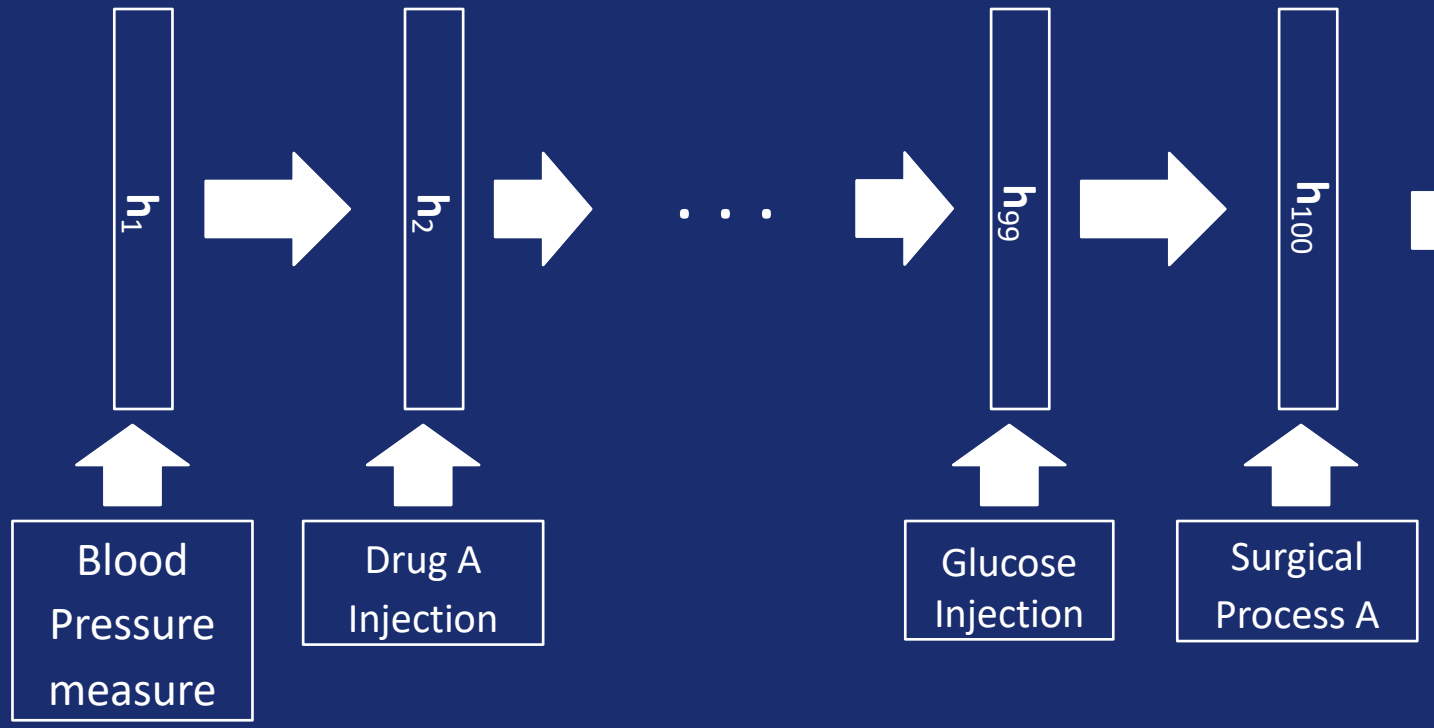
[L1003, 7712 , 531, 9871, L2123, 1033, D454, …]

Codes are recognized to just different code numbers

Time Series EHR & RNN



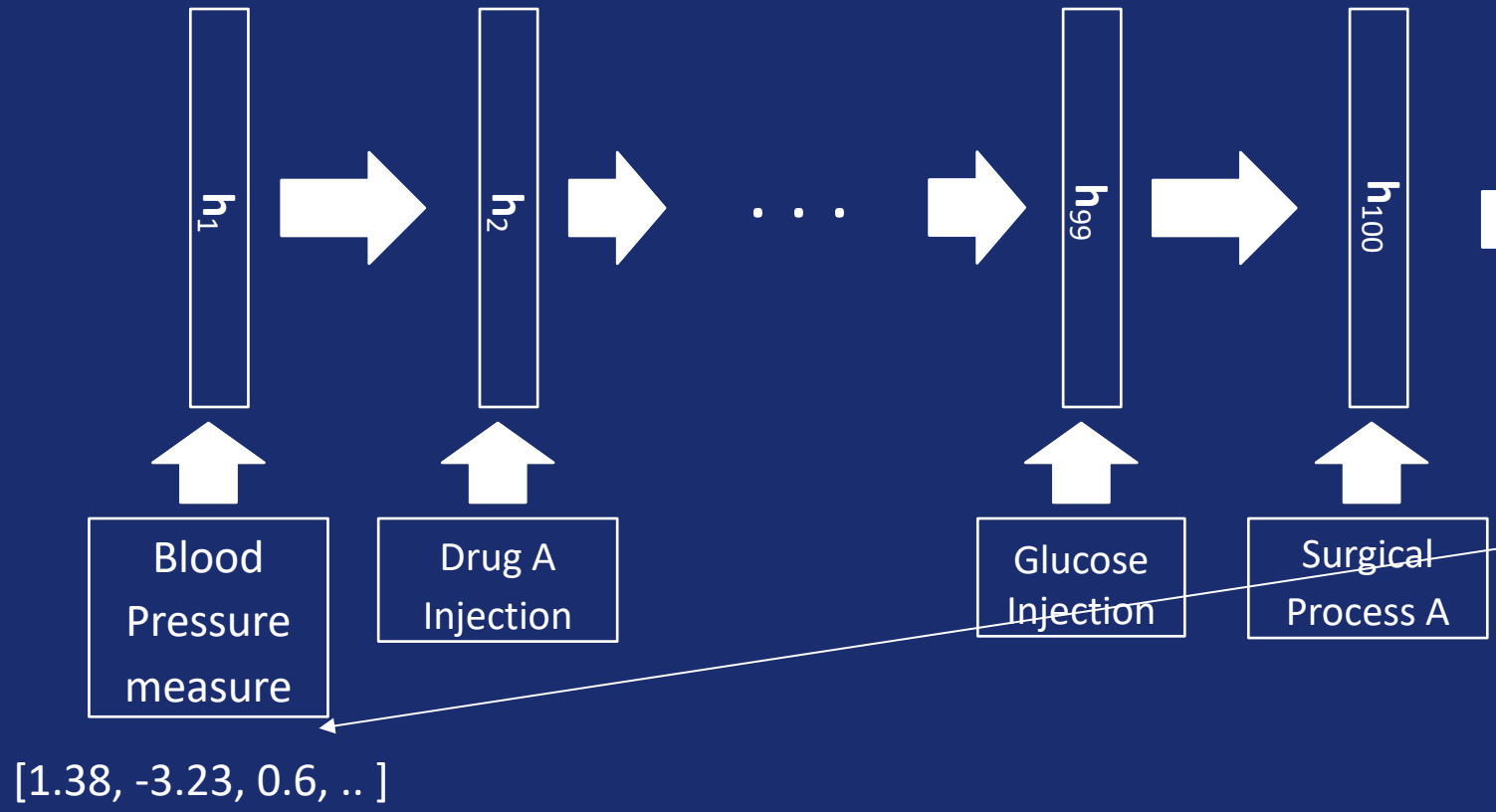
Time Series EHR & RNN



Vocabulary

Blood pressure	1
Drug A inject	2
Glucose inject	3
Surgical process A	4
…	…
…	…
…	…

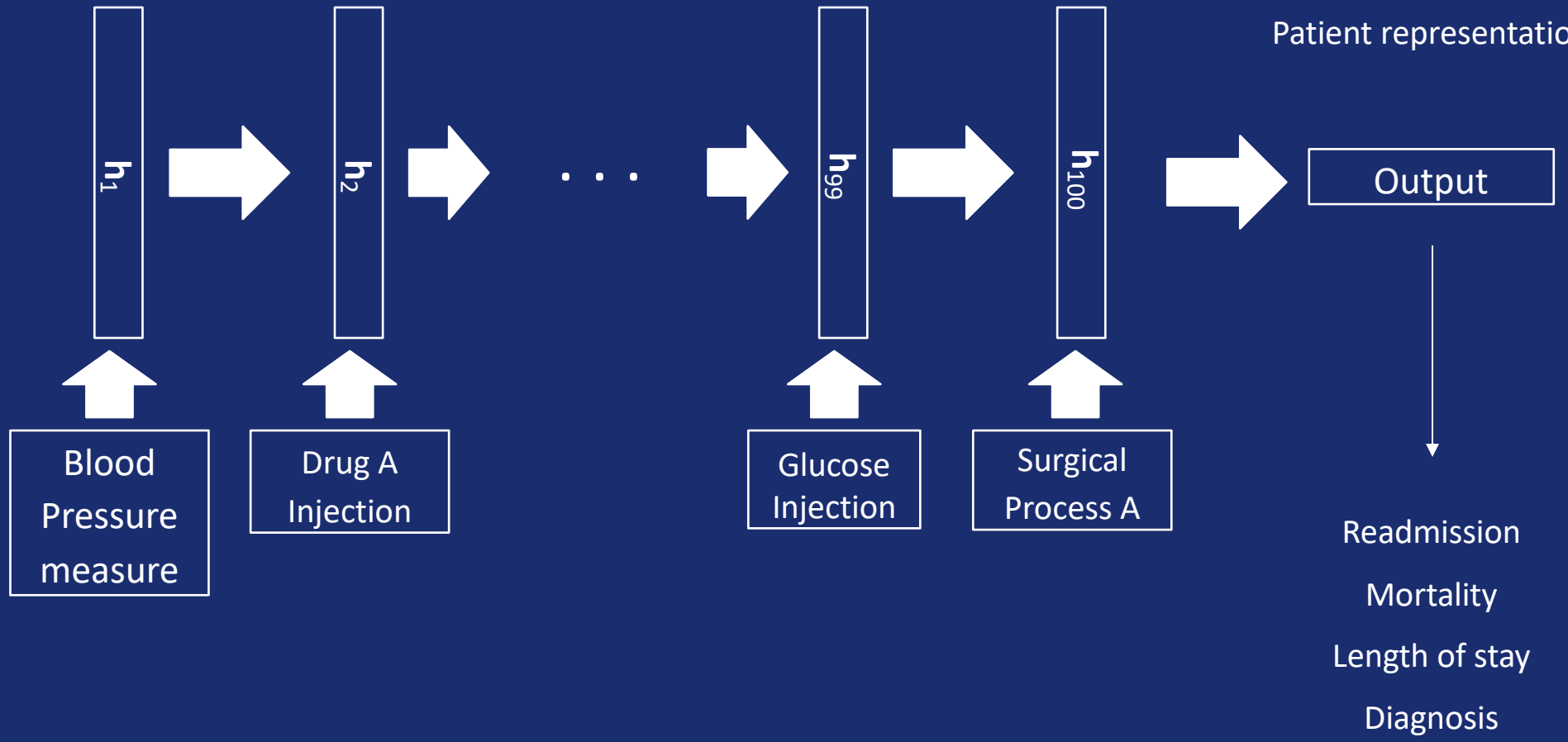
Time Series EHR & RNN



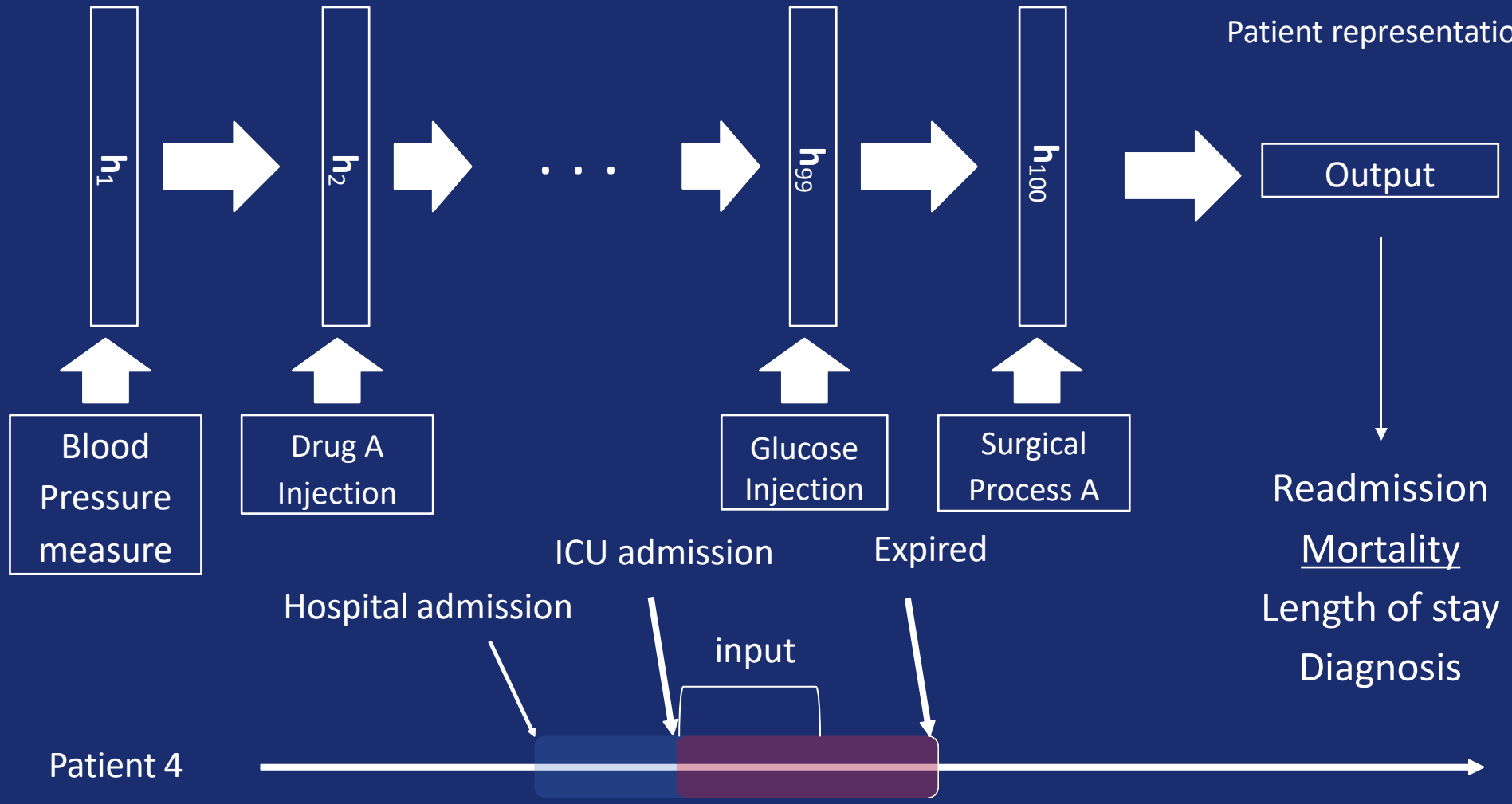
Vocabulary

Blood pressure	[1.38, -3.23, 0.6, ..]
Drug A inject	[0.52, -3.23, 4.1, ..]
Glucose inject	…
Surgical process A	[0.35, -0.23, 0.37, ..]
…	…
…	…
…	…

Time Series EHR & RNN



Time Series EHR & RNN



AGENDA

- 세션 소개
- Electronic Health Recoreds (EHR)
- Recurrent Neural Network (RNN)
- Time Series EHR & RNN
- Practice (Hands on session)

Hands on session

실습 목표 : 시계열 환자 데이터(MIMIC-III의 Chart event)를 가지고 사망 예측 (survive or die)

세부 설명 : 1)입원 기간이 24시간에서 48시간 사이인 ICU chart events 기록이 주어지면
(즉, 1일 <= 입원 기간 <= 2일)
2) 처음 3시간의 정보를 사용하여 환자가 24시간에서 48시간 동안 사망할 것인지 예측

환자의 입원 시간이 24시간 미만 또는 48시간 초과인 경우는 모두 제외함.

3) 시계열 이벤트의 최대 길이는 100 으로 제한

데이터셋 : 각 환자에는 하나 이상의 ICU 입원 기록이 있으며,
각 ICU 입원에는 ICUSTAY_ID (unique admission to ICU)로 표시된 고유 ID가 있음.

CHARTEVENTS.csv ADMISSION.csv ICUSTAY.CSV

Hands on session

1 Raw 데이터셋 다운로드	MIMIC-III Demo dataset
2 전처리	Pandas DataFrame 전처리
3 학습 및 테스트 데이터셋 구성	학습 데이터셋과 테스트 데이터셋을 Numpy로 구성
4 RNN 모델 구성 및 학습	Pytorch 라이브러리를 사용하여 RNN 학습 파이프라인 구성 (Demo가 아닌 실제 MIMIC-III 데이터로 학습)
5 학습된 모델로 테스트 셋에 예측	학습된 모델에 테스트셋 Inference

Hands on session

https://physionet.org/content/mimiciii-demo/1.4/

Hands on session

https://physionet.org/content/mimiciii-demo/1.4/

Files

Total uncompressed size: 103.0 MB.

Access the files

- Download the ZIP file (13.4 MB)
- Access the files using the Google Cloud Storage Browser here. Login with a Google account is required.
- Access the data using the Google Cloud command line tools (please refer to the gsutil documentation for guidance):
  `gsutil -m -u YOUR_PROJECT_ID cp -r gs://mimiciii-demo-1.4.physionet.org DESTINATION`
- Download the files using your terminal: `wget -r -N -c -np https://physionet.org/files/mimiciii-demo/1.4/`

Folder Navigation: <base>

Name	Size	Modified
ADMISSIONS.csv	26.2 KB	2019-10-16
CALLOUT.csv	13.5 KB	2019-10-16
CAREGIVERS.csv	174.0 KB	2019-10-16
CHARTEVENTS.csv	74.1 MB	2019-10-16
CPTEVENTS.csv	145.5 KB	2019-10-16
DATETIMEEVENTS.csv	1.7 MB	2019-10-16
DIAGNOSES_ICD.csv	47.8 KB	2019-10-16
DRGCODES.csv	22.6 KB	2019-10-16
D_CPT.csv	12.4 KB	2019-10-16
D_ICD_DIAGNOSES.csv	1.3 MB	2019-10-16
D_ICD_PROCEDURES.csv	283.1 KB	2019-10-16
D_ITEMS.csv	833.2 KB	2019-10-16

CHARTEVENTS.csv
ADMISSION.csv
ICUSTAY.CSV

데모 데이터셋으로 전처리 연습

실제 MIMIC-III 로 학습

Hands on session

Hands on session code:

- Github.
https://github.com/SuperSupermoon/KoSAIM2022

Please leave any issues if you have! Any time!