香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Learning Phenotypes and Dynamic Patient Representations via RNN Regularized Collective Non-negative Tensor Factorization

Kejing Yin[1], Dong Qian[1], William K. Cheung[1], Benjamin C. M. Fung[2], Jonathan Poon[3]

[1] Department of Computer Science, Hong Kong Baptist University
[2] School of Information Studies, McGill University
[3] Hong Kong Hospital Authority

Background: Phenotyping from EHR

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Electronic Health Records (EHR):


Patient demographics


Medication prescriptions


Diagnoses


Laboratory tests

...

☺ EHR data increasingly available
Providing opportunities for data-driven research

☹ Heterogeneous, inherently complex
Largely missing, heavily noisy, potentially biased
Difficult to be directly utilized

Hripcsak, George, and David J. Albers. "Next-generation phenotyping of electronic health records." *Journal of the American Medical Informatics Association* 20.1 (2012): 117-121.

Background: Phenotyping from EHR

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

- **Example:** Case patient identification for diabetes

Searching by diagnosis codes? Not accurate due to missing/ inaccurate records

Instead, use *the combination of diagnoses, medications, procedures, laboratory tests, etc.* to identify patients with certain conditions.

Phenotypes

Toy examples:

Diabetes Diagnoses?	✓	✓	✗	✓	✗
Diabetes Medications?	✓	✗	✓	✗	✗
High blood glucose?	✓	✗	✓	✓	✗
Case patient?	Yes	Probably Not	Yes	Yes	No

Background: Phenotyping from EHR

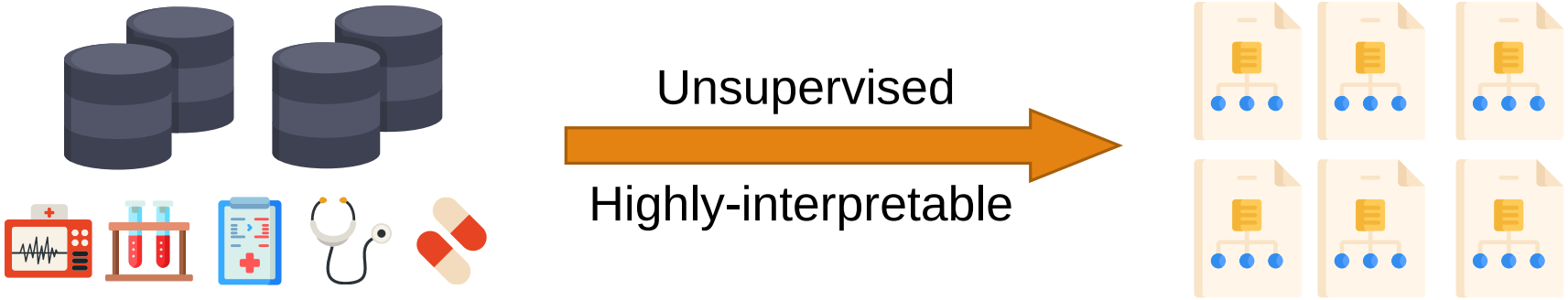
DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Phenotypes:

The combination of clinically meaningful items (e.g. diagnoses and medications) that reveals the true disease status.

Computational Phenotyping:

The process of discovering meaningful phenotypes from the raw EHR data without intensive supervision.

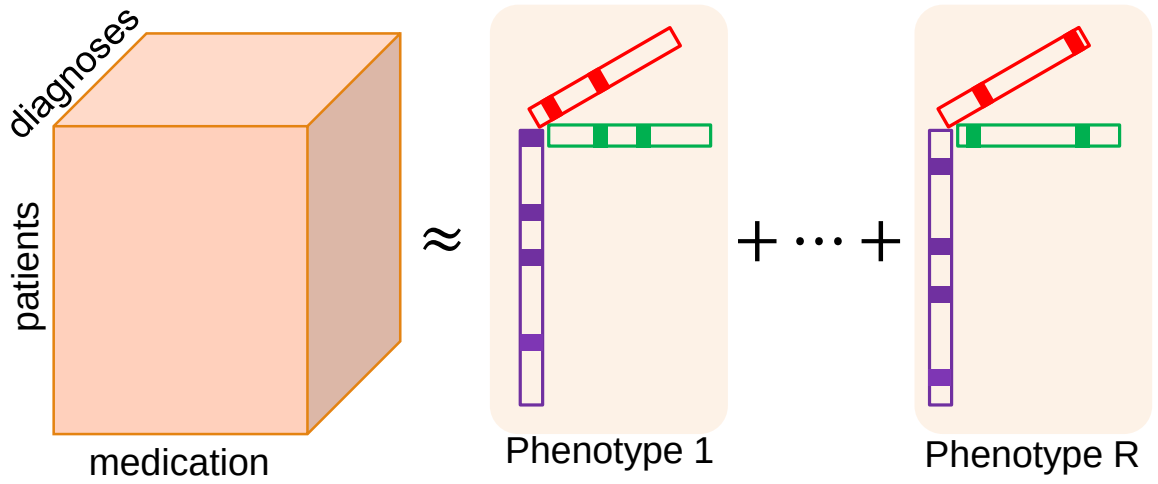


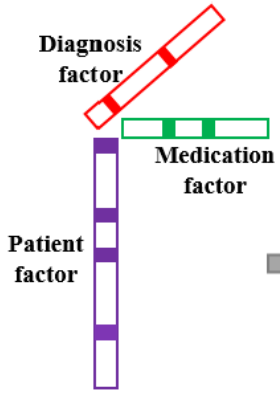
[1] Kirby, Jacqueline C., et al. "PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability." *Journal of the American Medical Informatics Association* 23.6 (2016): 1046-1052.
[2] Ho, Joyce C., et al. "Limestone: High-throughput candidate phenotype generation via tensor factorization." *Journal of biomedical informatics* 52 (2014): 199-211.
[3] Yang, Kai, et al. "TaGiTeD: Predictive Task Guided Tensor Decomposition for Representation Learning from Electronic Health Records." *AAAI*. 2017.

Tensor Factorization as a Data-Driven Method

DEPARTMENT OF COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

- Non-negative CP factorization has been intensively studied for computational phenotyping[1-7].





Diagnosis factor

Diagnosis entry	value
Cardiac dysrhythmias	**0.7**
Diabetes mellitus	0
Heart failure	**0.3**
Asthma	0
Pneumonitis	0

Medication factor

Medication entry	value
Insulin	0
Metoprolol	**0.55**
Furosemide	**0.45**
Albuterol	0
Diltiazem	0

Higher-order extension of Non-negative Matrix Factorization (NMF)

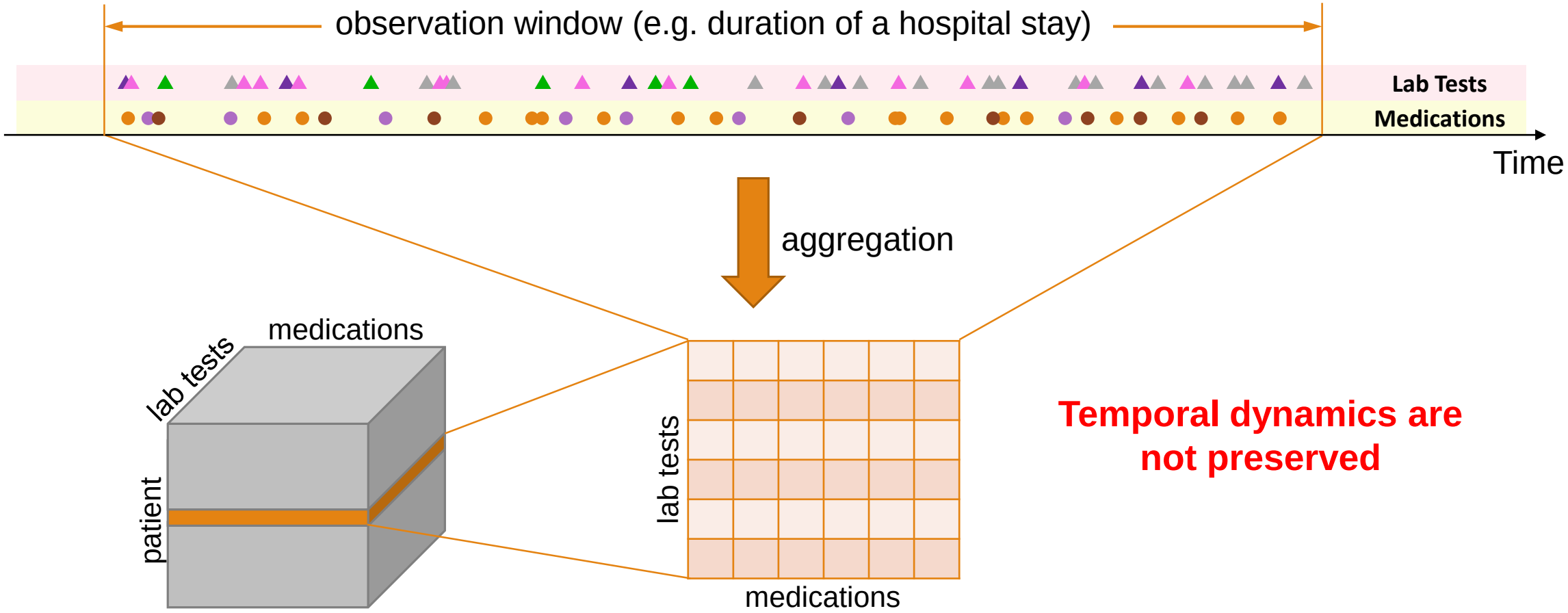
Learning "parts-of-object": highly interpretable

[1] Ho, Joyce C., et al. "Limestone: High-throughput candidate phenotype generation via tensor factorization." *Journal of biomedical informatics* 52 (2014): 199-211.
[2] Ho, Joyce C., Joydeep Ghosh, and Jimeng Sun. "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.
[3] Wang, Yichen, et al. "Rubik: Knowledge guided tensor factorization and completion for health data analytics." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.
[4] Kim, Yejin, et al. "Discriminative and distinct phenotyping by constrained tensor factorization." *Scientific reports* 7.1 (2017): 1114.
[5] Yang, Kai, et al. "TaGiTeD: Predictive Task Guided Tensor Decomposition for Representation Learning from Electronic Health Records." *AAAI*. 2017.
[6] Henderson, Jette, et al. "Granite: Diversified, Sparse Tensor Factorization for Electronic Health Record-Based Phenotyping." *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 2017.
[7] Yin, Kejing, et al. "Joint Learning of Phenotypes and Diagnosis-Medication Correspondence via Hidden Interaction Tensor Factorization." *Proceedings of the Twenty-Sixth International Joint Conference on Articial Intelligence*. 2018.

Limitations of Existing Models

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Temporal information not well considered



Temporal dynamics are not preserved

Limitations of Existing Models

DEPARTMENT OF COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

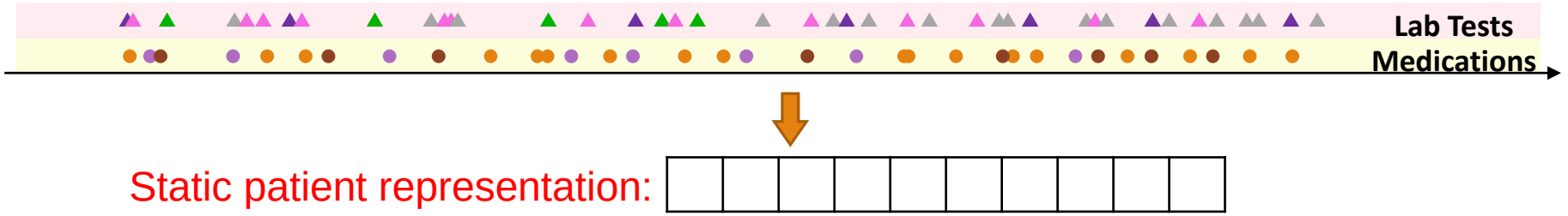
Consequently:

(1) Disease states appearing at different time are difficult to be well separated.

Diagnoses	Diabetes mellitus Other diseases of lung Acute kidney failure, …	Cardiac dysrhythmias Heart failure Other disease of lung	Other diseases of lung Cardiac dysrhythmias Heart failure, …
Medications	Insulin Insulin human regular	Amiodarone HCL Metoprolol Furosemide,…	Albuterol Diltiazem Fluticasone Propionate, …

Other diseases of lung (super class of acute respiratory failure) appear in almost every phenotype[1]

(2) Static patient representations do not reflect the dynamics within the observation window.



Considering temporal dynamics is important

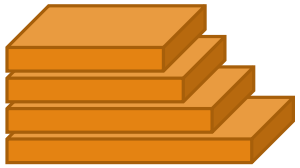
[1] Yin, Kejing, et al. "Joint Learning of Phenotypes and Diagnosis-Medication Correspondence via Hidden Interaction Tensor Factorization." *Proceedings of the Twenty-Sixth International Joint Conference on Articial Intelligence*. 2018.

Filling the Gap

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

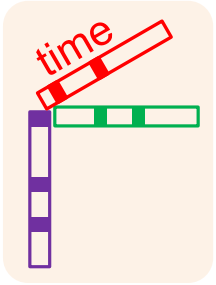
Adding time as a dimension?

1. Temporal alignment would be difficult



different length-of-stay (for inpatients): cannot form a tensor

2. Resulting phenotypes would be dynamic:



All patient share the same temporal dynamics with different magnitude.

A more natural solution:

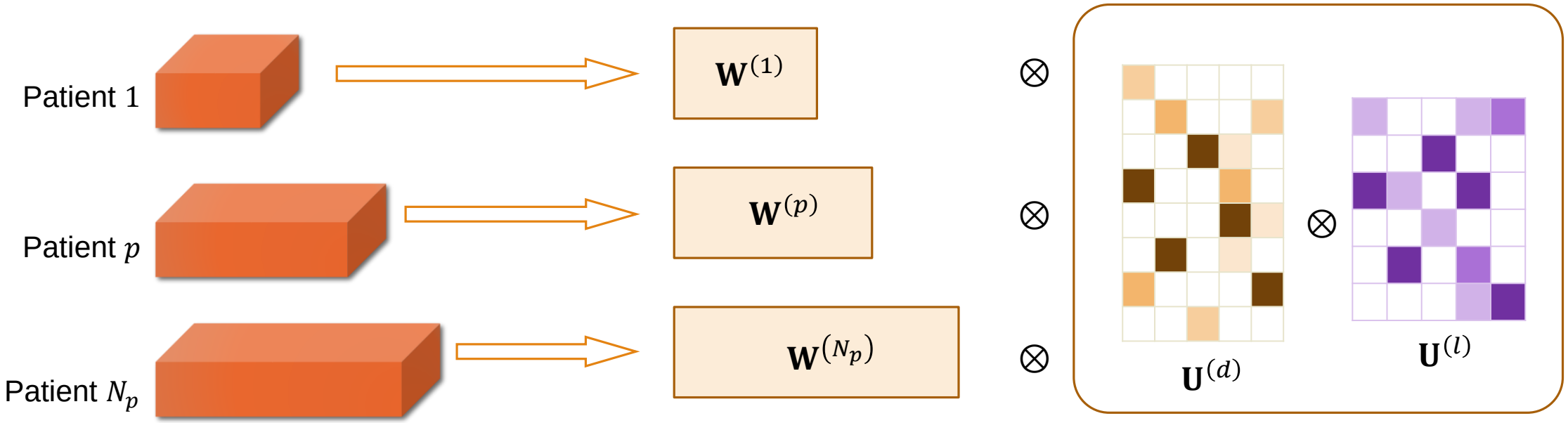
(1) Keep phenotype definitions static.
(2) Learn a dynamic representation for each patient.

[1] Perros, Ioakeim, et al. "SPARTan: Scalable PARAFAC2 for large & sparse data." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

Collective Non-negative Tensor Factorization

DEPARTMENT OF COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Collective Non-negative Tensor Factorization (CNTF)



Represent each patient with a temporal tensor

Dynamic patient representation

Phenotypes definitions shared across all the patients

No need of temporal alignment

$$f^{CNTF} \equiv \sum_{p=1}^{N_p} \frac{1}{T_p} \left(\sum_{ijk} \widehat{x}_{ijk}^{(p)} - x_{ijk}^{(p)} \log \widehat{x}_{ijk}^{(p)} \right)$$

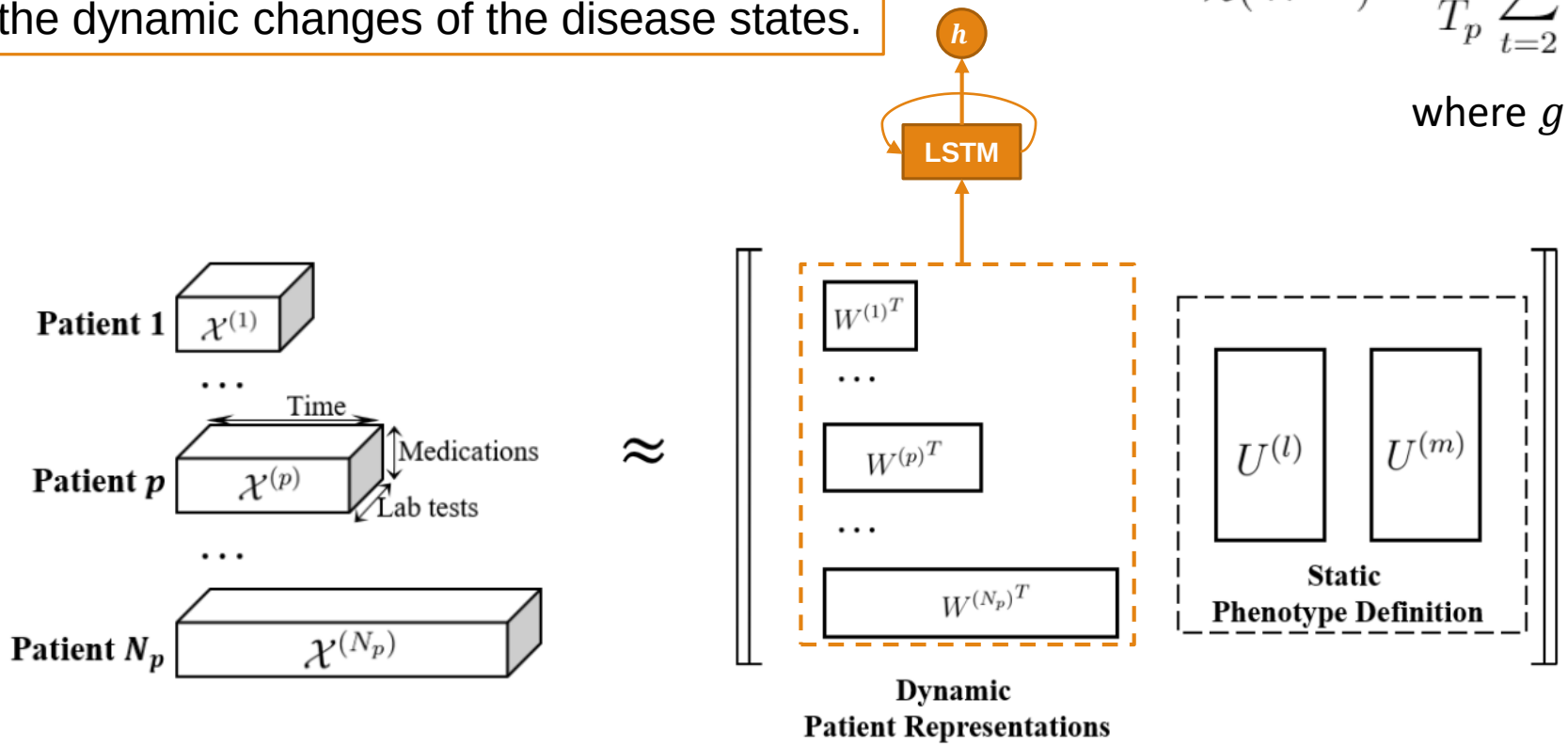
RNN Regularized CNTF

- **RNN as Regularization: modeling the temporal dependency**

Temporal representation: a *multi-variable time series* describing the dynamic changes of the disease states.

$$\mathcal{R}(\mathbf{W}^{(p)}) = \frac{1}{T_p}\sum_{t=2}^{T_p}\left\|g(\mathbf{w}_{t-1}) - \mathbf{w}_t\right\|_2^2,$$

where $g(\mathbf{w}_{t-1})$ is the LSTM output

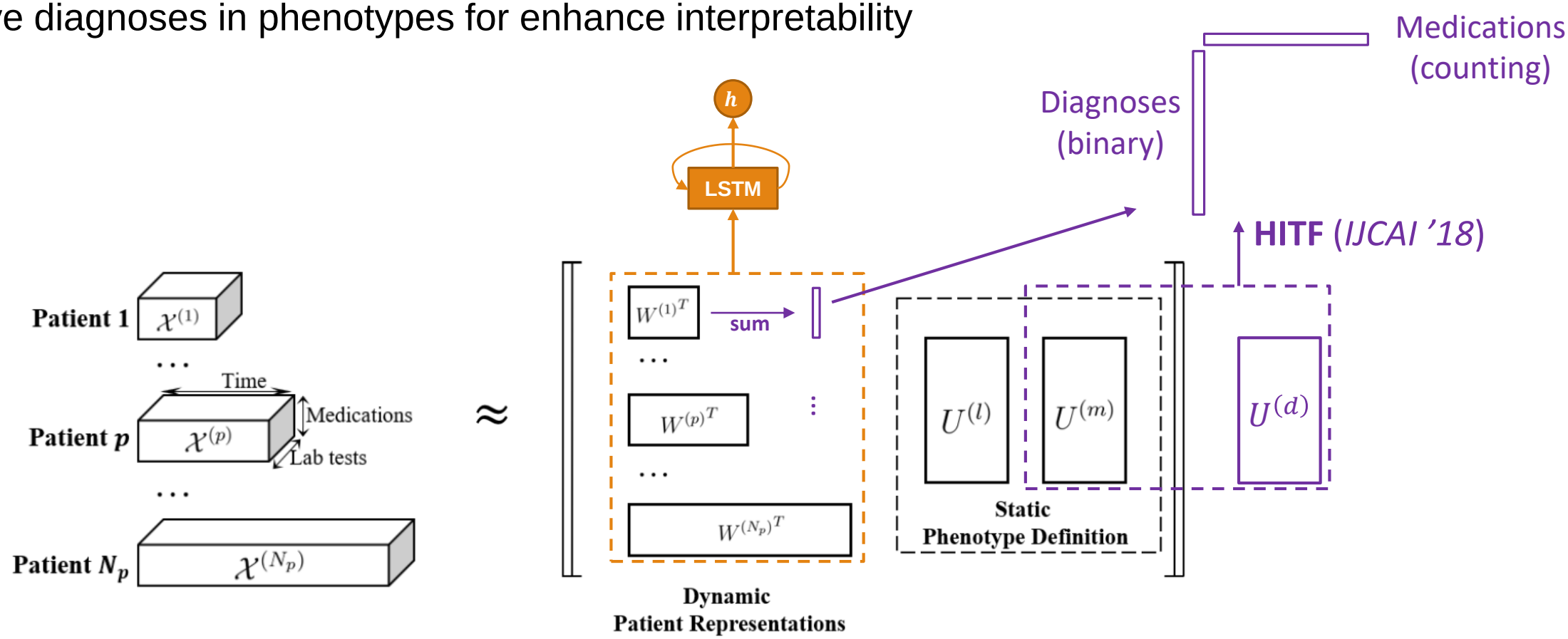


RNN Regularized CNTF

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

- **Incorporate non-temporal modalities (e.g. diagnoses for a hospital visit)**

To have diagnoses in phenotypes for enhance interpretability



[1] Yin, Kejing, et al. "Joint Learning of Phenotypes and Diagnosis-Medication Correspondence via Hidden Interaction Tensor Factorization." *Proceedings of the Twenty-Sixth International Joint Conference on Articial Intelligence*. 2018.

RNN Regularized CNTF

DEPARTMENT OF COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Formulation and Learning Algorithms

Objective function: $$\ell = \alpha_1 f^{CNTF} + \alpha_2 \sum_{p=1}^{N_p} f_p^{HITF} + \beta \sum_{p=1}^{N_p} \mathcal{R}(\mathbf{W}^{(p)}),$$

where $$f^{CNTF} \equiv \sum_{p=1}^{N_p} \frac{1}{T_p} \left(\sum_{ijk} \widehat{x}_{ijk}^{(p)} - x_{ijk}^{(p)} \log \widehat{x}_{ijk}^{(p)} \right)$$

$$f_p^{HITF} \equiv \sum_i \hat{d}_i^{(p)} - d_i^{(p)} \log(e^{\hat{d}_i^{(p)}} - 1) + \sum_j \hat{m}_j^{(p)} - m_j^{(p)} \log \hat{m}_j^{(p)}$$

$$\hat{\mathbf{d}}^{(p)} = \mathbf{e}^\top \mathbf{W}^{(p)} \operatorname{diag}(\mathbf{e}^\top \mathbf{U}^{(m)}) \mathbf{U}^{(d)\top}$$

$$\hat{\mathbf{m}}^{(p)} = \mathbf{e}^\top \mathbf{W}^{(p)} \operatorname{diag}(\mathbf{e}^\top \mathbf{U}^{(d)}) \mathbf{U}^{(m)\top}$$

$$\mathcal{R}(\mathbf{W}^{(p)}) = \frac{1}{T_p} \sum_{t=2}^{T_p} \left\| g(\mathbf{w}_{t-1}) - \mathbf{w}_t \right\|_2^2,$$

subject to non-negative constraints on $\{\mathbf{W}^{(p)}\}_{p=1}^{N_p}$ and $\{\mathbf{U}^{(n)}\}_{n=1}^{N}$.

Algorithm 1: Optimization Framework for Solving LSTM Regularized CNTF Model

Input : *time-labtest-medication* tensor collection: $\{\mathcal{X}^{(p)} | \mathcal{X}^{(p)} \in \mathbb{R}^{T_p \times I_l \times I_m}, p = 1, \ldots, N_p\}$, medication vectors: $\{\mathbf{m}^{(p)}, p = 1, \ldots, N_p\}$, diagnosis vectors: $\{\mathbf{d}^{(p)}, p = 1, \ldots, N_p\}$, model parameters: α_1, α_2 and β.

Output: patient representations: $\mathbf{W}^{(p)}$ $\forall p$, phenotype definitions: $\mathbf{U}^{(l)}$, $\mathbf{U}^{(m)}$ and $\mathbf{U}^{(d)}$.

```
initialization;
for each epoch do
    for each mini-batch do
        sample mini-batch of m tensors and vectors from
          input with indices ℒ;
        for X ∈ {U^(l), U^(m), U^(d)} do
            update X by descending its stochastic gradient;
            non-negative projection by X ← max(0, X);
        end
        for i ∈ ℒ do
            update W^(i) by descending its stochastic gradient;
            non-negative projection by
              W^(i) ← max(0, W^(i));
        end
        update LSTM model by back-propagation;
    end
end
```

Experiments and Results

- **Data Set**:
  - MIMIC-III: open-source, large-scale, de-identified, ICU related.
  - Medications and lab tests are with time stamps, but diagnosis codes are not.
  - We use 4,590 adult patients with length-of-stay longer than 7 days.
  - 50% of the patients deceased in the hospital.

- **Experiment setup**:
  - Fix α_1 to be one, determine α_2 and β by grid search.
  - Phenotypes are evaluated by clinical expert qualitatively.
  - Daily mortality prediction is used as quantitative evaluation.
  - Baseline: Rubik model.

[1] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." *Scientific Data* 3 (2016): 160035. https://mimic.physionet.org/
[2] Wang, Yichen, et al. "Rubik: Knowledge guided tensor factorization and completion for health data analytics." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2015.

Experiments and Results

DEPARTMENT OF COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Results: Phenotypes

Our proposed model

Clinically much more meaningful, evaluated by a medical expert.

Phenotype 1	Phenotype 4	Phenotype 9
Chronic kidney disease (CKD) (0.536)	Other forms of chronic ischemic heart disease (0.507) Cardiac dysrhythmias (0.372) Essential hypertension (0.024)	Other diseases of lung (0.876)
RBC (Urine) (0.200) Osmolality, Measured (Blood) (0.117) Protein/Creatinine Ratio (Urine) (0.069)	Hematocrit (Blood) (0.072) Red Blood Cells (Blood) (0.071) Hemoglobin (Blood) (0.070)	pO2 (Blood Gas) (0.253) pCO2 (Blood Gas) (0.237) pH (Blood Gas) (0.215)
Hydromorphone (0.336) (0.038) 33)		

"The disease state CKD is indeed associated with elevated RBC in urine due to renal tubular necrosis, elevated blood osmolality due to electrolyte retention in the vascular system, and elevated protein loss in the urine leading to an abnormal protein/creatinine ratio."

"Phenotype 9 corresponds to the diagnosis Other Disease of the Lung and abnormal laboratory tests pO2, pCO2, pH of the arterial blood gas. Again, this correlates well with the clinical context, where reduced oxygen levels and pH, and elevated carbon dioxide levels all indicate the presence of acute respiratory failure (which is classified under the "other disease of lung" in the ICD-9 coding system)."

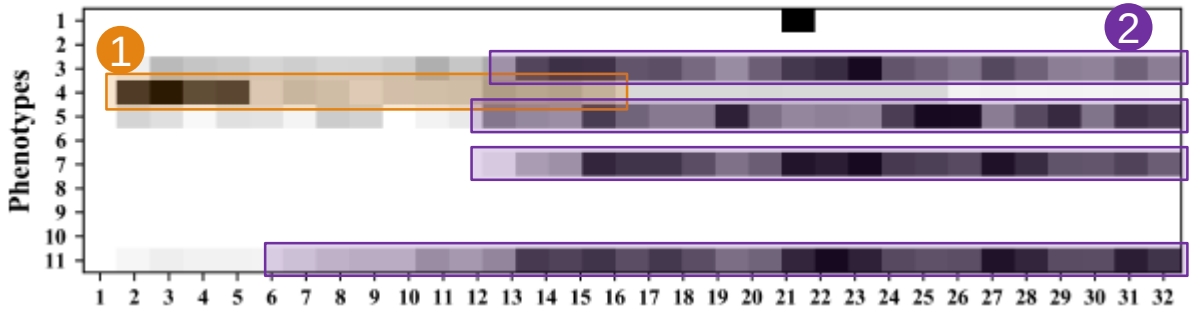
Baseline: Rubik

Phenotype 1	Phenotype 2	Phenotype 3
Other diseases of lung (0.045) Septicemia (0.040) Certain adverse effects not elsewhere classified (0.039)	Other diseases of lung (0.040) Acute kidney failure (0.036) Certain adverse effects not elsewhere classified (0.032)	Acute kidney failure (0.039) Other diseases of lung (0.037) Cardiac dysrhythmias (0.033)
Glucose(Blood) (0.019) Red Blood Cells(Blood) (0.019) Hematocrit(Blood) (0.019)	Hematocrit(Blood) (0.017) Red Blood Cells(Blood) (0.017) Glucose(Blood) (0.017)	Glucose(Blood) (0.018) Hematocrit(Blood) (0.018) Red Blood Cells(Blood) (0.018)
Vancomycin (0.017) Insulin (0.015) Potassium Chloride (0.015)	Vancomycin (0.013) Potassium Chloride (0.013) Pantoprazole Sodium (0.012)	Vancomycin (0.015) Potassium Chloride (0.014) Heparin (0.014)

Experiments and Results

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Visualization: Patient Representation



1. High value for phenotype 4 (Chronic Heart Disease) in the first several days.
2. High value for
   phenotype 3 (Other Disease of the Lung),
   phenotype 5 (Cardiac Dysrhythmias),
   phenotype 7 (Acute Kidney Failure),
   phenotype 11 (Cardiac Dysthymias with Heart Failure)

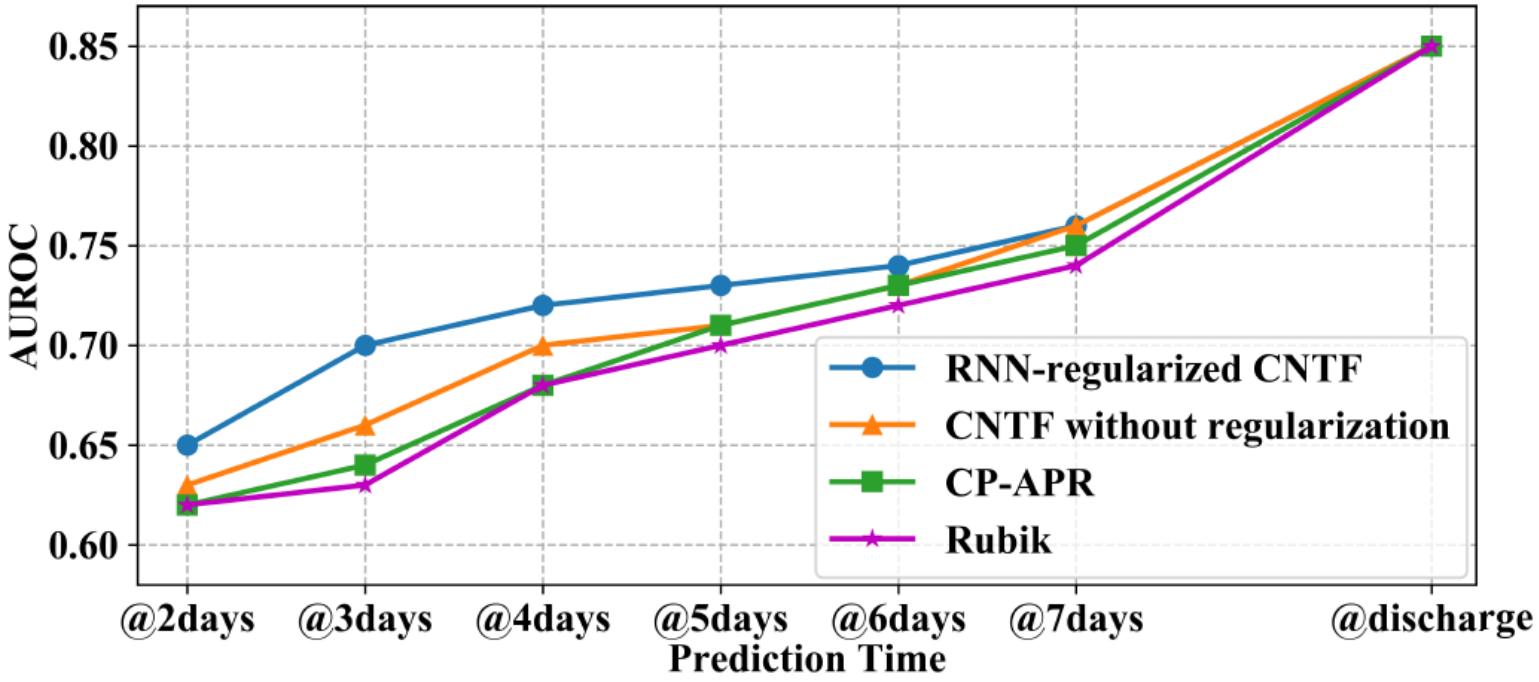
Patient admitted with existing condition, chronic heart disease, which is treated unsuccessfully, and eventually developed multiple organ failure. *(Supported by reviewing the clinical notes.)*

Health States appearing at different time can be separated by our proposed model.

Experiments and Results

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Results: Mortality Prediction



Temporal regularities are captured by the RNN regularization

Conclusion

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

- We proposed CNTF to jointly learn **the dynamic patient representations** and the **static globally shared phenotypes**.
- RNN-based regularization and HITF model are integrated to model the time dependency and incorporate the non-temporal modalities.
- The proposed model can derive clinically meaningful and interpretable phenotypes, and better separate the disease states appearing at different time.

Welcome to our poster tonight!

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Thank you!

All questions and comments are greatly appreciated!

Yin Kejing
cskjyin@comp.hkbu.edu.hk