Electronic Health Record data

opportunities, challenges and feature learning

Chi Zhang

2019-05-24, internal seminar @ OCBE

Outline

EHR data and challenges

Feature learning and application in EHR

2 methods for MIMIC III data

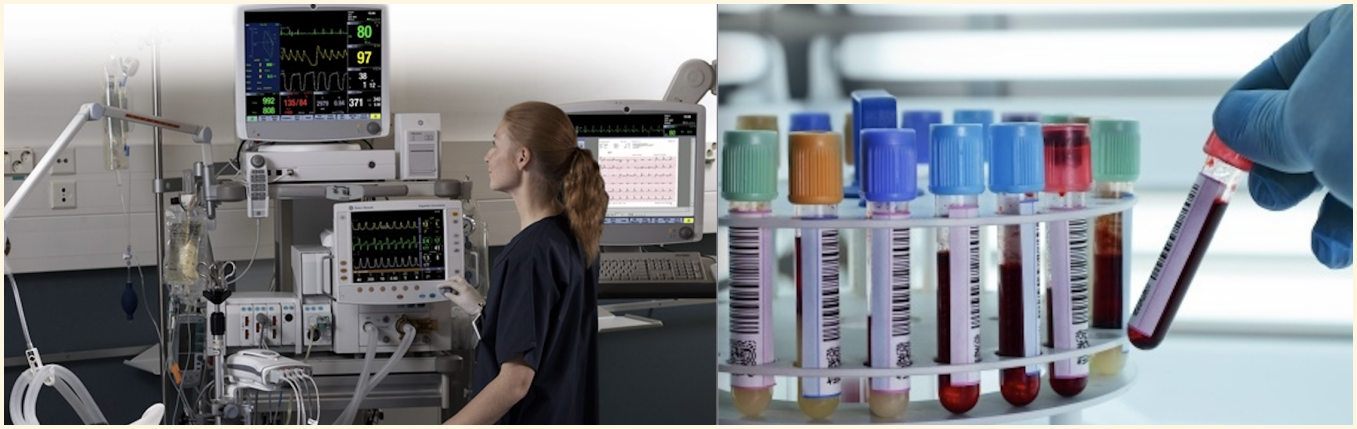
1. **LSTM autoencoder**
2. **Tensor decomposition**

Electronic Health Records

Initially used for billing and administrative purposes

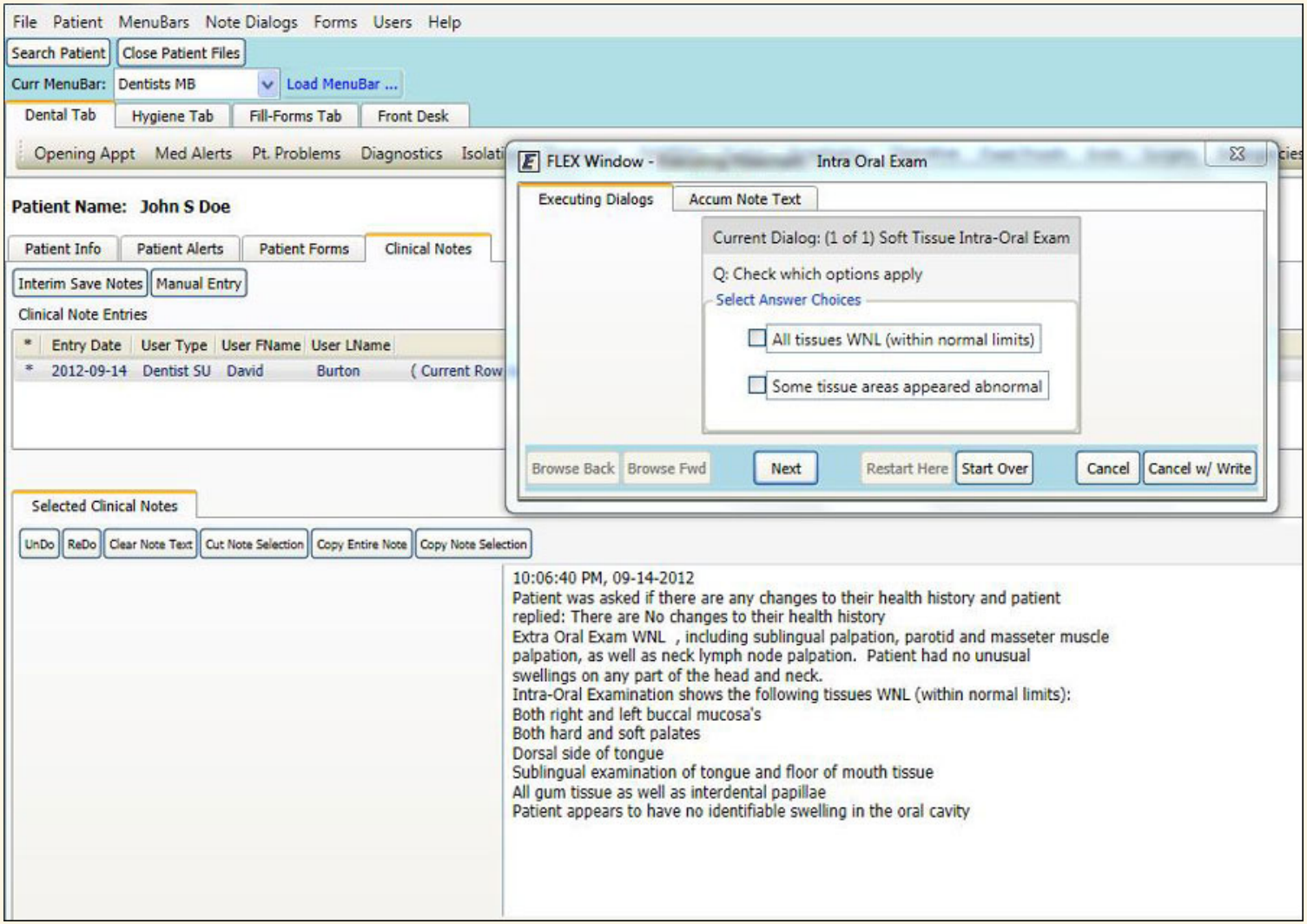
Routinely generated patient data from:

- bed side monitors, lab tests



- More specialised: images such as waveforms
- Administration: demographics

Clinical notes



Who benefits from EHR research?

Patients

- disease progression prediction / early detection
  - e.g. heart failure, (Choi 2016)
  - in-hospital mortality, (Suresh 2018)
  - Parkinson's disease stage (Che 2017)

Medical professionals

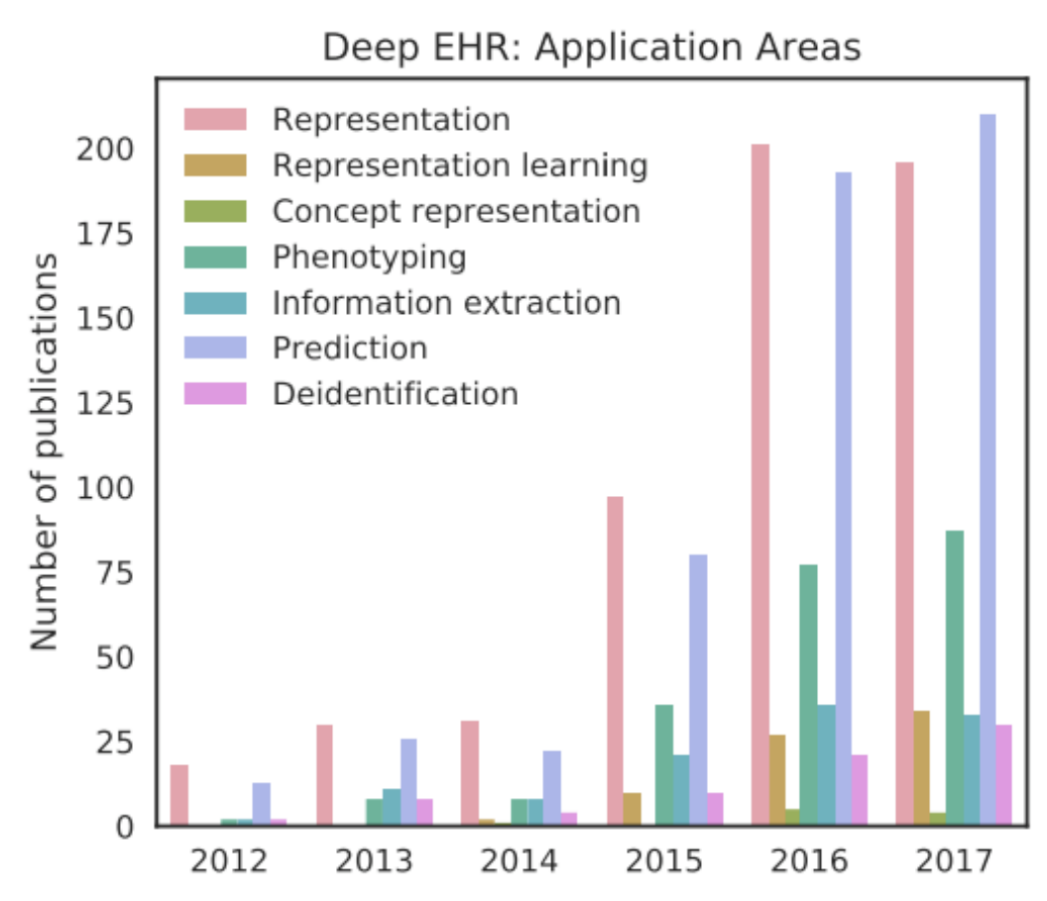
- automate diagnosis
  - image evaluation: detect hip fracture from X-ray (Gale 2017)
- automate routine process
  - summarize content from medical records (McCoy 2018)

Institution: hospitals

- cost reduction
- privacy protection

Research topics

Number of Google Scholar publications relating to **deep EHR** until 2017
(Shickel 2018 Survey)



Challenge 1: Availability

Privacy protection

- no data at all
- or: 'brutal' anonymisation

Lifesaver: MIMIC III data

ICU data

Pros: relatively complete

- demographics, vital sign measurements, lab test results,
- procedures, medications, imaging reports
- notes

Cons: severely ill, multiple diagnosis within one same subject

MIMIC III - Medical Information Mart for Intensive Care

- Critical care units, Beth Israel Deaconess Medical Center (Boston, US), 2001-2012
- over 50,000 records for 30,000+ patients
- De-identification: randomised time stamp

Static data (table **admissions**)

subject_id	hadm_id	admittime	deathtime	admission_type	religion	diagnosis
64	172056	2143-03-03 09:25:00	NULL	EMERGENCY	OTHER	FUNGAL MENINGITIS
65	143430	2132-08-10 07:13:00	NULL	EMERGENCY	UNOBTAINABLE	S/P STRUCK BY CAR
66	104518	2188-08-25 00:14:00	NULL	NEWBORN	NOT SPECIFIED	NEWBORN
67	186474	2155-02-25 12:45:00	NULL	URGENT	JEWISH	INCISIONAL HERNIA
67	155252	2157-12-02 00:45:00	2157-12-02 03:55:00	EMERGENCY	JEWISH	SUBARACHNOID HEMORRHAGE
68	170467	2173-12-15 16:16:00	NULL	EMERGENCY	PROTESTANT QUAKER	PNEUMONIA
68	108329	2174-01-04 22:21:00	NULL	EMERGENCY	PROTESTANT QUAKER	WEAKNESS
69	190201	2129-03-21 15:33:00	NULL	NEWBORN	UNOBTAINABLE	NEWBORN
70	178596	2185-09-01 10:42:00	NULL	NEWBORN	NOT SPECIFIED	NEWBORN
71	111944	2164-02-03 22:07:00	NULL	EMERGENCY	NOT SPECIFIED	OVERDOSE
72	156857	2163-09-22 23:52:00	NULL	NEWBORN	NOT SPECIFIED	NEWBORN

Dynamic data (table **chartevents**)

subject_id	icustay_id	itemid	charttime	value	valueuom
10006	206504	618	2164-10-25 00:00:00	20	BPM
10006	206504	211	2164-10-25 02:00:00	81	BPM
10006	206504	456	2164-10-25 02:00:00	85.333297729492188	mmHg
10006	206504	618	2164-10-25 02:00:00	21	BPM
10006	206504	646	2164-10-25 02:00:00	97	%
10006	206504	742	2164-10-25 02:00:00	1	kg
10006	206504	87	2164-10-25 06:00:00	19	number

Information scattering around: 26 such tables, possibly large (330 million rows in **chartevents**)

Connect PostgreSQL Database with R

```
library(RPostgreSQL)
drv ← dbDriver("PostgreSQL")        # PostgreSQL driver
con ← dbConnect(drv,
                dbname = "mimicbig",    # database name
                host = "localhost",
                port = 5432,
                user = "chizhang",
                password = pw)
query ← 'SELECT * FROM admissions WHERE subject_id = 10006'
record ← dbGetQuery(con, statement = query)
```

--

Comments:

Know what information you need;

Partition large tables for speed.

Challenge 2: data quality

Digital artifact corruption

- Undesired alteration in data due to technical reasons
- e.g. accidental removal of sensor

Inconsistency and duplications

- Various database systems within one hospital
- Example 1: Carevue and Metavision within MIMIC III, 2 sets of item IDs
- Example 2: measurement recorded by different staff into different systems

Missingness

- at random?
- not at random?

lack of labels Makes predictive models hard

Challenge 3: Multi-modality

Numeric measurements

Categorical information: diagnosis code, ethnicity

Free text

Time stamp

Images and signals: ECG waveforms

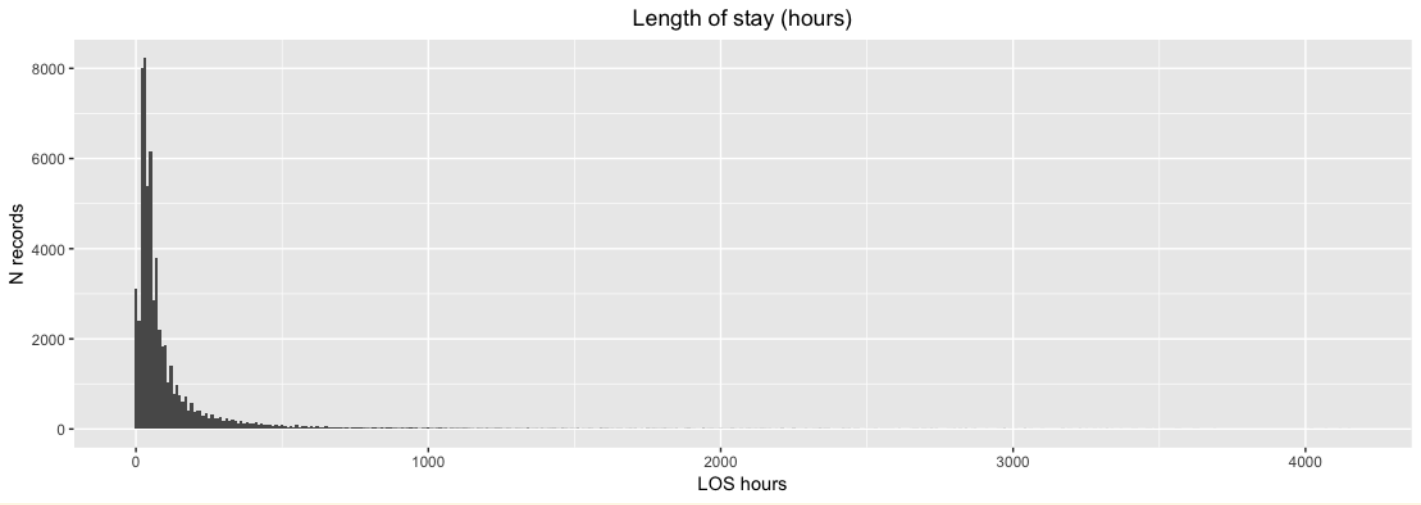
Genomic information

Challenge 4: Irregularity (length)

Unequal length of measurements: from a few hours to a few thousand hours

```
los ← read.csv('~/Documents/Data/los.csv')
loshours ← los*24
summary(loshours$los)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     0.002    26.593    50.214   118.031   107.596 4153.740
```

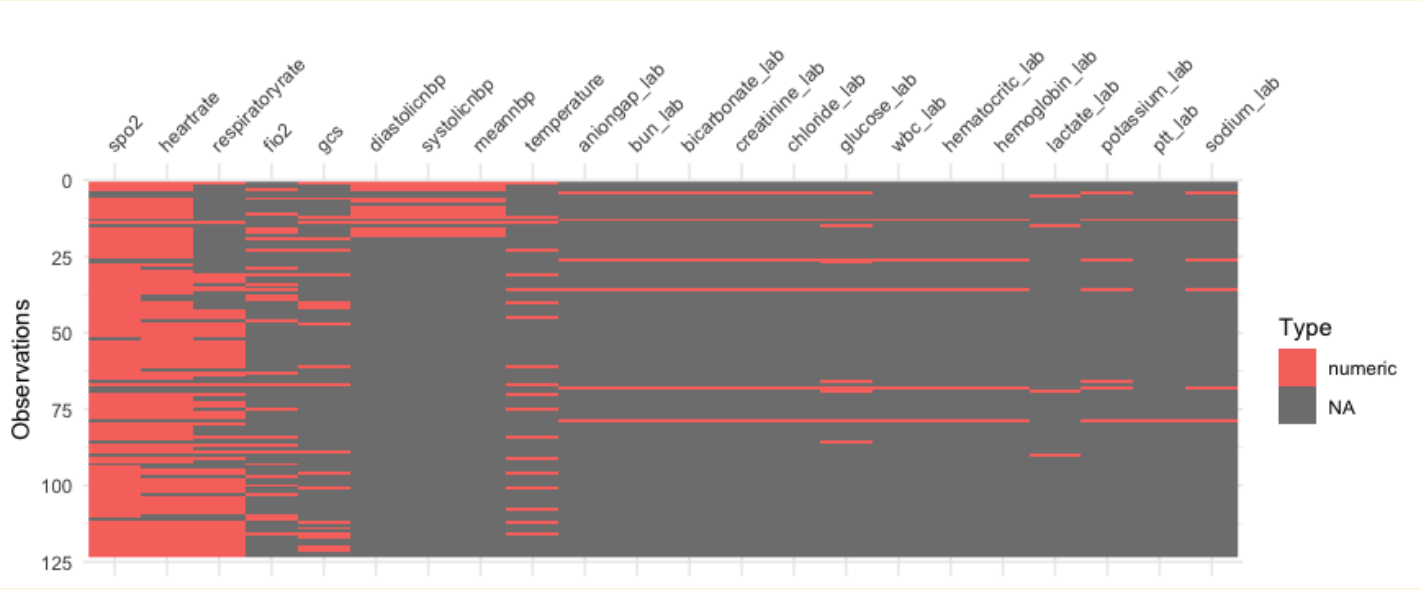


Challenge 5: Irregularity (frequency)

Various sampling frequency

- High: Electrocardiogram (ECG)
- Medium: vitals, input (e.g. medication), output (e.g. urination)
- Low: lab tests, clinical notes (done by order)
- Static: demographics, diagnosis

Example: patient in MICU for 124 hours, selected **vital** and **lab test results**



To sum up, challenges of EHR data

- data usability
- lack of labels
- multi-modality
- unequal length
- unequal frequency
- ...

Feature / Representation learning

Feature: a characteristic that helps with the modeling.

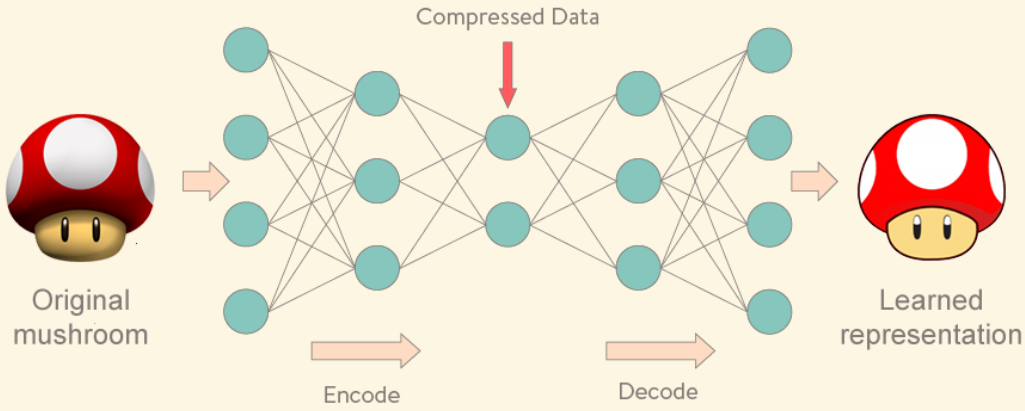
Feature / Representation learning: techniques that make it easier to extract information, eitehr to understand the data structure, or when building predictive models

Motivations

- Feature **Engineering** vs **Learning**
- Dimension reduction (*PCA*)
    - visualization (clustering, e.g. *K-means*)
    - memory saving (in the early days, e.g. *symbolic aggregation*)
- Predictive performance (*Lasso*, *Autoencoder*, *RNN*)
- Capture pattern from multiple modes such as time (*tensor*)

Example 1: Autoencoder

Non probablistic, direct encoding: parametric map from input to representation



- **Encoder**: $h = f_\theta(x) = s_f(b + Wx)$
- **Decoder**: $g_\theta(h) = s_g(d + W'h)$
  - Activation functions: linear, sigmoid, hyperbolic tangent
- Minimise **reconstruction** error: $L(x, g_\theta(f_\theta(x)))$
  - Squared loss, binary cross-entropy loss
- In some cases the 'bottle-neck' hidden layer is used as representation.
- Variations: Denoising AE, Sparse AE, Variational AE etc

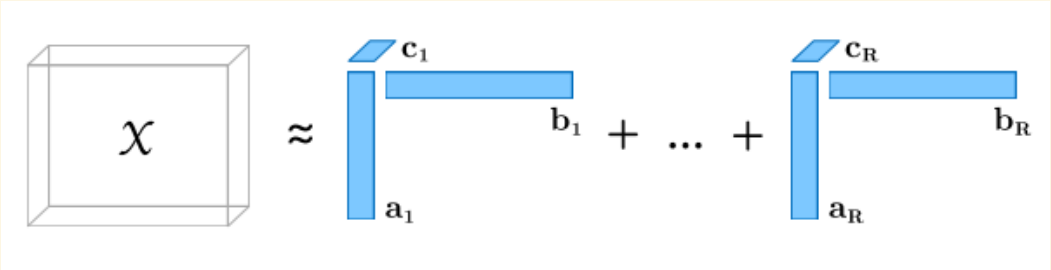
Example 2: Tensor decomposition

Tensor: multi-dimensional array

- order 0: scalar
- order 1: vector
- order 2: matrix
- order 3: cube

CP: CANDECOMP/PARAFAC

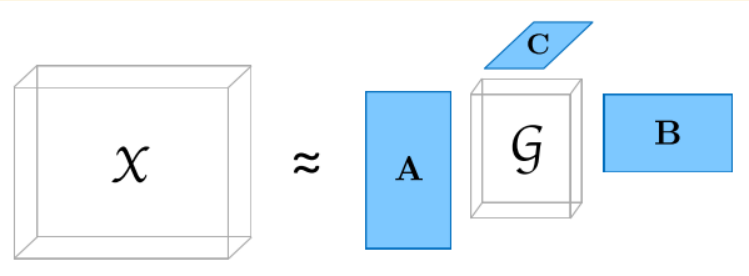
Express a tensor into a sum of finite number of rank-one tensors (i.e. can be written by the outer product of N vectors)



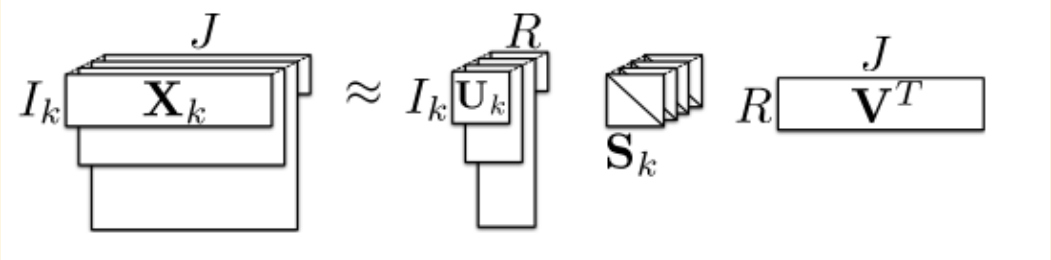
Factor matrices: combination of vectors $A = [\mathbf{a}_1 \quad \mathbf{a}_2 \ldots \mathbf{a}_R]$

Tucker decomposition

(Higher order SVD, N-mode PCA): a tensor into a core tensor multiplied by matrix along the sides; factor matrices are usually orthogonal



PARAFAC2

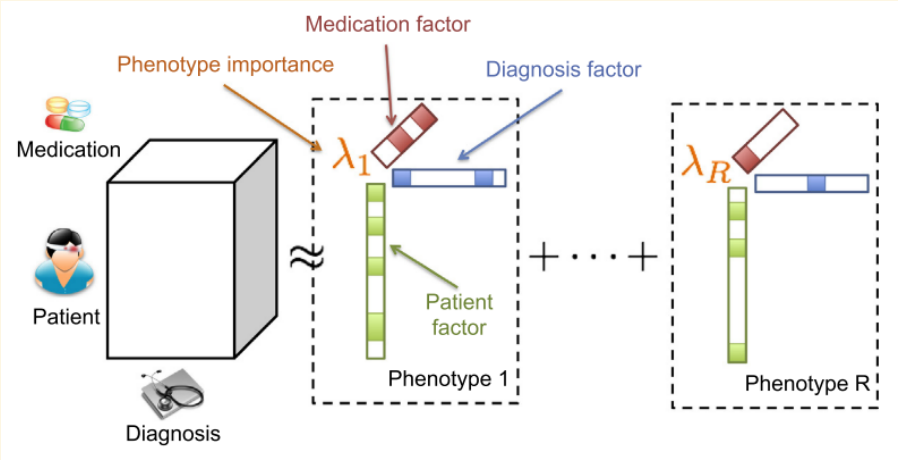


Applications: phenotyping with varying measurements

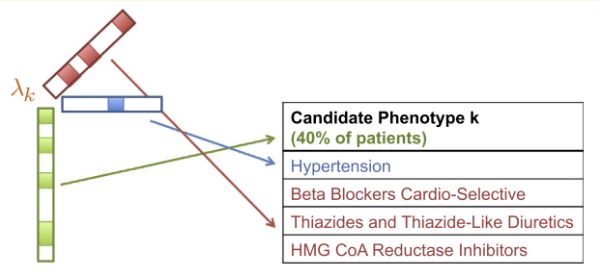
Application in computational phenotyping

Phenotyping: identify patients with certain characteristics of interest. As simple as Type 2 Diabetes

Example: (Ho 2014)



Phenotype interpretation:



Representation: Evaluation

What is a good representation?

(Bengio 2014) Sparsity, temporal and spatial coherent, smoothness, ...

More practical standards:

- Representation themselves
  - no standard metric, depend on the problems
  - custom metrics (e.g. 'Medical concept similarity measure'(Choi 2016))
  - visualization (e.g. heatmap)
- Classification performances
  - AUC, accuracy, precision, recall, ...

Current project (using MIMIC data)

Objective:

Investigate patient representation options, better representation means better accuracy in prediction of in-hospital mortality

> Backstory: we were inspired by Suresh 2018 study, hence the objective and choice of inputs

Options for representation

1. LSTM AE (Suresh 2018 paper)
2. Our experiment: feature similarity tensor decomposition

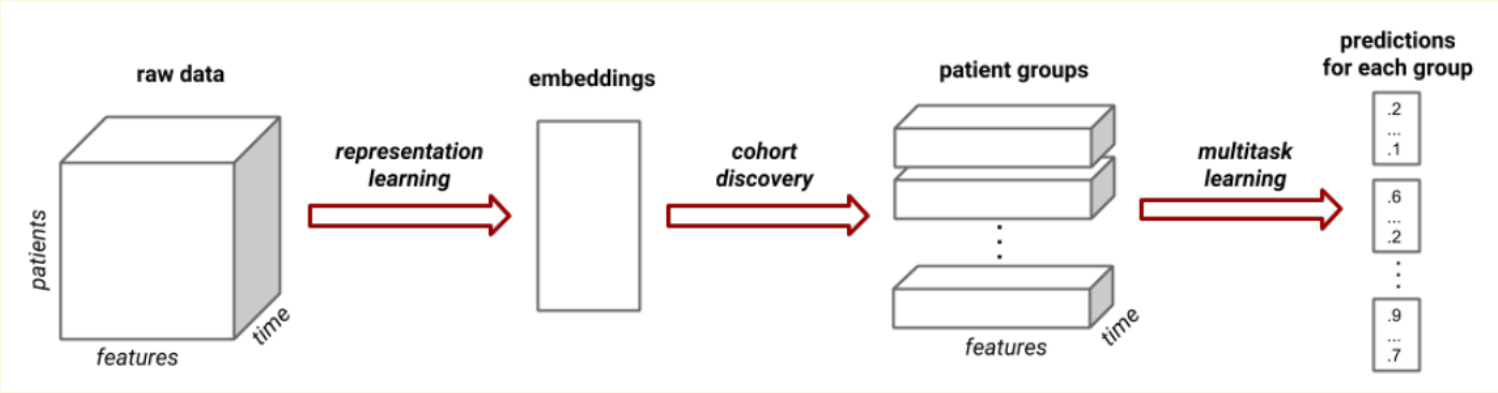
Long short term memory Autoencoder (LSTM AE)

Suresh, Gong, Guttag 2018 Paper @ KDD:

Learning Tasks for Multi-task Learning: Heterogenous Patient Population in ICU

Key idea:

- Input: 29 variables (static and dynamic)
- first 24 hours dynamic records
- use LSTM AE to embed patient data from 3 modes: time, features, patient
- predictive task: mortality at hour 36



Choice of inputs:

Static Variables	Gender	Age	Ethnicity
Vitals and Labs	Anion gap	Bicarbonate	blood pH
	Blood urea nitrogen	Chloride	Creatinine
	Diastolic blood pressure	Fraction inspired oxygen	Glascow coma scale total
	Glucose	Heart rate	Hematocrit
	Hemoglobin	INR*	Lactate
	Magnesium	Mean blood pressure	Oxygen saturation
	Partial thromboplastin time	Phosphate	Platelets
	Potassium	Prothrombin time	Respiratory rate
	Sodium	Systolic blood pressure	Temperature
	Weight	White blood cell count	

*International normalized ratio of the prothrombin time

Implementation:

- Keras
- Input dimension: `n_samples * n_timesteps * n_features`
- Embedded: `n_samples * n_latent`
  - (100, the elbow in reconstruction error curve on validation set)
- Clusters: Gaussian mixture model, 3 clusters(for best performance)
- Prediction:

Performance

AUC > 0.8 across different setups

Similarity tensor decomposition

(! testing stage!)

Key ideas

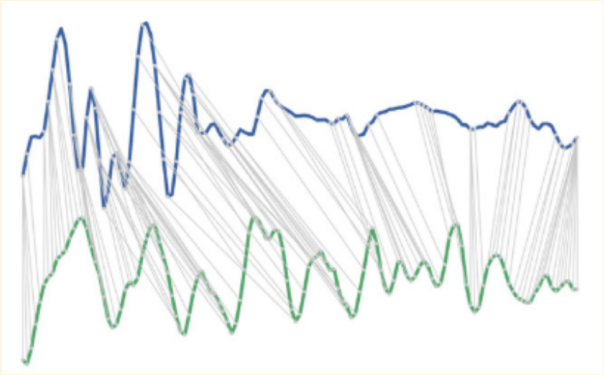
- utilise whole series, not only first 24 hours
- deal with irregularity and missingness: **use feature similarity instead of individual series**
    - natural cluster: better performance in personalized predictive tasks
- Tensor decomposition on feature-similarity tensor
- Representation: factor matrix

Feature similarity

Trying out 2 things:

SAX (Symbolic Aggregate Approximation): raw time series into discrete symbols

- Dimension reduction **Dynamic Time Warping**
- similarity measure for time series data
- computes the most optimal warping alignment path - unequal length!
- applications in TS clusterng / classification



Future plans

Investigate different configurations

Test predictive performance

Compare with Suresh 2018 study

...