Decoding and Analyzing Medical Data

Medical Insurance Fraud Detection

Sirui Li(sl4653) & Xiaoli Sun(xs2338)

Background

$3.65T

Health Care Cost

2018 CMS report US Health Care Cost

3-10%

Fraud Percentage

Investigated by FBI and Insurance Company

$110B

Medical Fraud

Conservatively Annual Estimate Lost

Goals

- Predict potential illness with similar patient retrieval
- Early-stage treatment recommendation for prevention
- Visualization platform for health providers
- (Potential) Anomaly Fraud Detection

Our Dataset

- MIMIC-III (Medical Information Mart for Intensive Care III) Critical Care Database
- Health data from over 40K ICU patients of the Beth Israel Deaconess Medical Center (2001-2012)
- Large dataset(60G), consisting of 40 tables, 534 columns and 700M rows
- Preliminary categorizing by diseases ICD-9 codes
- Google Cloud Platform

- List of ICD-9 codes 290–319: mental disorders
- List of ICD-9 codes 320–389: diseases of the nervous system and sense organs
- List of ICD-9 codes 390–459: diseases of the circulatory system
- List of ICD-9 codes 460–519: diseases of the respiratory system

Previous Research

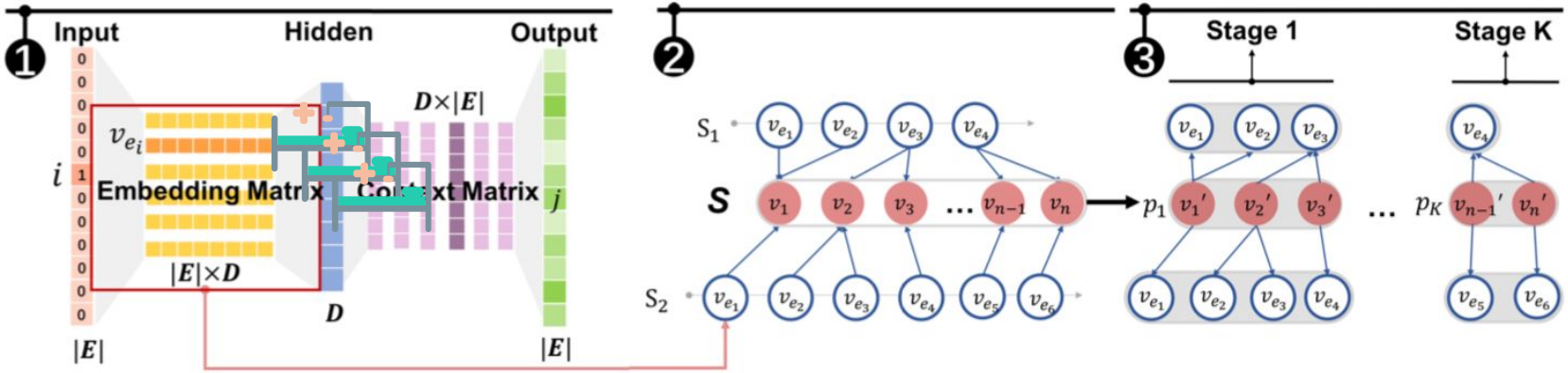
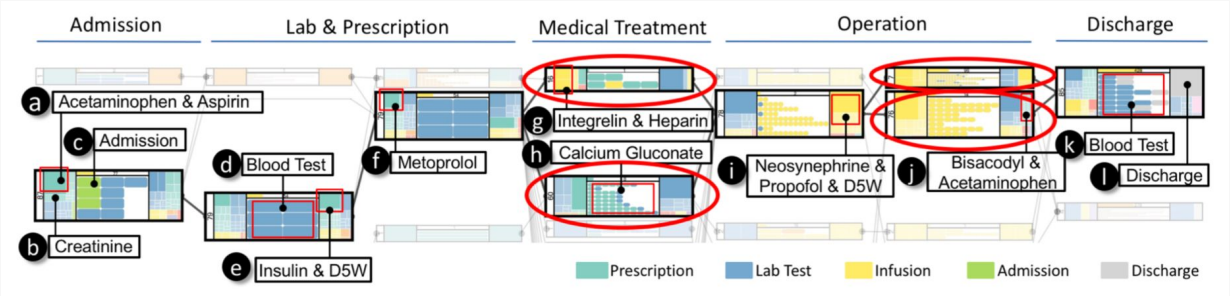


Fig. 3. Progression analysis includes three major steps: (1) event representation estimation, (2) sequence alignment, and (3) sequence segmentation.

- Provide concrete data-processing procedures
- Provide visualisation methods
- Apply the research to market
- Give algorithms for core steps

Our Project - Workflow



Dataset

ICD-9
NLP
Computable vector

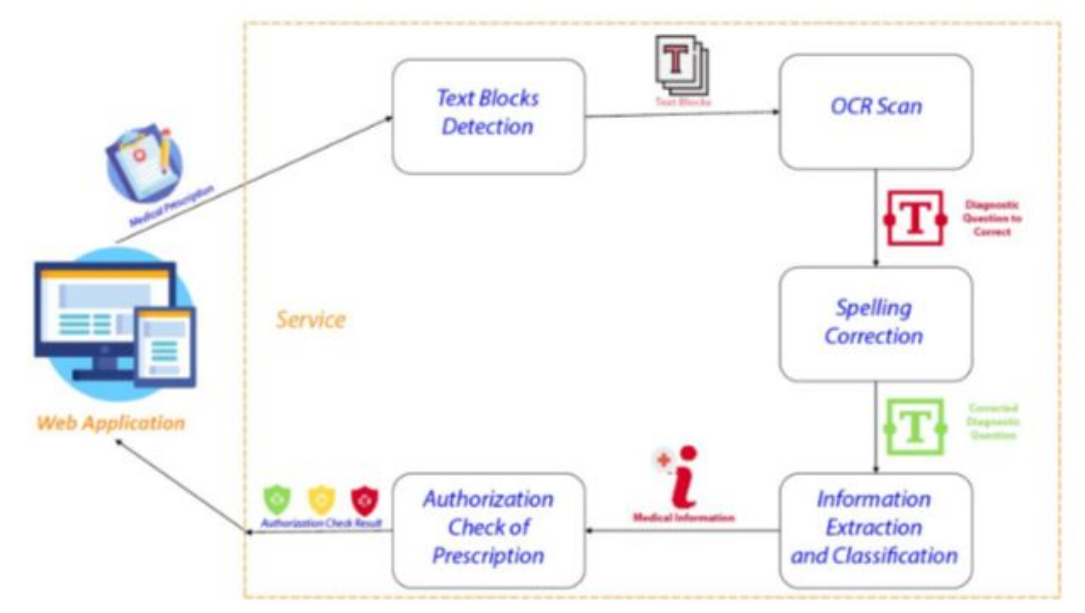
Preprocessing

Event Representation
Event Sequence Alignment
Event Sequence Segmentation

Event Sequence

Unsupervised Machine Learning Algorithms
RNN

Our Project - Data Cleaning



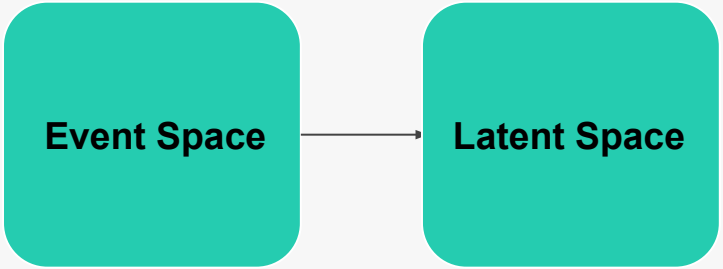
ICD-9 code

vector

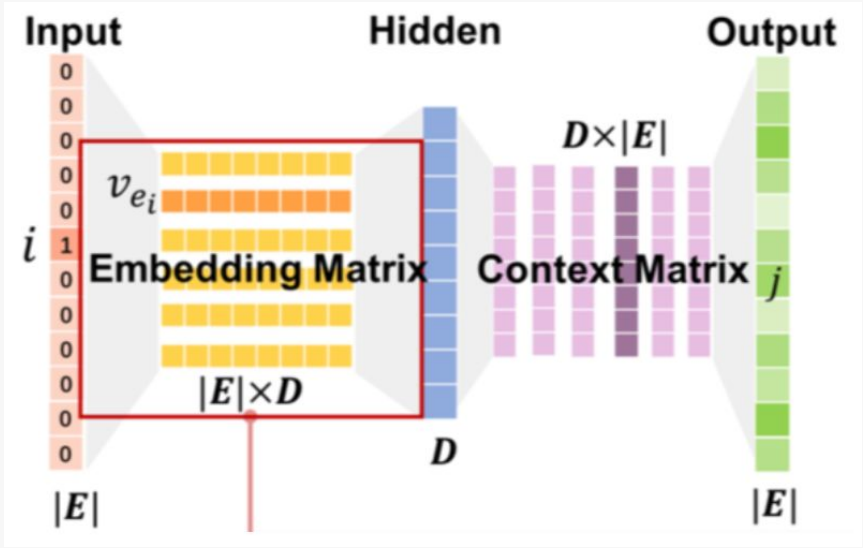
Related work: Medical prescription classification: a NLP-based approach

Our Project - Event Sequences Generation

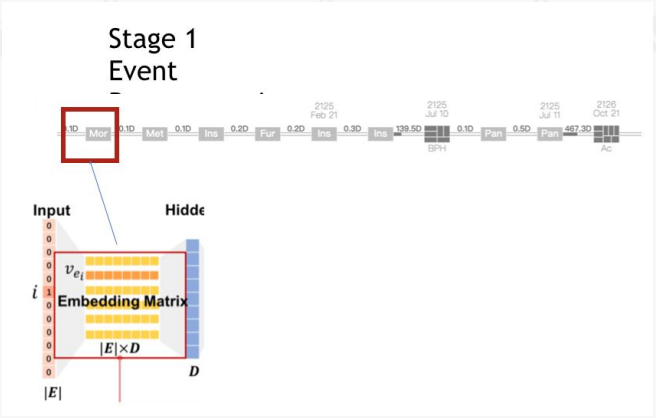
The Biggest **Challenge** :

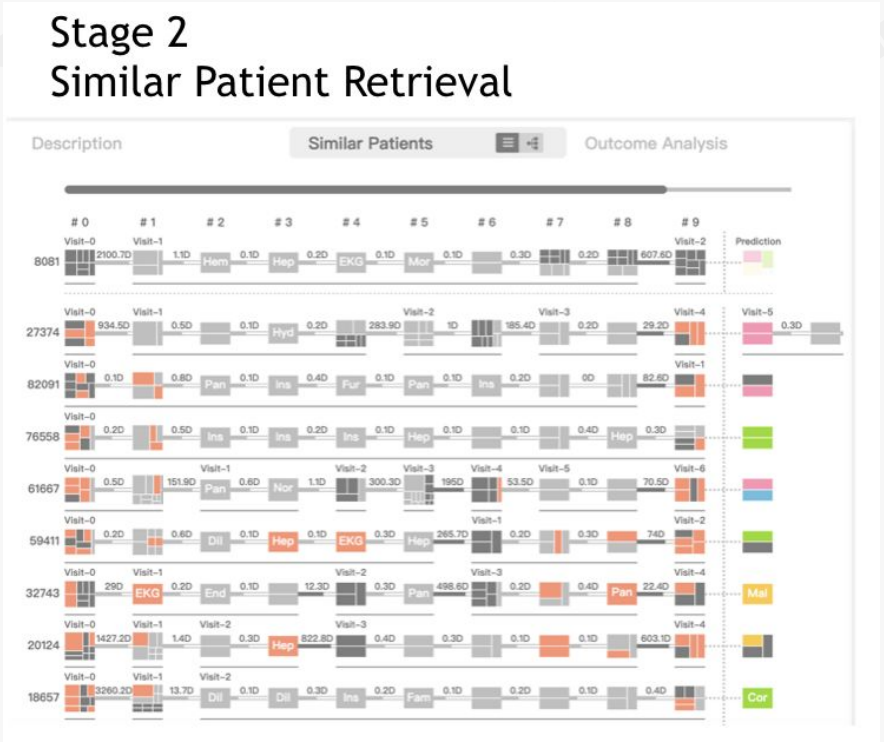


Challenge: It's hard to find such an embedding matrix. We're still working on it...



Our Project - Prediction & Recommendation





Challenge:

the reliance upon fixed-width time intervals

Technical Challenge

01

Data Cleaning

40 Tables
How to connect and clean

02

Medical Terminology

Crawl information from Wiki to create knowledge graph

03

Event Sequence Generation

Find latent space and embedding matrix

04

Anomaly Detection

Implementation of real-world fraud data

Novelty and Value

Insurance Company

Decrease Cost of Business

Health Provider

Visualization Platform

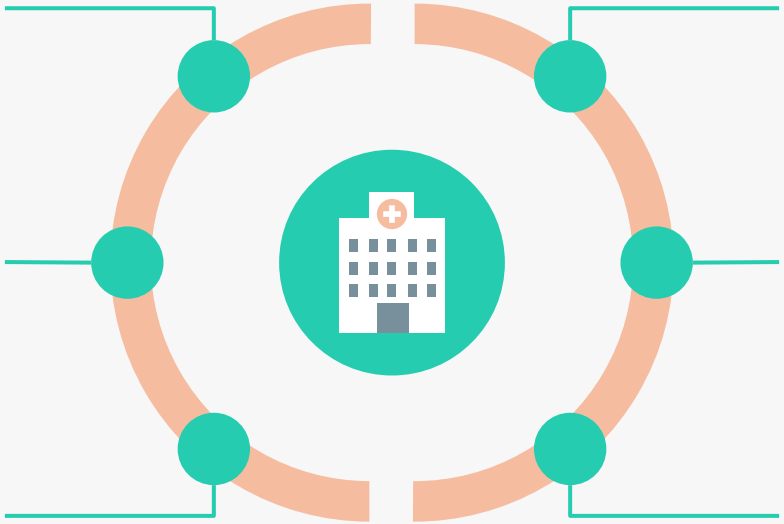
Prediction and Prevention

More Efficient Procedure

Government

Help to Fight Against Crimes

Automated Fraud Detection



Consumer

Avoid Financial Lost

Lower Insurance Premium

Higher Survival Rate

Scientist

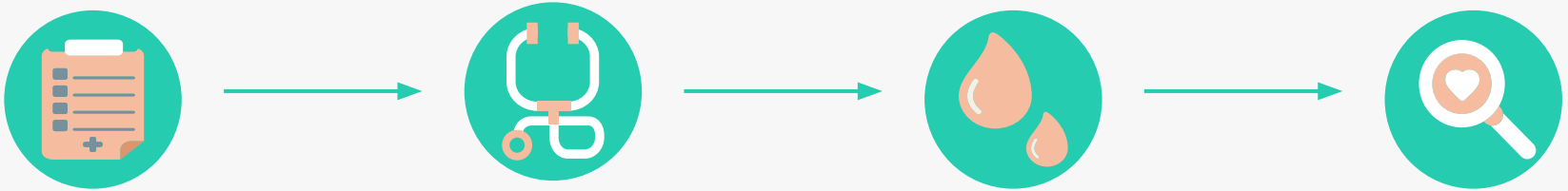
Apply Algorithms in Other Fields

More?

Longer Life Expectancy

OECD Better Life Index

Milestones Plan



Milestone 1

Related Work Study
Dataset Select
Proposal
Data Clean

Milestone 2

Model Build
Disease Predict
Treatment Recommend
(Anomaly Detect)

Milestone 3

Broaden Implementation
Accuracy Test

Final

Visualization Demo
Report & Video

Citation

- https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/ForecastSummary.pdf
- https://www.researchgate.net/publication/336071394_Medical_prescription_classification_a_NLP-based_approach
- https://en.wikipedia.org/wiki/List_of_ICD-9_codes
- http://gotz.web.unc.edu/files/2018/08/2018_VAST_EventThread_V2_Preprint.pdf
- http://www.nature.com/articles/sdata201635
- http://www.cs.umd.edu/hcil/trs/2012-06/2012-06.pdf