Patient Representation using Autoencoder

Cinyoung Hur
2018.7.7

I’m

Senior Researcher, Data Engineer
working at linewalks

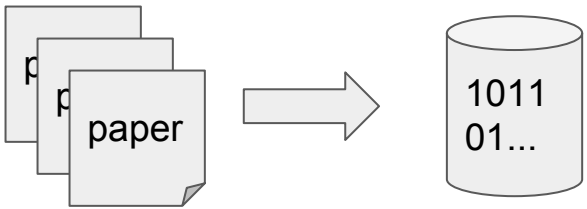
Co-organizer of 

Overview

- What is Electronic Health Record
- How to represent medical concepts and patient information
- Demo!

Electronic Health Records

- Patient Registry
- Disease Registry
- Drug Registry



80% adoption

- Electronic Health Record
  - A collection of health data from all clinicians involved in a patient's care
  - more patient-centric, powerful and useful for diagnosis and treatment

EHR data types

- Demographics
- Encounters
- Diagnosis
- Procedures
- Physical exams
- Sensor measurements
- Laboratory test results
- Prescribed or administered medications
- Clinical notes

EHR data types

- Numerical quantities
  - body mass index, height, blood pressure
- Datetime
  - data of birth, time of admission
- Categorical value
  - ethnicity, codes from controlled vocabularies
- Natural language free-text
  - progress notes, discharge summaries
- Derived time series
  - vital sign signals during the course of the operation

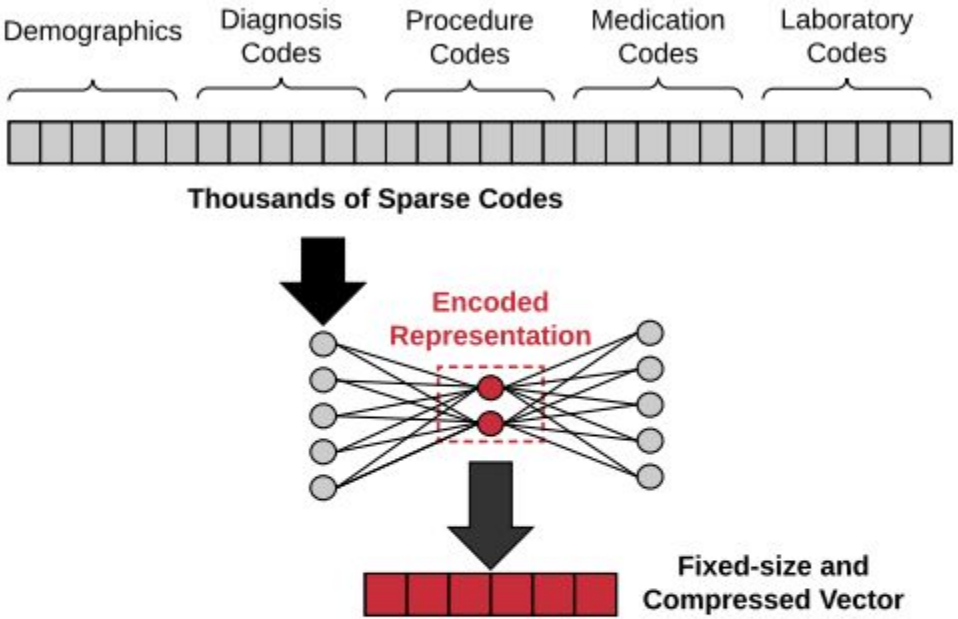
Example Classification Schema

- Demographics
- Encounters
- **Diagnosis***
- **Procedures***
- Physical exams
- Sensor measurements
- **Laboratory test results***
- **Prescribed or administered medications***
- Clinical notes

Schema	Number of Codes	Examples
ICD-10 *(Diagnosis)*	68,000	- J9600: Acute respiratory failure - I509: Heart failure - I5020: Systolic heart failure
CPT *(Procedures)*	9,641	- 72146: MRI Thoracic Spine - 67810: Eyelid skin biopsy - 19301: Partial mastectomy
LOINC *(Laboratory)*	80,868	- 4024-6: Salicylate, Serum - 56478-1: Ethanol, Blood - 3414-0: Buprenorphine Screen
RxNorm *(Medications)*	116,075	- 161: Acetaminophen - 7052: Morphine - 1819: Buprenorphine

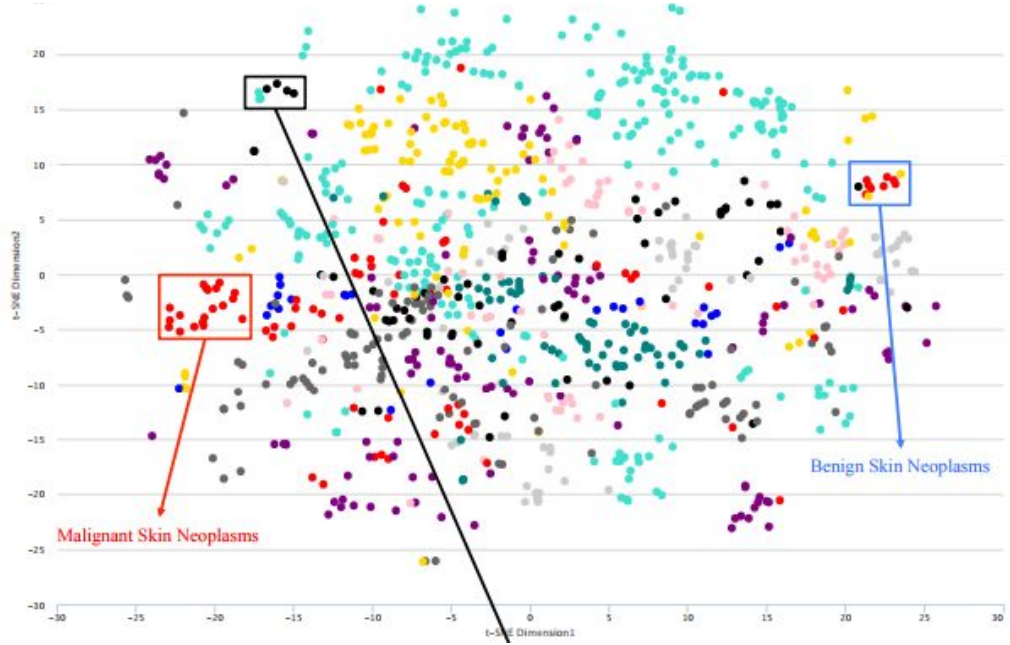
EHR Representation Learning

- Project discrete codes into vector space



EHR Representation Learning
: Concept Representation

- Evaluation of Medical Concept Representation Learning



https://arxiv.org/pdf/1602.03686.pdf

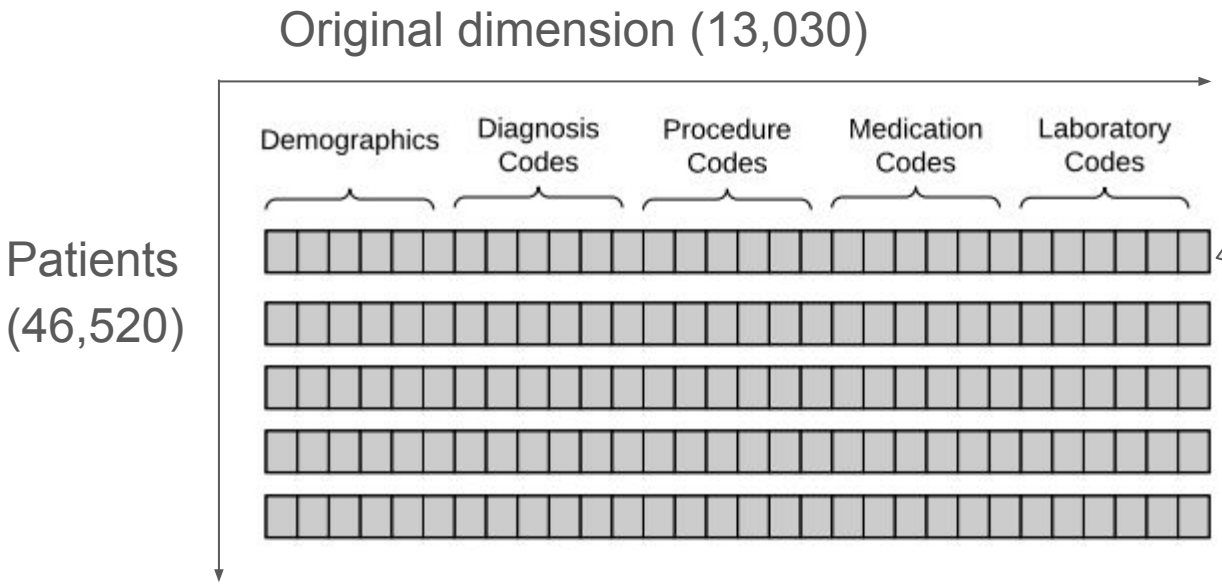
Demo!

MIMIC III dataset

- Medical Information Mart for Intensive Care III
- a large, freely-available database comprising de-identified health-related data
- over forty thousand patients who stayed in critical care units in the Beth Israel Deaconess Medical Center between 2001 and 2012.

https://github.com/MIT-LCP/mimic-code

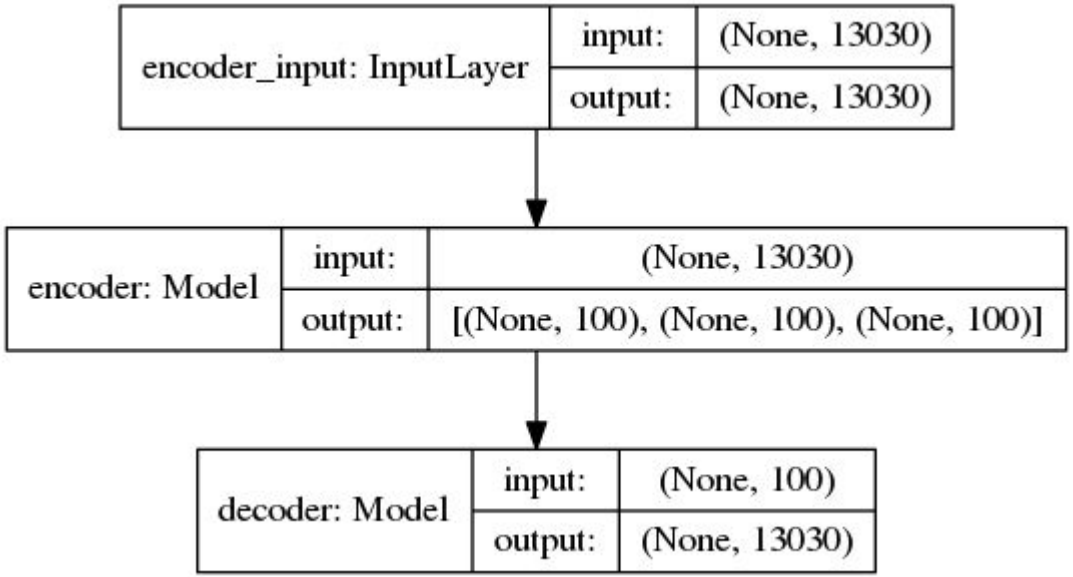
MIMIC III Preprocessing



- One-hot-encoded all codes in each visit
- Sum by patient id
- Normalization

Variational Autoencoder

- Keras
  - original_dim = 13030
  - intermediate_dim = 2048
  - batch_size = 100
  - latent_dim = 100
  - epochs = 10



We are looking for colleagues! Join us.

linewalks

Linewalks is a healthcare big data startup that develops data services using machine learning and visualization techniques in large-scale data processing systems.

2016. Angel investing from KITE Entrepreneurship Foundation
2018. Series A investing from Kakao Investment (19 billion won)

- Positions
  - Frontend Developer
  - Backend Developer
  - Machine Learning Developer
- For more information, visit http://bit.ly/lw-tech-recruit

Thanks!