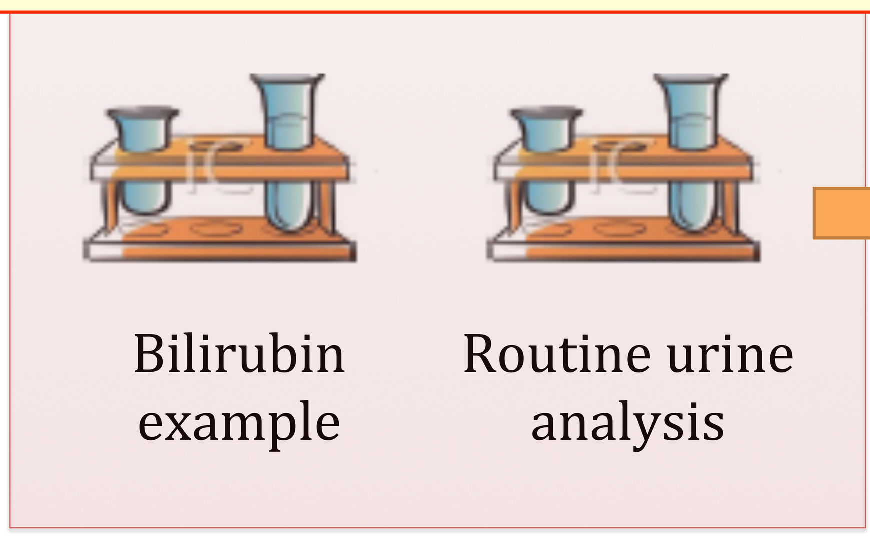
Forecasting Potential Diabetes Complications

Yang Yang, Walter Luyten, Lu Liu, Marie-Francine Moens, Jie Tang, Juanzi Li

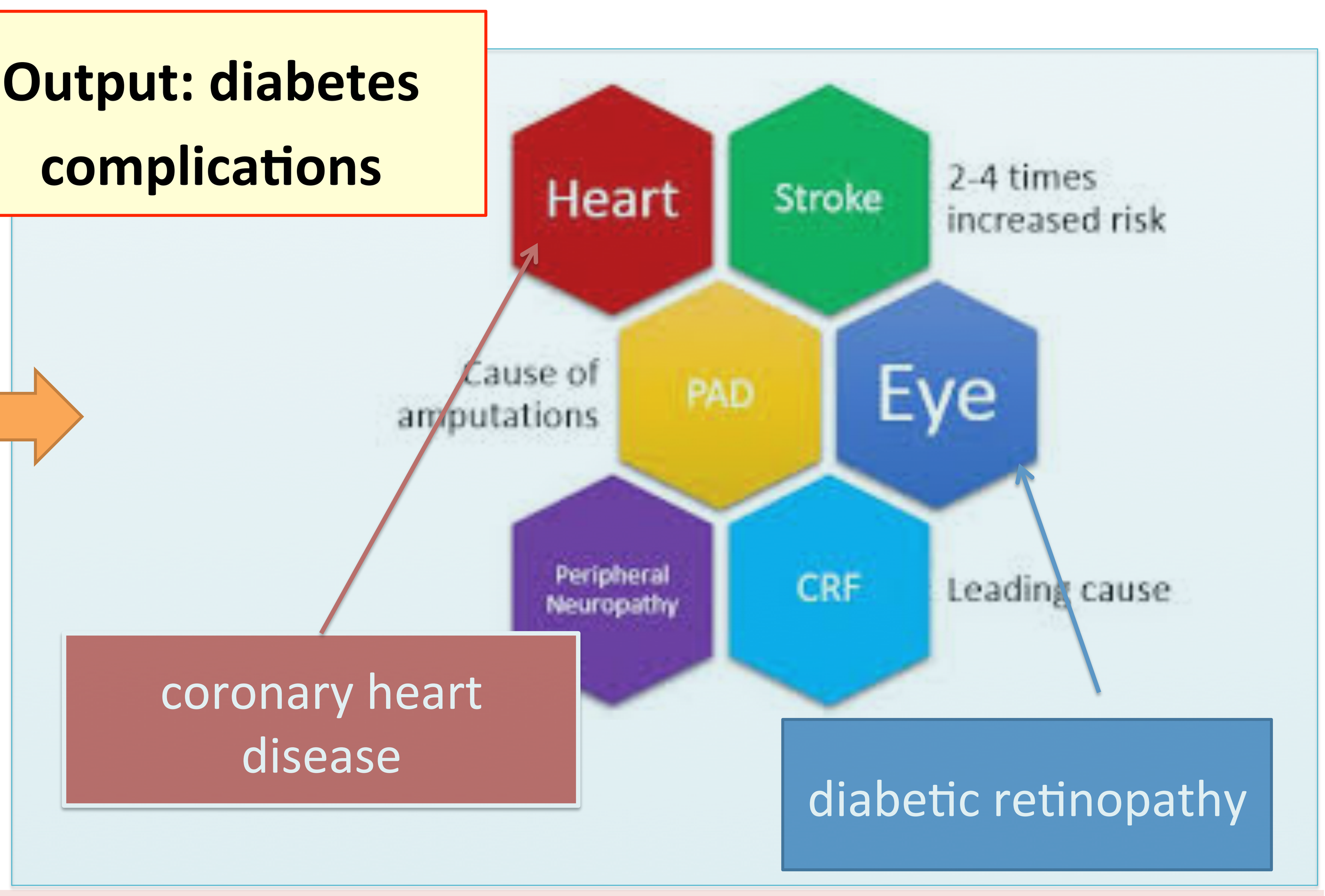
Tsinghua Univeristy, Katholieke Universiteit Leuven, Northwestern University

- **Diabetes are major causes of early death in most countries and over 68% of diabetes-related mortality is caused by diabetes complications.**
- **471 billion USD were spent on healthcare for 371 million diabetes patients world-wide in 2012, still 4.8 million people died in 2012 due to diabetes.**

Input: a patient's lab test results



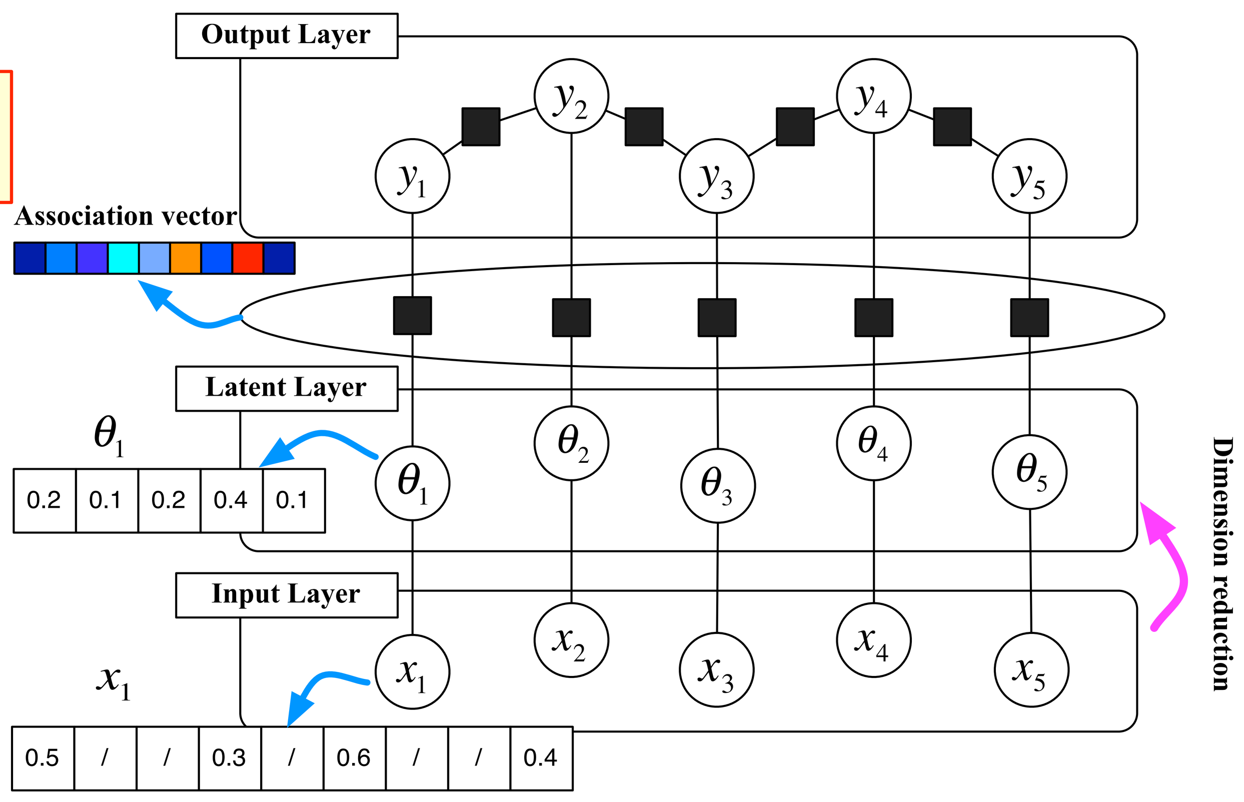
Output: diabetes complications



Proposed Model

Key challenge: *feature sparseness*

- **Averagely each clinical record only contains 1.26% of lab tests on average.**
- **65.5% types of lab tests are recorded in less than 0.0054% of clinical records.**



The model alleviates the sparseness issue by projecting feature space into a low-dimensional latent space.

Model the joint distribution of a given set of lab tests X over complication labels Y as

$$P(y_n|\theta_n, x_n) = P(y_n|\theta_n)\prod_l \left(\sum_{k=1}^{K} \theta_{nk} \cdot \Omega_{x_{nl}k}\right)$$

The feature factor is defined as

$$P(y_n|\theta_n) = \frac{1}{Z_1}\exp\{\alpha \cdot f(\theta_n, y_n)\}$$

The correlation factor is defined as

$$P(y_n, y_{n'}) = \frac{1}{Z_2}\exp\{\beta \cdot g(y_n, y_{n'})\}$$

The correlation factor is defined as

$$\Omega_{x_{nl}k} = \begin{cases} N(x_{nl}|\mu_{kl}, \delta_k) & x_{nl} \text{ is numerical} \\ \phi_{klx_{nl}} & x_{nl} \text{ is categorical} \end{cases}$$

Forecasting Performance

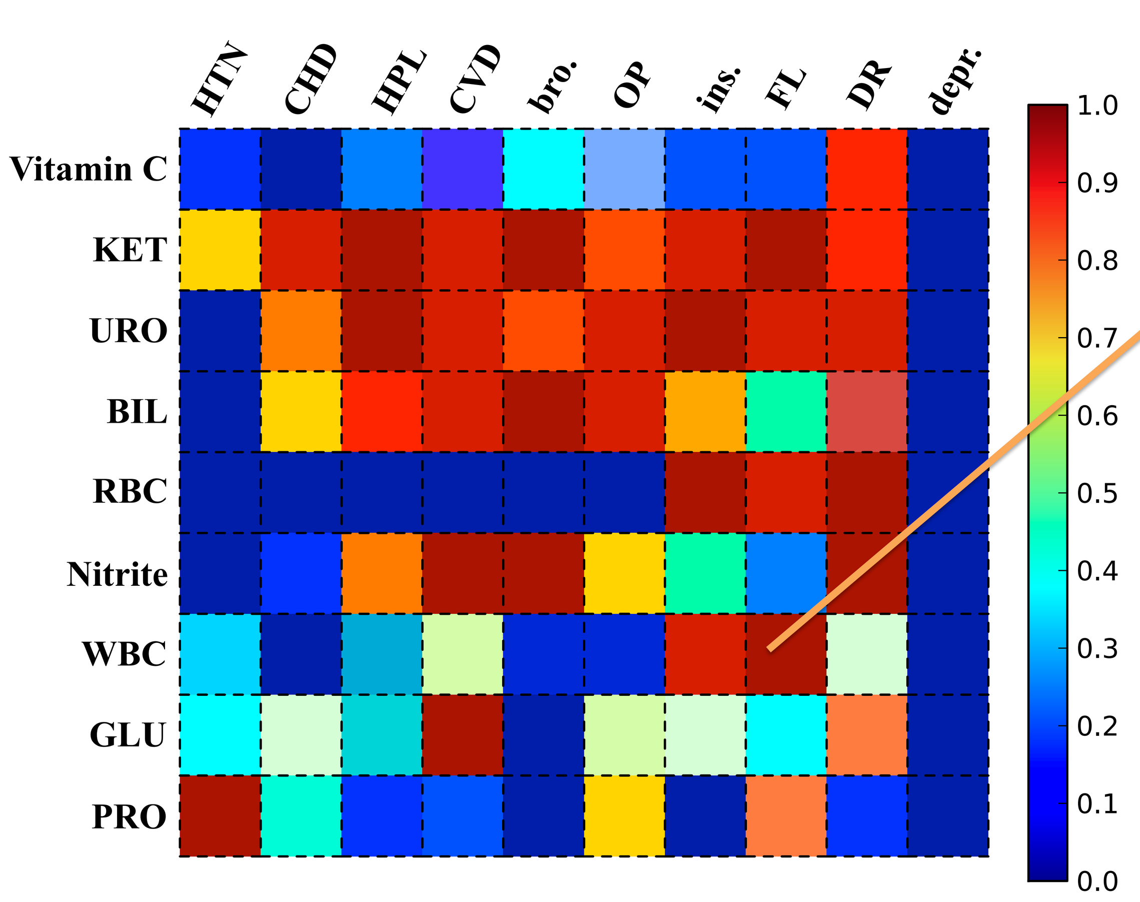
- **Dataset: a collection of real medical records from a famous geriatric hospital:**
  - **181,933 medical records, 35,525 patients and 1,945 kinds of lab tests.**
- **On average each clinical record contains 24.43 lab tests (1.26% of all lab tests).**
- **We consider 3 complications:**
  - **hypertension (HTN), coronary heart disease (CHD), hyperlipidemia (HPL).**

Table 2: Performance of diabetes complication forecasting.

Complication	Method	Precision	Recall	F1
HTN	SVM	0.3804	0.4789	0.4241
	FGM	0.5666	0.4959	0.5075
	FGM+PCA	0.5741	0.3284	0.4178
	SparseFGM	0.4714	0.6319	**0.5400**
CHD	SVM	0.2132	0.0636	0.0980
	FGM	0.6264	0.1369	0.2247
	FGM+PCA	0.2425	0.8367	0.3761
	SparseFGM	0.2522	0.7972	**0.3832**
HPL	SVM	0.2208	0.0460	0.0761
	FGM	0.6557	0.0591	0.1084
	FGM+PCA	0.2047	0.8035	0.3262
	SparseFGM	0.2796	0.8396	**0.4195**

- **SVM and traditional factor graph model suffer from the feature sparseness (-59.9% compared with our method by recall)**
- **PCA improve the performance. However, it separates sparse coding and classification into two processes.**
- **Our approach (SparseFGM) outperforms the baselines 20% by F1**

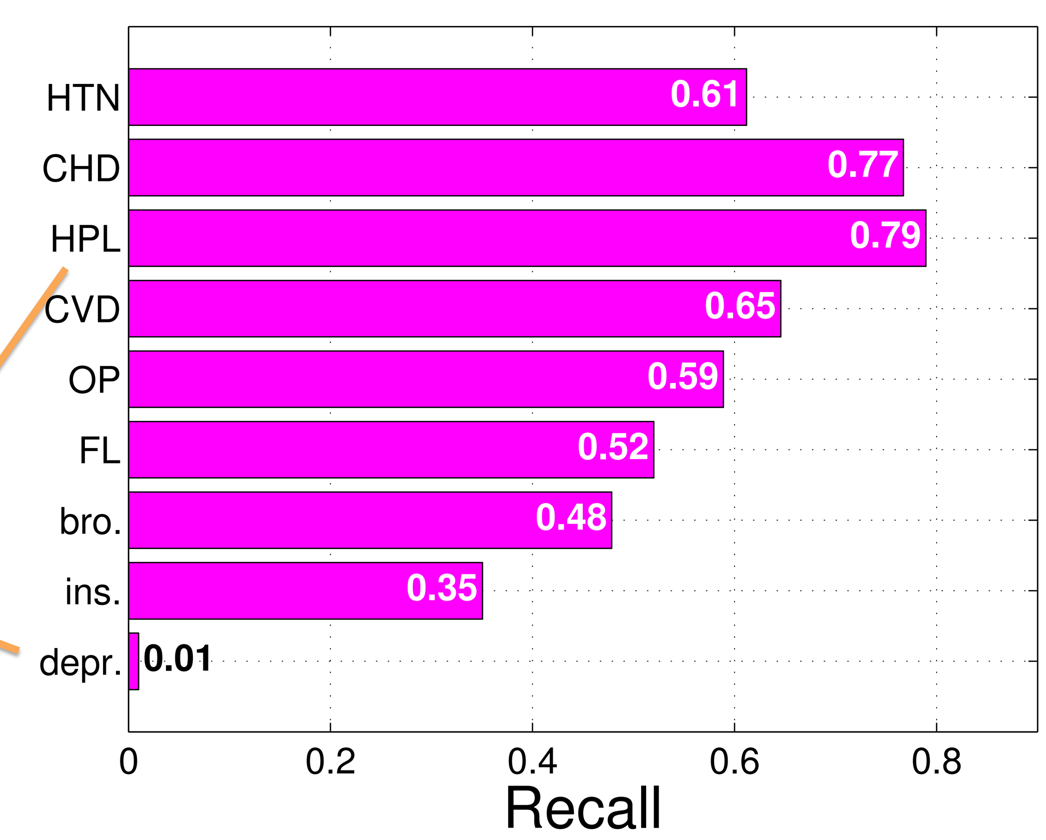
Association Analysis



Micro-level: association between 10 complications and 9 parameters of routine urine anlaysis discovered by the proposed mode.

WBC in the urine typically is found in urinary tract infections which cause frequent voiding, which causes insomnia.

Hyperlipidemia (HPL) can be diagnosed more precisely, while depression (depr.) is usually recognized from psychological investigation instead of physiological lab tests.



Macro-level: study how each complication is diagnosable from lab test results.

Contact: SherlockBourne@gmail.com Code & Data: http://aminer.org/diabetes Tsinghua University