Phenotyping via Bayesian Nonparametric Tensor Factorization using Markov Chain Monte Carlo

Hanjun Dai
Yiting Xiao
Yue Peng

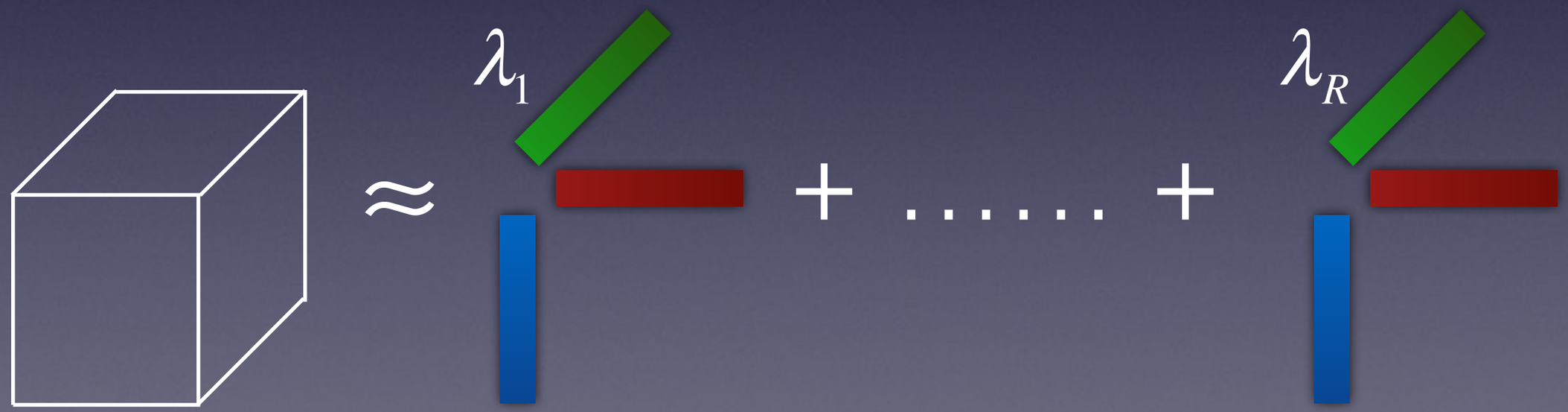
- Problem Definition
- Challenges
- Model & Solution
- Data
- Preliminary Results
- Future Work

Problem Definition

- Do phenotyping via tensor factorization

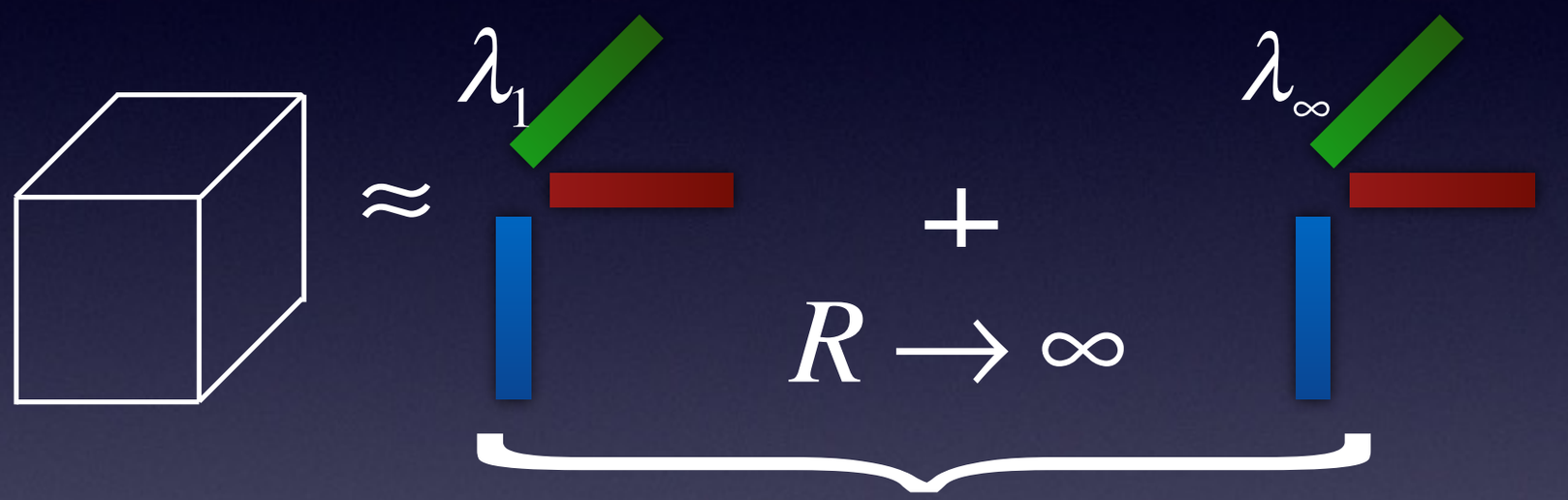
Problem Definition

- Do phenotyping via tensor factorization
  - R: number of phenotypes
  - Each vector: multinomial distribution



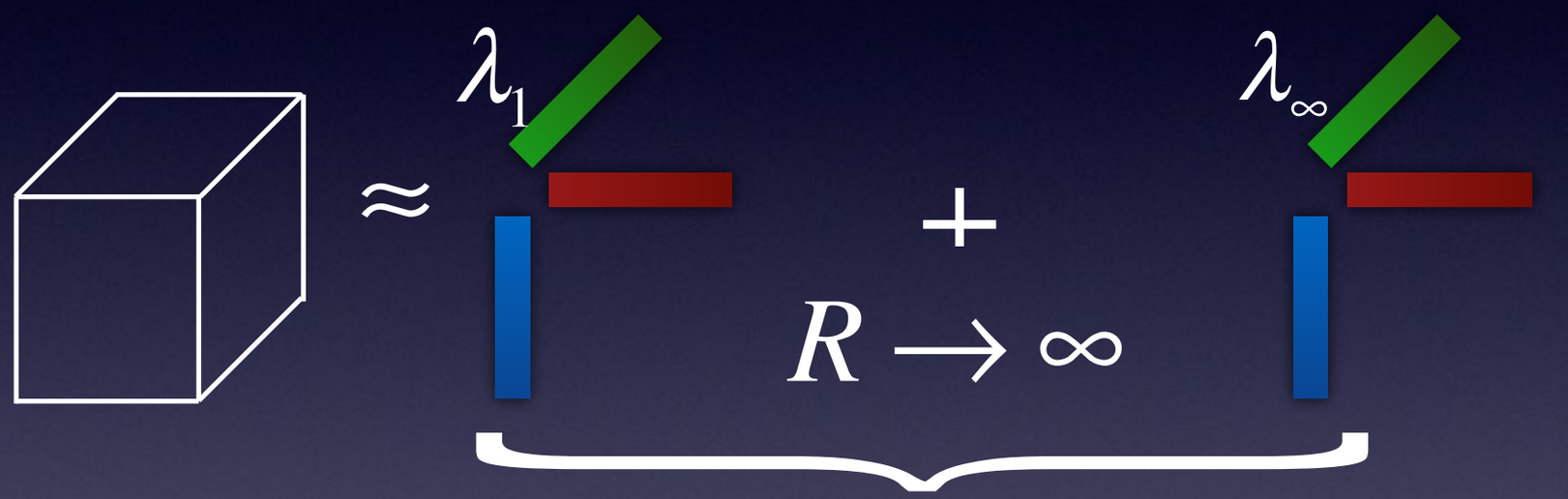
Problem Definition

- Automatically learn ‘R’ from data



Problem Definition

- Automatically learn ‘R’ from data



- Treat parameters in a Bayesian way

Problem Definition

- Possible Advantages:

Problem Definition

- Possible Advantages:
  - Find ‘R’ which fits the data best
  - Averaging over all the possibilities
  - Prevent over fitting

Challenges

- How to model the problem in a probabilistic way
- How to solve the model efficiently
- How to deal with large scale tensor data (future work)

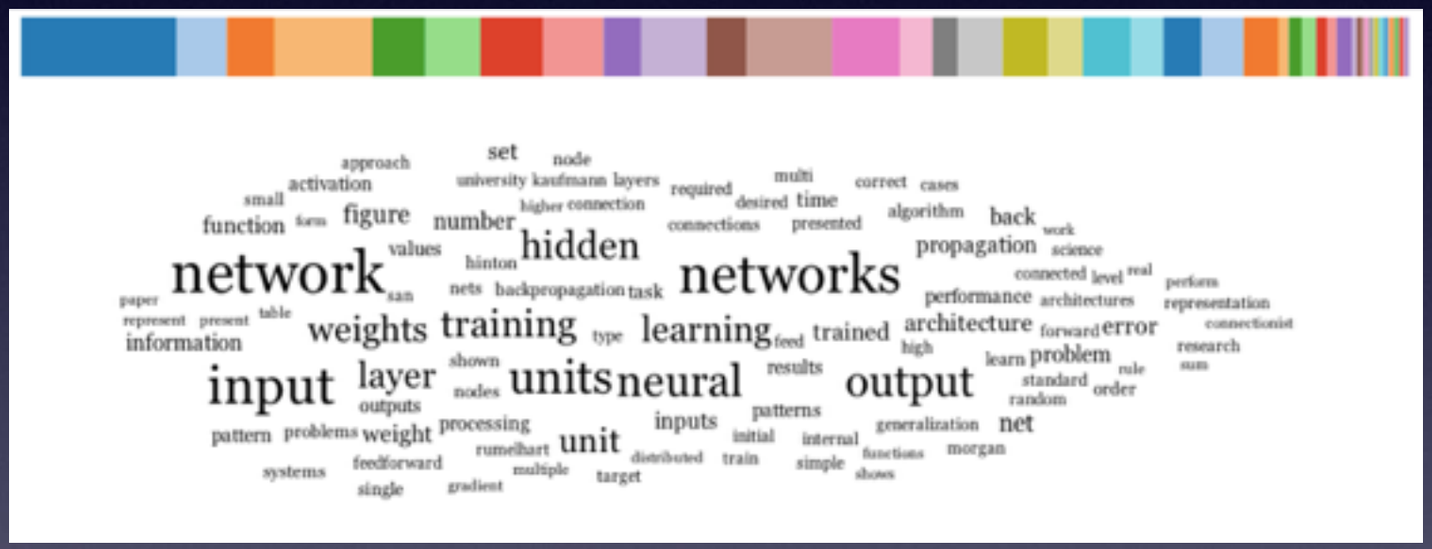
Model & Solution

- How to model the problem in a probabilistic way

Model & Solution

- How to model the problem in a probabilistic way

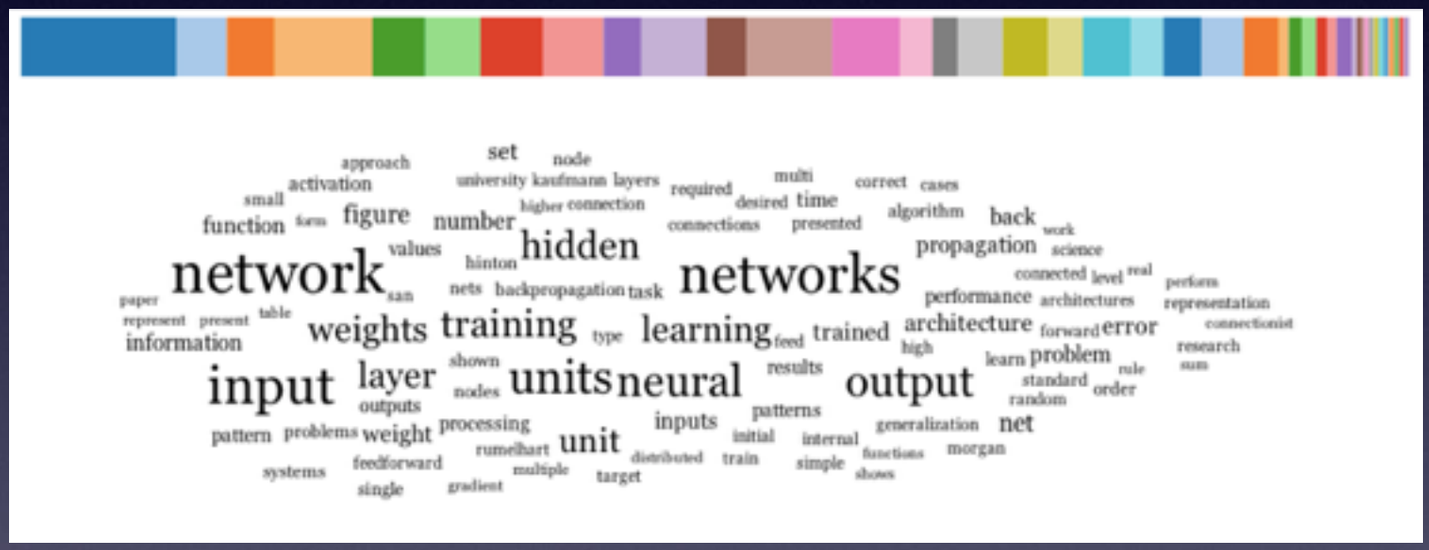
Inspired by LDA:



Model & Solution

- How to model the problem in a probabilistic way

Inspired by LDA:



Topic-Words Distribution ↔ Phenotype-Medication Distribution

topic of med

Model & Solution

- **Dirichlet Process over λ**

$$\lambda_k = \beta_k \prod_{i=1}^{k-1} (1-\beta_i)$$

$$\beta_i \sim Beta(1,\alpha)$$

Model & Solution

- **Dirichlet Process over λ**

$$\lambda_k = \beta_k \prod_{i=1}^{k-1} (1-\beta_i)$$

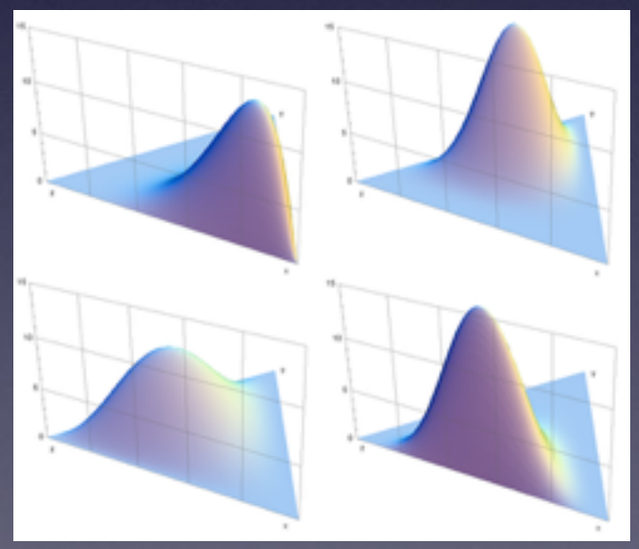
$$\beta_i \sim Beta(1,\alpha)$$

- **Dirichlet Prior over each multinomial distribution**

$$Patient\ \ Column \sim Dir(\gamma_p = 1/\#patients)$$

$$Medicine\ Column \sim Dir(\gamma_m = 1/\#medicine)$$

$$Diagnosis\ Column \sim Dir(\gamma_d = 1/\#diagnosis)$$



Model & Solution

- **Dirichlet Process over λ**

$$\lambda_k = \beta_k \prod_{i=1}^{k-1} (1-\beta_i)$$

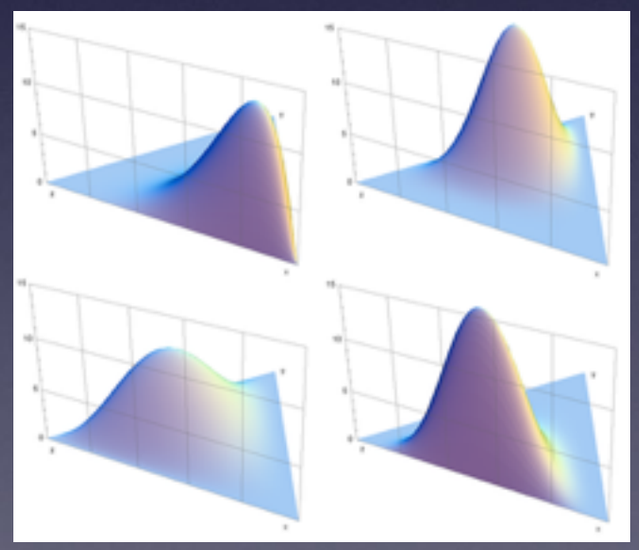
$$\beta_i \sim Beta(1,\alpha)$$

- **Dirichlet Prior over each multinomial distribution**

$$Patient\ \ Column \sim Dir(\gamma_p = 1/\#patients)$$

$$Medicine\ Column \sim Dir(\gamma_m = 1/\#medicine)$$

$$Diagnosis\ Column \sim Dir(\gamma_d = 1/\#diagnosis)$$



- **Likelihood:**

$$p(T_{i,j,k} \mid \alpha,\gamma_p,\gamma_m,\gamma_d) = \sum_{r=1}^{\infty} p(r \mid \alpha)\, p(patient_i \mid r,\gamma_p)\, p(med_j \mid r,\gamma_m)\, p(diag_k \mid r,\gamma_d)$$

Model & Solution

- How to solve the model efficiently

Model & Solution

- How to solve the model efficiently
  - **Truncation technique:**
    - The CP-rank of tensor is limited
  - **Inference: Sampling from posterior distribution**
    - $p(patient_i \mid r, \gamma_p)$
    - $p(med_i \mid r, \gamma_m)$
    - $p(diag_i \mid r, \gamma_d)$

  - **Sequential MC:**
    - Do sampling in an online (stochastic) way

Data

- Data set: MIMIC2
- Total number of diagnostics records: 314,648
- Total number of medication records: 4,206,941
- 5675 distinct ICD9 code, 63 distinct medicines

Data

- Data set: MIMIC2
- Total number of diagnostics records: 314,648
- Total number of medication records: 4,206,941
- 5675 distinct ICD9 code, 63 distinct medicines

Diagnostics

subject_id	code	date	binary
7	V30.01	2666-06-22 00:00:00	1

Medications

subject_id	med_name	date	boolean
6	Heperin	3389-07-08 09:43:00	1

Data

How to pair up diagnosis with medicine?

Data

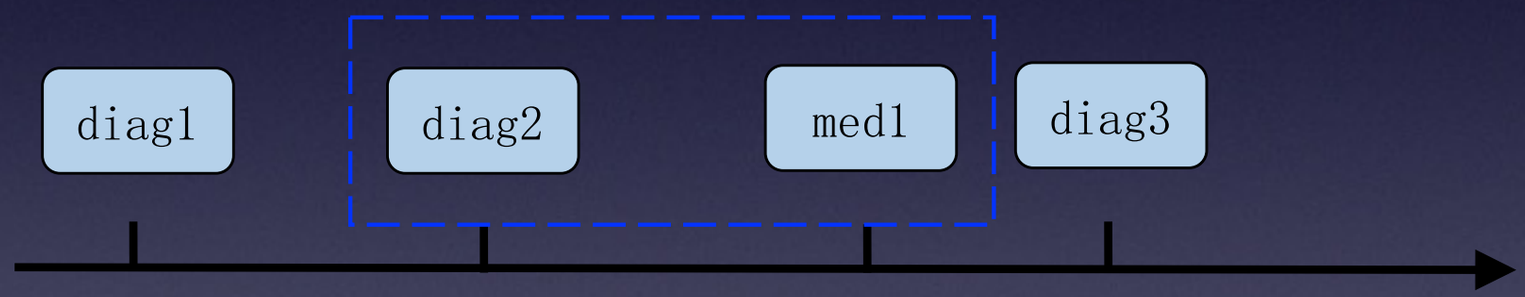
How to pair up diagnosis with medicine?

- Most recent diagnosis prior to medication record
- Set threshold to 7

Data

How to pair up diagnosis with medicine?

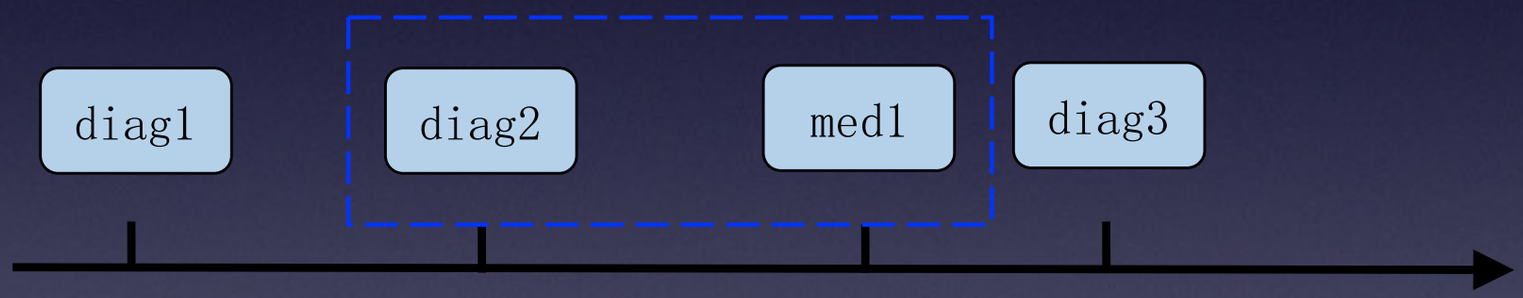
- Most recent diagnosis prior to medication record
- Set threshold to 7



Data

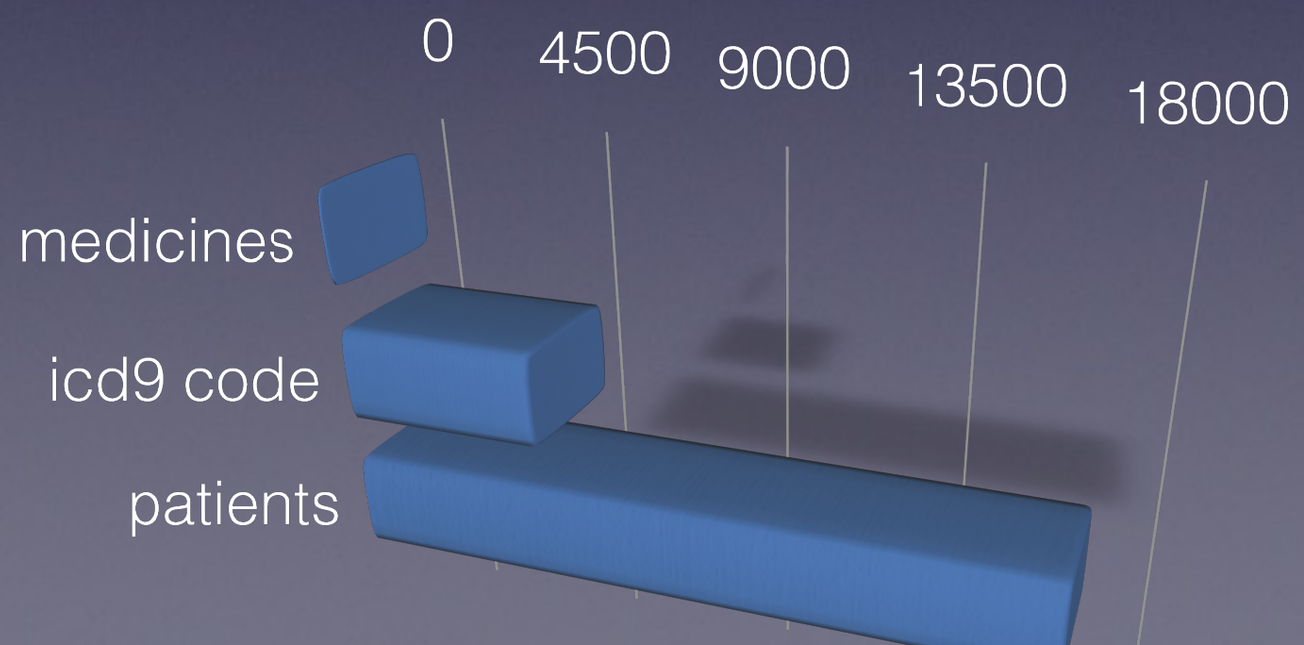
How to pair up diagnosis with medicine?

- Most recent diagnosis prior to medication record
- Set threshold to 7



Result tensor:

- 16619 patients;
- 4729 ICD9 code;
- 61 medicines;



Preliminary Results

- Phenotype Study: MIMIC2

Preliminary Results

- Phenotype Study: MIMIC2

Phenotype #1:

Top 5 medcine:	**diagnosis:**
Neosynephrine-k	414.01
Propofol	272.4
Insulin	413.9
Precedex	401.9
Epinephrine-k	410.71

Phenotype #2:

Top 5 medcine:	**diagnosis:**
Propofol	414.01
Levophed-k	410.71
Fentanyl	412
Nitroprusside	428.0
Dopamine	272.0

Preliminary Results

- Phenotype Study: MIMIC2

Phenotype #1:

Top 5 medcine:	**diagnosis:**
Neosynephrine-k	414.01
Propofol	272.4
Insulin	413.9
Precedex	401.9
Epinephrine-k	410.71

anesthetic related

Phenotype #2:

Top 5 medcine:	**diagnosis:**
Propofol	414.01
Levophed-k	410.71
Fentanyl	412
Nitroprusside	428.0
Dopamine	272.0

pain reliever related

Preliminary Results

- Synthetic experiment: factorizing black-white image

Preliminary Results

- Synthetic experiment: factorizing black-white image

Original Binary

Preliminary Results

- Synthetic experiment: factorizing black-white image

Original Binary

Recovered Prob

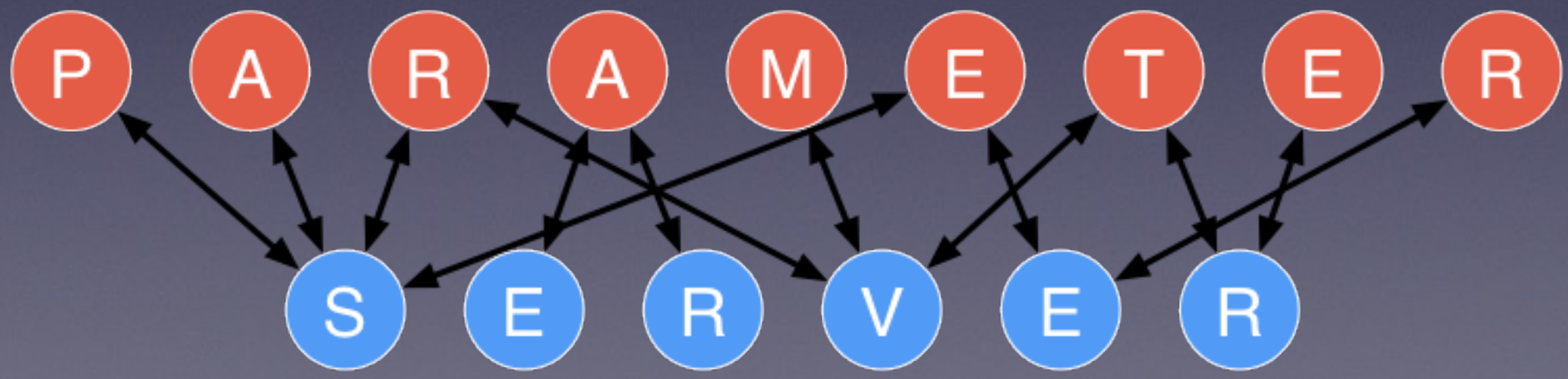
Recovered Binary

Future Work

- Deploy the algorithm in Spark system

Future Work

- Deploy the algorithm in Spark system
- In the future:
  - using stochastic variational inference (more efficient)
  - using data parallelism (downpour SGD)



Thanks!

Q&A