Phenotyping from Electronic Health Records using Tensor Factorization

Jimeng Sun

jsun@cc.gatech.edu

Georgia Tech College of Computing
Computational Science and Engineering

PHENOTYPING FROM EHR WITH TENSOR FACTORIZATION

Project: SCH: INT: Collaborative Research: High-throughput Phenotyping on Electronic Health Records using Multi-Tensor Factorization

Principal Investigators

Jimeng Sun
Associate Professor
College of Computing
Georgia Tech

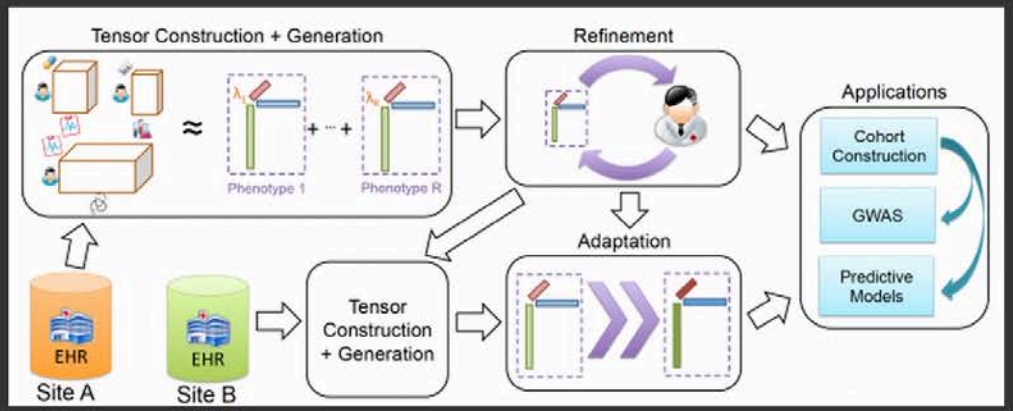
Bradley Malin
Associate Professor
Biomedical Informatics
and Computer Science
Vanderbilt University

Joshua Denny
Associate Professor
Biomedical Informatics
and Medicine
Vanderbilt Utnvierty

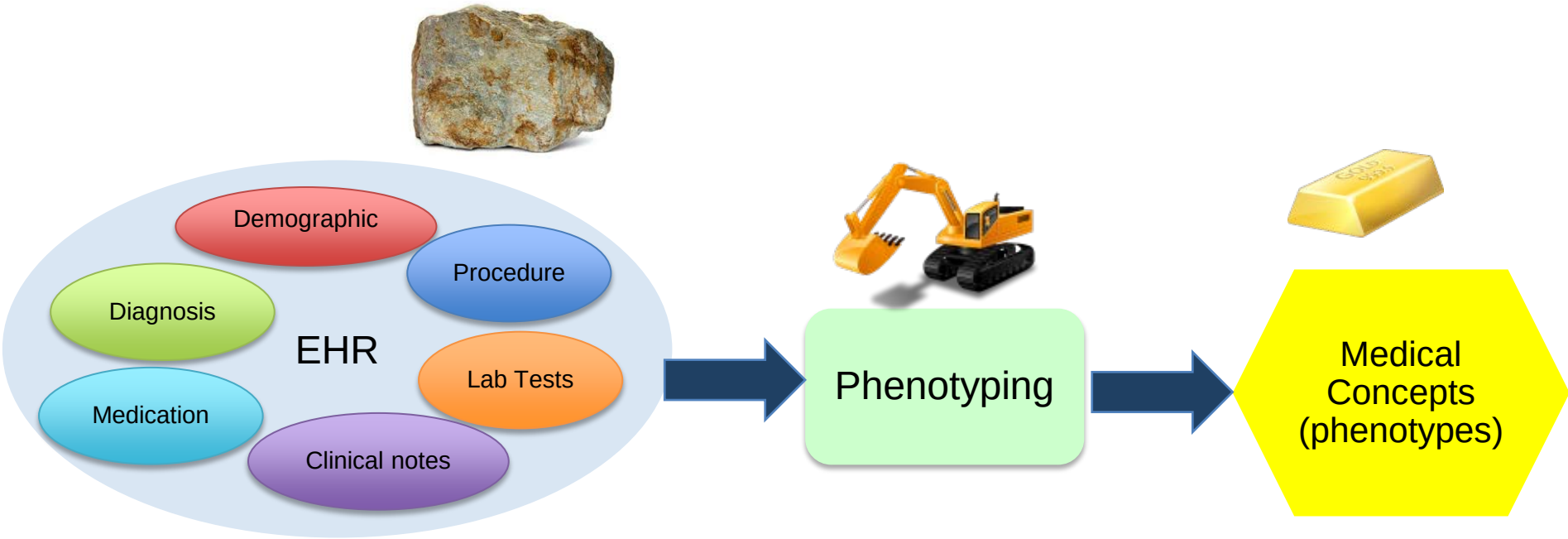
Joydeep Ghosh
Professor
Electrical & Computer
Engineering
Univ of Texas, Austin

Abel Kho
Assistant Professor
Medicine - Biomedical
Informatics
Northwestern Univ

Funding Source: NSF Smart Connect Health Integrated Grant: Award Number 1418511



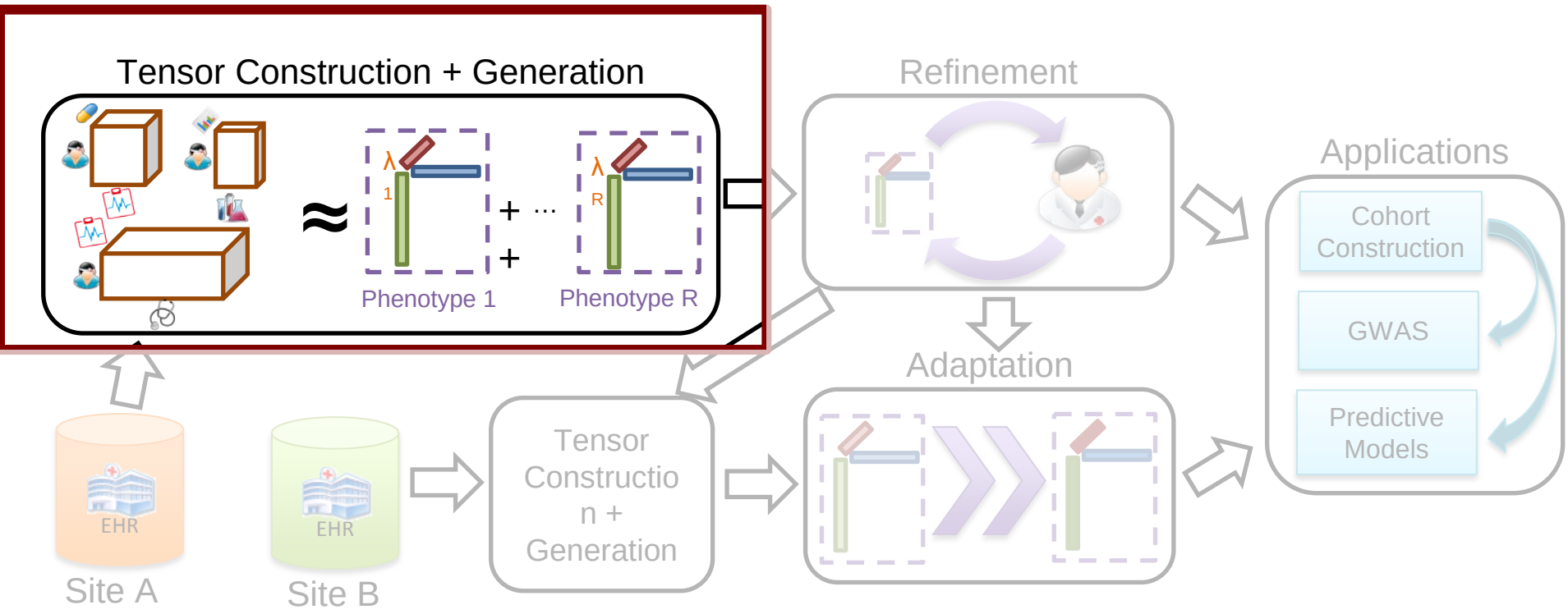
Phenotyping from Electronic Health Records



- Limitations of existing phenotyping methods
  - Labor intensive
  - Unable to discover novel phenotypes

Our Project on Phenotyping using Tensor Factorization

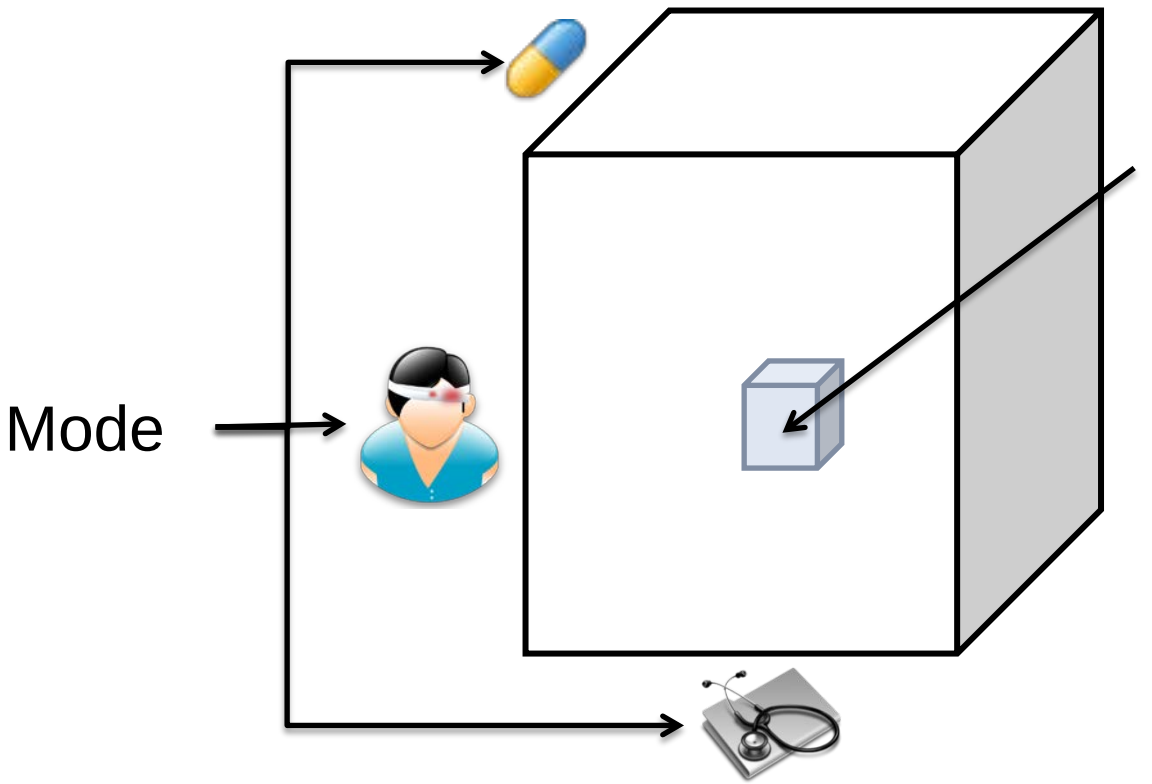
This talk



- NSF SCH INT project between Georgia Tech, Vanderbilt, UT Austin, Northwestern

Constructing Feature Tensor

- Tensor is a generalization of matrix
  - Matrix is a 2nd order tensor
- Tensors can better capture interactions among concepts



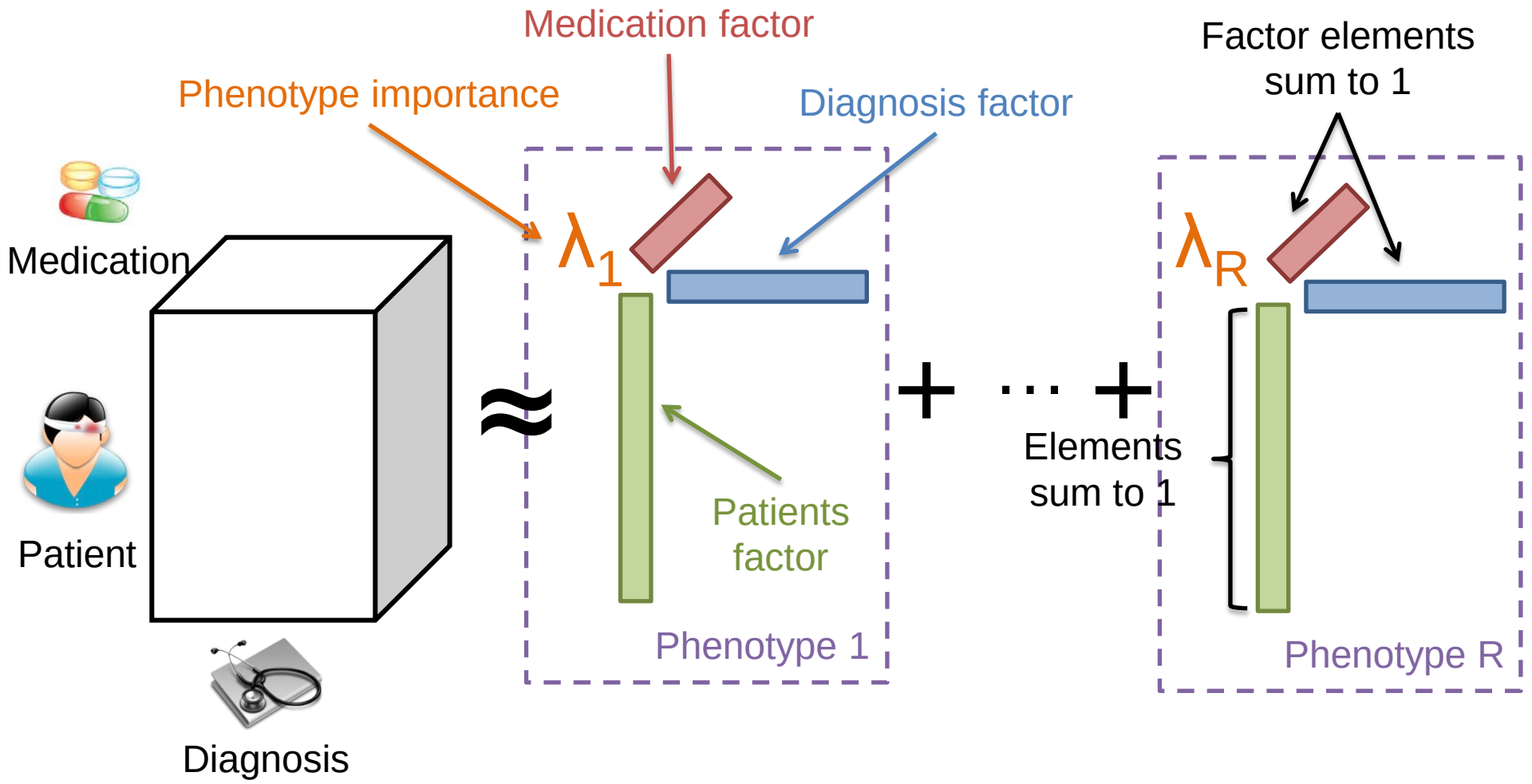
Data element types:

- Binary
- Count (integer)
- Continuous (numeric)

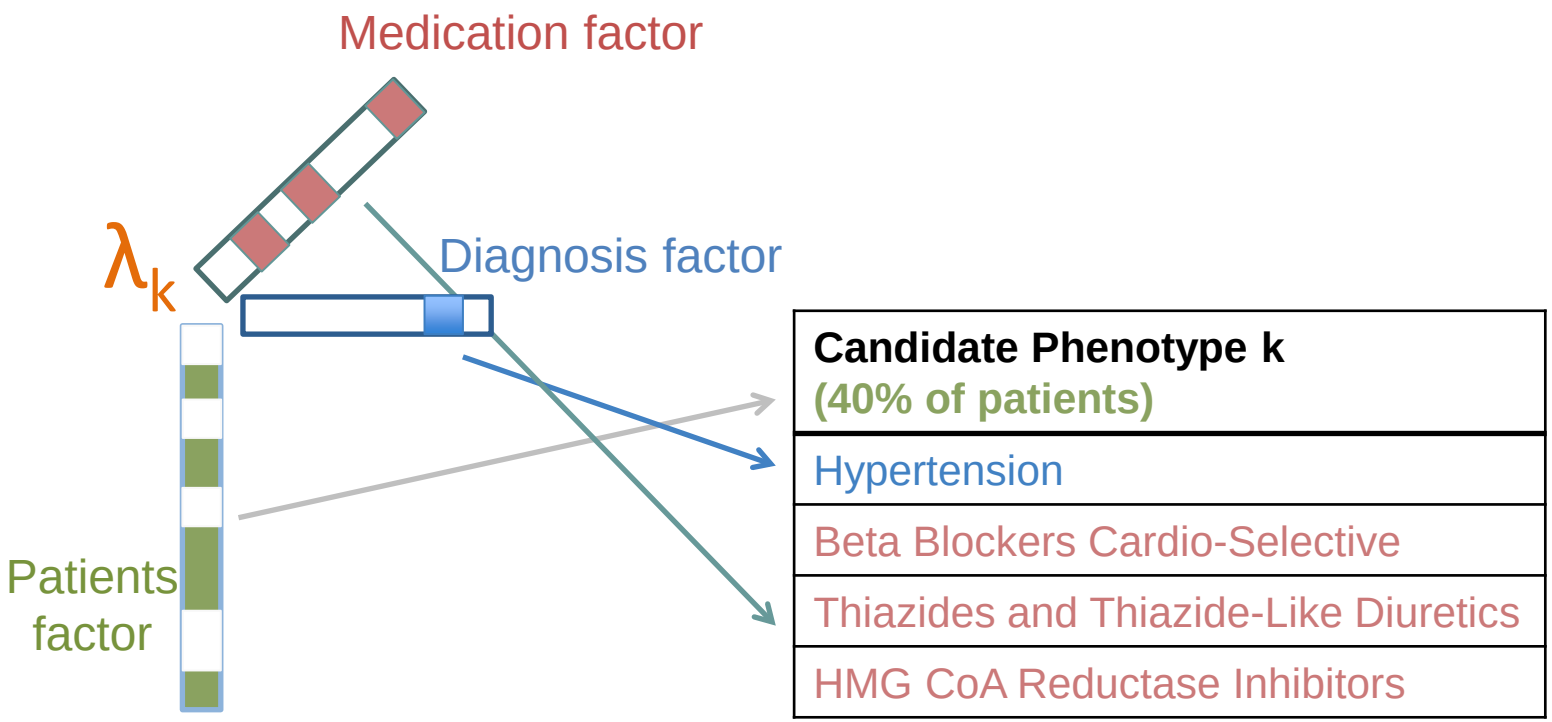
Limestone

Ho, Joyce C., Joydeep Ghosh, Steve R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. “Limestone: High-Throughput Candidate Phenotype Generation via Tensor Factorization.” Journal of Biomedical Informatics. 2014

Phenotyping through Tensor Factorization



Example Phenotype



a phenotype = a group of patients that share common characteristics (e.g. diagnosis, medication)

Limestone

- Nonnegative input tensor
- Nonnegative constraints
- Stochastic column constraints on factor matrices
- Hard thresholding on elements in factor matrices

$$\min f(\mathcal{M}) \equiv \min \sum_{\mathbf{i}} [m_{\mathbf{i}} - x_{\mathbf{i}} \log m_{\mathbf{i}}]$$

$$\text{s.t } \mathcal{M} = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \ldots; \mathbf{A}^{(N)}]\!] \in \Omega$$

$$\Omega = \Omega_\lambda \times \Omega_1 \times \cdots \times \Omega_N$$

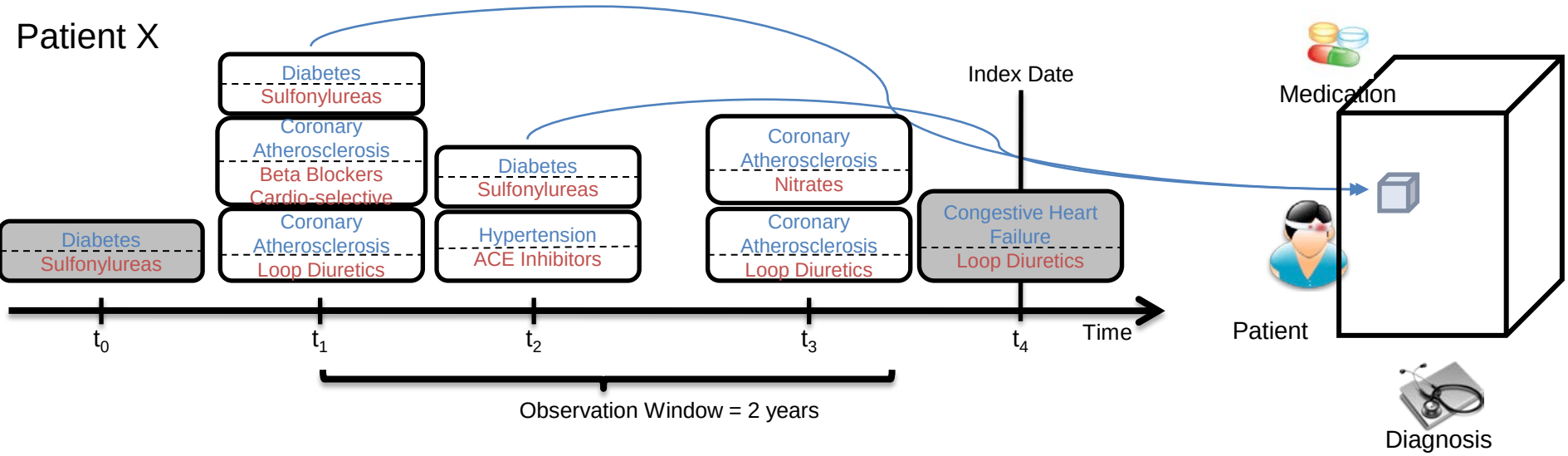
$$\Omega_\lambda = [0, +\infty)^R$$

$$\Omega_{An} = \{\mathbf{A} \in \{0, [\gamma_n, 1]\}^{I_{n \times R}} \mid ||\mathbf{a}_{:r}||_1 = 1 \quad \forall r\}$$

Hard thresholding constraints

Ho, Joyce C., Joydeep Ghosh, Steve R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. "Limestone: High-Throughput Candidate Phenotype Generation via Tensor Factorization." *Journal of Biomedical Informatics*. 2014

Quantitative Experiment Setup



- Medication orders from Geisinger dataset
- Diagnosis codes aggregated into HCC codes
- Medications are defined as pharmacy subclass
- 31,816 patients x 169 diagnoses x 471 medications

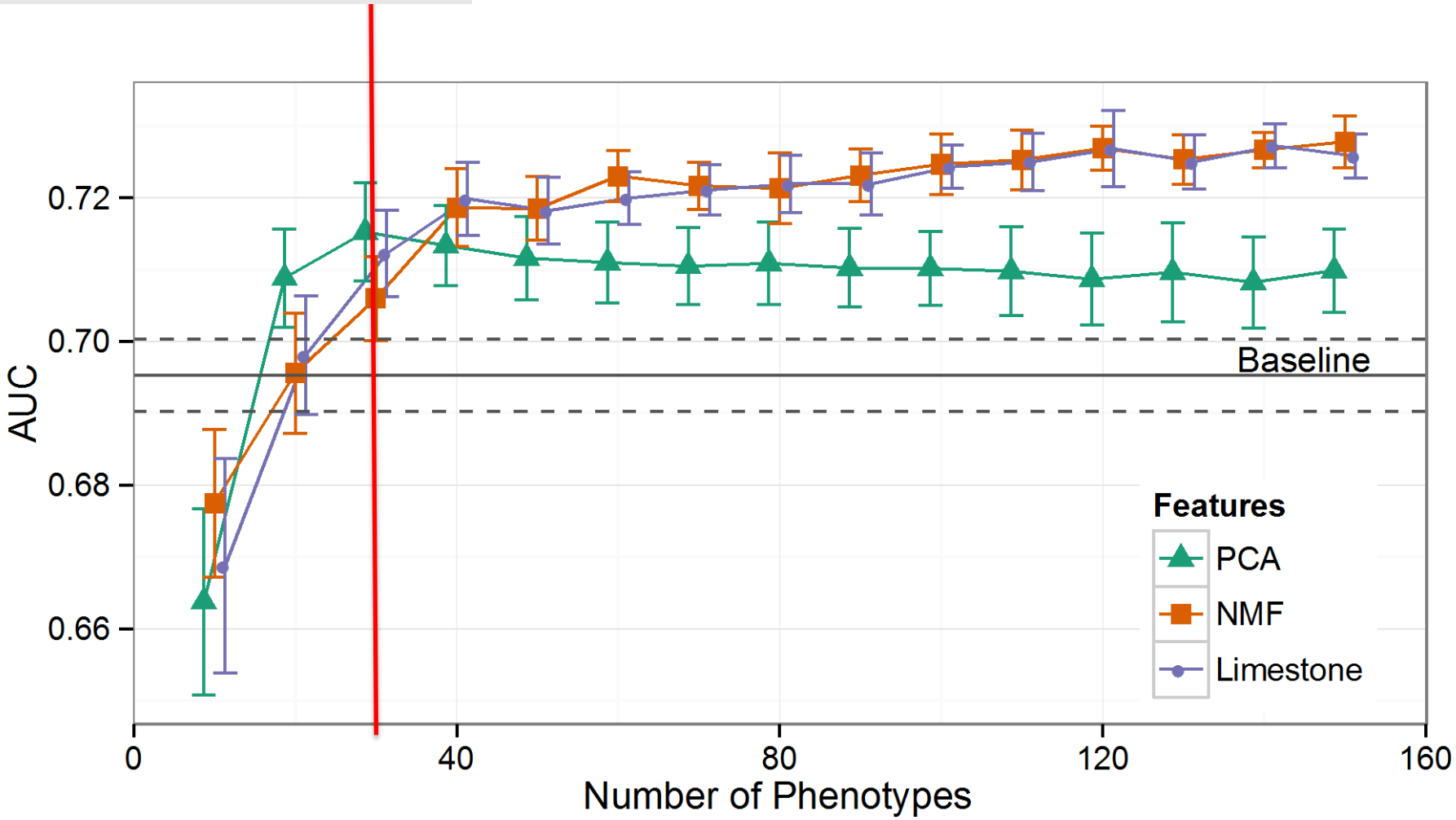
Quantitative Evaluation: HF Prediction

- Task: predict patients with heart failure (HF)
- Model: logistic regression with ℓ_1 regularization
- 10 random even splits of the dataset (50% training)

- **Comparison methods** for feature construction:
  1. Baseline using source independence matrix
  2. Principal Component Analysis (PCA)
  3. Nonnegative Matrix Factorization (NMF)
  4. **Limestone**

Predictive Performance

Small # of features outperforms 640 features



Qualitative Evaluation: Major disease phenotypes can be identified

Uncomplicated Diabetes

Phenotype 3 (17.6% of patients)
Diabetes with No or Unspecified Complications
Sulfonylureas
Biguanides
Diagnostic Tests
Insulin Sensitizing Agents
Diabetic Supplies
Meglitinide Analogues
Antidiabetic Combinations

Mild Hypertension

Phenotype 4 (31.1% of patients)
Hypertension
ACE Inhibitors
Thiazides and Thiazide-Like Diuretics

Chronic Respiratory Inflammation/Infection

Phenotype 5 (36.7% of patients)
Other Ear, Nose, Throat, and Mouth Disorders
Viral and Unspecified Pneumonia, Pleurisy
Significant Ear, Nose, and Throat Disorders
Cough/Cold/Allergy Combinations
Azithromycin
Fluoroquinolones
Sympathomimetics
Penicillin Combinations
Antitussives
Glucocorticosteroids
Tetracyclines
Anti-infective Misc. - Combinations
Clarithromycin
Cephalosporins - 2nd Generation
Cephalosporins - 1st Generation
Expectorants

Qualitative Evaluation: Disease subtypes can be identified

Mild Hypertension

Phenotype 4 (31.1% of patients)
Hypertension
ACE Inhibitors
Thiazides and Thiazide-Like Diuretics

Moderate Hypertension

Phenotype 2 (31.5% of patients)
Hypertension
Beta Blockers Cardio-Selective
Angiotensin II Receptor Antagonists
Loop Diuretics
Potassium
Nitrates
Alpha-Beta Blockers
Vasodilators

Severe Hypertension

Phenotype 6 (24.3% of patients)
Hypertension
Calcium Channel Blockers
Antihypertensive Combinations
Antiadrenergic Antihypertensives
Potassium Sparing Diuretics

Over 80% phenotype factors are clinically meaningful

Limestone vs. NMF

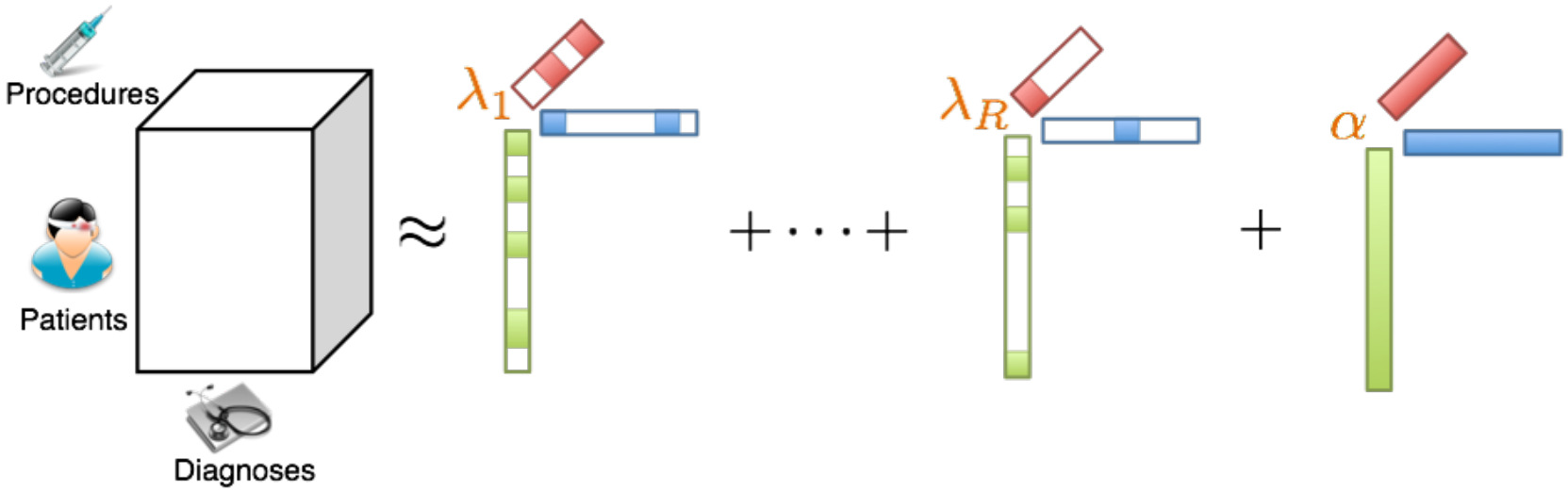
Limestone Phenotype	
Hypertension	0.94
	0.06
	0.51

9

… 1,549 total combinations

Limestone provides more concise phenotype representation than NMF

Summary: Phenotyping using Tensor Factorization



- **Unsupervised:** Sparse Nonnegative tensor factorization can be used to learn phenotypes without supervision

- **Predictive**: Resulting phenotypes outperforms features from raw EHR data in predictive modeling tasks.

Phenotyping from Electronic Health Records using Tensor Factorization

Jimeng Sun

jsun@cc.gatech.edu

Georgia Tech College of Computing
Computational Science and Engineering

1. Ho, Joyce C., Joydeep Ghosh, Steve R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. "Limestone: High-Throughput Candidate Phenotype Generation via Tensor Factorization." *Journal of Biomedical Informatics*. 2014
2. Ho, Joyce C., Joydeep Ghosh, and Jimeng Sun. "Marble: High-Throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization." In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 115–24. KDD '14.