# Extracting Phenotypes from Patient Claim Records Using Nonnegative Tensor Factorization

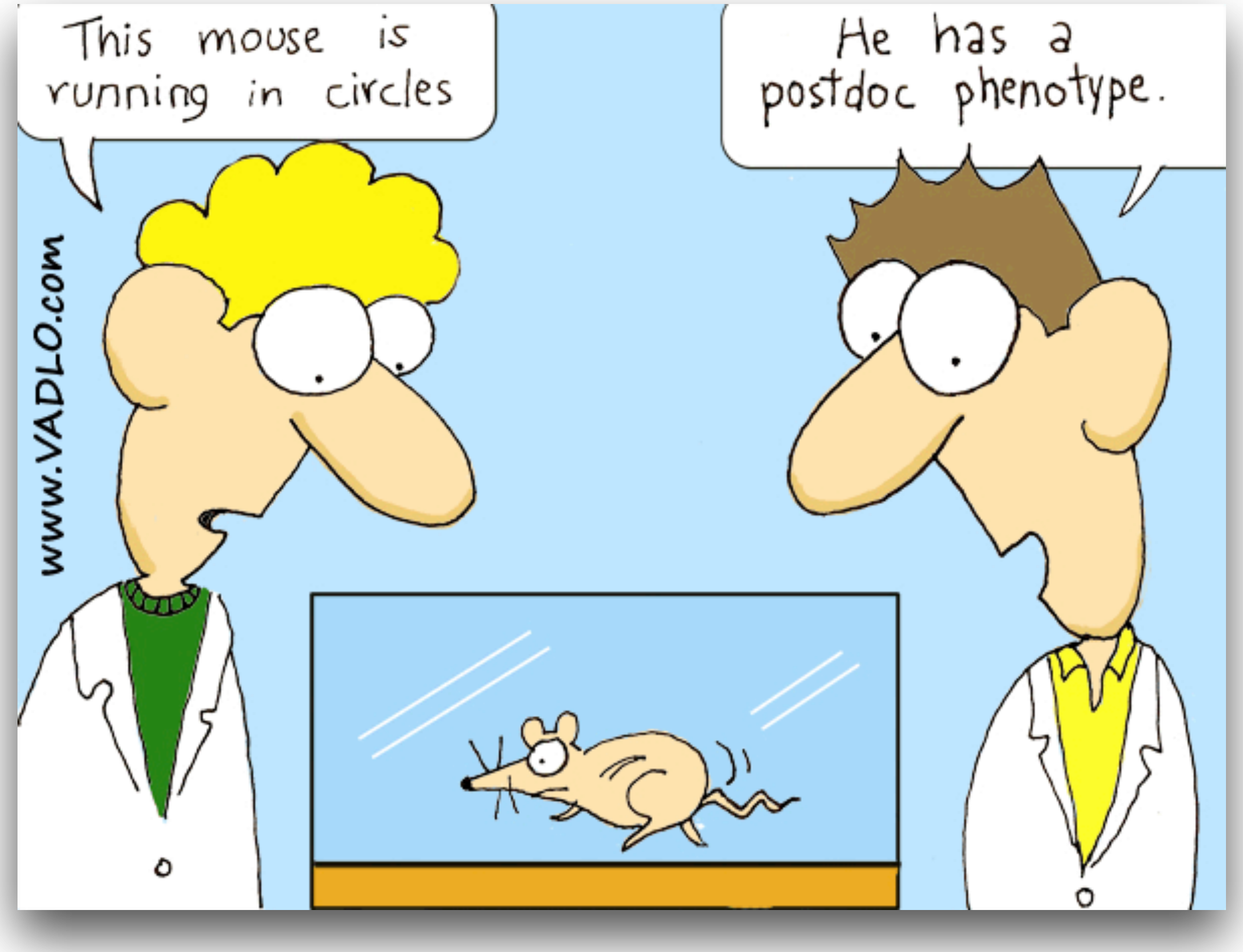
Joyce C. Ho[1], **Joydeep Ghosh**[1], Jimeng Sun[2]

[1]University of Texas at Austin

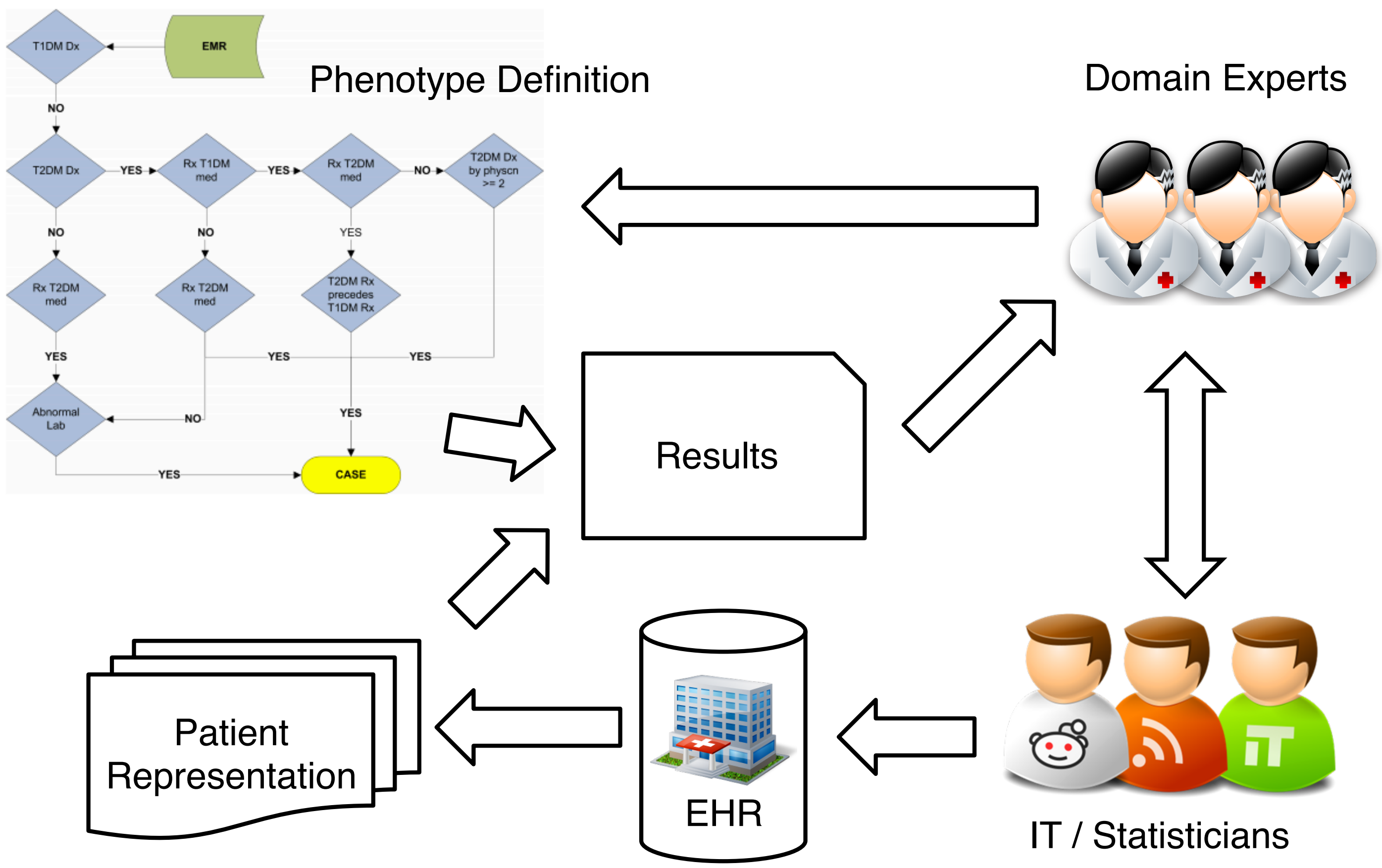
[2]Georgia Institute of Technology

# EHR-Based Phenotyping

- Map EHR data to meaningful medical concepts
- Learn medically relevant data characteristics



# CURRENT PHENOTYPING PROCESS



Phenotype Definition

Domain Experts

Results

Patient Representation

EHR

IT / Statisticians

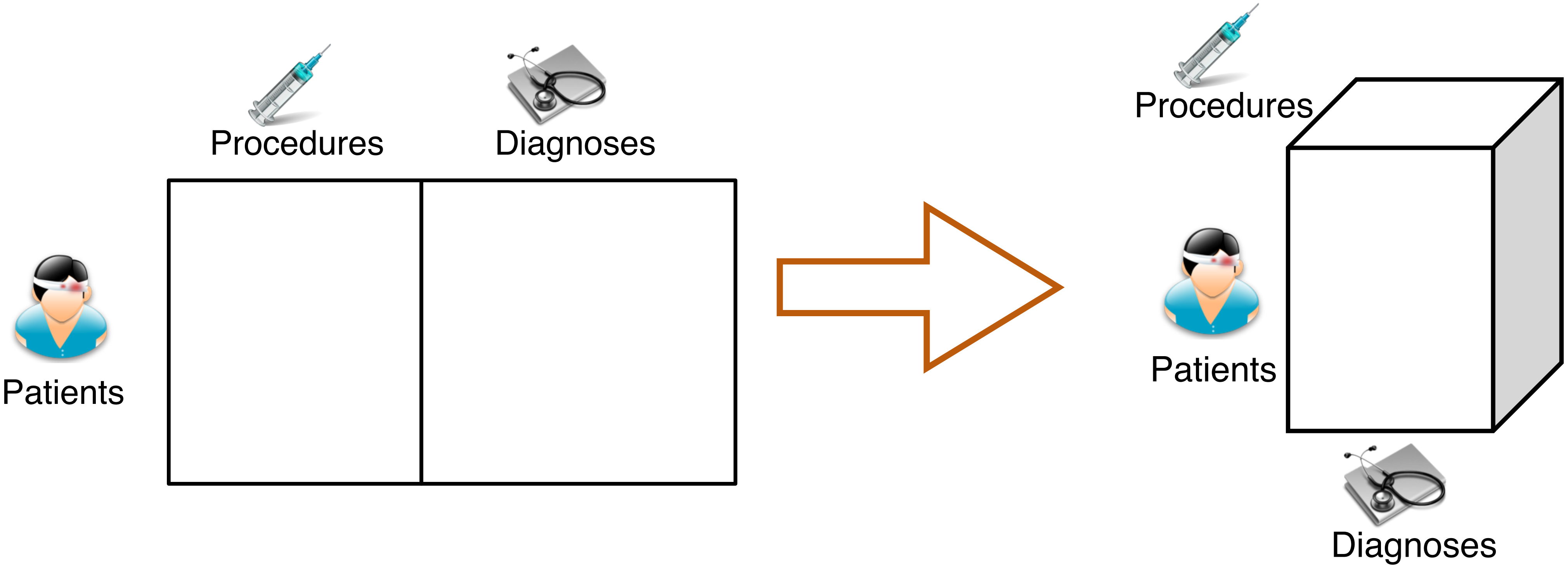
- Iterative process with significant time, effort, and expert involvement

- Existing high-throughput methods require human annotated samples

# LIMESTONE: OVERVIEW

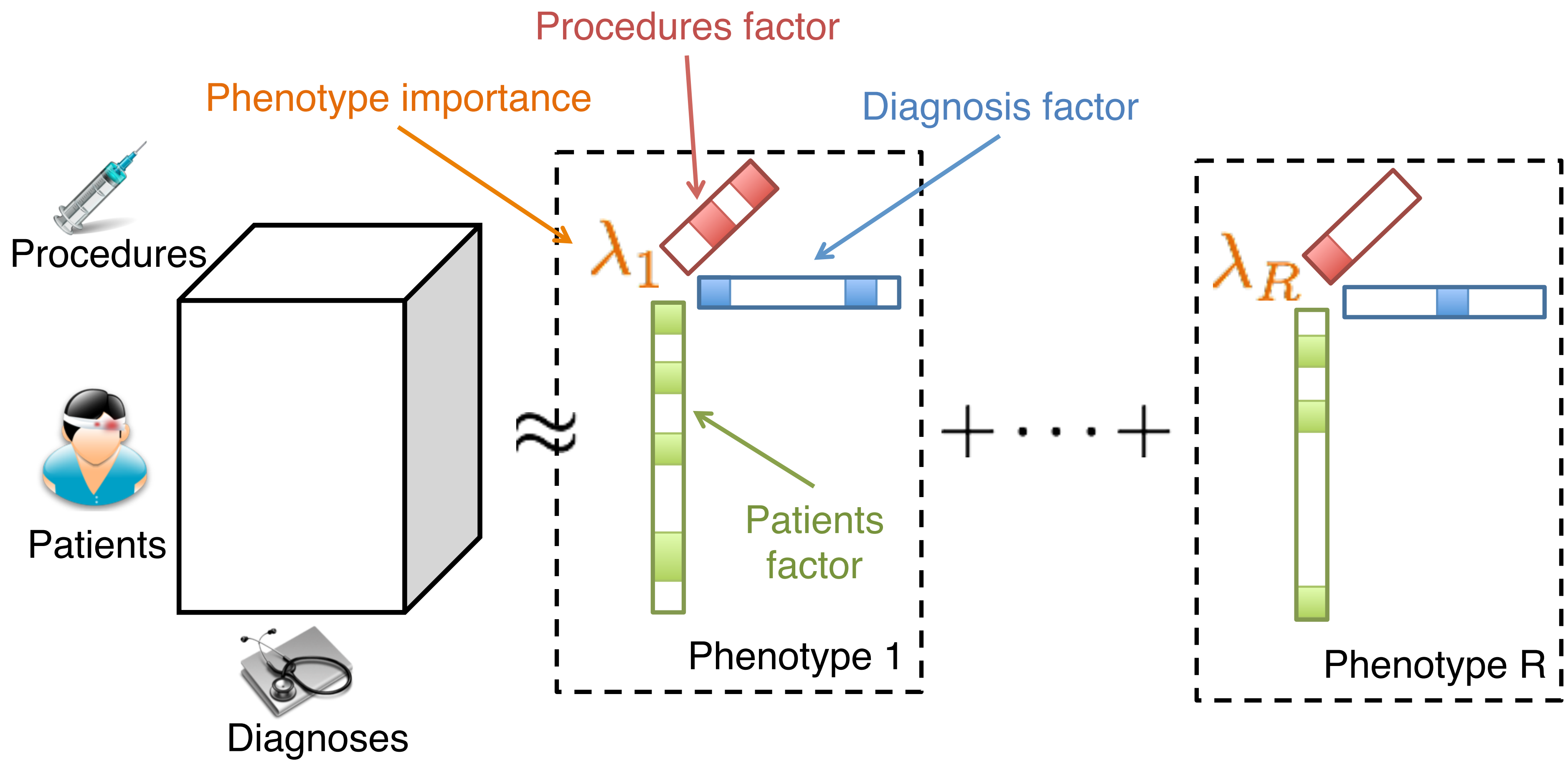
- Phenotyping is similar to dimensionality reduction
- Developed a tensor factorization model to achieve high-throughput phenotyping
  - Tensor representation to capture source interactions
  - Nonnegative and sparsity constraints yield concise, interpretable results

# Motivation for Tensors



- Generalization of matrix to multi-way data
- Natural representation to capture structured source interactions (e.g. group of procedures to treat a disease)

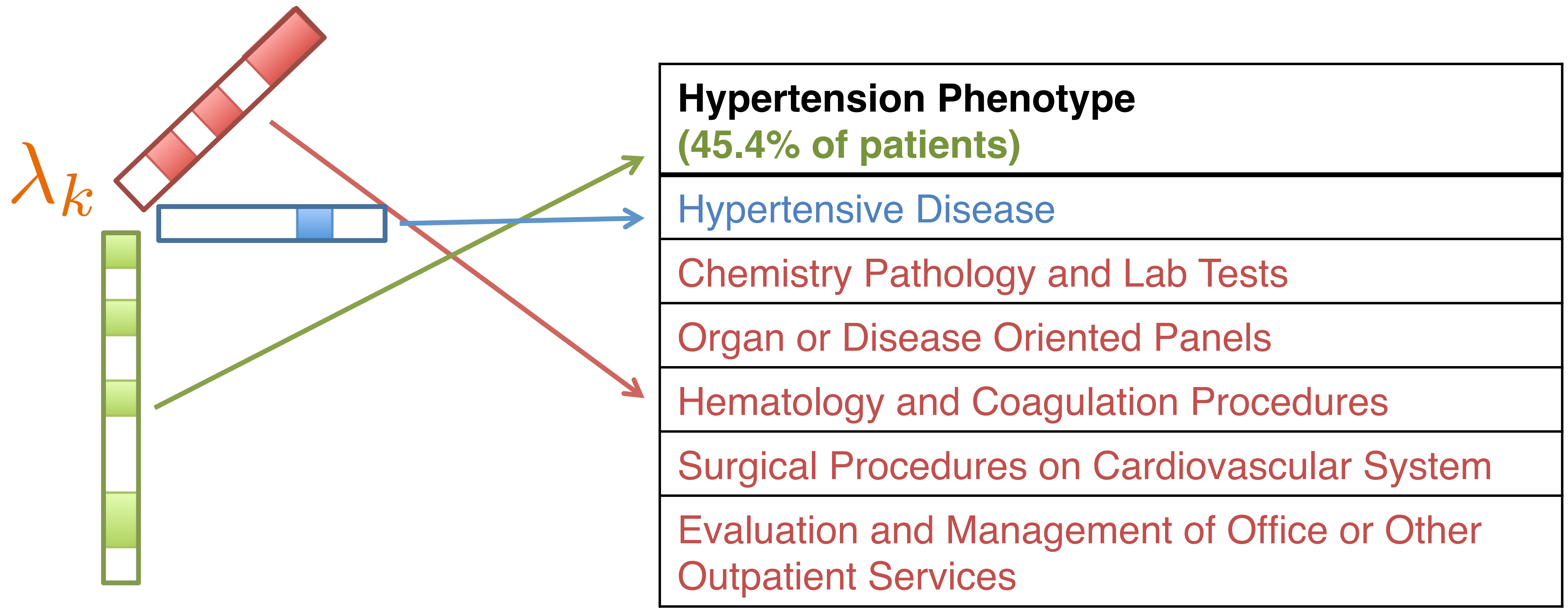
# LIMESTONE: PHENOTYPE GENERATION



- Extension of common tensor decomposition (CANDECOMP/PARAFAC)

- A candidate phenotype is a **single** rank-1 tensor

# LIMESTONE: Candidate Phenotype



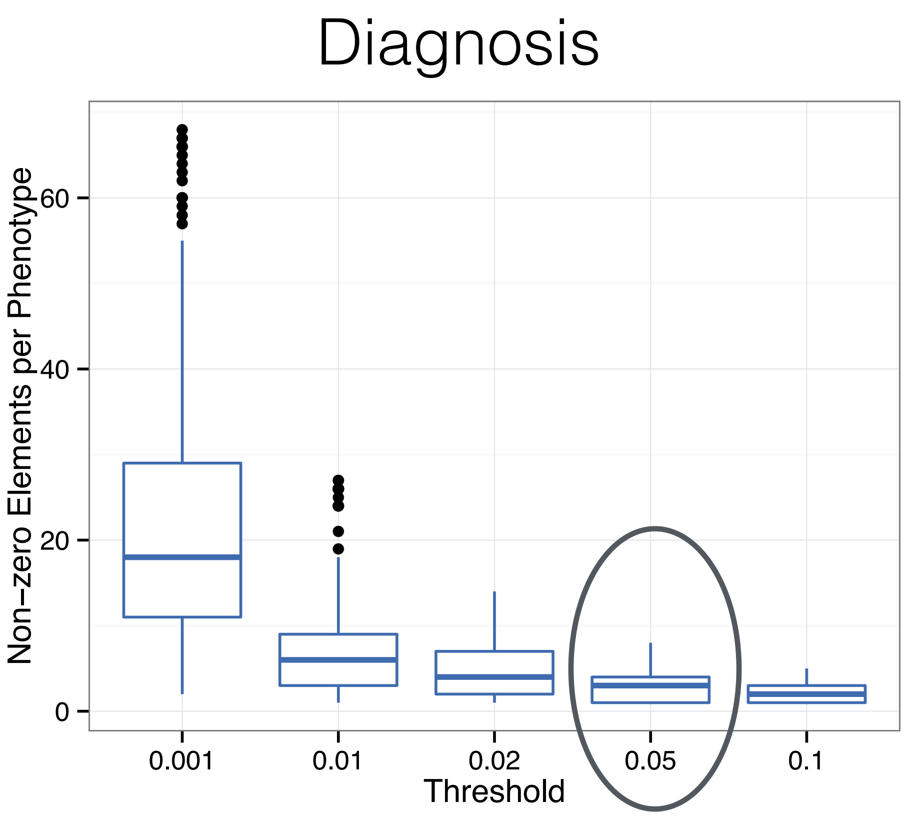
| **Hypertension Phenotype**<br>**(45.4% of patients)** |
|---|
| Hypertensive Disease |
| Chemistry Pathology and Lab Tests |
| Organ or Disease Oriented Panels |
| Hematology and Coagulation Procedures |
| Surgical Procedures on Cardiovascular System |
| Evaluation and Management of Office or Other Outpatient Services |

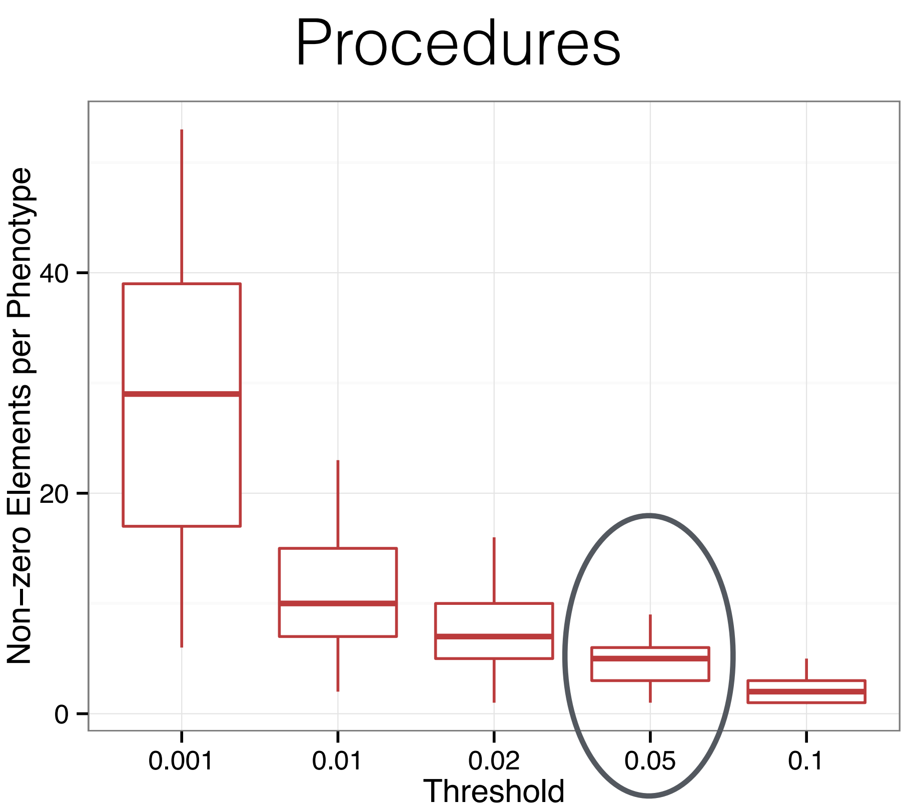
- Non-zero elements are the clinical characteristics of a candidate phenotype
- Each element represents conditional probability given the phenotype and mode

# Experiment Data

- CMS 2008-2010 Medicare Data Entrepreneurs' Synthetic Public Use File
    - Inpatient, outpatient, carrier and prescription drug claims for 5% of Medicare population
    - Synthesized to protect privacy of beneficiaries
- Constructed tensor from random subset of patients in carrier claims records
- Tensor size:10,000 patients x 129 diagnoses x 115 procedures

# Concise Phenotype Definitions





2-3 diagnosis / phenotype

4-6 procedures / phenotype

- Tunable knob to adjust threshold
- Higher threshold values improves interpretability (and conciseness)

# Results: Example Phenotypes

**Diabetes Phenotype 1**
(34.8% of patients)

Diseases of other endocrine glands
Other metabolic and immunity disorders

Eval. and Mgmt. of Office or Other Outpatient Svcs.
Surgical Procs. on the Cardiovascular System
Ophthalmology Procs.
Cardiovascular Procs.
Urinalysis Procs.
Diagnostic/Screening Processes or Results

**Arthritis Phenotype 1**
(29.1% of patients)

Arthropathies and related disorders

Physical Medicine and Rehabilitation Procs.
Eval. and Mgmt. of Office or Other Outpatient Svcs.

**Diabetes Phenotype 2**
(33.1% of patients)

Diseases of other endocrine glands

Chemistry Pathology and Laboratory Tests
Organ or Disease Oriented Panels
Hematology and Coagulation Procedures
Surgical Procs. on the Cardiovascular System
Eval. and Mgmt. of Office or Other Outpatient Svcs.

**Arthritis Phenotype 2**
(38.6% of patients)

Arthropathies and related disorders
Rheumatism, excluding the back

Eval. and Mgmt. of Office or Other Outpatient Svcs.
Surgical Procs. on the Musculoskeletal System
Surgical Procs. on the Cardiovascular System
Cardiovascular Procs.
Hematology and Coagulation Procs.

Phenotypes are concise and interpretable

# Results: Arthritis Phenotypes

- Phenotypes can represent disease subtypes (or severity levels)
- Rapidly characterize and manage diverse population

| **Heart Failure Phenotype 1**<br>(36.7% of patients) |
|---|
| Other forms of heart disease |
| Complications of surgical and medical care<br>Hematology and Coagulation Procs.<br>Eval. and Mgmt. of Office or Other Outpatient Svcs.<br>Surgical Procs. on the Cardiovascular System<br>Chemistry Pathology and Laboratory Tests<br>Cardiovascular Procs.<br>Organ or Disease Oriented Panels |

| **Heart Failure Phenotype 2**<br>(30.9% of patients) |
|---|
| Other forms of heart disease<br>Ischemic heart disease |
| Hospital Inpatient Svcs.<br>Eval. and Mgmt. of Office or Other Outpatient Svcs. |

higher degree of severity as procedures involves inpatient services

# CONCLUSION

- Data-driven solution to generate multiple phenotypes simultaneously from diverse population
- Minimal human intervention (no expert supervision)
- Derived phenotypes are concise and interpretable
- Future work:
  - Multi-relational tensors to incorporate multiple data sources
  - Improve computational speed