MARBLE: High-Throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization

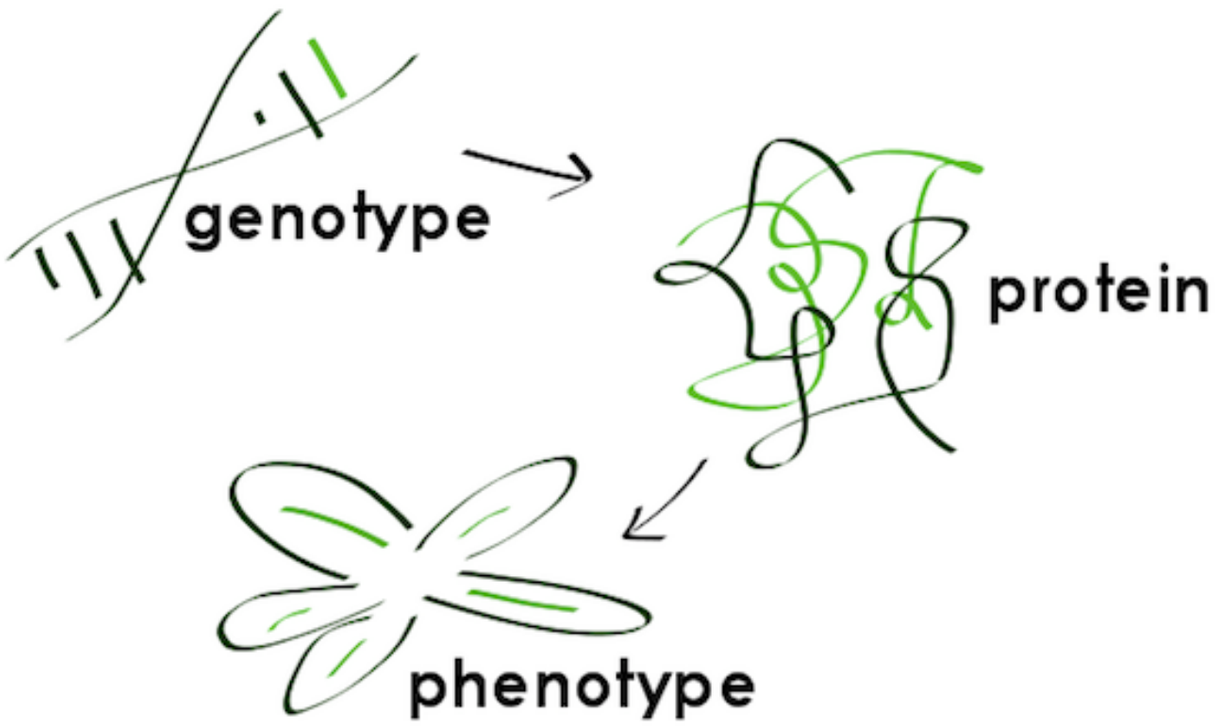
Joyce C. Ho[1], Joydeep Ghosh[1], Jimeng Sun[2]
[1]University of Texas at Austin
[2]Georgia Institute of Technology

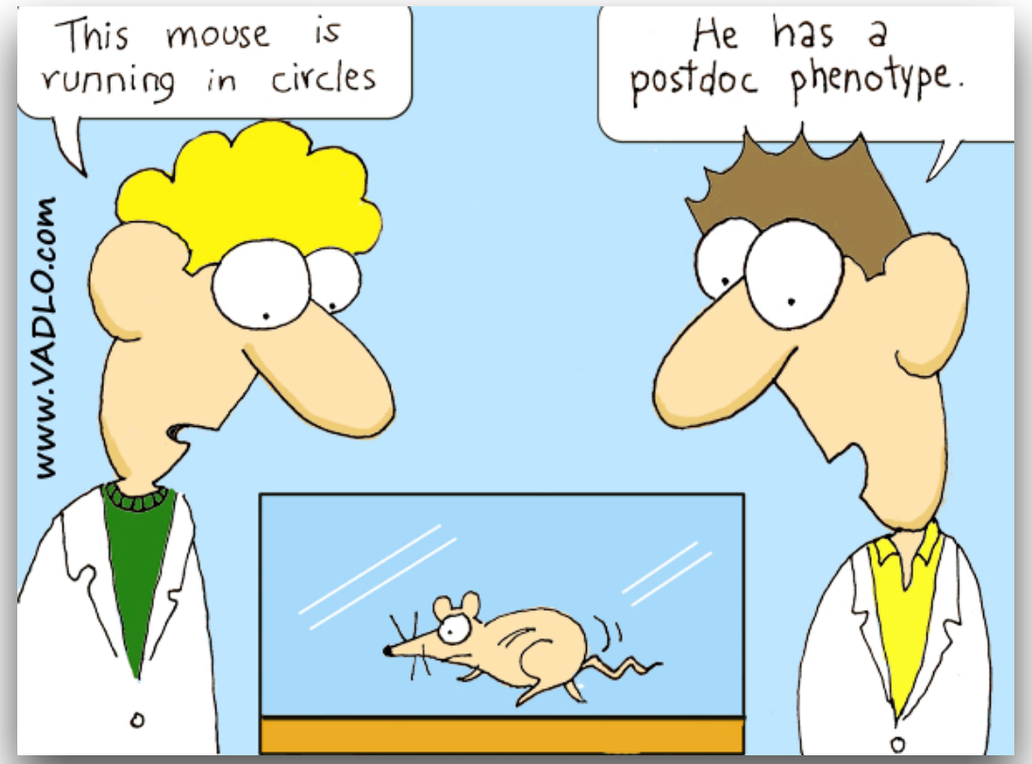
Phenotype

American Heritage Dictionary:

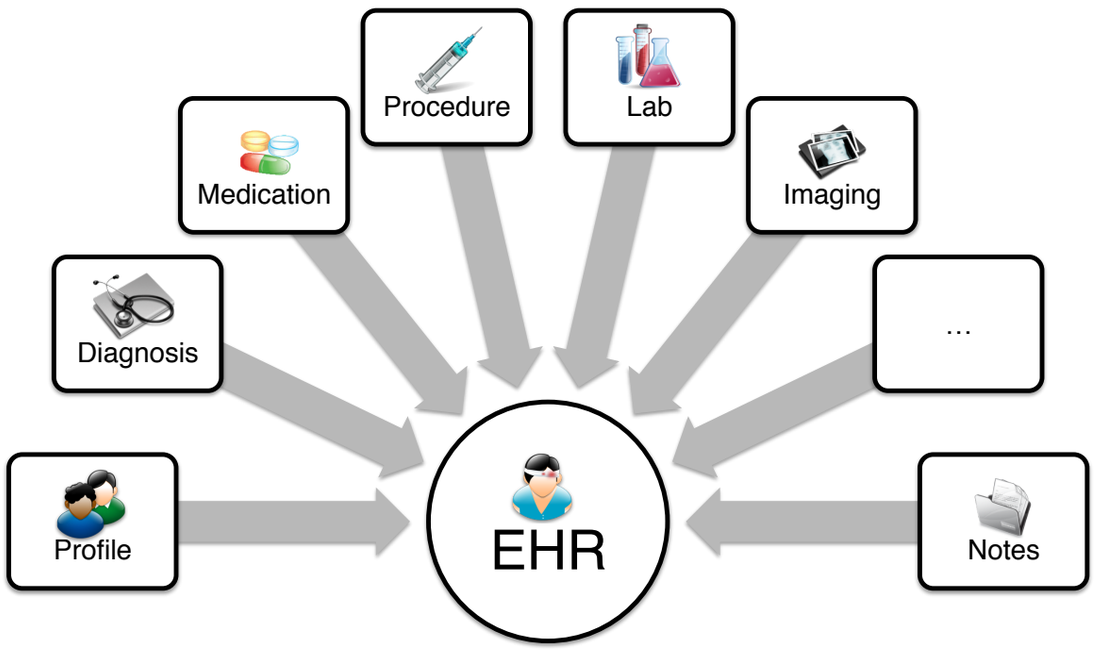
A. The **observable** physical or biochemical **characteristics** of an organism, as determined by both genetic makeup and environmental influences.

B. The **expression of a specific trait**, such as stature or blood type, based on genetic and environmental influences.





Electronic Health Records (EHR)



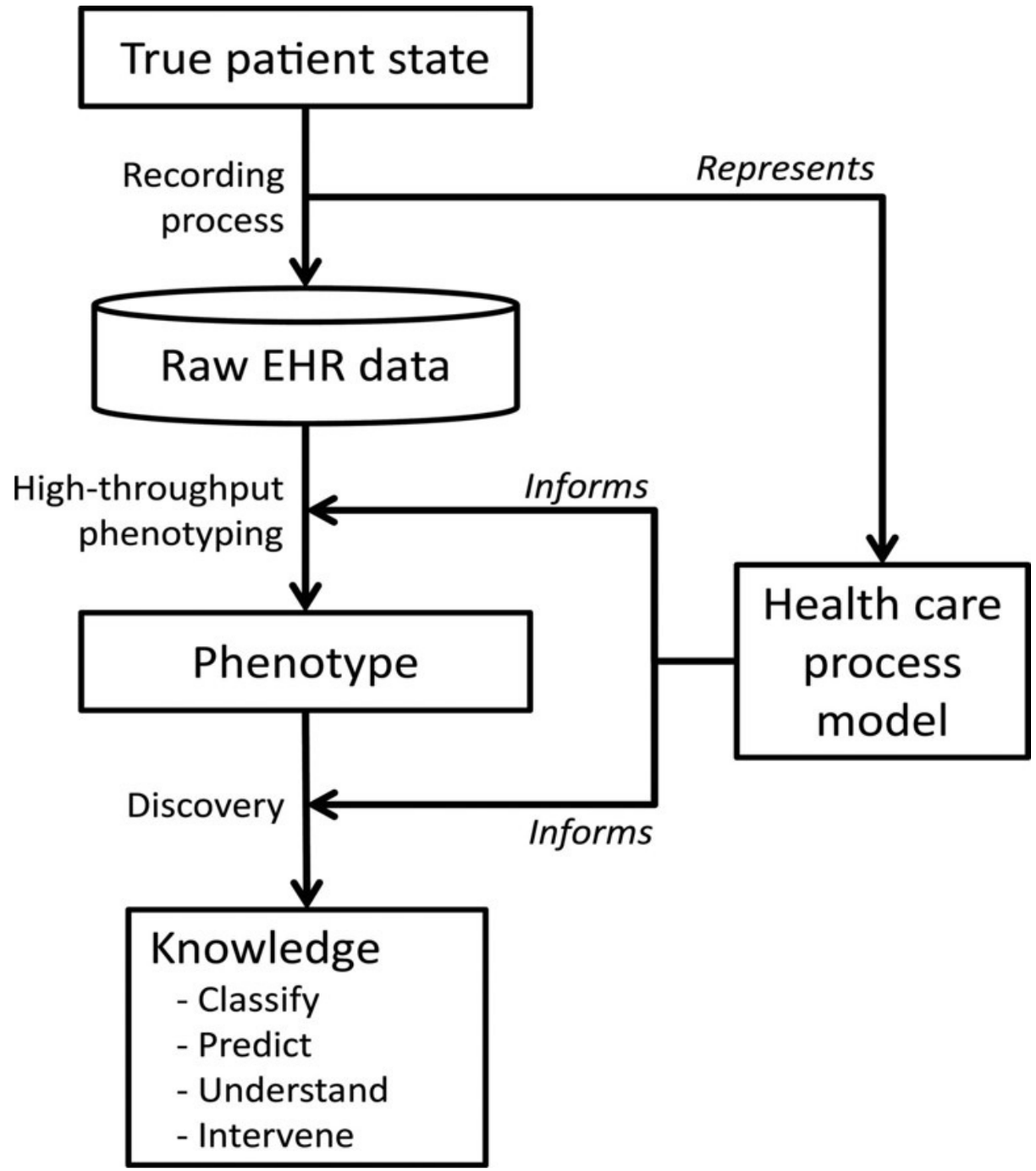
Patient's health history in one place

Challenges:

- Diverse population
- Noisy, incomplete, and potentially inaccurate patient representation
- Medical professionals accustomed to medical concepts

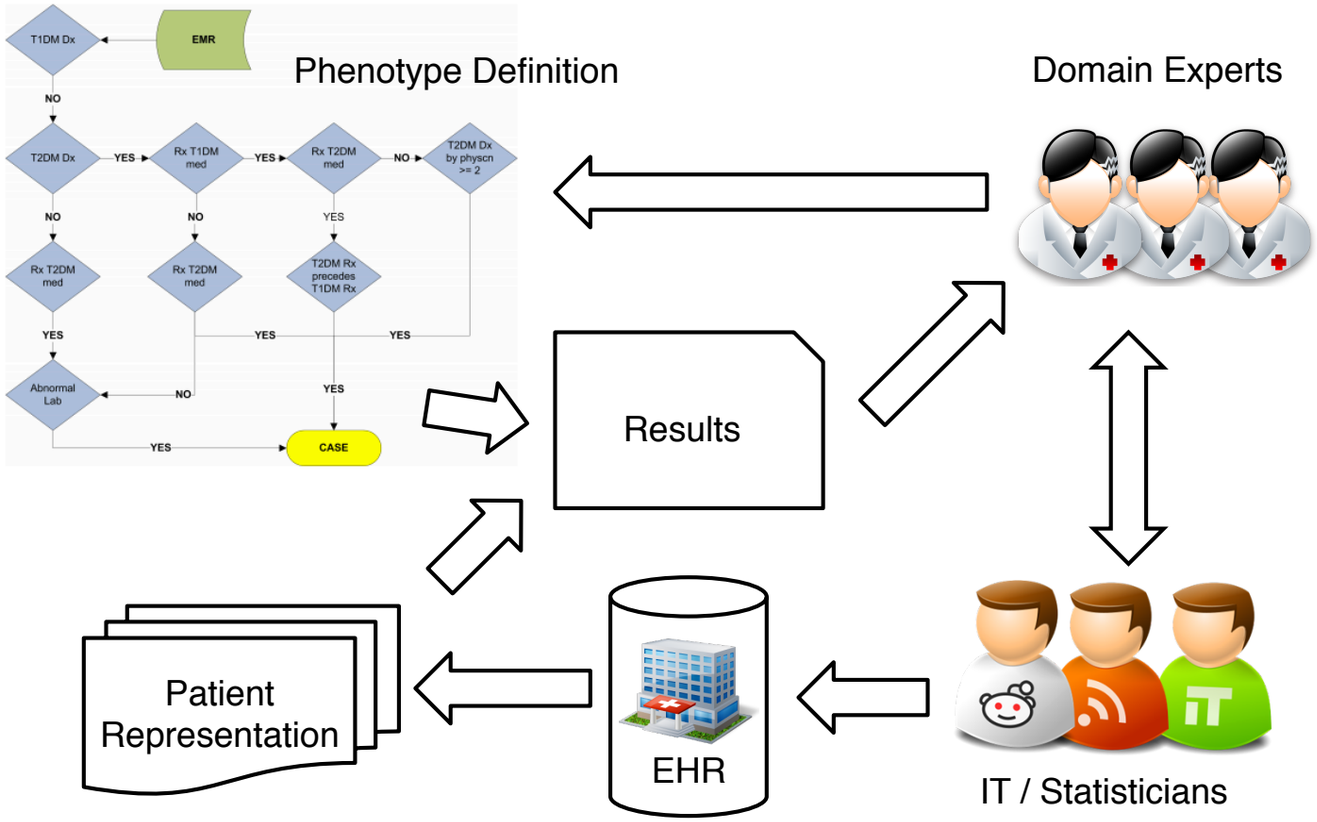
EHR-Driven Phenotypes

- Learn set of clinically relevant features (characteristics)
- Mapping data to meaningful medical concepts
- Identify cohorts to conduct genome and phenome-wide association studies



Hripcsak, G., & Albers, D. J. (2012). Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association, 20(1), 117–121.

CURRENT PHENOTYPING PROCESS



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

- Existing high-throughput methods require human annotated samples

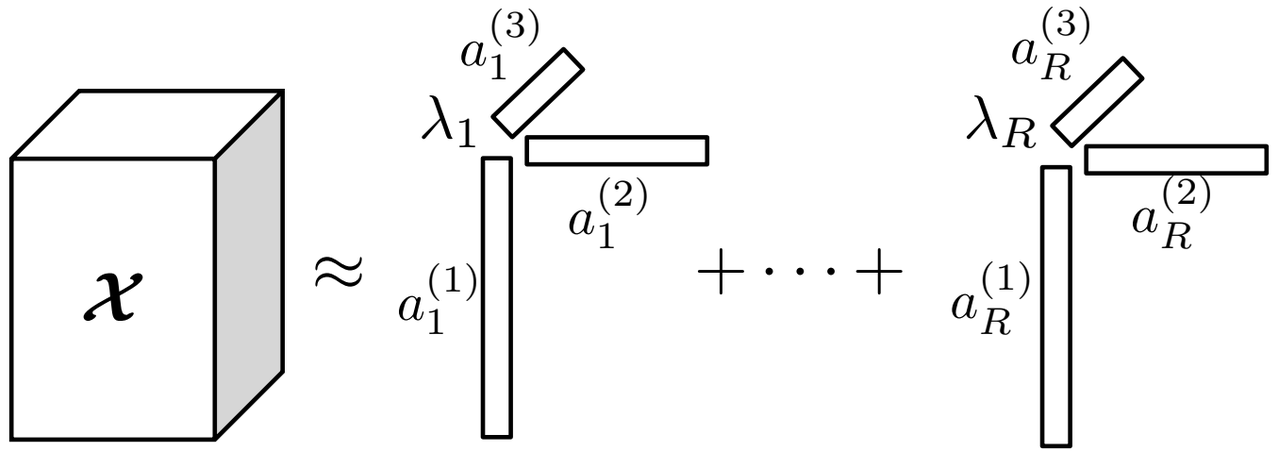
Talk by Dr. Jyoti Pathak - High-Throughput Phenotyping (HTP) at AMIA 2013 Annual Symposium

Prior Work: Limestone*

- Phenotyping is similar to dimensionality reduction
- Pilot study (Limestone):
  - Tensor representation captures source interactions
  - Tensor decomposition using CANDECOMP/ PARAFAC Alternating Poisson Regression model (CP-APR) by Chi & Kolda

*Joyce C. Ho, Joydeep Ghosh, Steven R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization in *Journal of Biomedical Informatics (2014).*

CP-APR: Tensor Decomposition



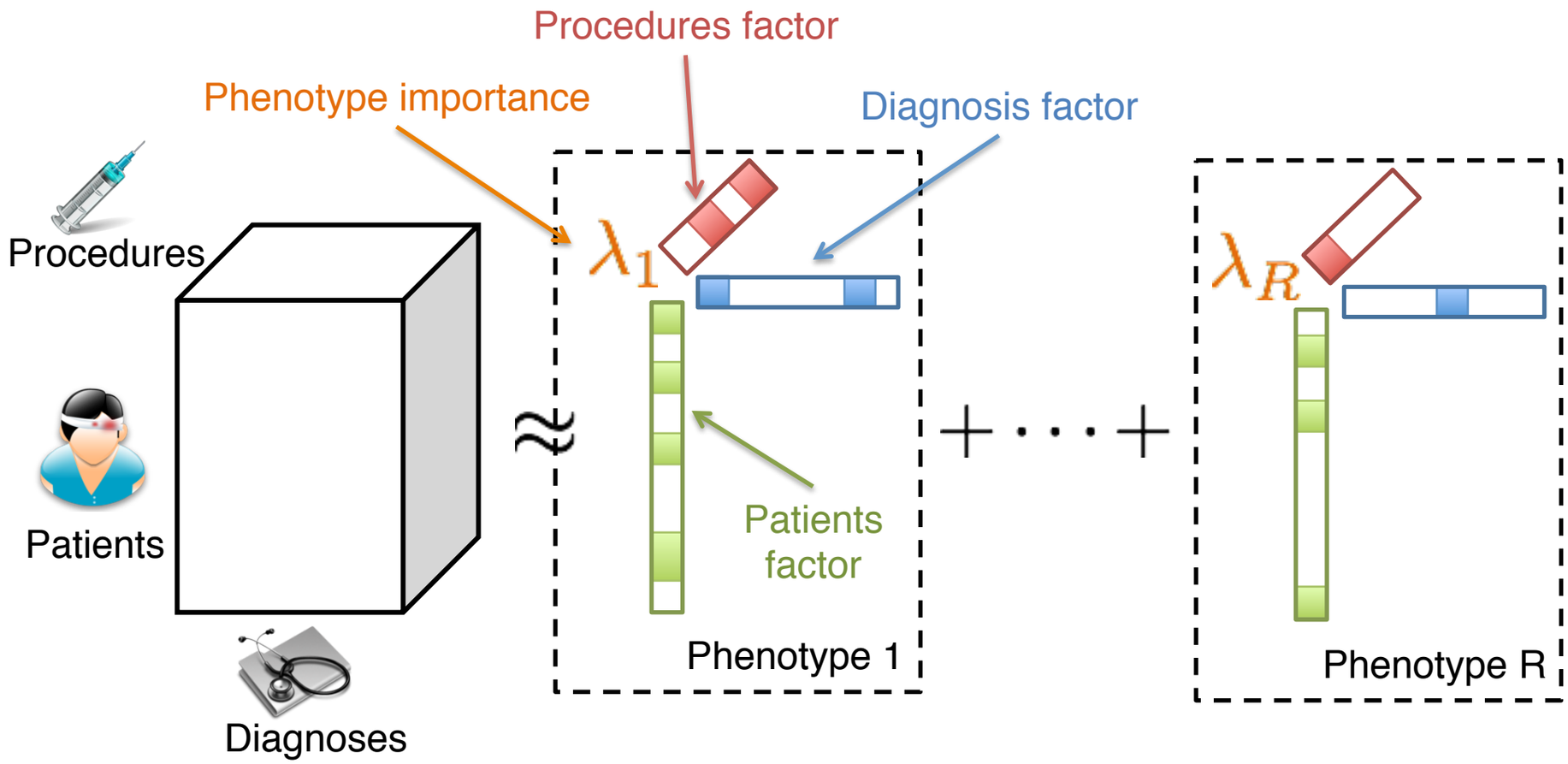
$$\min f(\mathcal{M}) \equiv \sum_{\vec{i}} m_{\vec{i}} - x_{\vec{i}} \log m_{\vec{i}}$$

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

- Generalized KL divergence for count (Poisson) data
- Nonnegative constraints on weights and factors
- Stochastic factor constraints

Eric C. Chi and Tammy G. Kolda.
On tensors, sparsity, and nonnegative factorizations in *SIAM Journal on Matrix Analysis and Applications* (2014).

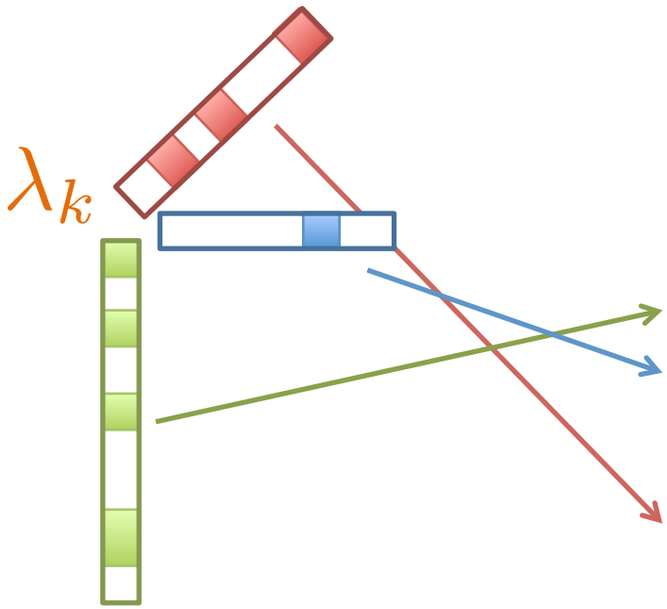
LIMESTONE: PHENOTYPE GENERATION



- Non-zero elements are the clinical characteristics of a candidate phenotype

- Each element represents conditional probability given the phenotype and mode

LIMESTONE: RESULTING PHENOTYPES

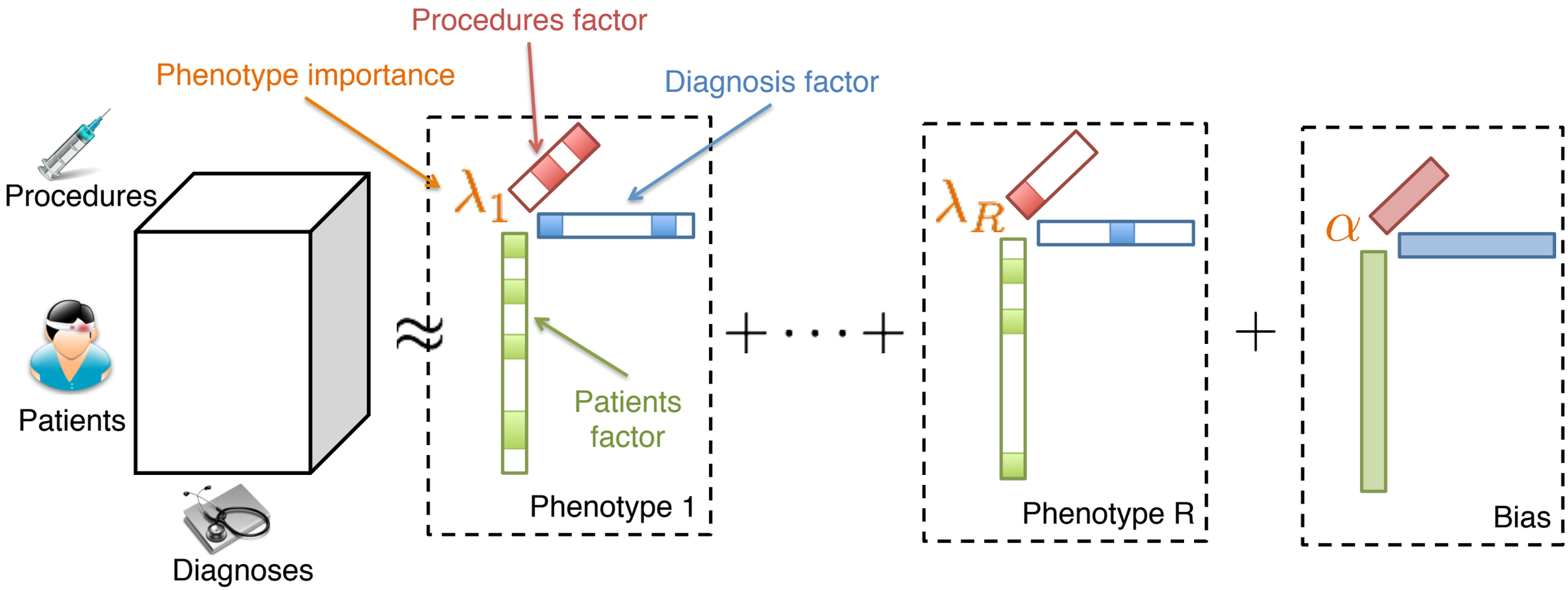


Hypertension Phenotype (30% of patients)
Hypertensive Disease
Chemistry Pathology and Lab Tests
Evaluation and Management of Office or Other Outpatient Services
Surgical Procedures on Cardiovascular System
Hematology and Coagulation Procedures

- Post-process factors to obtain concise results
- 82% of phenotypes deemed clinically meaningful

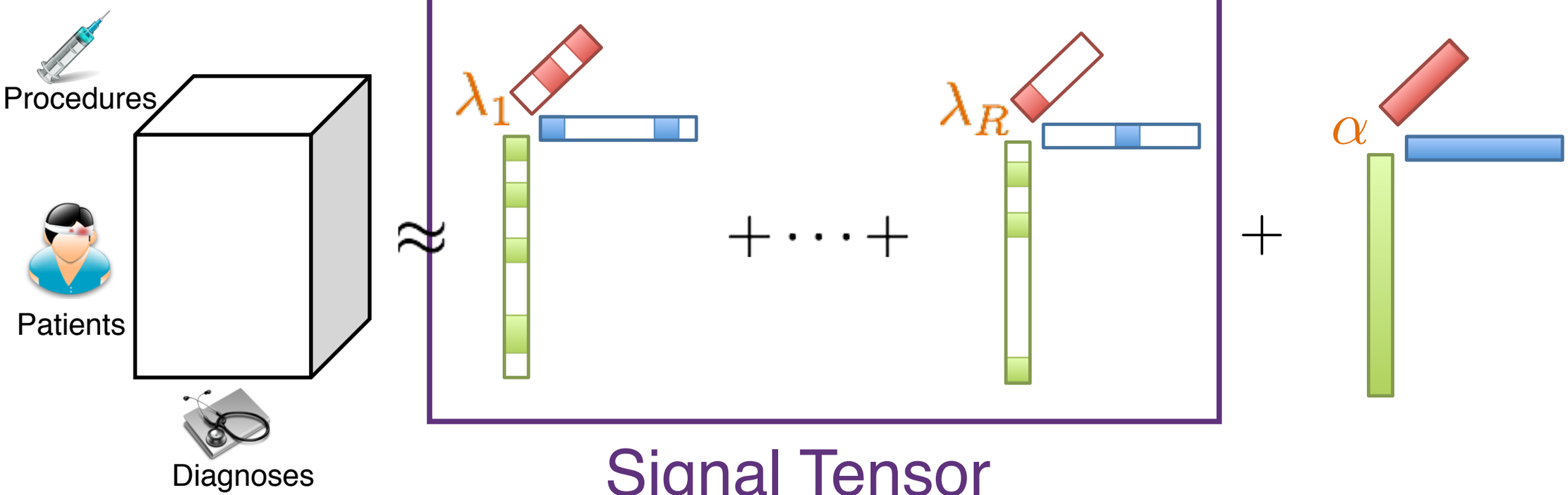
What about baseline characteristics and computational stability (inadmissible zero problem)?

MARBLE: PHENOTYPE GENERATION



- Phenotypes are defined on the signal tensor
- Baseline characteristics defined by bias tensor

MARBLE: TENSOR DECOMPOSITION



Signal Tensor

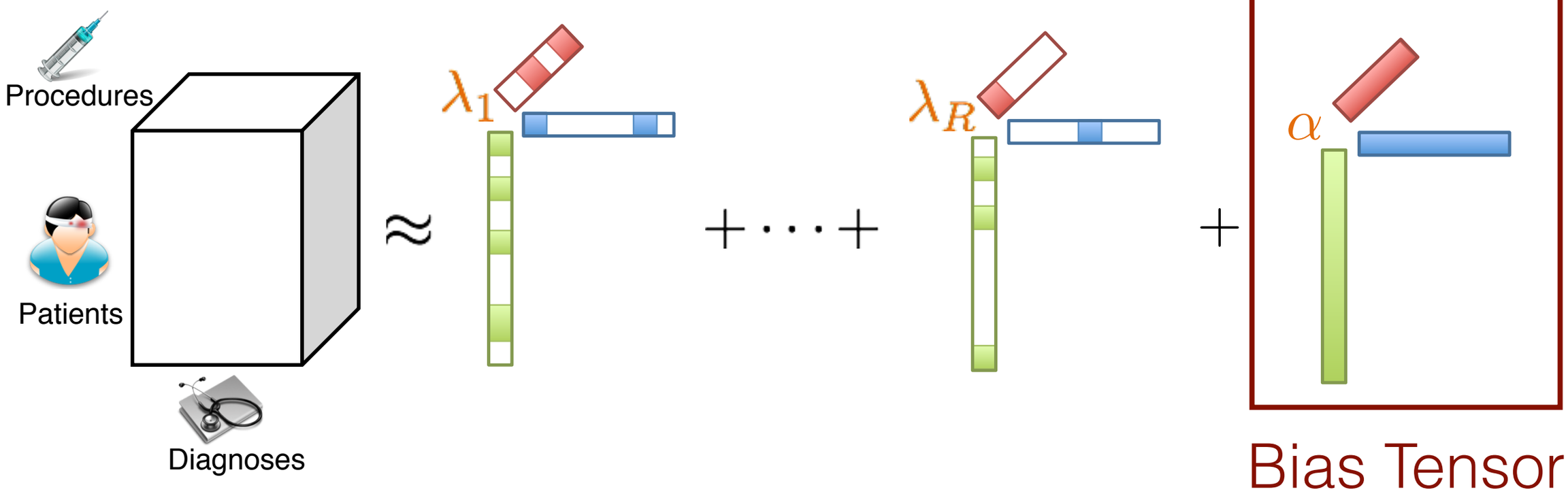
$$\boldsymbol{\mathcal{V}} = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \cdots ; \mathbf{A}^{(N)}]\!]$$

$\lambda_n \in [0, +\infty)$ non-negative weights

$\mathbf{A}^{(n)} \in \{0, [\gamma_n, 1]\}^{I_n \times 1}$ sparsity constraints

$||\mathbf{A}^{(n)}||_1 = 1$ stochastic constraints

MARBLE: TENSOR DECOMPOSITION



- captures data bias (offset)
- avoids inadmissible zeros problem $\left(\log(m_{\vec{i}} = 0)\right)$
- computationally stable

$$\boldsymbol{\mathcal{C}} = [\![\alpha; \mathbf{u}^{(1)}; \cdots ; \mathbf{u}^{(N)}]\!]$$

$$\alpha \in (0, +\infty)$$

$$\mathbf{u}^{(n)} \in (0, 1]^{I_n \times 1}$$

$$||\mathbf{u}^{(n)}||_1 = 1$$

MARBLE: Optimization Problem

$$\min f(\boldsymbol{\mathcal{M}}) \equiv \sum_{\vec{i}} (m_{\vec{i}} - x_i \log m_{\vec{i}})$$

$$\text{s.t } \boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{V}}$$

$$\boldsymbol{\mathcal{C}} = [\![\alpha; \mathbf{u}^{(1)}; \cdots; \mathbf{u}^{(N)}]\!] \in \Omega_C$$

$$\boldsymbol{\mathcal{V}} = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \cdots; \mathbf{A}^{(N)}]\!] \in \Omega_V$$

$$\Omega_C = \Omega_\alpha \times \Omega_{u1} \times \cdots \times \Omega_{uN}$$

$$\Omega_\alpha = (0, +\infty)$$

$$\Omega_{un} = \{\mathbf{u} \in (0,1]^{I_{n\times 1}} \mid ||\mathbf{u}||_1 = 1\}$$

$$\Omega_V = \Omega_\lambda \times \Omega_{A1} \times \cdots \times \Omega_{AN}$$

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

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

Algorithm:

- Alternating minimization to solve for each mode
- Gradual projection to threshold factor matrices
- Sparse implementation performs calculation only for non-zero elements

MARBLE: ALGORITHM OVERVIEW

$$\mathbf{B}^{(n)} = \text{argmin}_{\mathbf{B}}\ \mathbf{e}^{\intercal}\left[\mathbf{C}_{(n)} + \mathbf{B}\mathbf{\Pi}^{(n)} - \mathbf{X}_{(n)} * \log\left(\mathbf{C}_{(n)} + \mathbf{B}\mathbf{\Pi}^{(n)}\right)\right]\mathbf{e}$$

while *not converged* **do**
  foreach *mode* n **do**
    Solve the nth interaction factor matrix;
    Project onto sparse factors;
    Solve nth bias vector;
  end
  Calculate gradual projection penalty;
end

$$a_{gh}^{(n)} > \xi^{(k)}\gamma_n$$

$$\kappa^{(k)} = 1 - \frac{|f(\boldsymbol{\mathcal{M}}^{(k-1)}) - f(\boldsymbol{\mathcal{M}}^{(k)})|}{|f(\boldsymbol{\mathcal{M}}^{(k-1)})|}$$

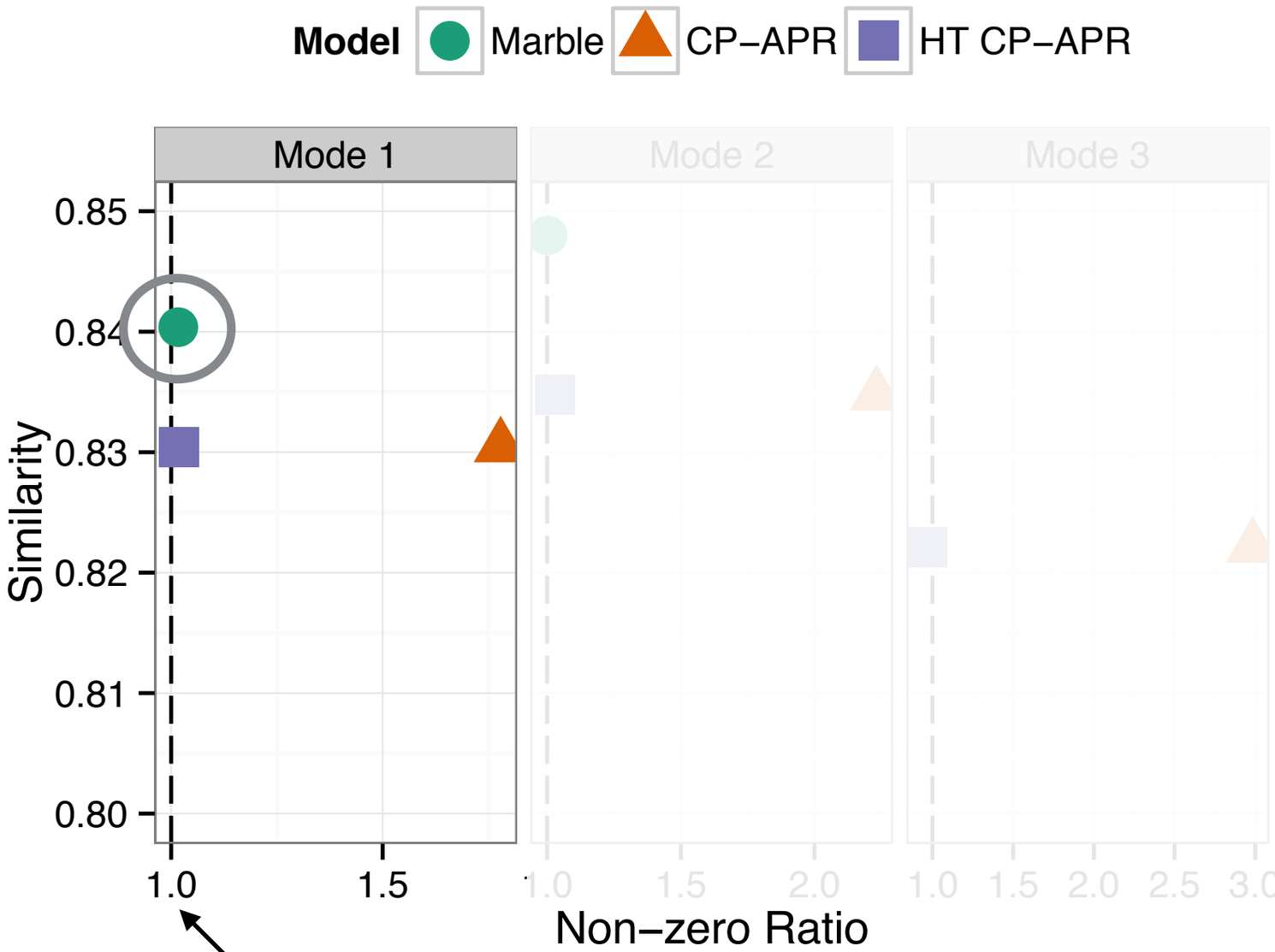
$$\xi^{(k+1)} = \max(\xi^{(k)}, \frac{1}{2}\xi^{(k)} + \frac{1}{2}\kappa^{(k)})$$

MARBLE: RECAP

- Model extends CP-APR
  - Observation tensor decomposes into signal (interaction) tensor and bias tensor
  - Sparsity constraints on signal factors minimize number of non-zero elements
  - Bias term provides baseline characteristics of population and computational stability

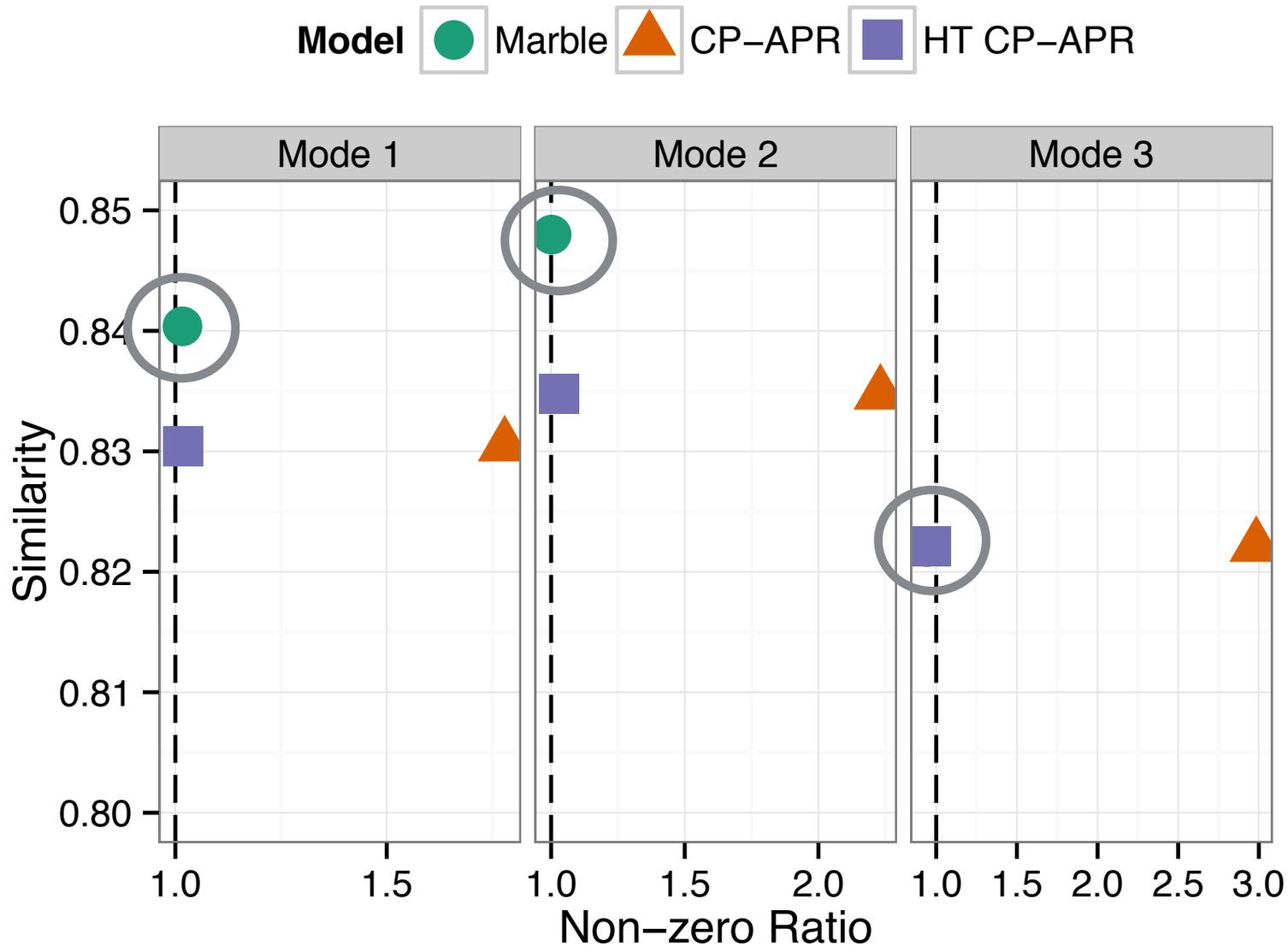
Simulation Results: Model Comparison

higher => closer to known solution



dotted line = correct support (# non-zeros)

Simulation Results: Model Comparison



Marble recovers the original solution and the sparsity pattern better than the others

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

Results: Bias (Top 10 Elements)

Diagnosis Mode
Symptoms
Complications of surgical and medical care
Arthropathies and related disorders
Other forms of heart disease
Dorsopathies
Disorders of the human eye
Diseases of other endocrine glands
Hypertensive disease
Other metabolic and immunity disorder
Other diseases of urinary system

Procedure Mode
Evaluation and Management of Other Outpatient Services
Diagnostic Radiology Procedures
Hospital Inpatient Services
Chemistry Pathology and Laboratory Tests
Physical Medicine and Rehabilitation Procedures
Surgical Procedures on the Cardiovascular System
Cardiovascular Procedures
Emergency Department Services
Nursing Facility Services
Hematology and Coagulation Procedures

Common chronic diseases amongst elderly (e.g., hypertension, arthritis, heart disease, and diabetes)

Patients generally visit clinics because of various symptoms and complications

RESULTS: CHRONIC DISEASES

Diabetes Phenotype
Diseases of other endocrine glands
Complications of surgical and medical care
Chemistry Pathology and Laboratory Tests
Organ or Disease Oriented Panels
Hematology and Coagulation Procedures
Surgical Procedures on the Cardiovascular System

Arthritis Phenotype
Arthropathies and related disorders
Physical Medicine and Rehabilitation Procedures
Evaluation and Management of Other Outpatient Services
Surgical Procedures on the Musculoskeletal System
Diagnostic Radiology Procedures

Phenotype descriptions map to known characteristics of chronic diseases

Results: Disease Subtypes

Heart Failure Phenotype
Other forms of heart disease
Complications of surgical and medical care
Symptoms
Cardiovascular Procedures
Hematology and Coagulation Procedures
Evaluation and Management of Other Outpatient Services
Surgical Procedures on the Cardiovascular System
Chemistry Pathology and Laboratory Tests

Severe Heart Failure Phenotype
Other forms of heart disease
Pneumoconioses and other lung diseases
Ill-defined and unknown causes of morbidity and mortality
Hospital Inpatient Services
Cardiovascular Procedures

Inpatient services and mortality suggest higher degree of severity

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