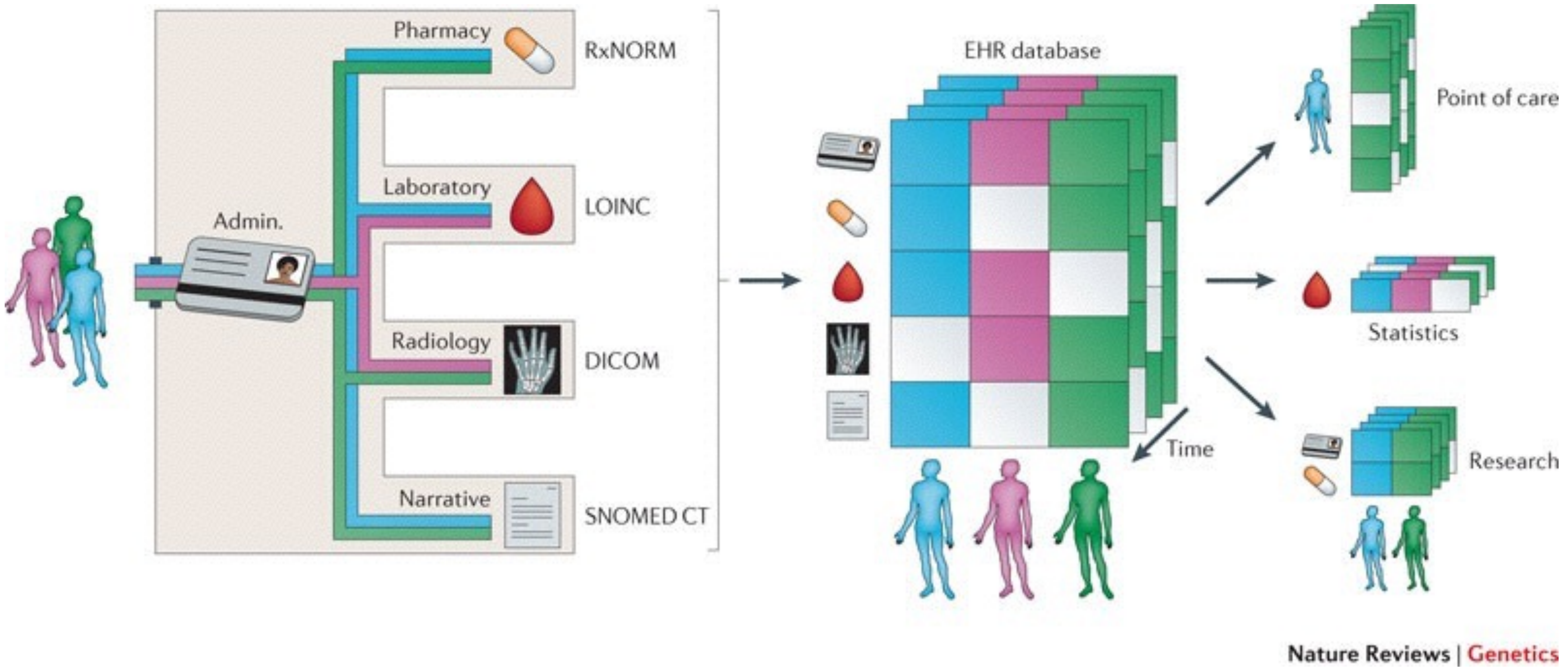
GRANITE: DIVERSIFIED, SPARSE TENSOR FACTORIZATION FOR ELECTRONIC HEALTH RECORD-BASED PHENOTYPING

Jette Henderson[*], Joyce C. Ho[†], Abel N. Kho[‡], Joshua C. Denny[§], Bradley A. Malin[§], Jimeng Sun[¶], Joydeep Ghosh[*]

[*]University of Texas at Austin, [†]Emory University, [‡]Northwestern University
[§]Vanderbilt University, [¶]Georgia Institute of Technology

ELECTRONIC HEALTH RECORD (EHR)



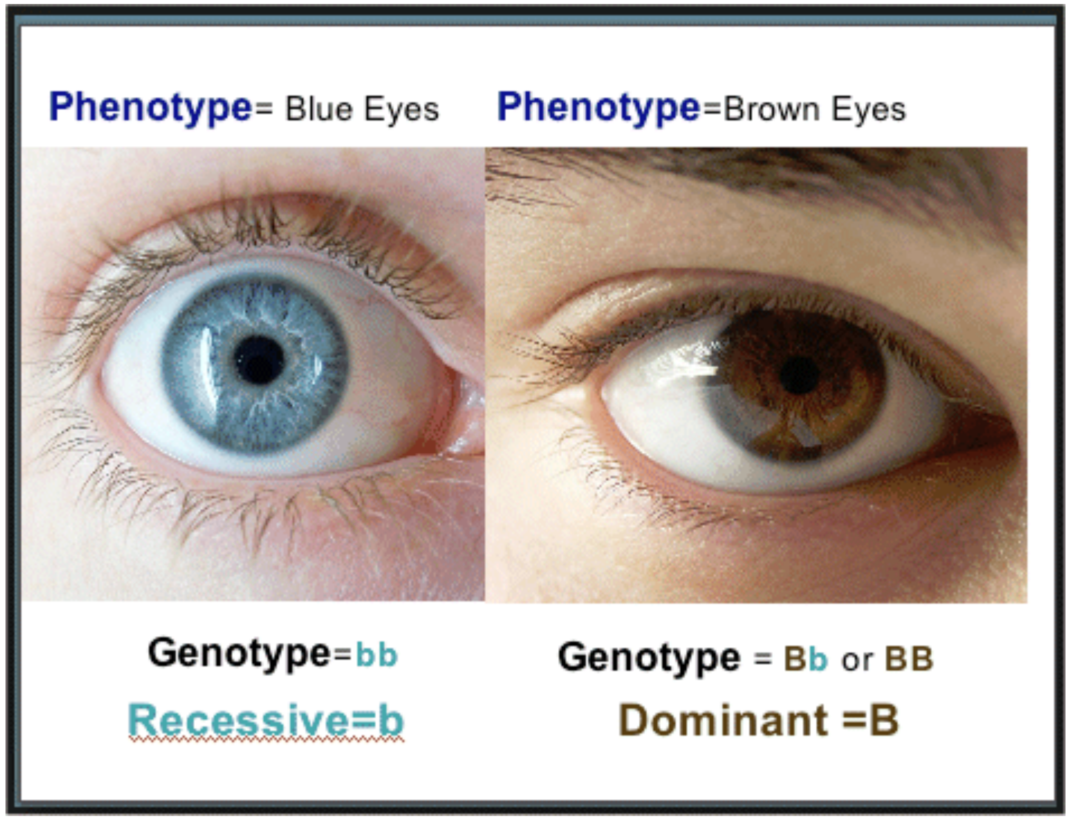
Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews: Genetics, 13(6), 395–405.

EHR: CHALLENGES

- Data
  - Diverse patient population
  - Heterogenous data types
  - Noisy & varying time scales
- Application
  - Good performance
  - Medical interpretability

PHENOTYPE

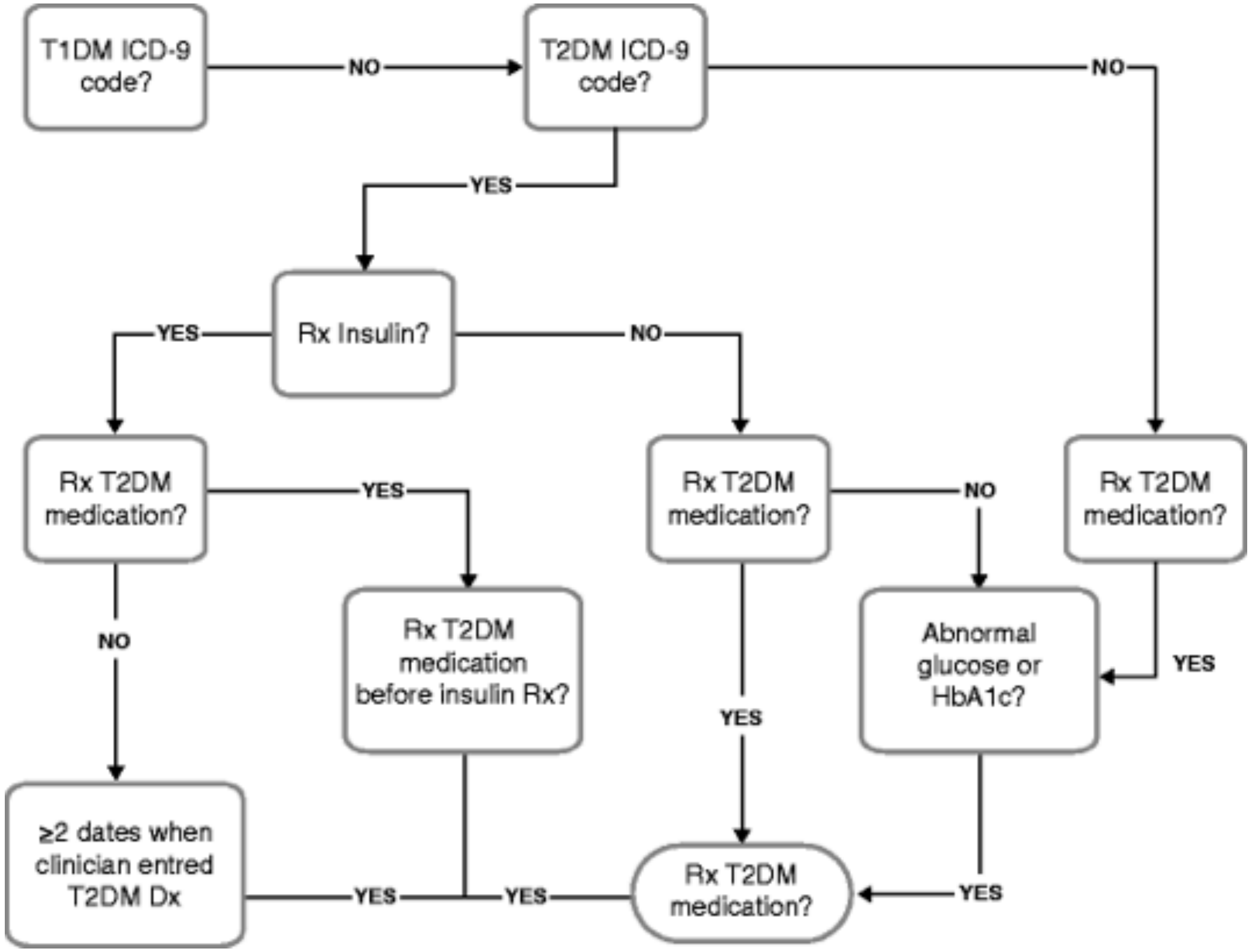
- **Observable characteristics** of an organism determined by both genetic makeup and environmental influences
- Usage
  - Retrospective research
  - Clinical trial
  - Epidemiology/ population health



Pathak, J., Kho, A. N., & Denny, J. C. (2013). Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. Journal of the American Medical Informatics Association, 20(e2), e206–e211.

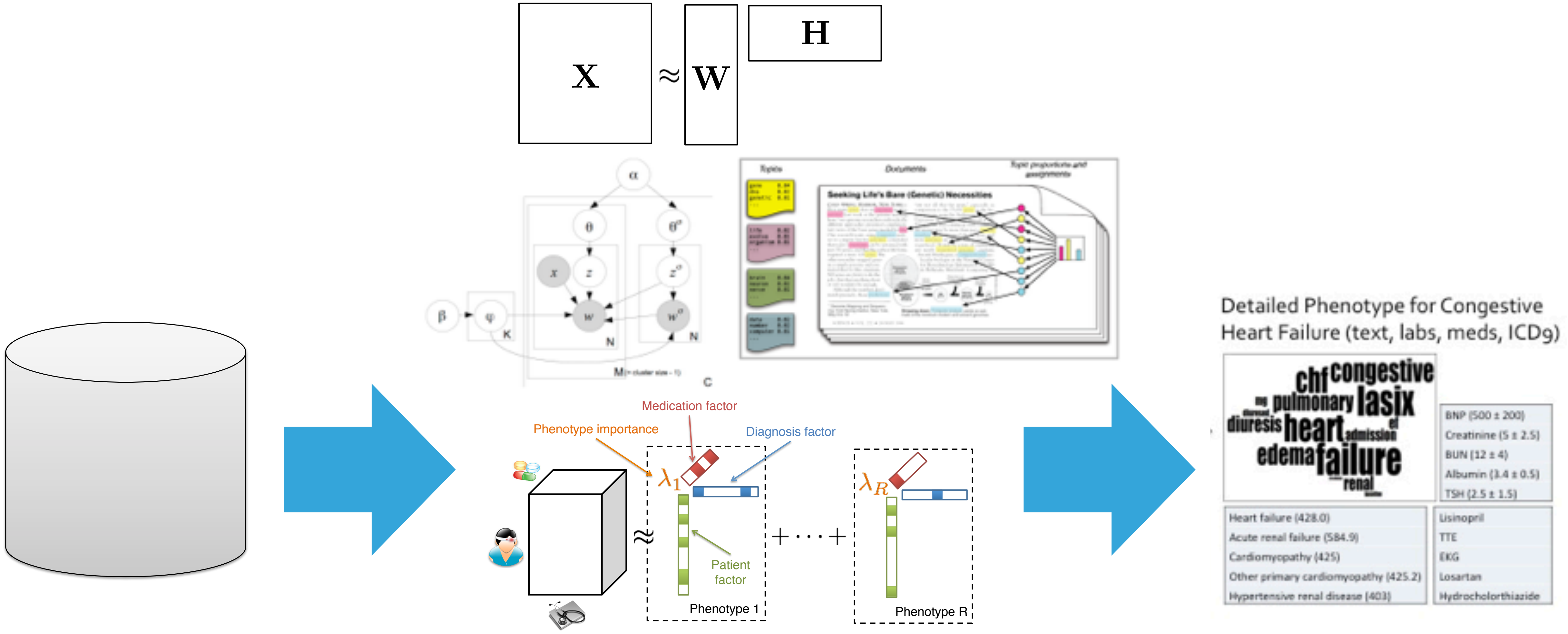
MODERN INTERPRETATION: EHR-BASED PHENOTYPING

- Specifications for identifying patients with a given condition of interest
- Concept representation easily understood (and therefore actionable) by clinicians



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

HIGH-THROUGHPUT PHENOTYPING: RECENT DEVELOPMENTS

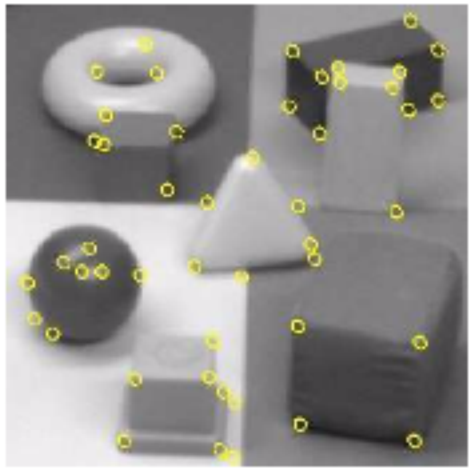

X
≈
W
H
α
θ
β
φ
K
N
M
C
Topics
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities
Medication factor
Phenotype importance
Diagnosis factor
λ_1
λ_R
≈
+ · · · +
Patient factor
Phenotype 1
Phenotype R
Detailed Phenotype for Congestive Heart Failure (text, labs, meds, ICD9)
chf congestive pulmonary lasix diuresis heart admission edema failure renal
BNP (500 ± 200)
Creatinine (5 ± 2.5)
BUN (12 ± 4)
Albumin (3.4 ± 0.5)
TSH (2.5 ± 1.5)
Heart failure (428.0)
Acute renal failure (584.9)
Cardiomyopathy (425)
Other primary cardiomyopathy (425.2)
Hypertensive renal disease (403)
Lisinopril
TTE
EKG
Losartan
Hydrocholorthiazide

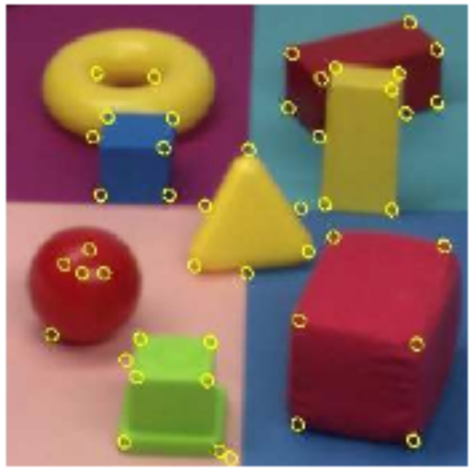
EHR database Machine learning algorithms Phenotypes

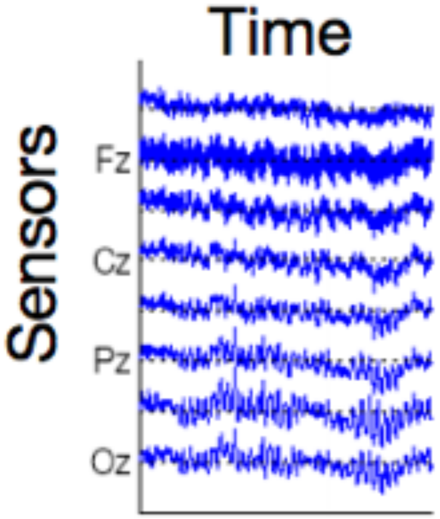
- These methods do not focus on generating sparse, diverse phenotypes with minimal supervision

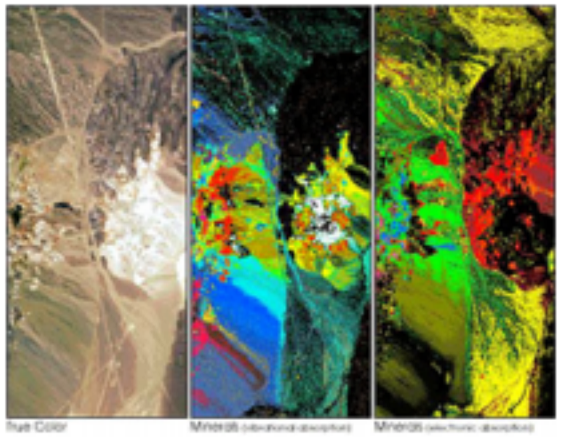
TENSORS (MULTIWAY ARRAYS)

- Generalization of matrices to multidimensional array
- Representation of an n-way interaction
- Captures hierarchical information in the structure
- Used in many domains



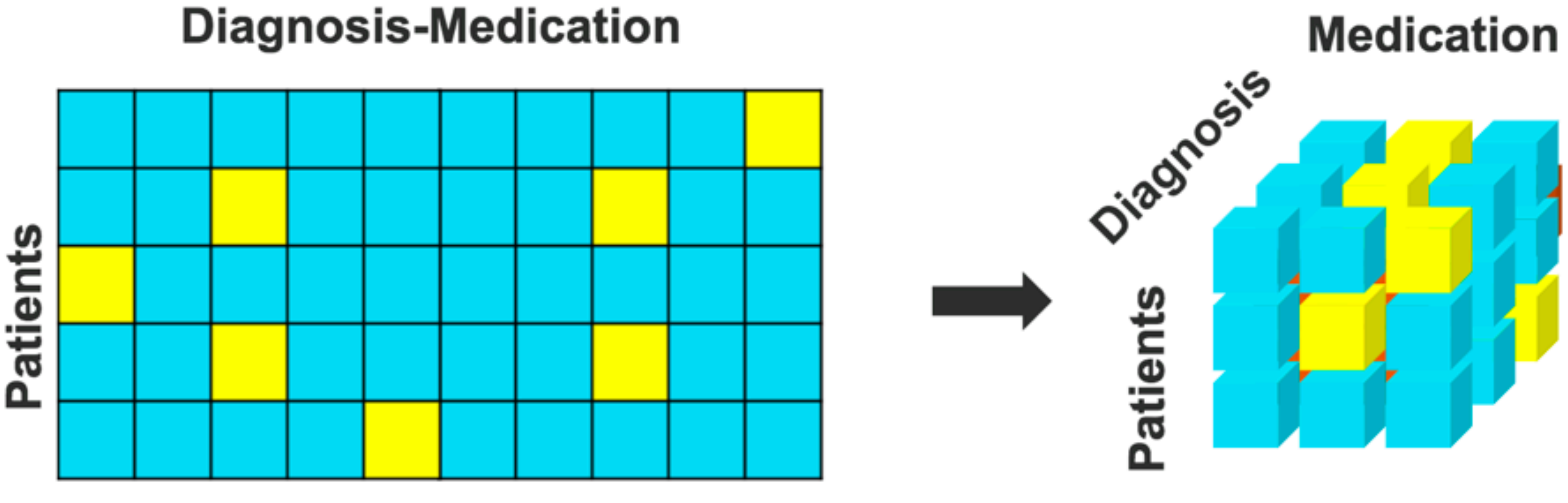





True Color
Minerals (vibrational absorption)
Minerals (electronic absorption)

Adapted from Ryoto Tomioka's Atalk @ MLSS 2015 talk on Tensor decompositions: old, new, and beyond

TENSORS



Interaction matrix of medication for specific disease

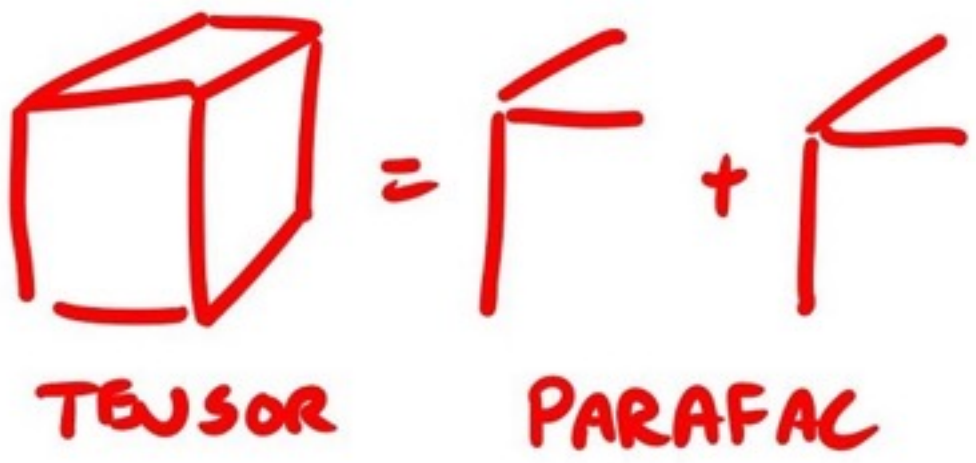
3-mode Feature Tensor

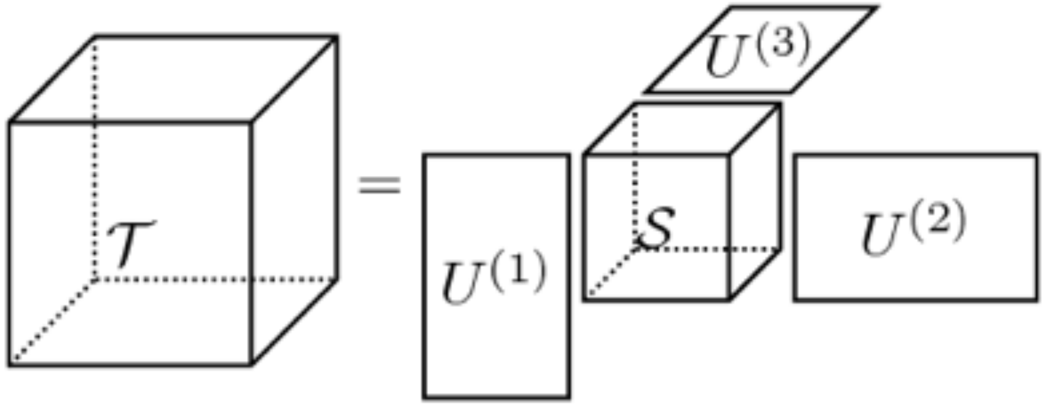
Each element represents # times a patient receives the medication to treat a specific diagnosis

Adapted from Fei Wang's Talk @ ICHI 2016

TENSOR FACTORIZATION

- Generalization of matrix factorization
- Multiway structure information utilized during decomposition process
- Many decomposition models: CANDECOMP / PARAFAC (CP), Tucker, etc.





STANDARD CP ALTERNATING LEAST SQUARES (CP-ALS)

$$
\begin{aligned}
\min \quad & \sum_{\vec{i}} (x_{\vec{i}} - m_{\vec{i}})^2 \\
\text{s.t.} \quad & \boldsymbol{\mathcal{M}} = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)}]\!]
\end{aligned}
$$

- Objective function assumes Gaussian distribution for numeric data
- Can be altered to be nonnegative
- May not be suitable for count data

CP ALTERNATING POISSON REGRESSION (CP-APR)

- Poisson distribution for nonnegative, discrete data
- Nonnegative constraints
- Stochastic column constraints

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

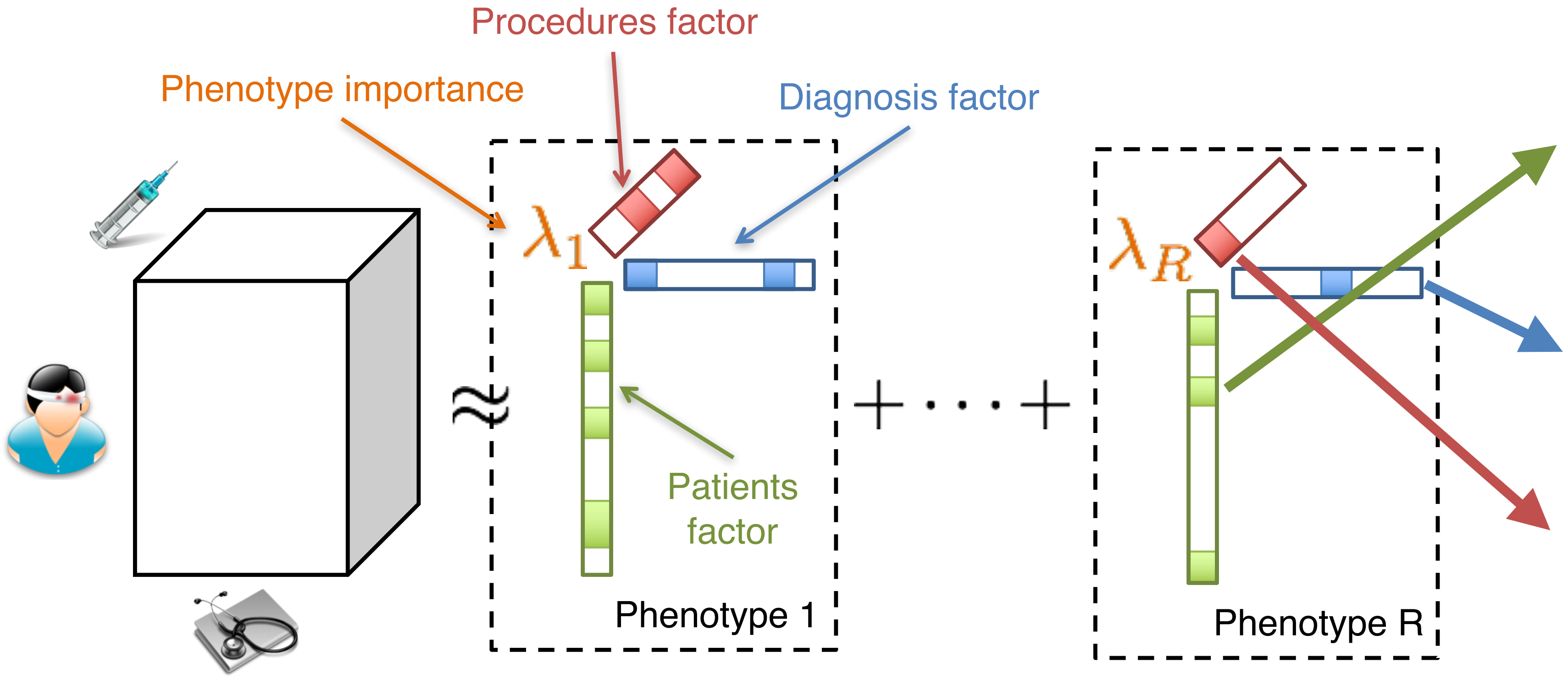
$$\text{s.t } \boldsymbol{\mathcal{M}} = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \ldots; \mathbf{A}^{(N)}]\!] \in \Omega$$

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

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

$$\Omega_n = \{\mathbf{A} \in [0,1]^{I_n \times R} \;\mid\; ||\mathbf{a}_r||_1 = 1 \;\forall r\}$$

LIMESTONE: PHENOTYPING VIA TENSOR FACTORIZATION



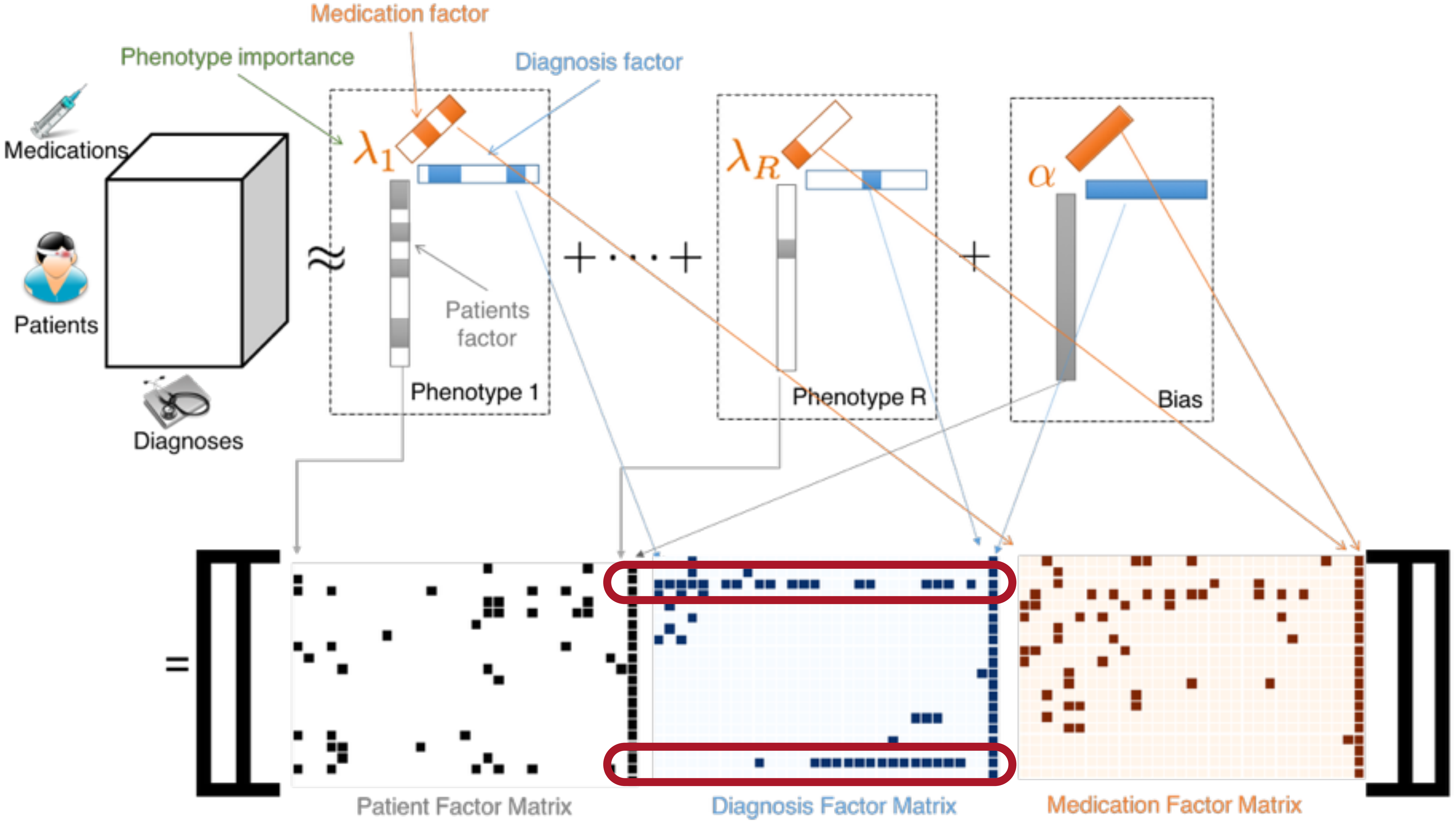
Hypertension Phenotype
(22% of patients)

- Bone/Joint/Muscle Infections/ Necrosis
- Major Symptoms, Abnormalities
- Central Nervous System Infection
- Urinary Obstruction and Retention
- Surgical Procedures on the Female Genital System
- Microbiology Procedures

Nonzero elements are clinical characteristics with the conditional probability given the phenotype and mode

Ho, J. C., Ghosh, J., Steinhubl, S. R., Stewart, W. F., Denny, J. C., Malin, B. A., & Sun, J. (2014). Limestone: High-throughput candidate phenotype generation via tensor factorization. Journal of Biomedical Informatics, 52, 199–211.

MARBLE: MOTIVATION FOR DIVERSE PHENOTYPES



OVERLAPPING ELEMENTS CAN BE DIFFICULT TO INTERPRET

GRANITE: DIVERSIFIED, SPARSE TENSOR FACTORIZATION

- Poisson model for count data
- Angular and ridge terms to reduce overlapping factors
- Simplex projection for better sparsity control
- Projected gradient descent to fit decomposition

PUSH ELEMENTS TO BE SMALL

$$\min \left(\sum_{\vec{i}} (z_{\vec{i}} - x_{\vec{i}} \log z_{\vec{i}}) + \frac{\beta_1}{2} \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{p=1}^{r} (\max\{0, \boxed{\frac{(\mathbf{a}_p^{(n)})^\top \mathbf{a}_r^{(n)}}{||\mathbf{a}_p^{(n)}||_2 ||\mathbf{a}_r^{(n)}||_2} - \theta_n\}})^2 + \frac{\beta_2}{2} \sum_{n=1}^{N} \sum_{r=1}^{R} \boxed{||\mathbf{a}_r^{(n)}||_2^2} \right)$$

$$\begin{aligned} \text{s.t } \mathcal{Z} &= [\![\sigma; \mathbf{u}^{(1)}; \cdots ; \mathbf{u}^{(N)}]\!] + [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \cdots ; \mathbf{A}^{(N)}]\!] \\ & \sigma > 0, \lambda_r \geq 0, \ \forall r \\ & \mathbf{A}^{(n)} \in [0,1]^{I_n \times R}, \mathbf{u}^{(n)} \in (0,1]^{I_n \times 1}, \ \forall n \\ & \boxed{||\mathbf{a}_r^{(n)}||_1 = ||\mathbf{u}^{(n)}||_1 = 1} \ \forall n \end{aligned}$$

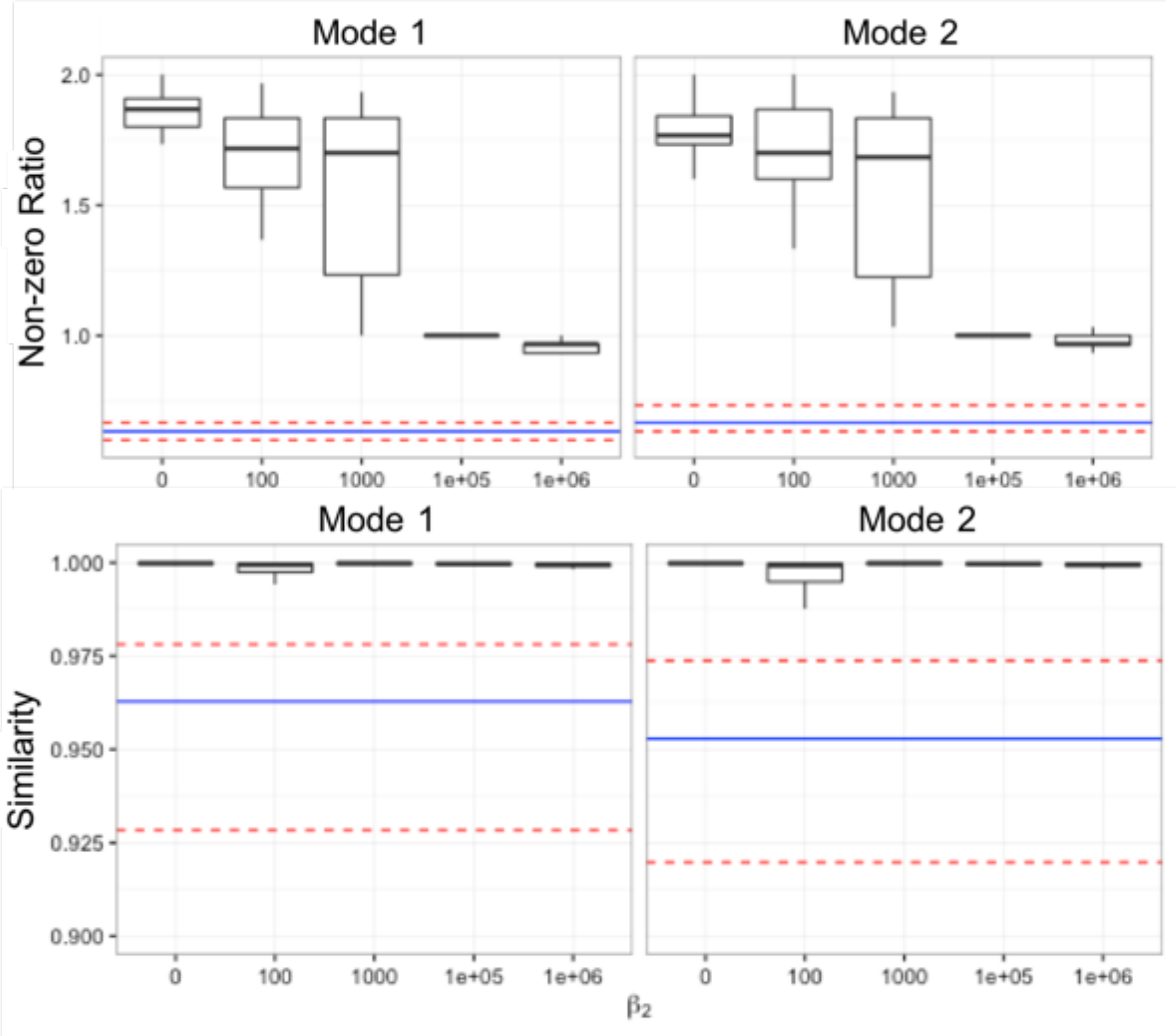
REDUCE COSINE SIMILARITY FOR INTRA-PHENOTYPE DIVERSITY

SPARSITY CONTROL

Henderson, J., Ho, J. C., Kho, A.K., Denny, J. C., Malin, B. A., Sun, J., & Ghosh, J. (2017). Granite: Diversified, Sparse Tensor Factorization for Electronic Health Record-Based Phenotyping. Proceedings of ICHI 2017.

SIMULATED TENSORS: ACCURATE RECOVERY

- Simulated 50 third-order tensor of size 40 x 20 x 20 with rank of 5 with cosine similarity threshold set to .3
- Fit Granite and Marble decompositions

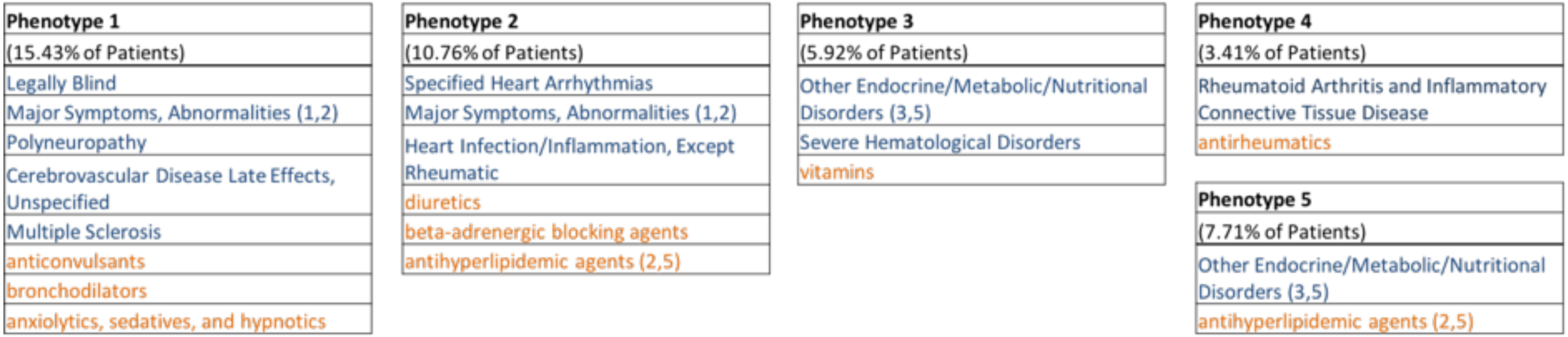


DATA: VANDERBILT UNIVERSITY SYNTHETIC DERIVATIVE

- Inpatient and outpatient billing and medication codes for nearly 2 million patients
- Focus on resistant hypertension
  - 1394 patients (33% cases) - manually identified by domain experts
  - 177 diagnoses (HCC categories)
  - 149 medications (MeSH PA)
- Compare Granite, Marble, CP-APR, CP-ALS, NMF

RESULTS: TOP 5 RESULTING PHENOTYPES

Granite



Phenotype 1
(15.43% of Patients)
Legally Blind
Major Symptoms, Abnormalities (1,2)
Polyneuropathy
Cerebrovascular Disease Late Effects, Unspecified
Multiple Sclerosis
anticonvulsants
bronchodilators
anxiolytics, sedatives, and hypnotics

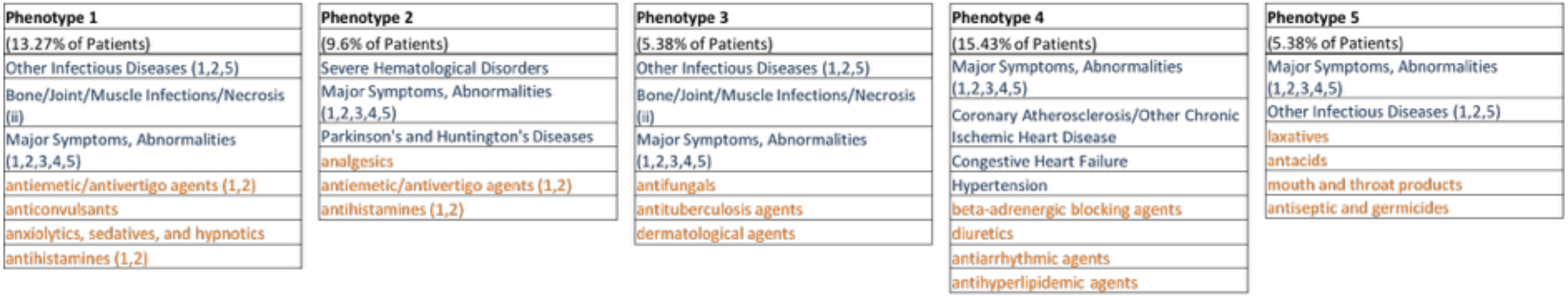
Phenotype 2
(10.76% of Patients)
Specified Heart Arrhythmias
Major Symptoms, Abnormalities (1,2)
Heart Infection/Inflammation, Except Rheumatic
diuretics
beta-adrenergic blocking agents
antihyperlipidemic agents (2,5)

Phenotype 3
(5.92% of Patients)
Other Endocrine/Metabolic/Nutritional Disorders (3,5)
Severe Hematological Disorders
vitamins

Phenotype 4
(3.41% of Patients)
Rheumatoid Arthritis and Inflammatory Connective Tissue Disease
antirheumatics

Phenotype 5
(7.71% of Patients)
Other Endocrine/Metabolic/Nutritional Disorders (3,5)
antihyperlipidemic agents (2,5)

Marble



Phenotype 1
(13.27% of Patients)
Other Infectious Diseases (1,2,5)
Bone/Joint/Muscle Infections/Necrosis (ii)
Major Symptoms, Abnormalities (1,2,3,4,5)
antiemetic/antivertigo agents (1,2)
anticonvulsants
anxiolytics, sedatives, and hypnotics
antihistamines (1,2)

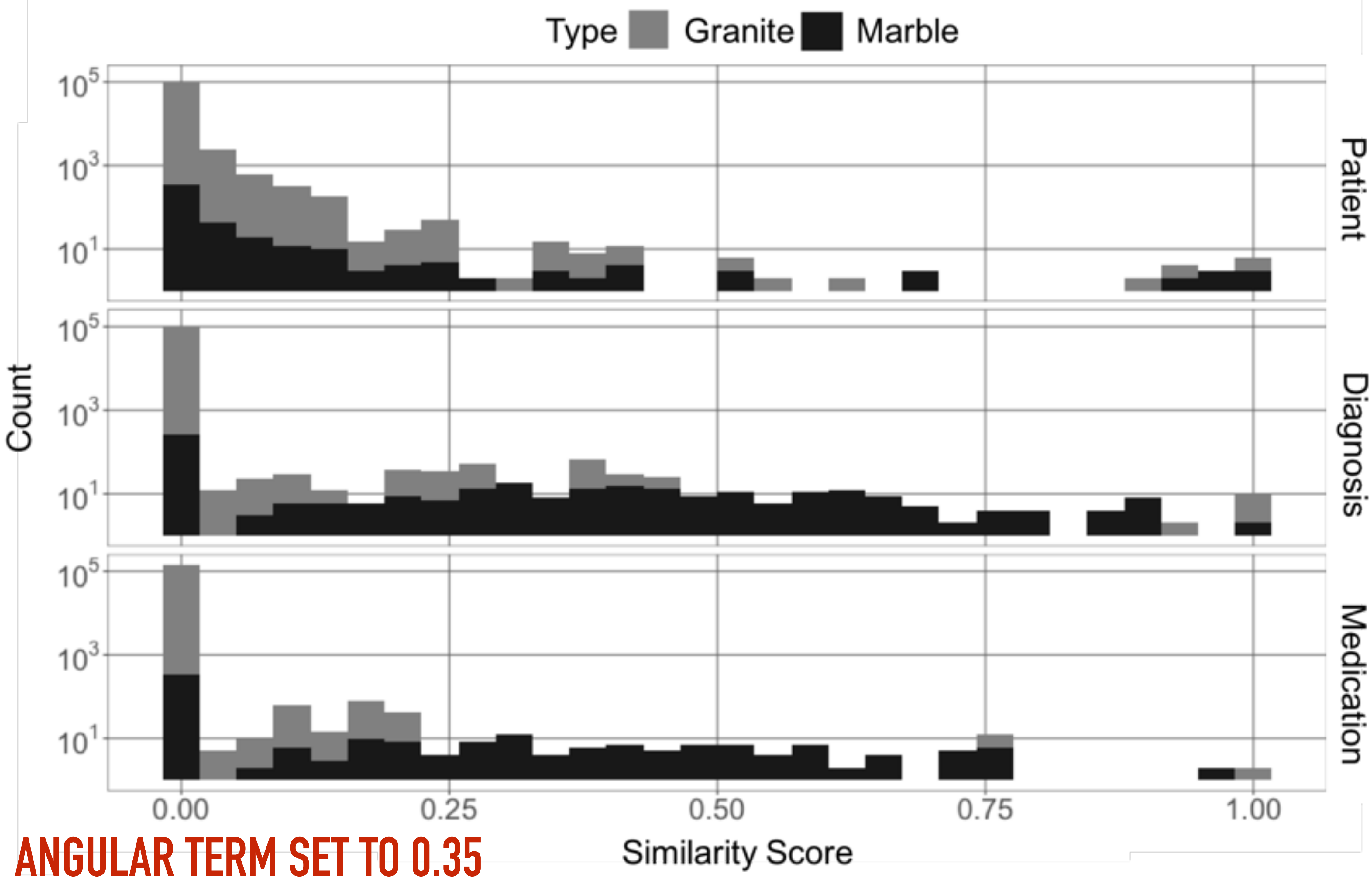
Phenotype 2
(9.6% of Patients)
Severe Hematological Disorders
Major Symptoms, Abnormalities (1,2,3,4,5)
Parkinson's and Huntington's Diseases
analgesics
antiemetic/antivertigo agents (1,2)
antihistamines (1,2)

Phenotype 3
(5.38% of Patients)
Other Infectious Diseases (1,2,5)
Bone/Joint/Muscle Infections/Necrosis (ii)
Major Symptoms, Abnormalities (1,2,3,4,5)
antifungals
antituberculosis agents
dermatological agents

Phenotype 4
(15.43% of Patients)
Major Symptoms, Abnormalities (1,2,3,4,5)
Coronary Atherosclerosis/Other Chronic Ischemic Heart Disease
Congestive Heart Failure
Hypertension
beta-adrenergic blocking agents
diuretics
antiarrhythmic agents
antihyperlipidemic agents

Phenotype 5
(5.38% of Patients)
Major Symptoms, Abnormalities (1,2,3,4,5)
Other Infectious Diseases (1,2,5)
laxatives
antacids
mouth and throat products
antiseptic and germicides

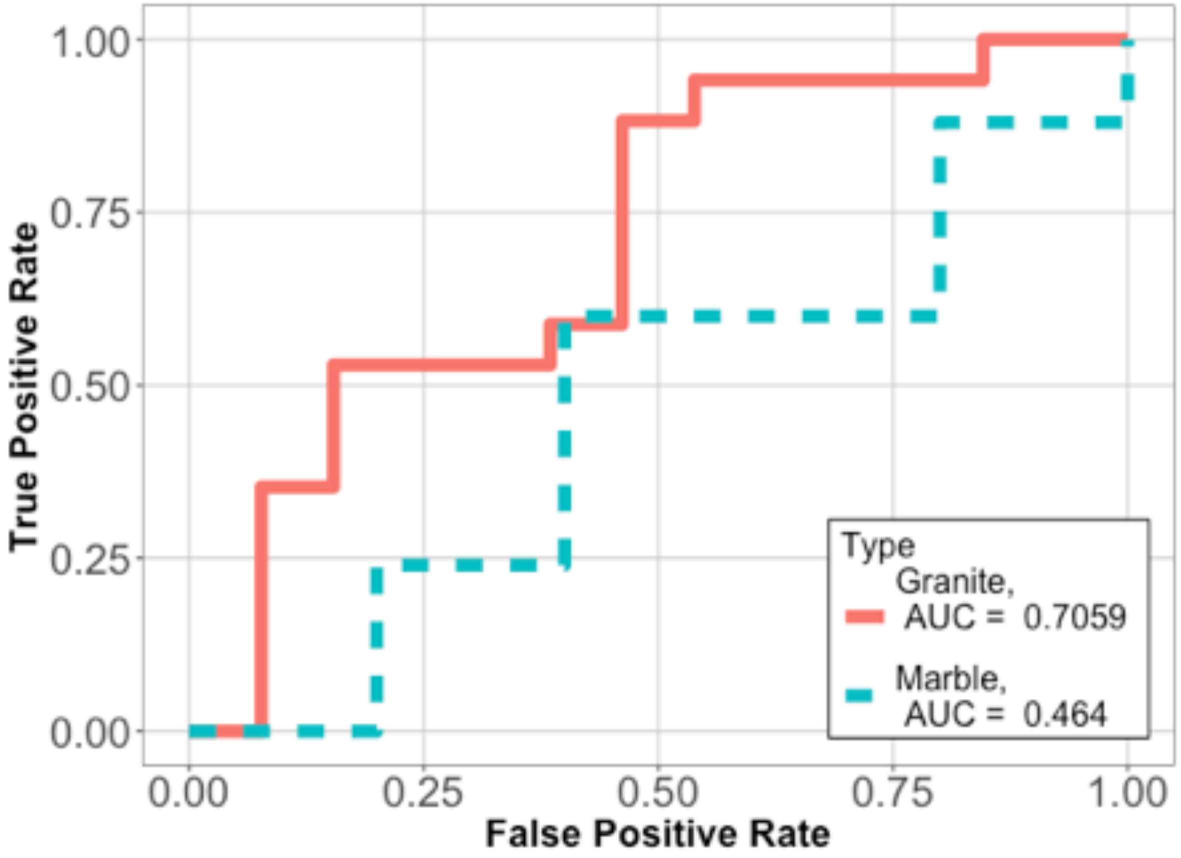
RESULTS: COSINE SIMILARITY



ANGULAR TERM SET TO 0.35

RESULTS: IMPORTANCE OF PHENOTYPE WEIGHTS

- Domain expert annotated phenotypes into 3 categories
  - Clinically relevant
  - Possibly clinically relevant
  - Not relevant
- Granite generated fewer clinically relevant ones than Marble



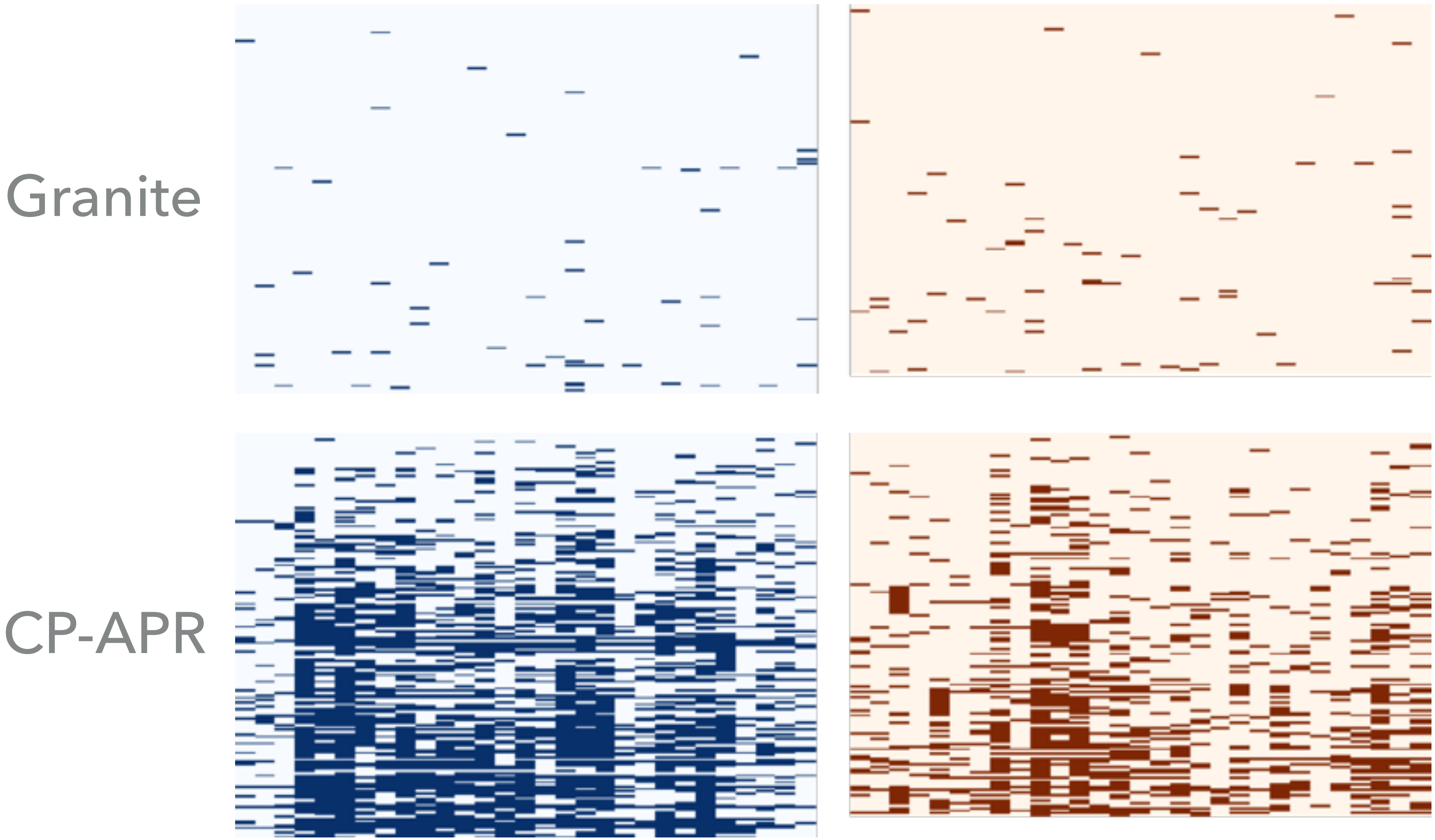
HIGH CORRELATION BETWEEN WEIGHTS AND CLINICAL RELEVANCY

RESULTS: RESISTANT HYPERTENSION PREDICTION

- Task: Predict case vs controls
- 5 80-20 train/test splits with stratified sampling
- Logistic regression with Lasso
    - 10-fold CV to learn weight
    - Train on loadings (patient) matrix with R = 30

Model	AUC	NNZ / Phenotype
Granite	0.7298	**4.63**
Marble	0.7197	5.3330
CP-APR	**0.7406**	111.0000
CP-ALS	0.6765	113.1522
NMF	0.7203	N/A

RESULTS: NON-ZERO ELEMENTS



CONCLUSION

- Granite provides an unsupervised framework to extract concise and diverse phenotypes that retain predictive power

FUTURE WORK

- Provide weak supervision using outside data sources to increase the number of clinically relevant phenotypes

COLLABORATORS

- Emory University: Joyce C. Ho
- UT-Austin: Joydeep Ghosh
- GaTech: Jimeng Sun
- Vanderbilt: Joshua Denny & Bradley A Malin
- Northwestern: Abel N Kho

CONTACT INFORMATION

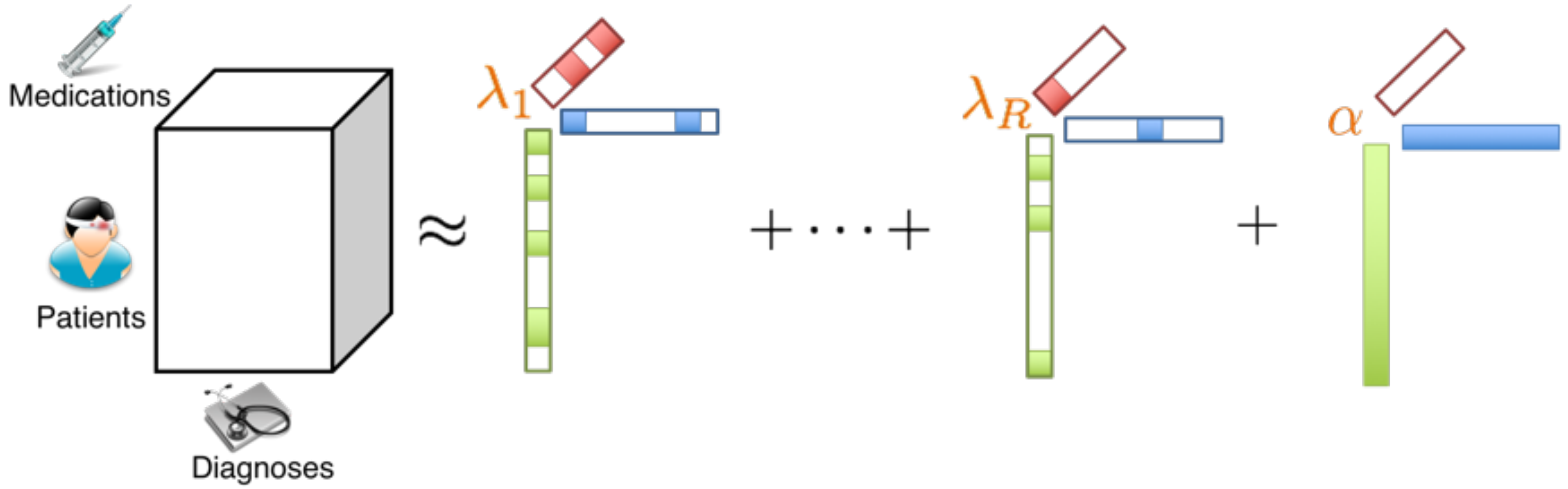
- Email: jette@ices.utexas.edu
- Website: http://ejette.github.io/

TEXAS

The University of Texas at Austin

FEATURE MATRIX

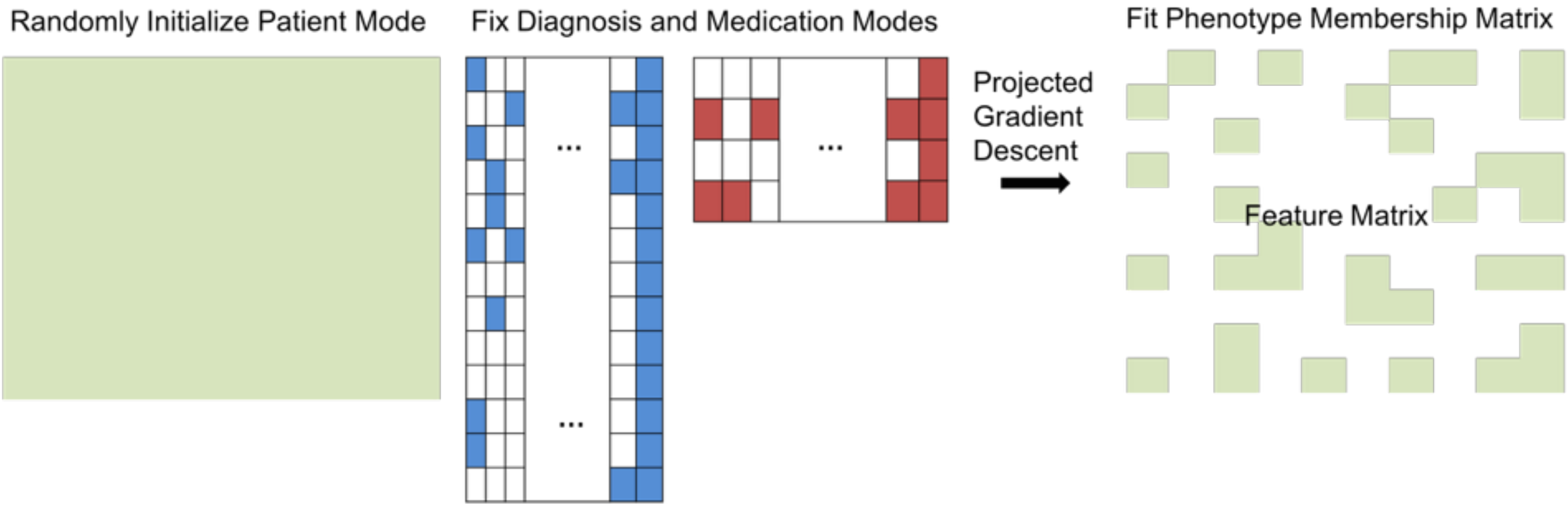
- Fit a decomposition on a set of patients X



- For a new set of patients X_{test}, fix diagnosis and medication modes and use projected gradient descent to fit a new patient mode
- Row normalize new patient mode to find a patient's membership to phenotypes

FEATURE MATRIX

- Fit a decomposition on a set of patients X_{train}


Randomly Initialize Patient Mode
Fix Diagnosis and Medication Modes
...
...
...
Projected Gradient Descent
Fit Phenotype Membership Matrix
Feature Matrix

- For a new set of patients X_{test}, fix diagnosis and medication modes and use projected gradient descent to fit a new patient mode
- Row normalize new patient mode to find a patient's membership to phenotypes