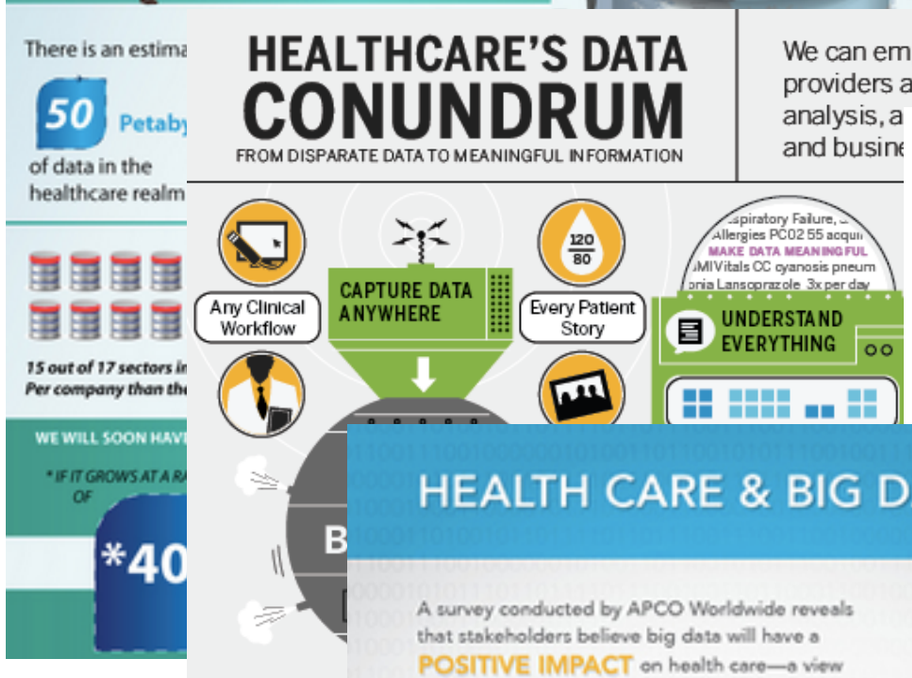
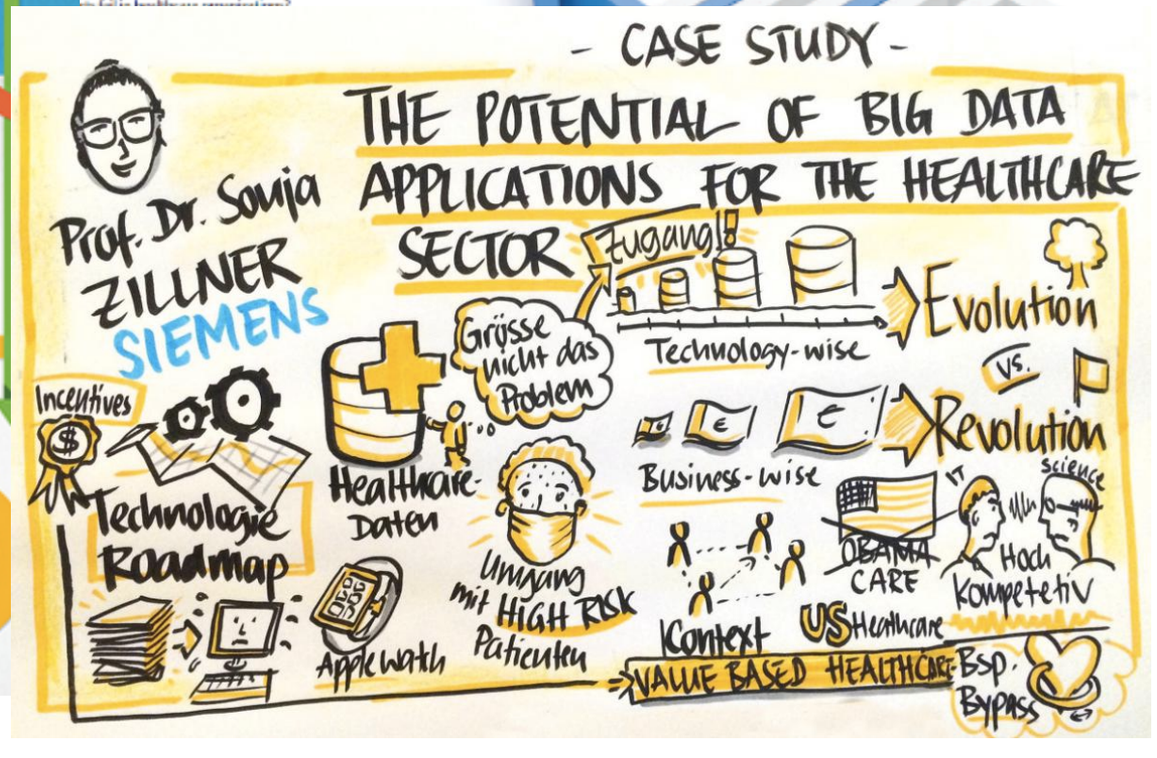
Clinically Interpretable Models for Healthcare Data

Joyce Ho
Emory University

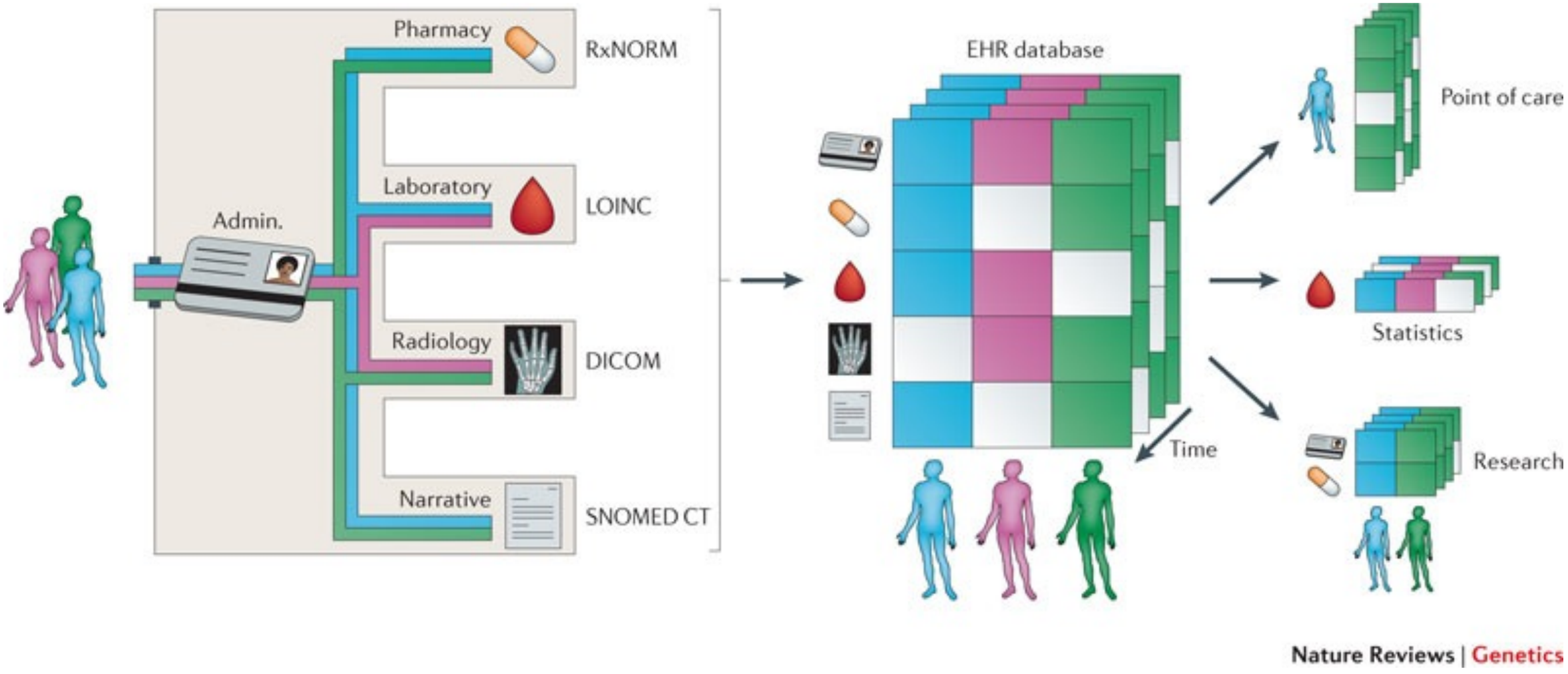
April 1, 2016

Electronic Health Data: Hype or Hope





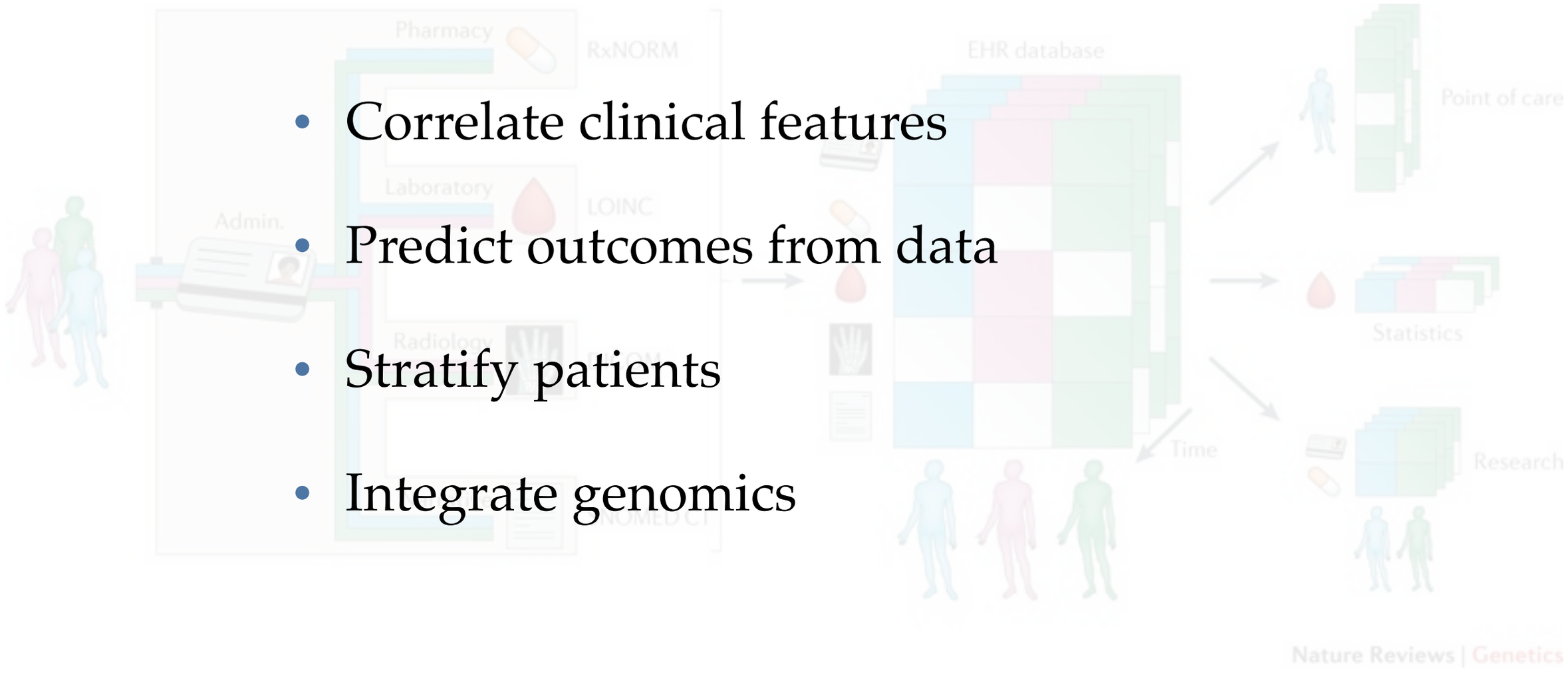
Electronic Health Records (EHRs)



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.

Electronic Health Records (EHRs)

- Correlate clinical features
- Predict outcomes from data
- Stratify patients
- Integrate genomics



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.

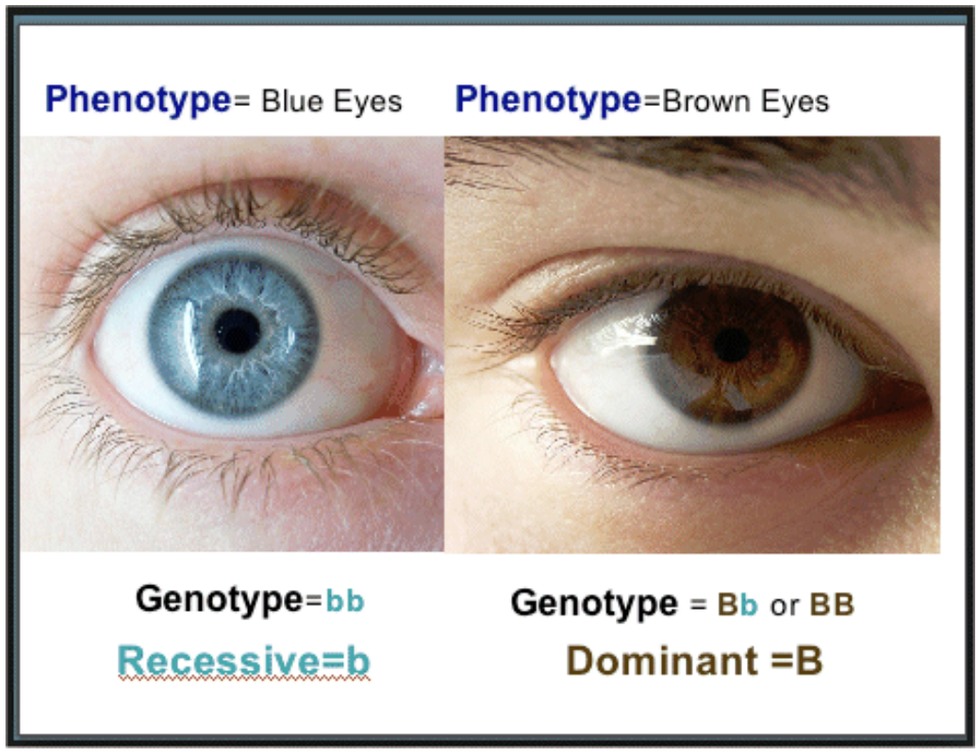
EHR Challenges

- Diverse patient population
- Heterogenous data types
- Noisy data
- Varying time scales

Also, medical interpretability is important!

Phenotype

- **Observable characteristics** of an organism determined by both genetic makeup and environmental influences

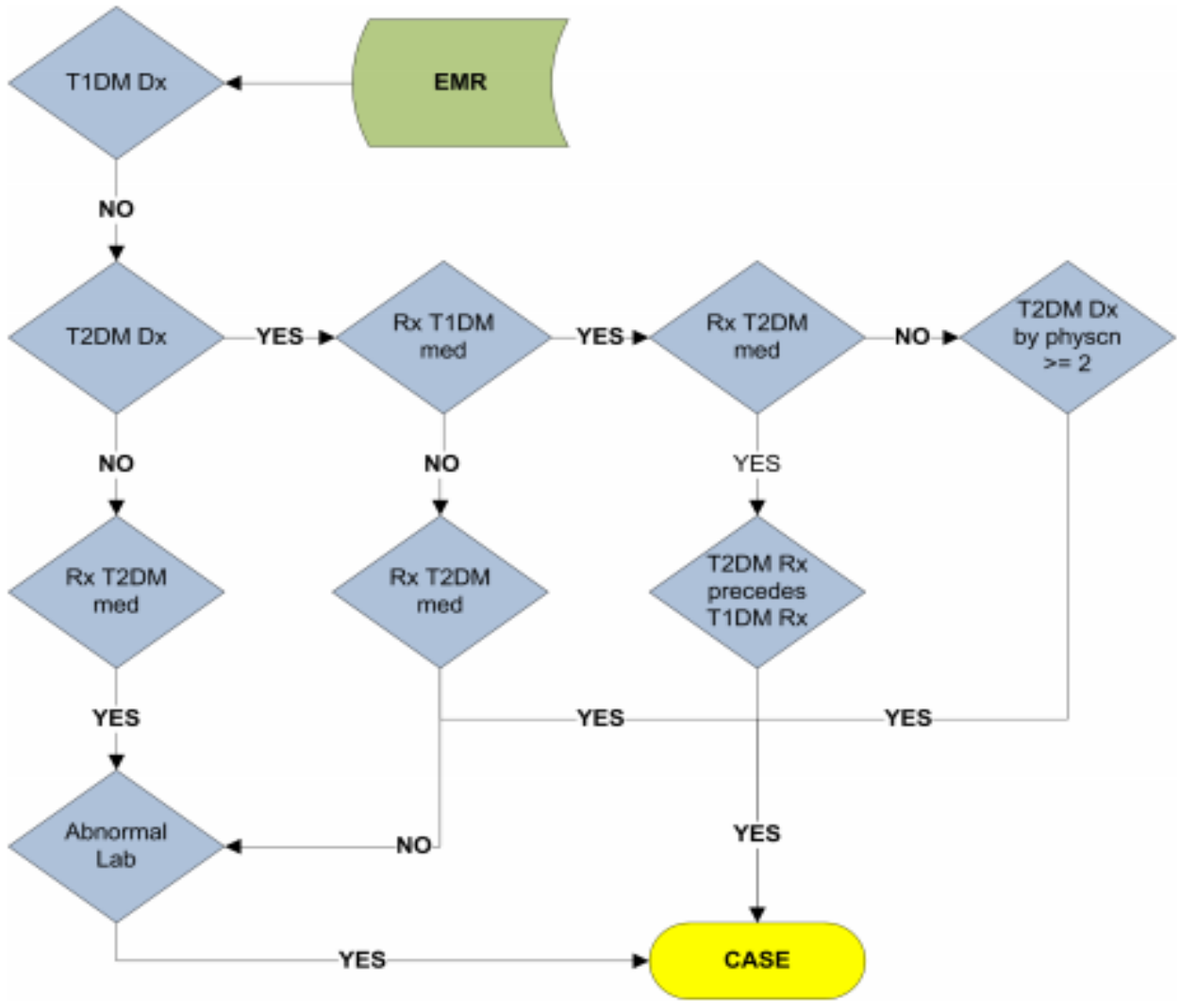


- Uses
  - Retrospective research
  - Clinical trial
  - Epidemiology / population health

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.

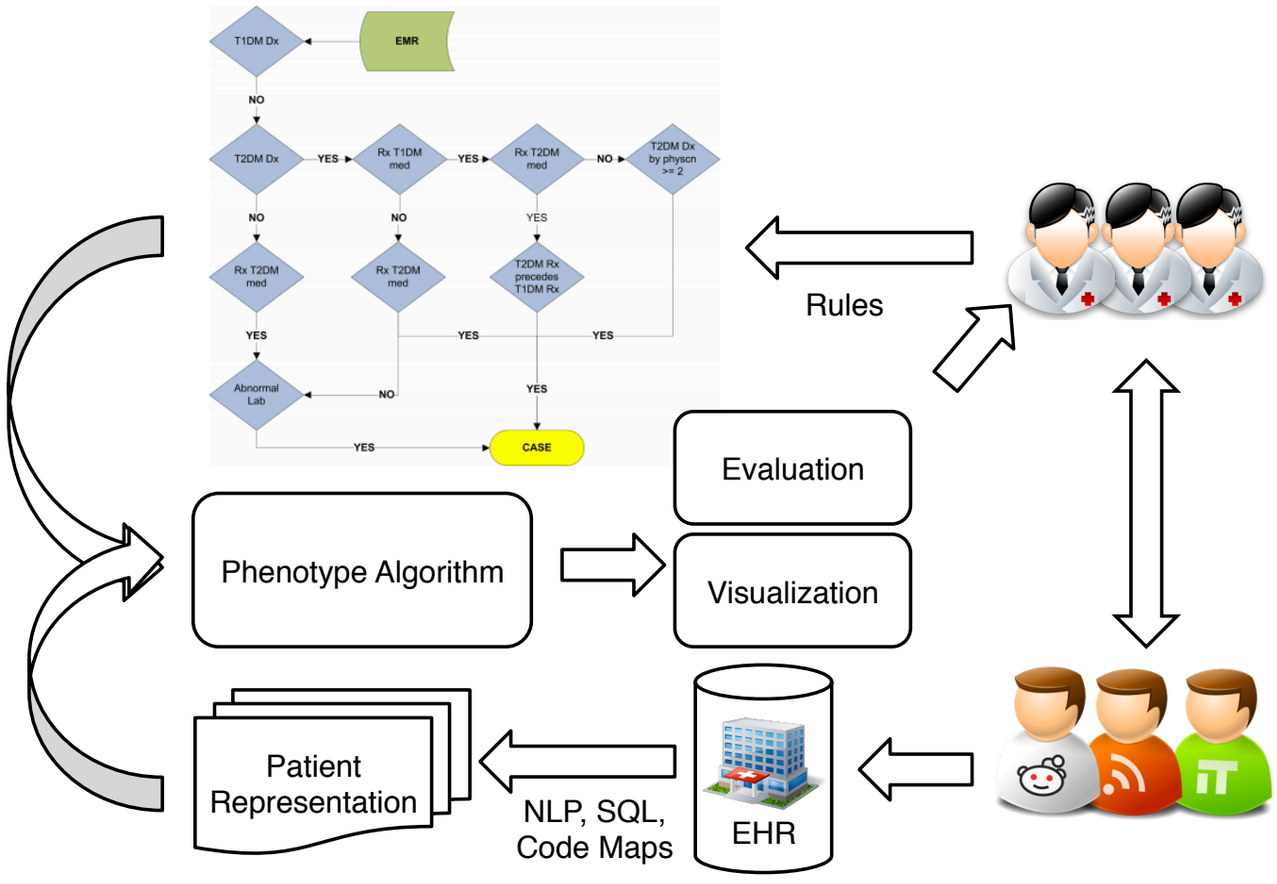
Phenotype: Modern Interpretation

- Specifications for identifying patients with a given characteristic (or condition) of interest
- Concept representation that is 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.

Current Phenotyping Process



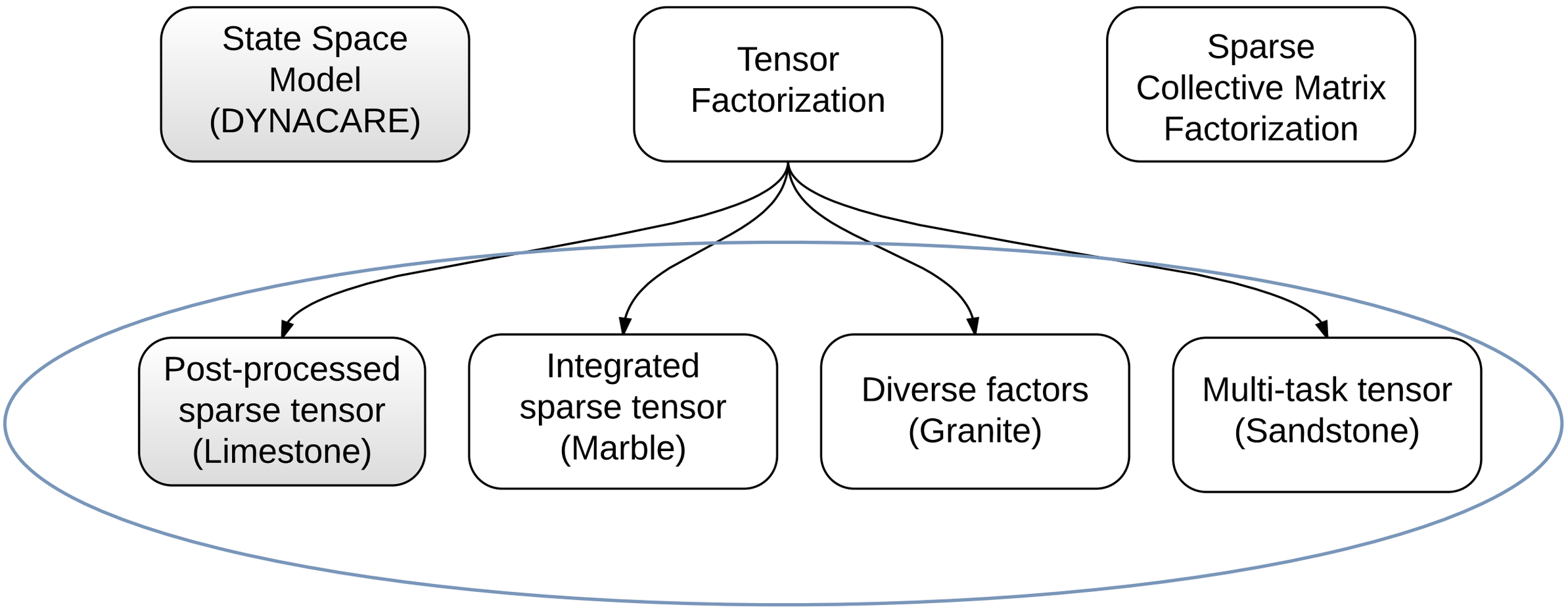
- Iterative & time-consuming
- Single disease-specific phenotype
- Human annotated samples necessary
- Not easily portable (cross-institutional)

Newton, K. M., Peissig, P. L., Kho, A. N., Bielinski, S. J., Berg, R. L., Choudhary, V., et al. (2013). Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association, 20(e1), e147–e154.

Our Approaches

- Derivation of medical concepts can be viewed as a form of dimensionality reduction
- Development of latent variable models that are:
  - succinct and concise
  - easily understood by medical professional
  - require minimal human supervision

Our Approaches



Collaborators







Suriya Gunasekar (UT), Jette Henderson (UT),
Joydeep Ghosh (UT), Jimeng Sun (GaTech),
Brad Malin, Josh Denny (Vanderbilt), Abel Kho, (NW)

Vector Representation

Each patient is summarized via a single vector

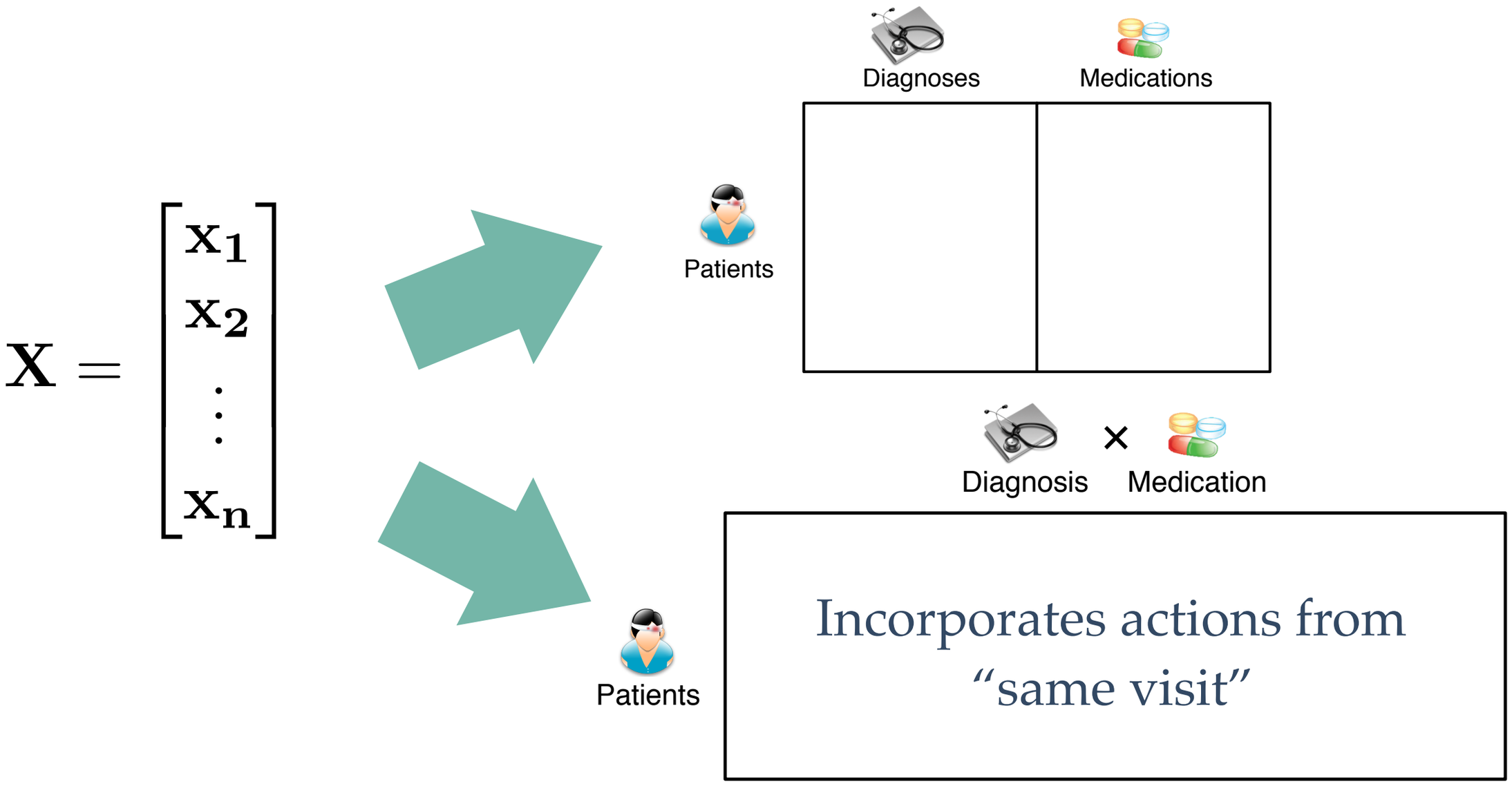
$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix}$$

Type	Value
Diagnosis	frequency of code
Medication	number of prescription
Lab	recent test result
Physiological	summary statistic of measurements

Feature Matrix Representation

Patient vectors are stacked together to obtain a matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_n} \end{bmatrix}$$

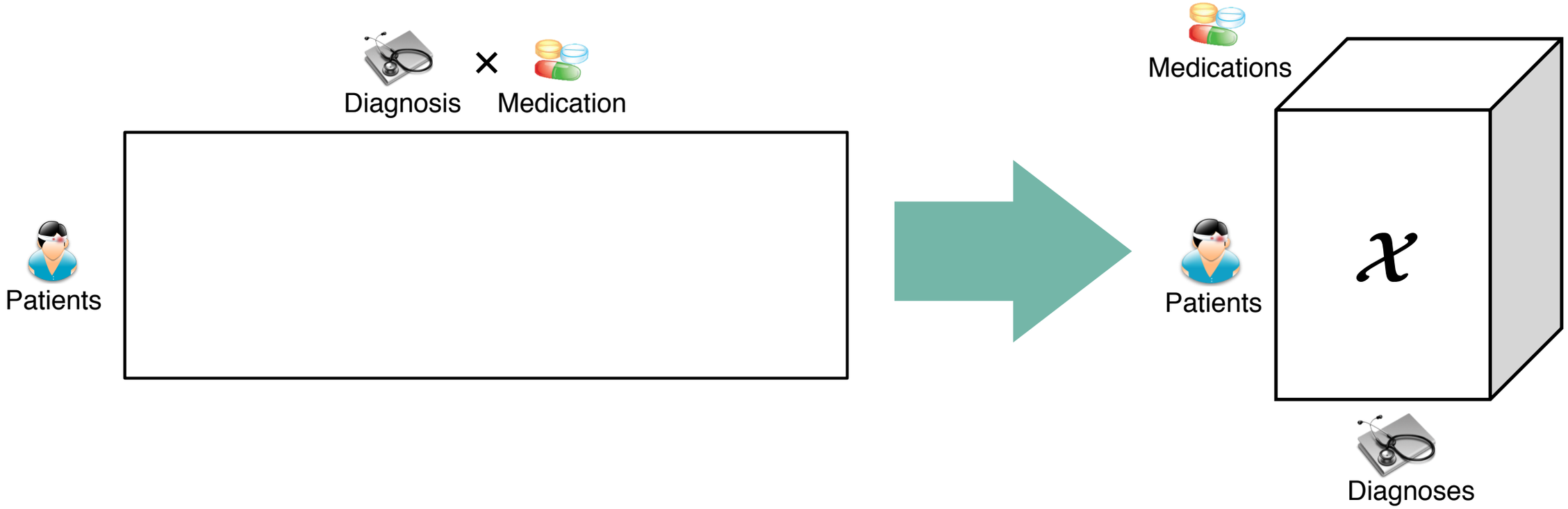


Incorporates actions from "same visit"

Tensors (Multiway Arrays)

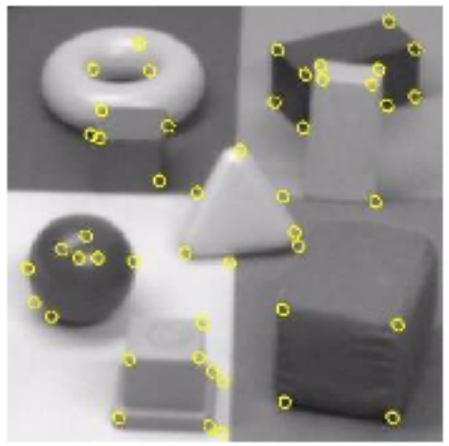
- Generalization of scalars, vectors, and matrices to multidimensional array
- Representation of an n-way interaction
- Captures hierarchical information in the structure
- Used in lots of places (e.g., chemistry, neuroimaging, bioinformatics, text mining, psychology, etc.)

Tensors (Multiway Arrays)

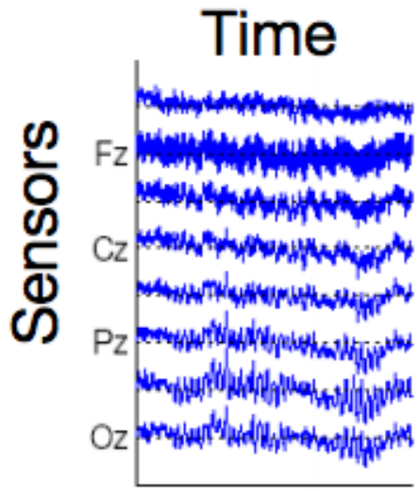


Tensors are Everywhere

Matrices



Black and White



Multivariate time series

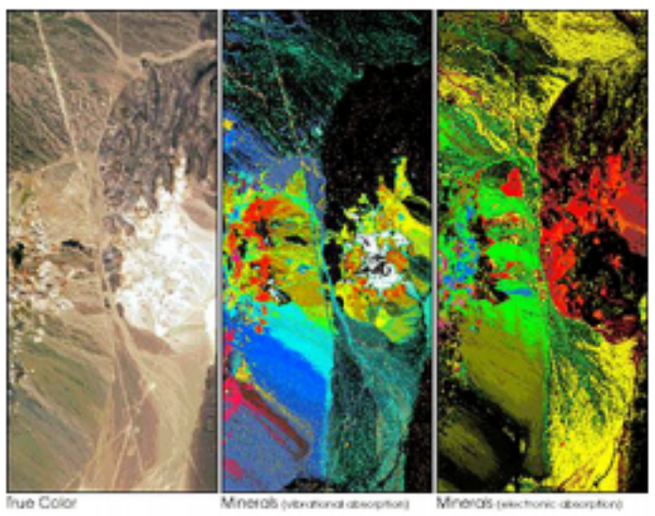
Movies

Users	Star Wars	Titanic	Blade Runner
User 1	5	2	4
User 2	1	4	2
User 3	5	?	?

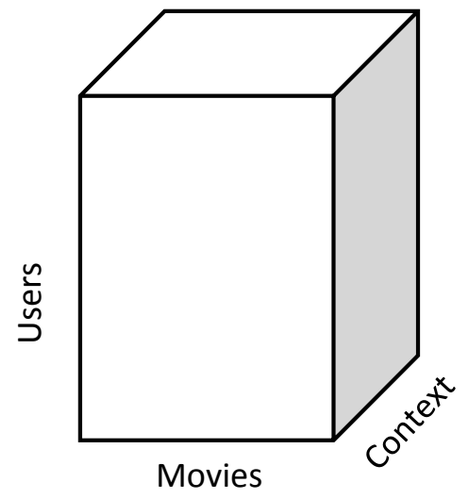
Movie recommendation

Tensors

Color



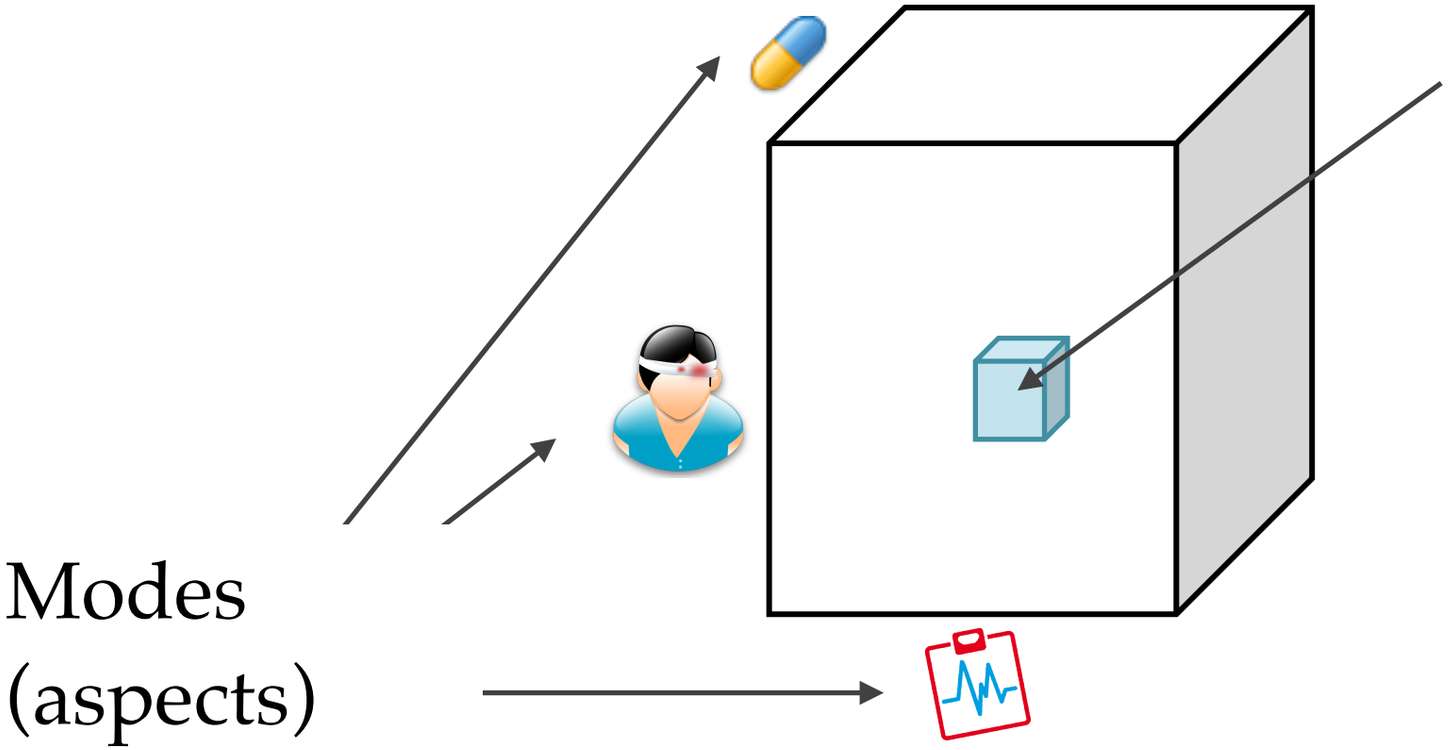
Spatio-temporal data



Multiple relations

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

3-D Mode Tensor



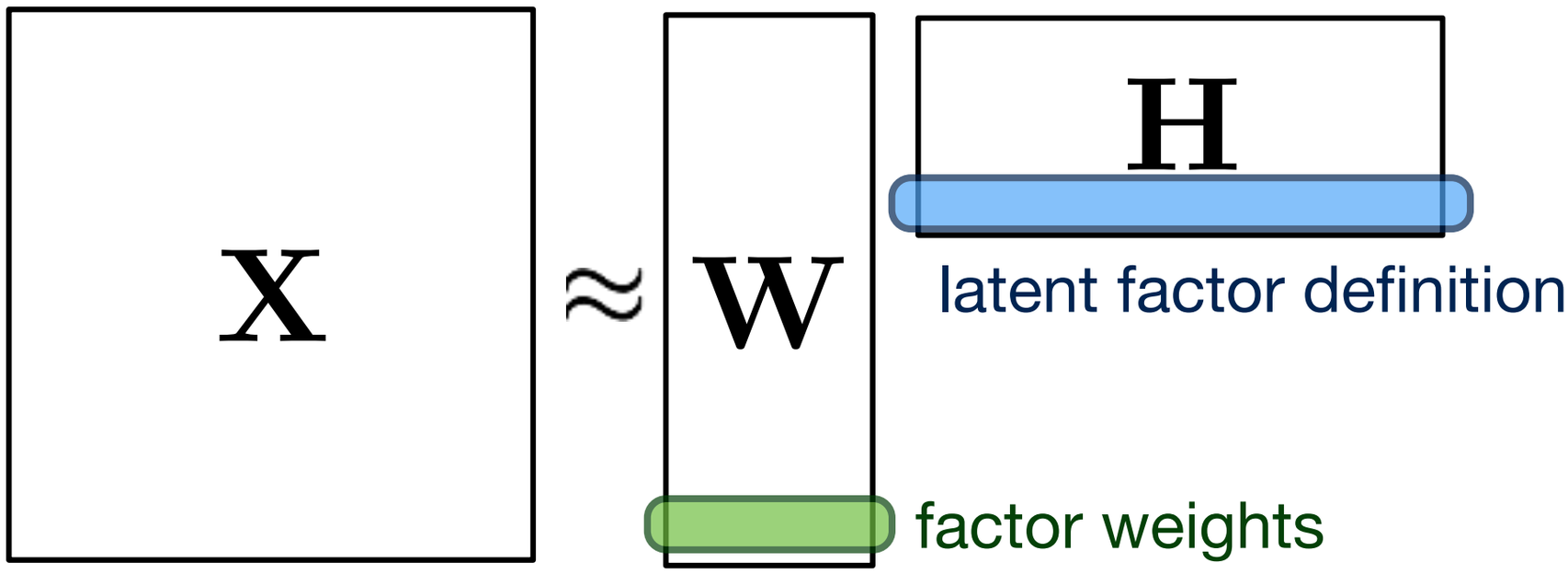
Each element represents the relationship between patient, medication, and diagnosis

Modes (aspects) of the tensor

Dimensionality Reduction

- Generate a low-dimensional encoding of a high-dimensional space
- Purposes:
  - Data compression / visualization
  - Robustness to noise and uncertainty
  - Potentially easier to interpret

Matrix Factorization



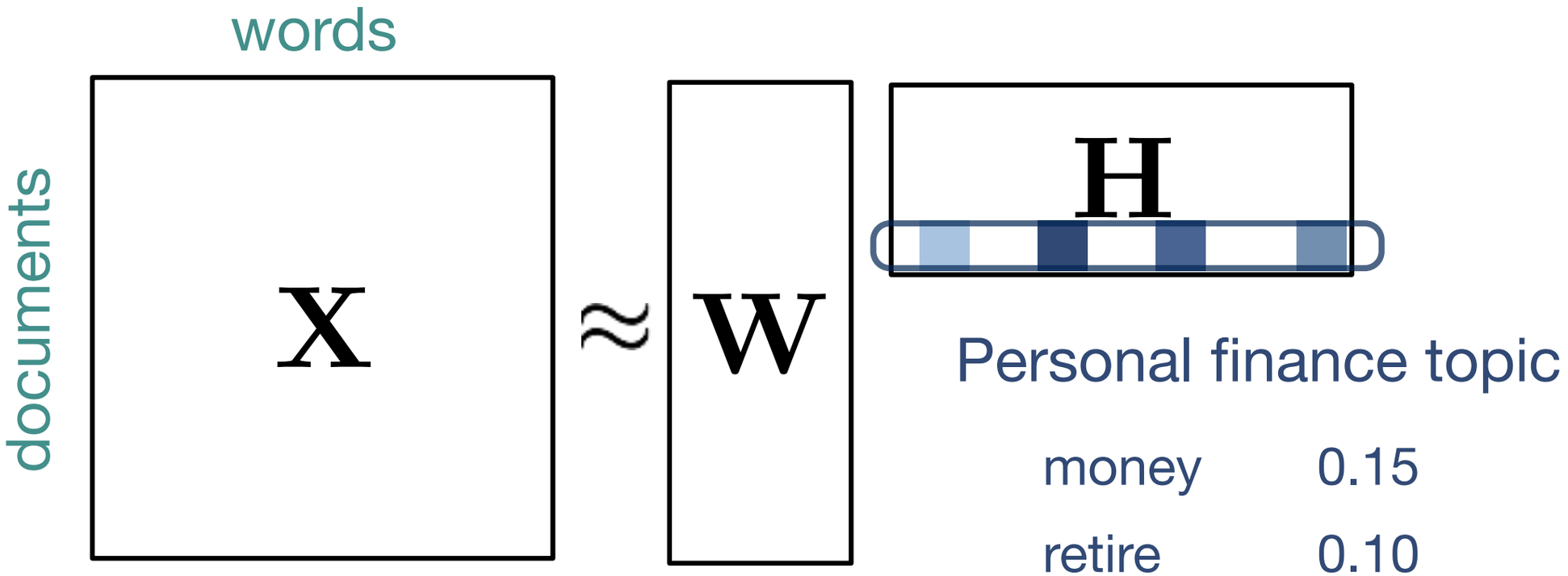
- Low rank approximation to original matrix
- Common dimensionality reduction (recommendation systems, clustering, etc.)
- Can uncover latent relations

Nonnegative Matrix Factorization (NMF)

- Both **W** and **H** are nonnegative
- Empirically induces sparsity
- Improved interpretability (sum of parts representation)
- Popularized by Lee and Seung (1999) for "learning the parts of objects"

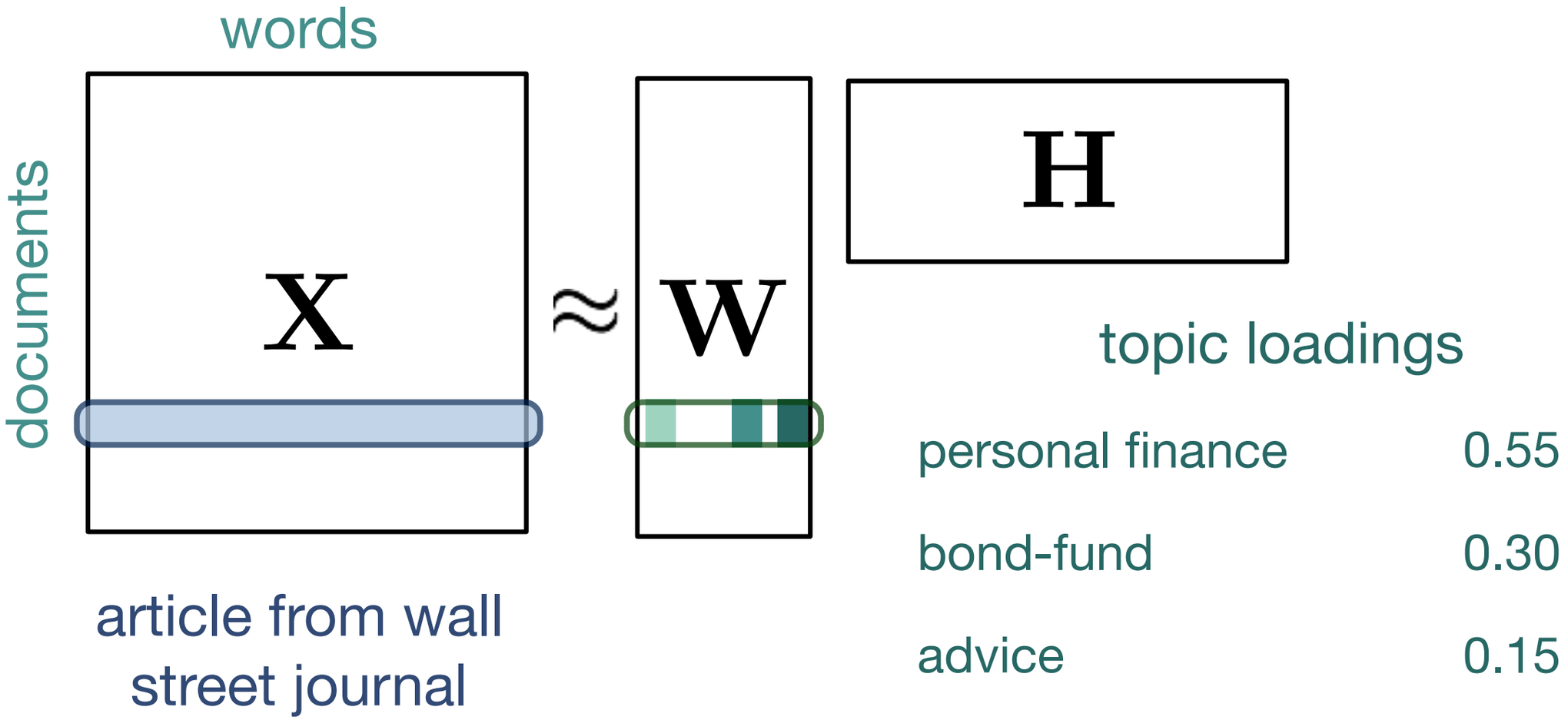
$$\mathbf{X} \approx \mathbf{W}\mathbf{H}^\top$$
$$\text{s.t. } \mathbf{W} \geq 0, \ \mathbf{H} \geq 0$$

NMF (Example)



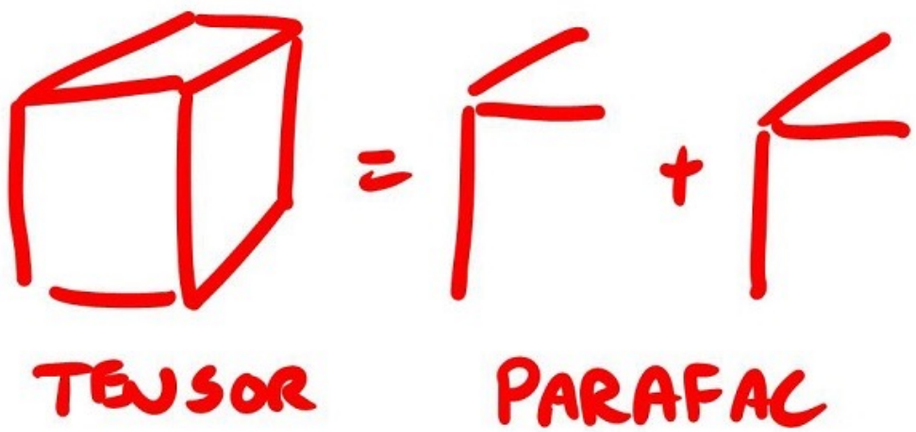
WLOG, assume columns of $\mathbf{W}$ and $\mathbf{H}$ sum to 1

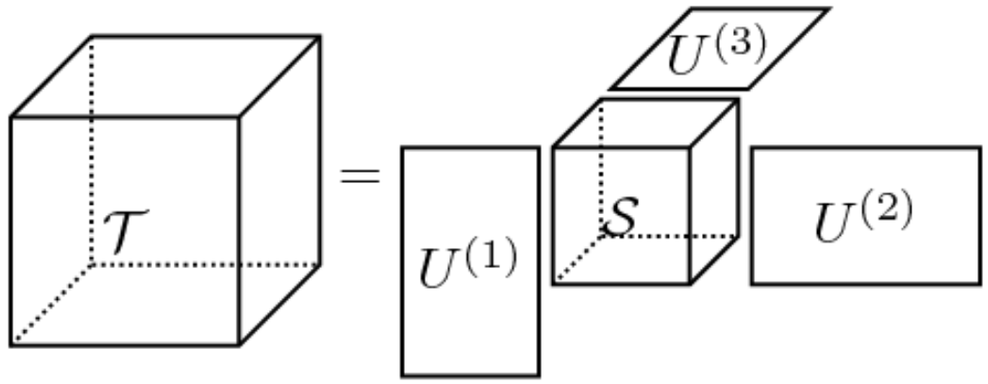
NMF (Example)

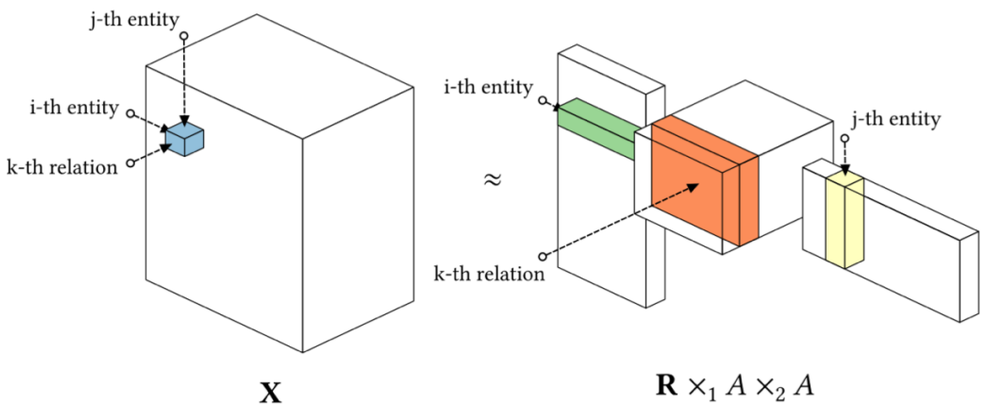


WLOG, assume columns of $\mathbf{W}$ and $\mathbf{H}$ sum to 1

Tensor Factorization

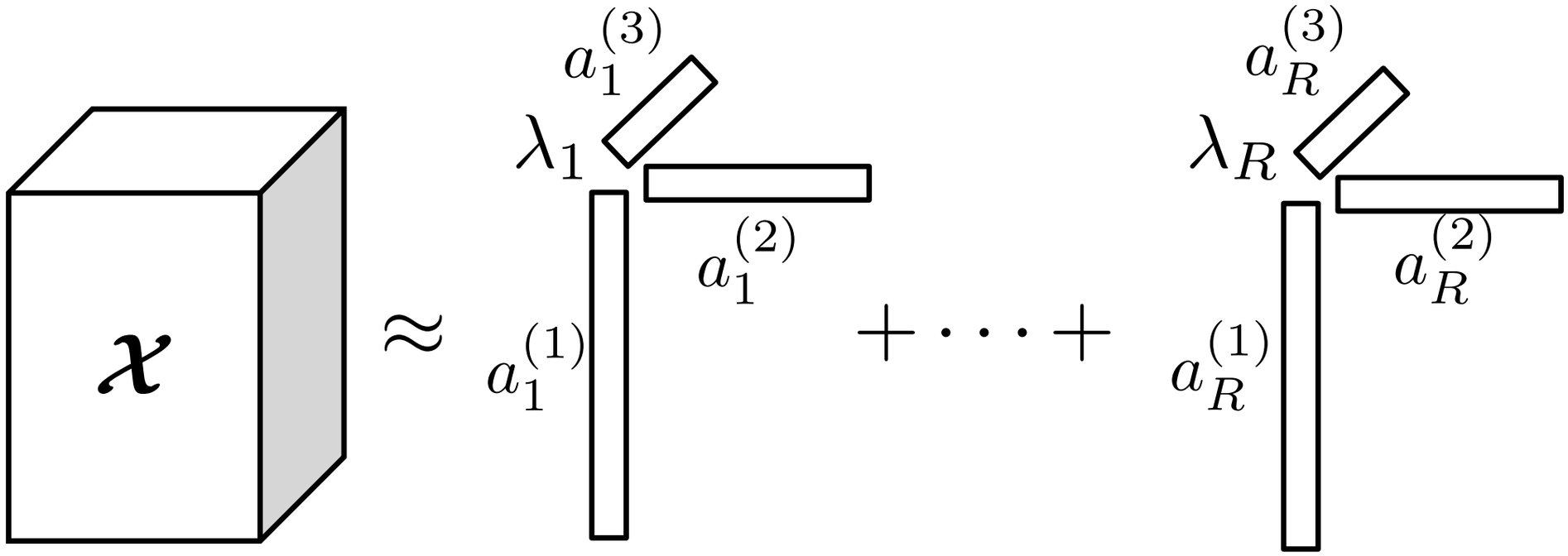






- Generalization of matrix factorization
- Multiway structure information utilized during decomposition process
- Many decomposition models: CANDECOMP / PARAFAC (CP), Tucker, Rescal

CP Decomposition



$$\boldsymbol{\mathcal{X}} \approx \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \ldots \circ \mathbf{a}_r^{(N)}$$ sum of rank one tensors

$$= [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \ldots; \mathbf{A}^{(N)}]\!]$$ shorthand notation

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: models and conditions for an“ explanatory” multimodal factor analysis. UCLA Working Papers in Phonetics, 16, 1–84.
Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition. Psychometrika, 35(3), 283–319.

Standard CP Algorithm

$$\min \sum_{\vec{i}} (x_{\vec{i}} - m_{\vec{i}})^2$$

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

- Objective function assumes Gaussian distribution for numeric data
- Almost all existing CP work focuses on squared loss
- What about nonnegative integer data?

CP Alternating Poisson Regression (CP-APR)

$$\min \sum_{\vec{i}} (x_{\vec{i}} - m_{\vec{i}})^2$$

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

Poisson distribution for nonnegative, discrete data

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

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

Nonnegative constraints

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

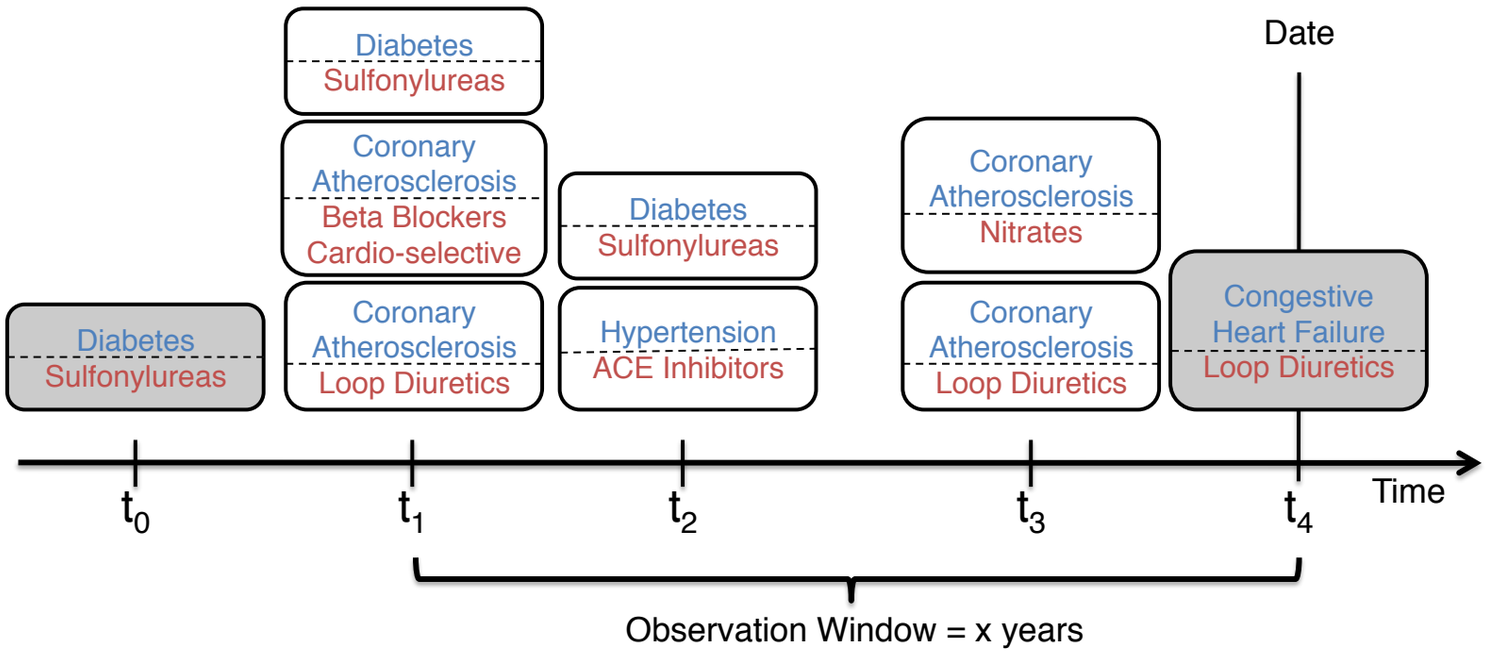
$$\Omega_\lambda = [0, +\infty)^R$$

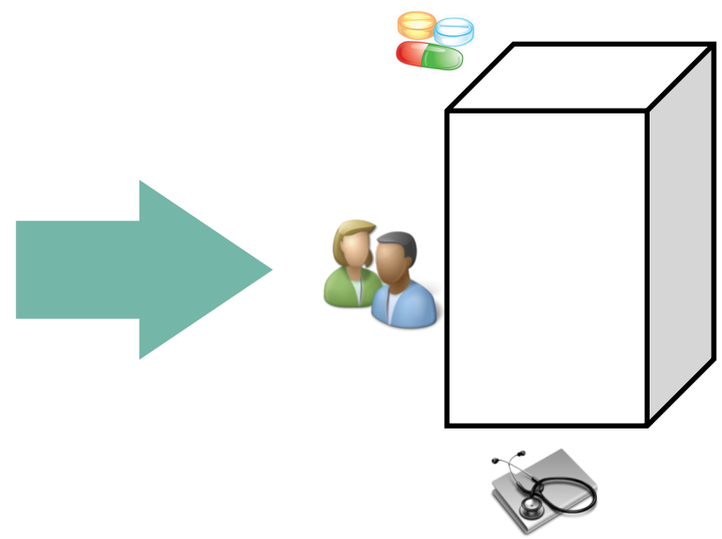
Stochastic column constraints

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

Chi, E. C., & Kolda, T. G. (2012). On tensors, sparsity, and nonnegative factorizations. SIAM Journal on Matrix Analysis and Applications, 33(4), 1272–1299.

LIMESTONE: Tensor Generation

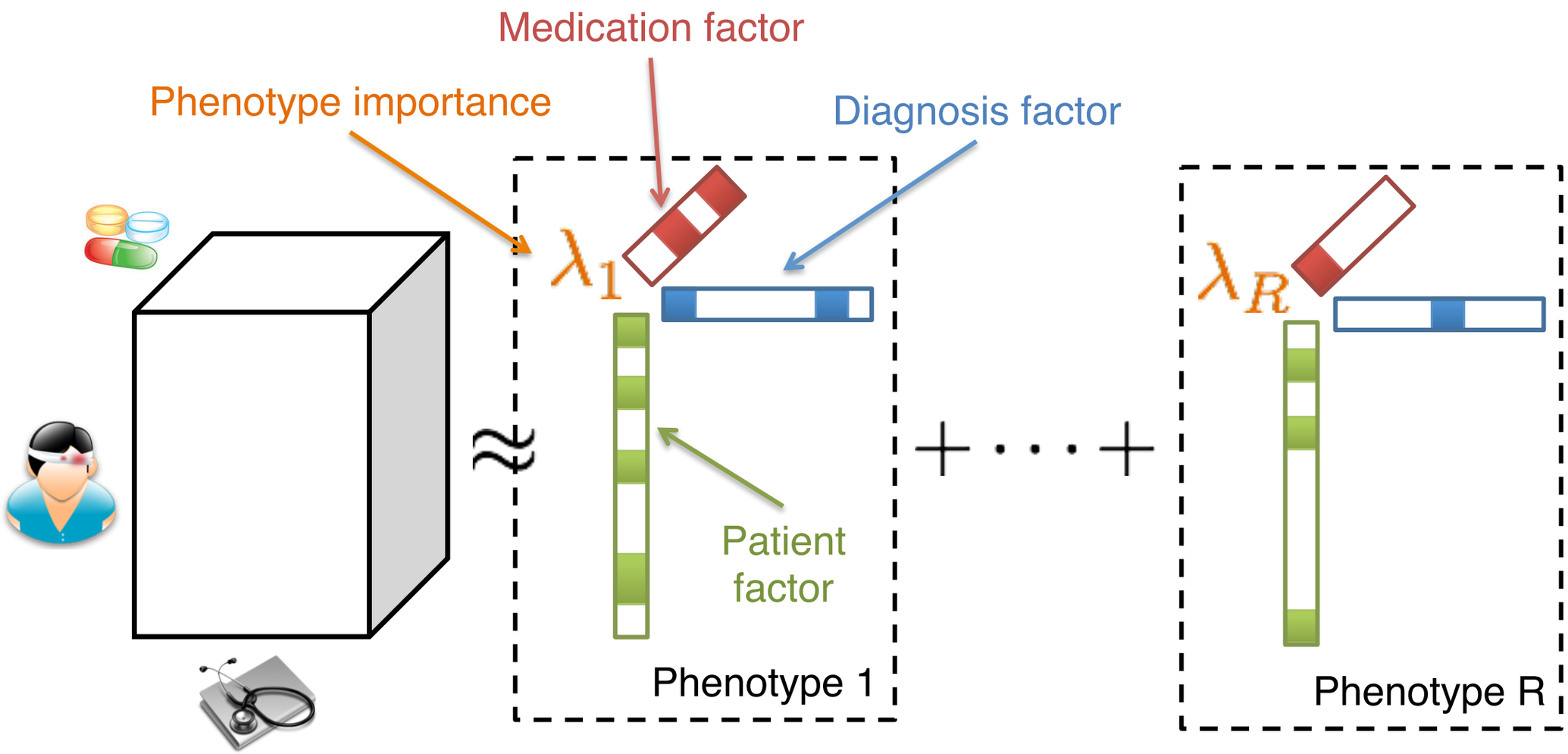




Construct EHR tensor

JBI '14, KDD '14, BIH '14

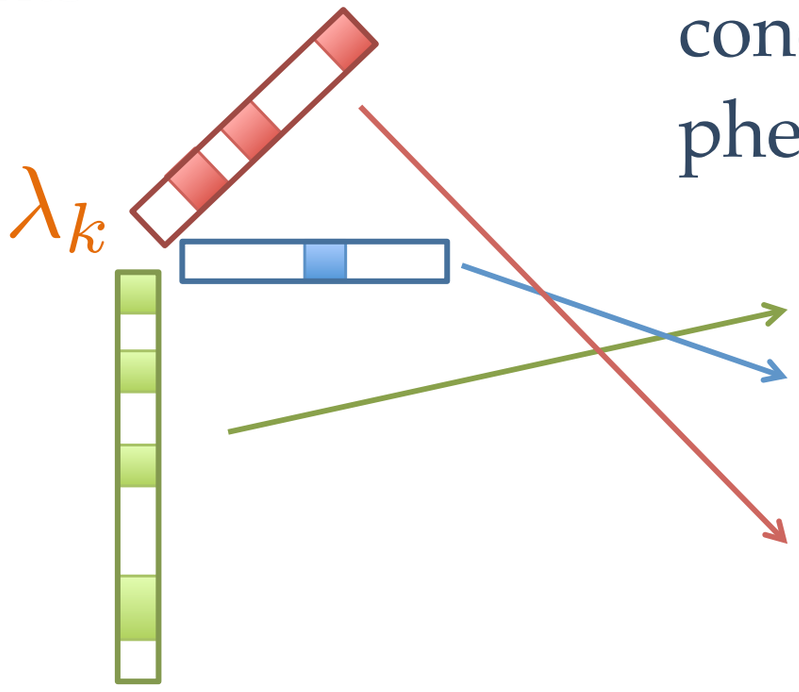
LIMESTONE: Phenotype Generation



- CP-APR decomposition of EHR feature tensor
- A candidate phenotype is a single rank-one tensor

LIMESTONE: Candidate Phenotype

Nonzero elements are clinical characteristics



Each element value represents conditional probability given the phenotype and mode

Candidate Phenotype k (40% of patients)
Hypertension
Beta Blockers Cardio-Selective
Thiazides and Thiazide-Like Diuretics
HMG CoA Reductase Inhibitors

Mode elements ordered in decreasing importance

LIMESTONE: Experimental Results

- Real EHR data from Geisinger Health System over a span of 7 years

- Focus on medication orders
  (medication type and associated diagnoses)

- 31,815 patients x 169 diagnoses x 471 medications
  (< 1% of non-zero elements)

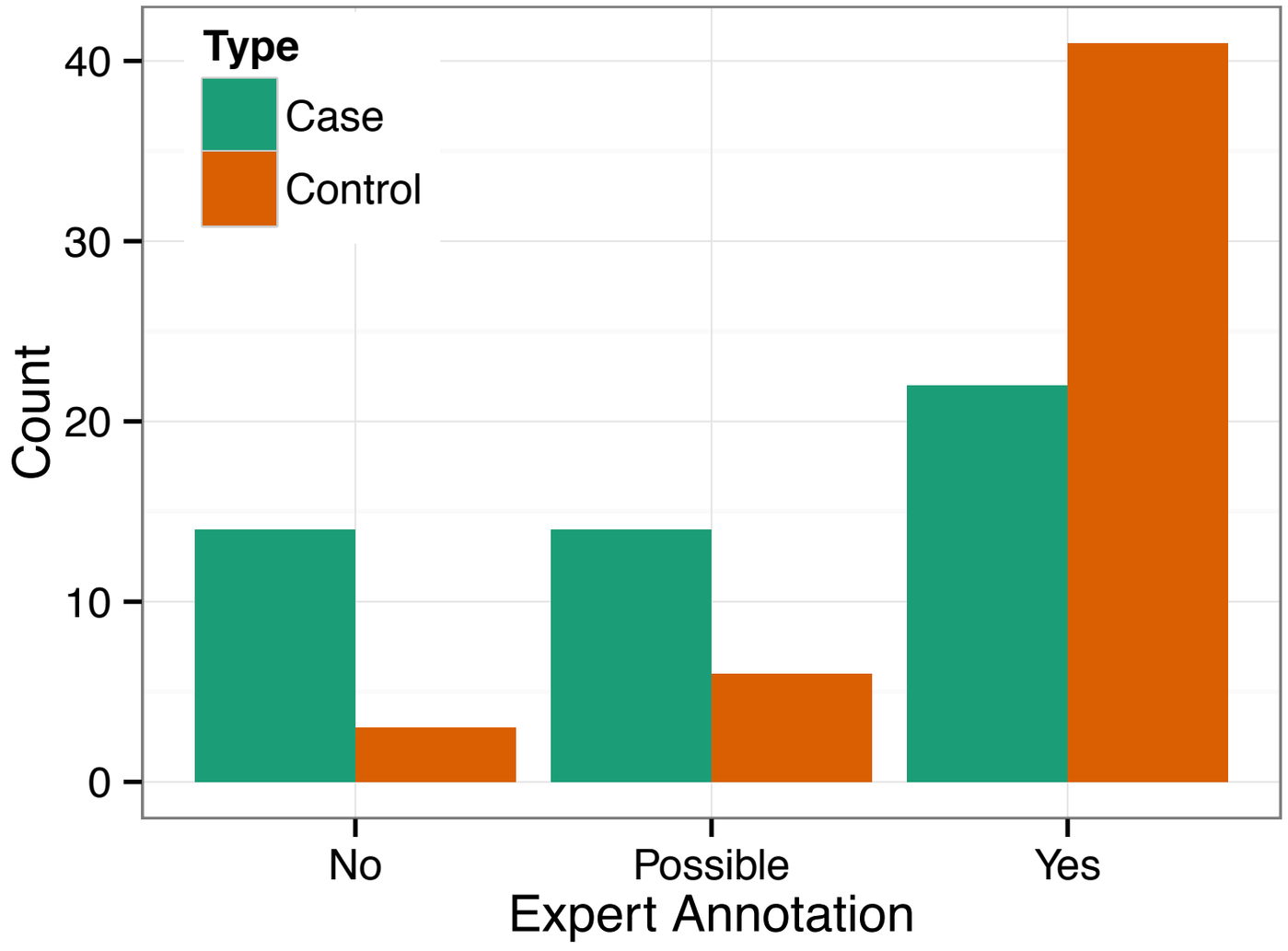
LIMESTONE: Interpretability

Limestone Phenotype
Hypertension
Hypertensive Heart Disease
Beta Blockers Cardio-Selective
Calcium Channel Blockers
Diuretic Combinations
Nitrates
HMG CoA Reductase Inhibitors
Vasodilators
Cardiac Glycosides

NMF Phenotype
Hypertension – Sympathomimetics
Hypertension – Beta Blockers Cardio-Selective
Hypertension – HMG CoA Reductase Inhibitors
Hypertension – Insulin
Hypertension – Potassium
Major Symptoms, Abnormalities – Sympathomimetics
Major Symptoms, Abnormalities – Insulin
Major Symptoms, Abnormalities – Sodium
Major Symptoms, Abnormalities – Potassium
Major Symptoms, Abnormalities – Coumarin Anticoagulants
Vascular Disease – Sympathomimetics
Other Gastrointestinal Disorders – Sympathomimetics
Other Endocrine/Metabolic/Nutritional Disorders – Sympathomimetics
History of Disease – Sympathomimetics
Other Dermatological Disorders – Sympathomimetics
Other Infectious Diseases – Sympathomimetics
… 2,728 total combinations

Limestone phenotypes are more succinct

LIMESTONE: Clinical Relevance



Domain expert confirmed 82% as clinically meaningful

LIMESTONE: Classification Task

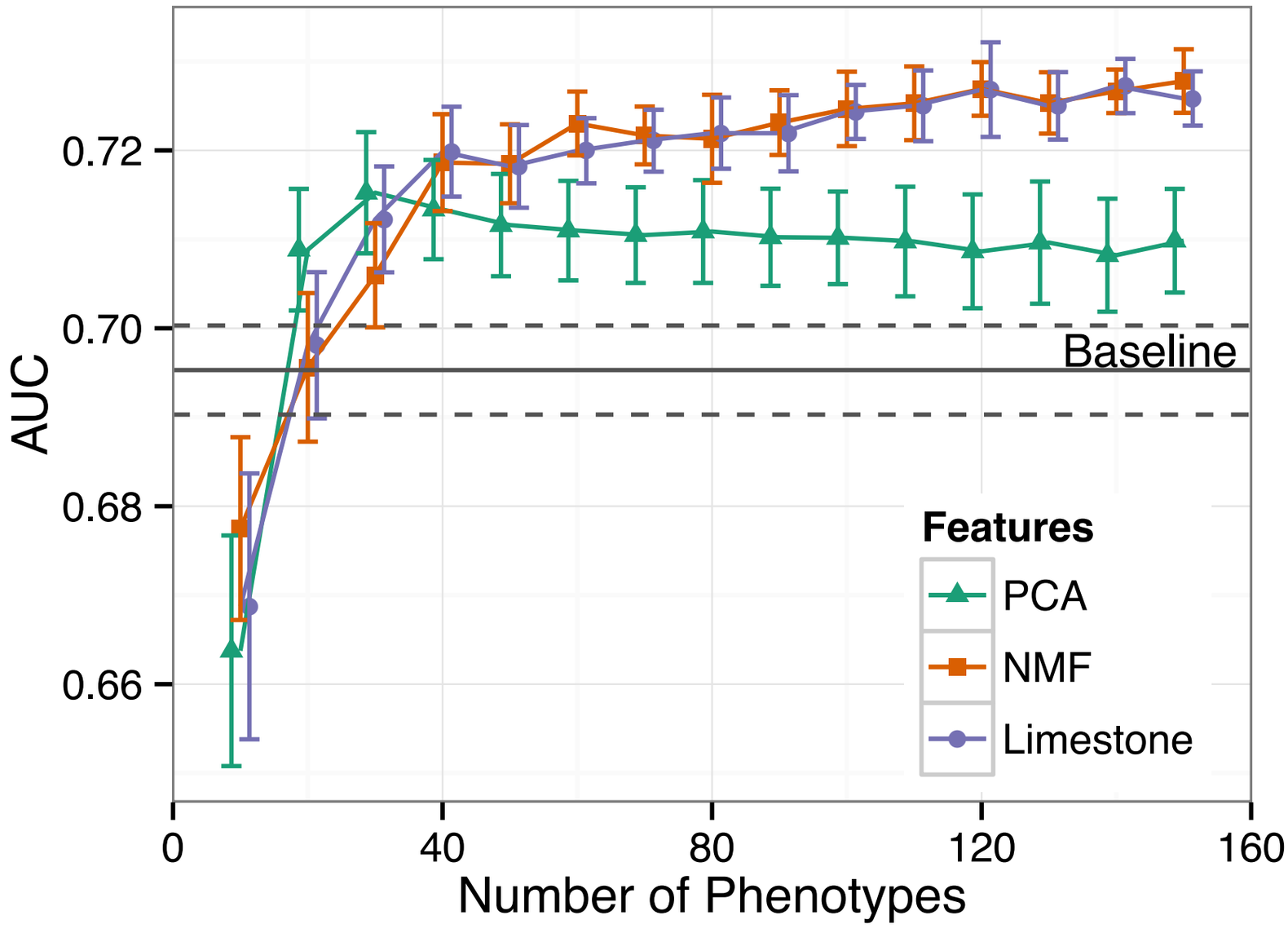
- Predict patients with heart failure
- Evaluate on four feature sets

1. Raw 640 features (no interaction)
2. PCA
3. NMF
4. Limestone

Using patient factor matrix (loadings on phenotypes)

- Logistic regression model

LIMESTONE: Predictive Power



40 phenotypes outperforms original 640 features!

LIMESTONE: Summary

- Phenotype definitions are consistent with regards to initialization parameters and noise
- Limestone-derived phenotypes are more succinct and easier to interpret compared to NMF
- 82% of the phenotypes were confirmed as clinically meaningful by a domain expert
- Reduced phenotype representation retains predictive information from full dataset to predict heart failure

LIMESTONE: Potential Problems

"inadmissible zero problem"

$$\min f(\boldsymbol{\mathcal{M}}) \equiv \sum_{\vec{i}} m_{\vec{i}} - \boxed{x_{\vec{i}} \log m_{\vec{i}}} \qquad \lim_{i \to 0} \log i = -\infty$$

$$\begin{aligned}
\text{s.t } \boldsymbol{\mathcal{M}} &= [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; ...; \mathbf{A}^{(N)}]\!] \in \Omega \\
\Omega &= \Omega_\lambda \times \Omega_1 \times \cdots \times \Omega_N \\
\Omega_\lambda &= [0, +\infty)^R \\
\Omega_n &= \{\mathbf{A} \in [0,1]^{I_n \times R} \;\; | \;\; ||\mathbf{a}_r||_1 = 1 \; \forall r\}
\end{aligned}$$

What about baseline characteristics, computational stability, and local optimum guarantee?

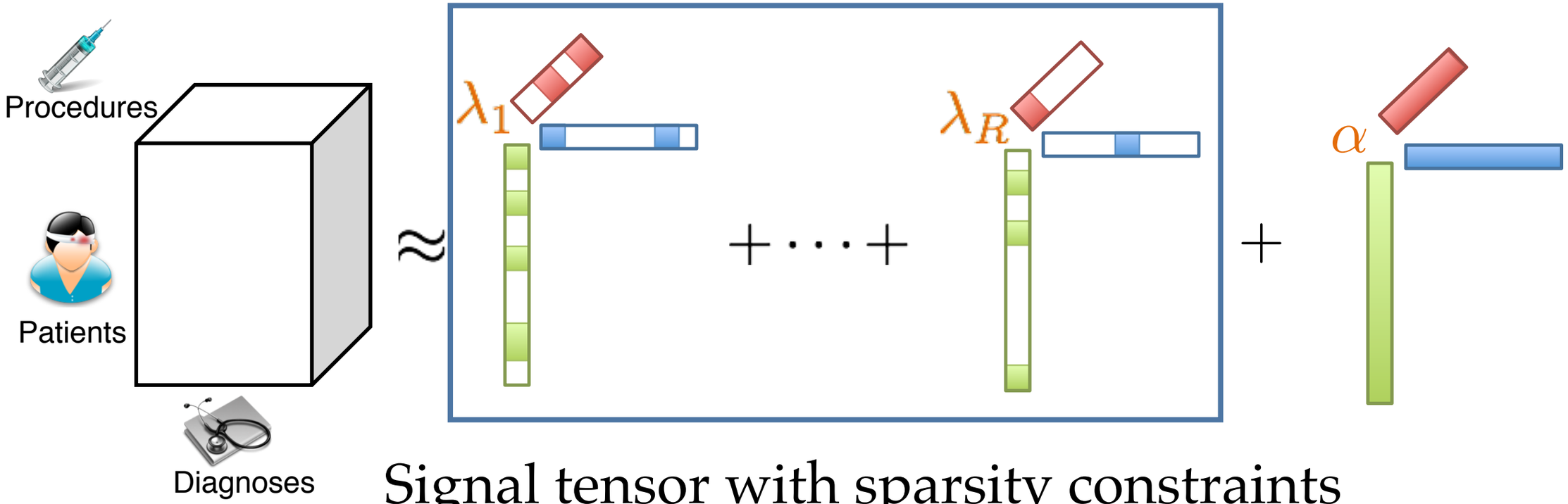
MARBLE & GRANITE

Integrated Sparse Tensor Factorization

Integrated Sparse Tensor Factorization

- Sparsity constraints on factor matrices to minimize non-zeros
  - Threshold “probabilistically unlikely” elements
  - Simplex projection
- Rank-one tensor augmentation (aka bias tensor)
  - Provides computational stability
  - Captures baseline characteristics

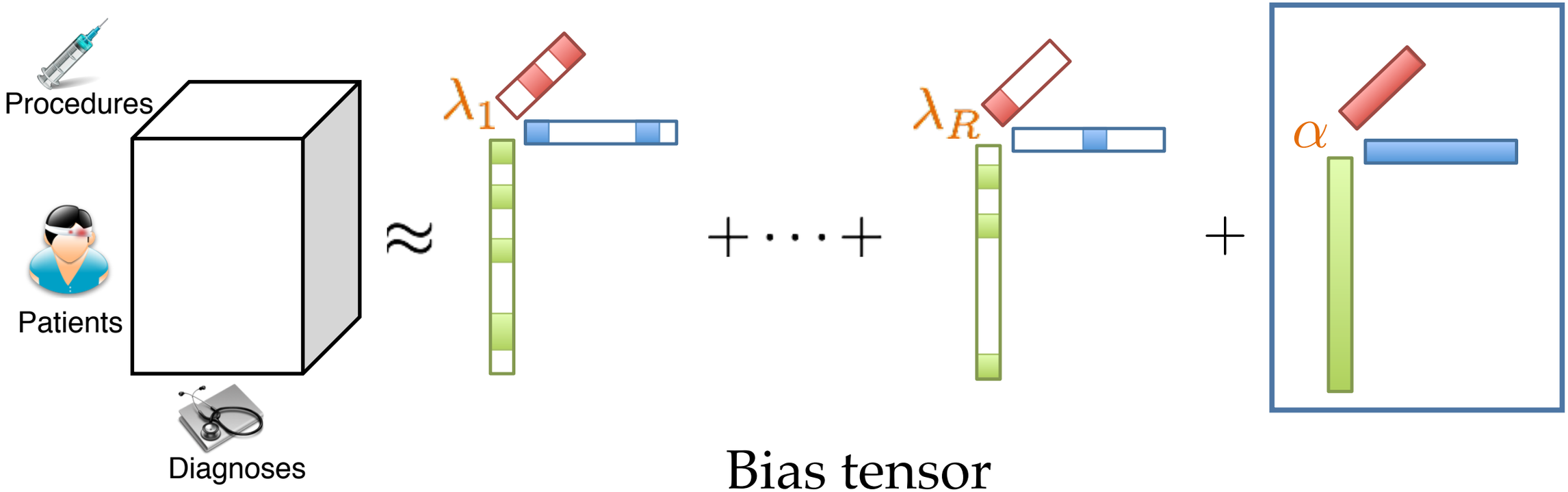
Integrated Sparse Tensor Factorization



Signal tensor with sparsity constraints

- captures concise phenotypes
- sparsity built into algorithm and not post-processed

Integrated Sparse Tensor Factorization



Bias tensor

- captures baseline characteristics
- absorbs the data offset
- offers computational stability

MARBLE: Optimization Problem

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

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

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

bias tensor

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

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

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

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

bias has positive constraints

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

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

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

minimizes “unlikely” elements

MARBLE: Algorithm

- Alternating minimization to solve for each mode
- Multiplicative update to enforce nonnegative constraints
- Gradual projection to threshold factor matrices

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

MARBLE: Empirical Study

- 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
- Focus on carrier claims with random subset of patients
- 10,000 patients x 129 diagnoses x 115 procedures

MARBLE: Top 10 Bias 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

- 2 out of 3 have multiple chronic conditions
- 55.7% had hypertension
- 47% had hyperlipidemia
- 28.5% had arthritis
- 27% had diabetes

MARBLE: Top 10 Bias Elements

- Top procedure includes doctor visits (regular checkups)
- Other procedures fit the top 5 chronic disease conditions

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

MARBLE: Chronic Diseases

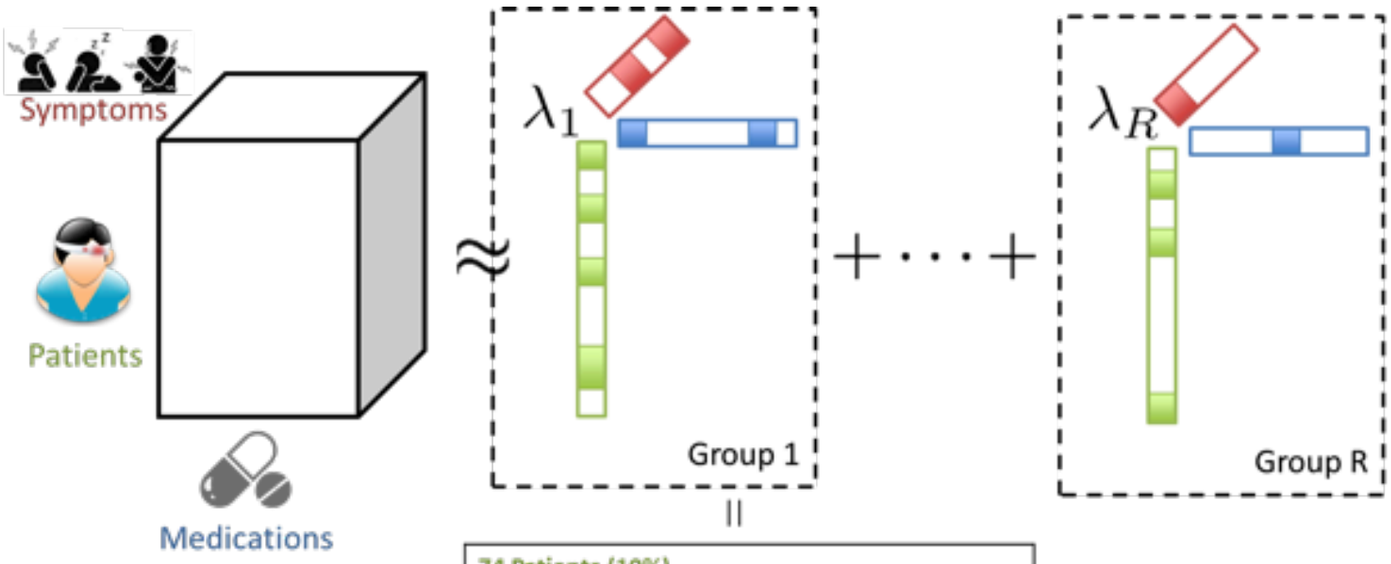
Diabetes Phenotype
Diseases of other endocrine glands
Complications of surgical and medical care
Chemistry Pathology and Laboratory Tests
Organ or Disease Oriented Panels
Hematology and Coagulation Procedures
Surgical Procedures on the Cardiovascular System

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

MARBLE: Chronic Fatigue Syndrome

- Complex, devastating illness
- Managing symptoms is complicated
- Uncover common medication patterns with similar symptom severities



74 Patients (19%)
Analgesics
Anticonvulsants
Anxiolytics, sedatives, and hypnotics
Unusual Fatigue (13.6)
Unrefreshing Sleep (13.1)
Muscle Pain (11.5)
Joint Pain (9.5)
+ 9 medications and 8 symptoms

MARBLE: Diversity

Phenotype 1 (37.95% of patients)
Coronary Atherosclerosis/ Other Chronic Ischemic Heart Disease (i)
Other Endocrine/Metabolic/ Nutritional Disorders (ii)
Diabetes with No or Unspecified Complications
Hypertension (iii)
Angina Pectoris/Old Myocardial Infarction
Surgical Procedures on the Musculoskeletal System (i)

Phenotype 2 (35.68% of patients)
Major Symptoms, Abnormalities (iv)
Heart Arrhythmias (v)
Other Significant Endocrine and Metabolic Disorders
Hypertension (iii)
Other Endocrine/Metabolic/ Nutritional Disorders (ii)
Surgical Procedures on the Musculoskeletal System (i)

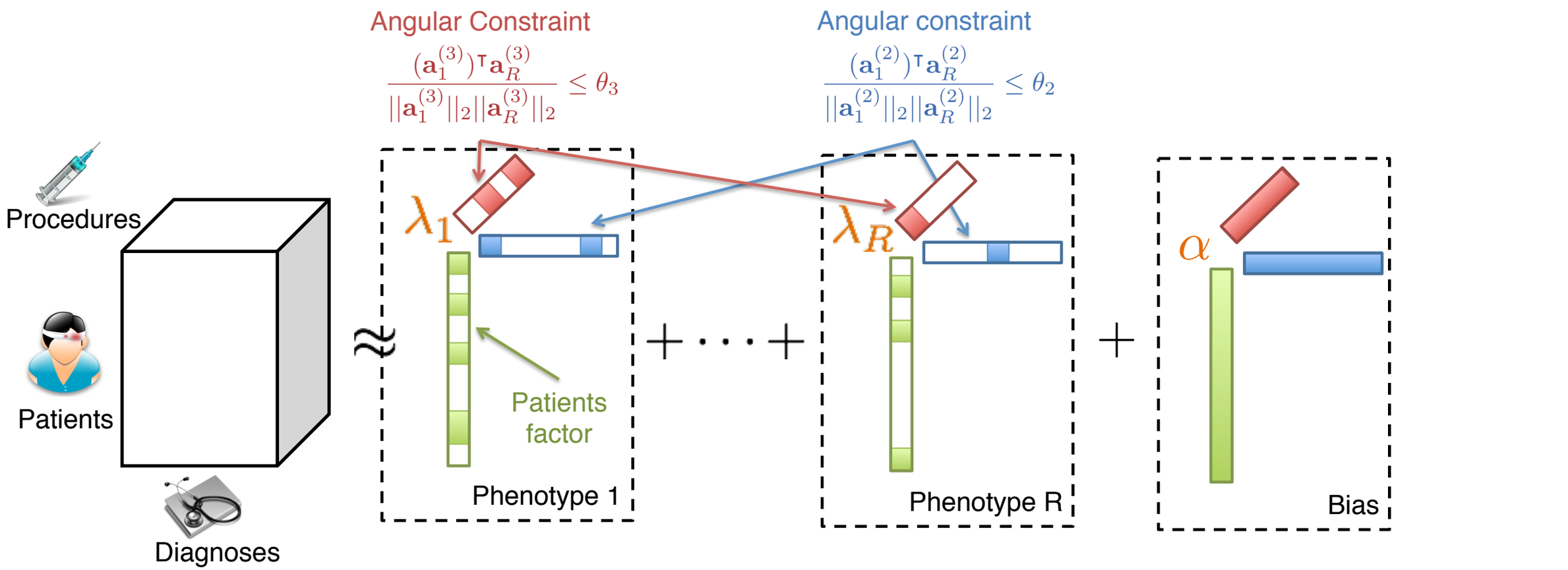
Phenotype 3 (19.94% of patients)
Major Symptoms, Abnormalities (iv)
Coronary Atherosclerosis/ Other Chronic Ischemic Heart Disease (i)
Other Endocrine/Metabolic/ Nutritional Disorders (ii)
Hypertension (iii)
Heart Arrhythmias (v)
Iron Deficiency and Other/ Unspecified Anemias and Blood Disease (vi)
Peptic Ulcer, Hemorrhage, Other Specified Gastrointestinal Disorders
Therapeutic, Preventive or Other Interventions Codes for Performance Measurement

Phenotype 4 (23.73% of patients)
Congestive Heart Failure
Septicemia/Shock
Cardio-Respiratory Failure and Shock
Viral and Unspecified Pneumonia, Pleurisy
Acute Renal Failure
Chronic Kidney Disease, Mild or Unspecified (Stages 1-2 or Unspecified)
Surgical Procedures on the Musculoskeletal System (i)

Phenotype 5 (27.19% of patients)
Iron Deficiency and Other/ Unspecified Anemias and Blood Disease (vi)
Chronic Obstructive Pulmonary Disease
Other Gastrointestinal Disorders
Osteoarthritis of Hip or Knee
Hypertension (iii)
Surgical Procedures on the Musculoskeletal System (i)

Sparsity parameter can be difficult to tune and lead to poor diversity

GRANITE: Diverse Phenotypes



Discourage overlapping factor vectors by adding angular constraints on the signal factors

GRANITE: Optimization Problem

Angular penalty term

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

$$\text{s.t } \boldsymbol{\mathcal{Z}} = [\![\sigma; \mathbf{u}^{(1)}; \cdots; \mathbf{u}^{(N)}]\!] + [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \cdots; \mathbf{A}^{(N)}]\!]$$

$$\sigma > 0, \lambda_r \geq 0, \ \forall r$$

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

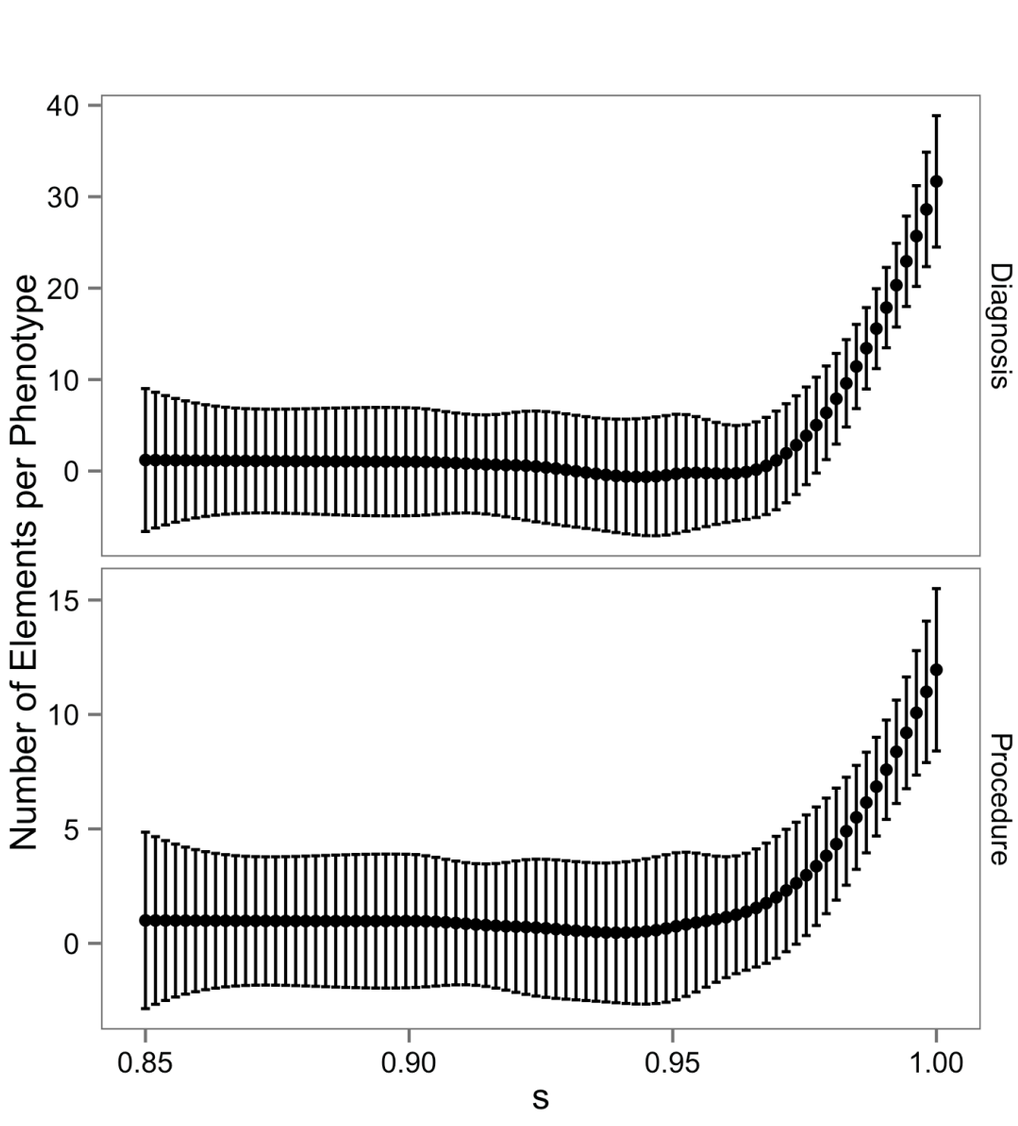
$$||\mathbf{a}_r^{(n)}||_1 = ||\mathbf{u}^{(n)}||_1 = 1, \ \forall n$$

- Penalty term softly imposes diversity by relaxing the angular constraints
- Only penalizes factor vectors whose cosine angle with other vectors are greater than θ_n

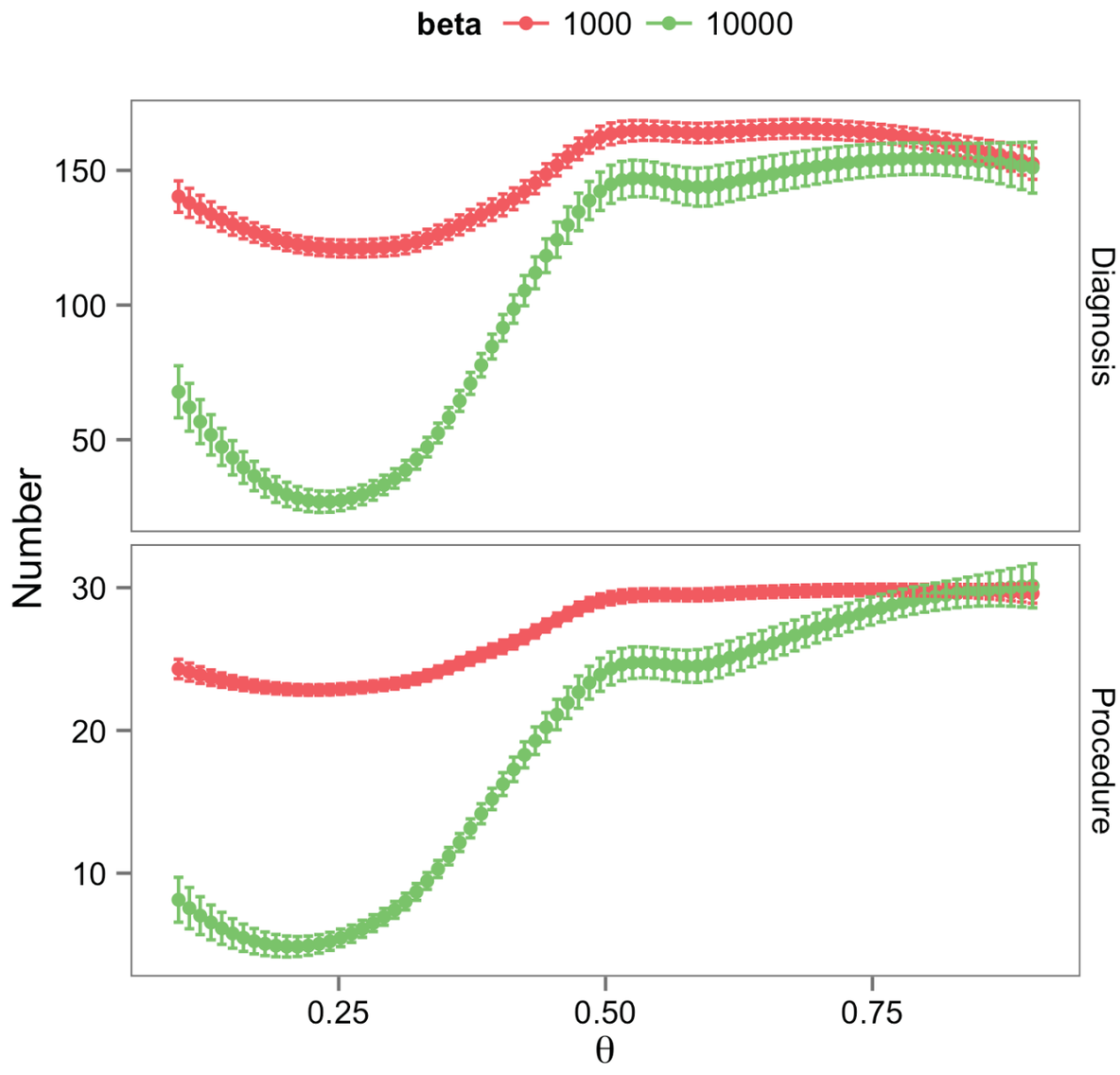
GRANITE: Algorithm

- Projected gradient descent approach to solve bias and interaction factor matrices simultaneously
  - Avoids multiplicative update problem of zeroing out components too early
- Simplex projection instead of non-convex space
  - Sparser factors can be achieved by setting s to be smaller value (< 1)

GRANITE: Sparsity Tuning



Simplex parameter provides better sparsity control



Angular constraints also offers improved sparsity

GRANITE: Improved Sparsity + Diversity

Phenotype 1 (4.12% of patients)
Coagulation Defects and Other Specified Hematological Disorders
Pleural Effusion/ Pneumothorax
Radiation Oncology Procedures

Phenotype 2 (1% of patients)
Dementia With Complications
Major Depressive, Bipolar, and Paranoid Disorders
Schizophrenia
Physical Examination Codes for Performance Measurement (i)

Phenotype 3 (0.36% of patients)
Major Fracture, Except of Skull, Vertebrae, or Hip
Severe Head Injury
Major Head Injury
Ischemic or Unspecified Stroke
Osteoporosis and Other Bone/Cartilage Disorders
Bone/Joint Studies

Phenotype 4 (0.54% of patients)
Cerebral Hemorrhage
Therapeutic, Preventive or Other Interventions Codes for Performance Measurement
Chemistry Pathology and Laboratory Tests

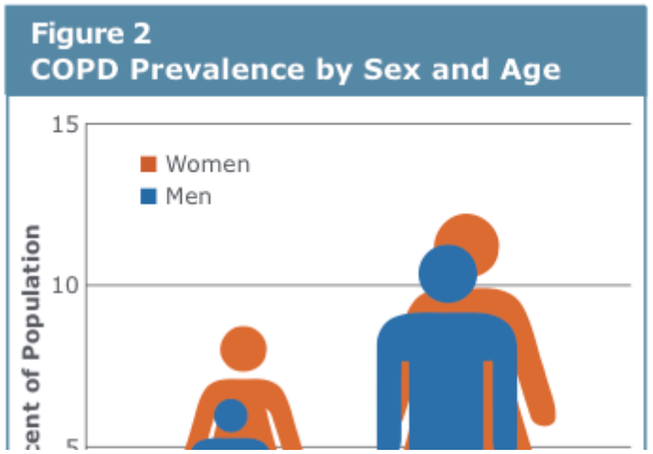
Phenotype 5 (2.24% of patients)
Nephritis
Physical Examination Codes for Performance Measurement (i)

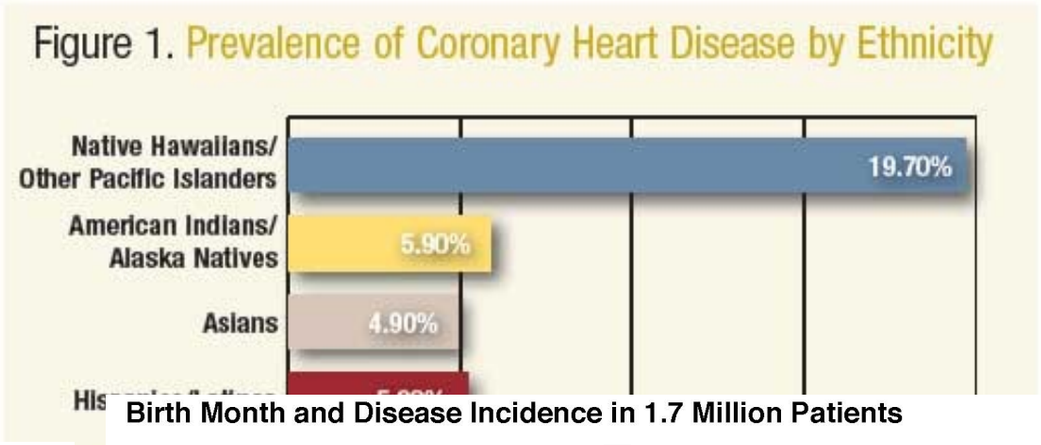
- Overlapping factor elements minimized via the angular regularization penalty
- Tradeoff between discovery of rare phenotypes and diversity

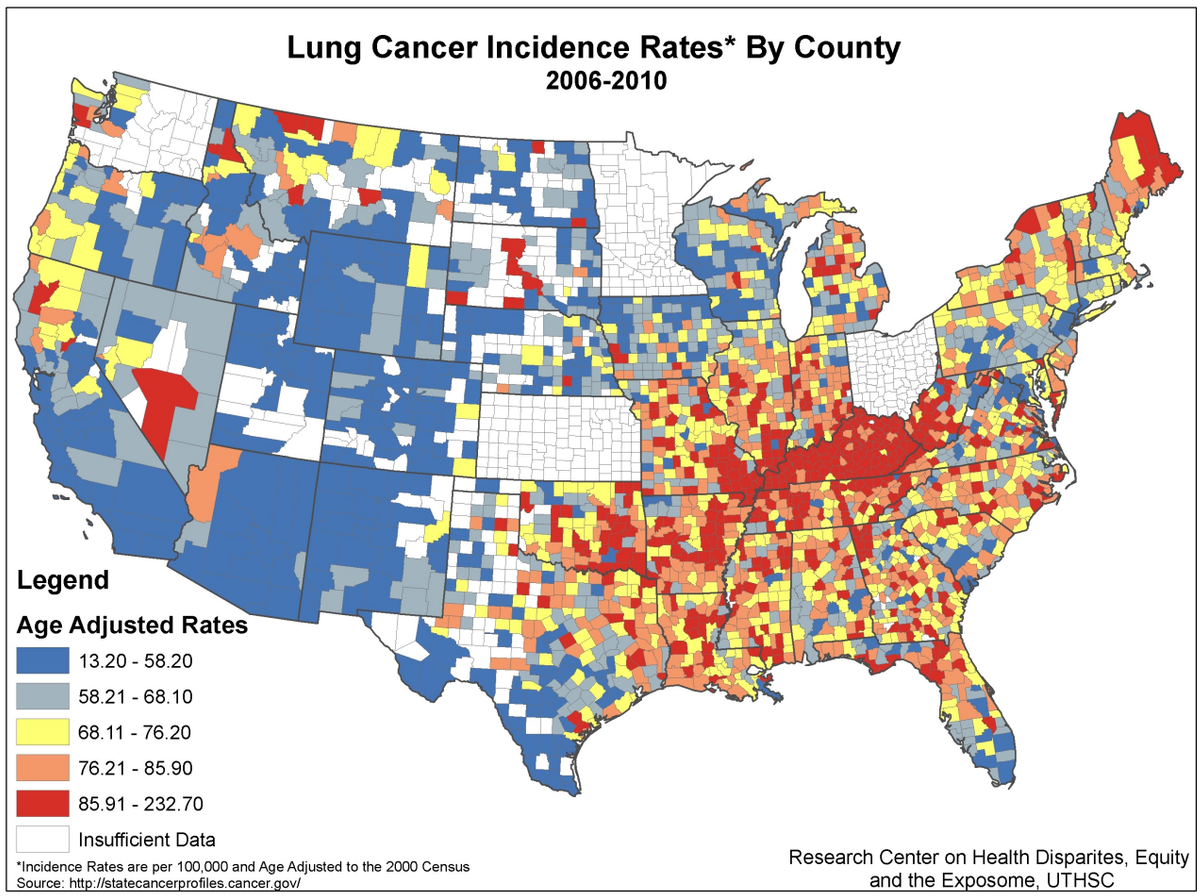
SANDSTONE

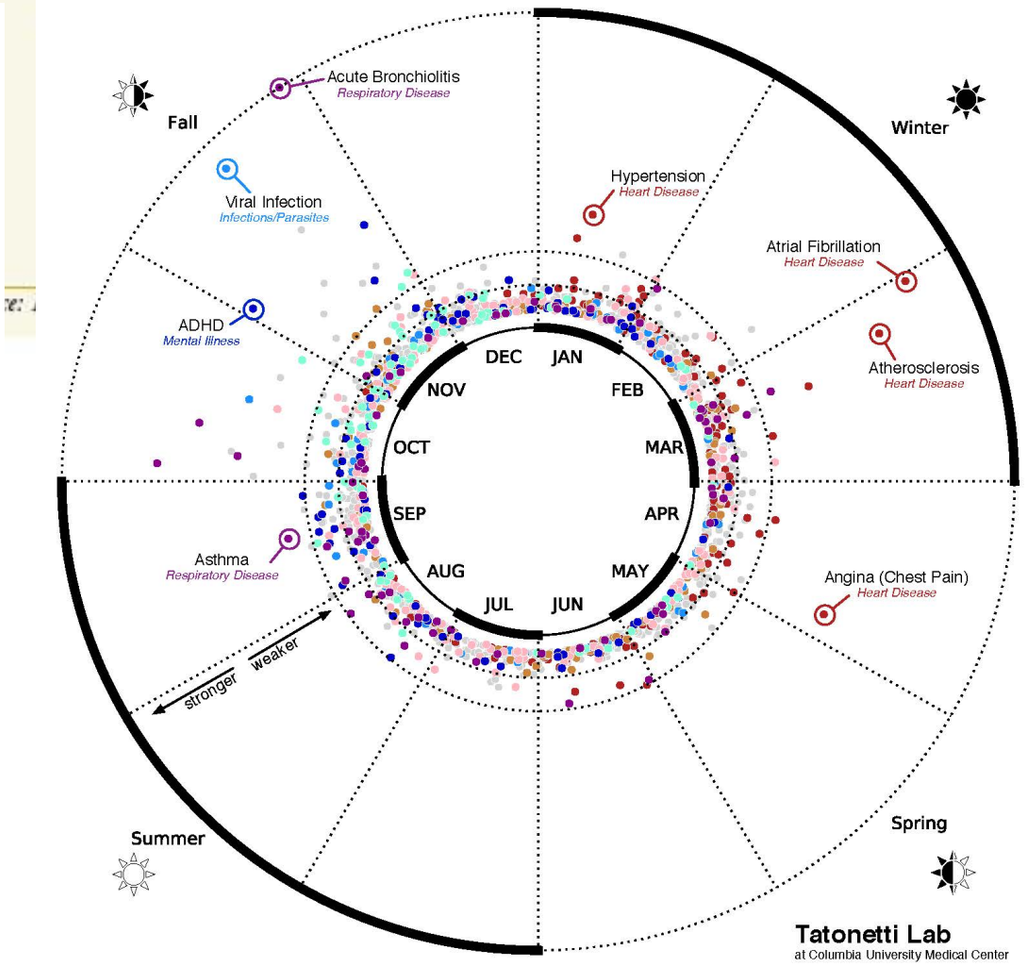
Multi-Task Tensor Factorization

Splitting Patients into Subgroups





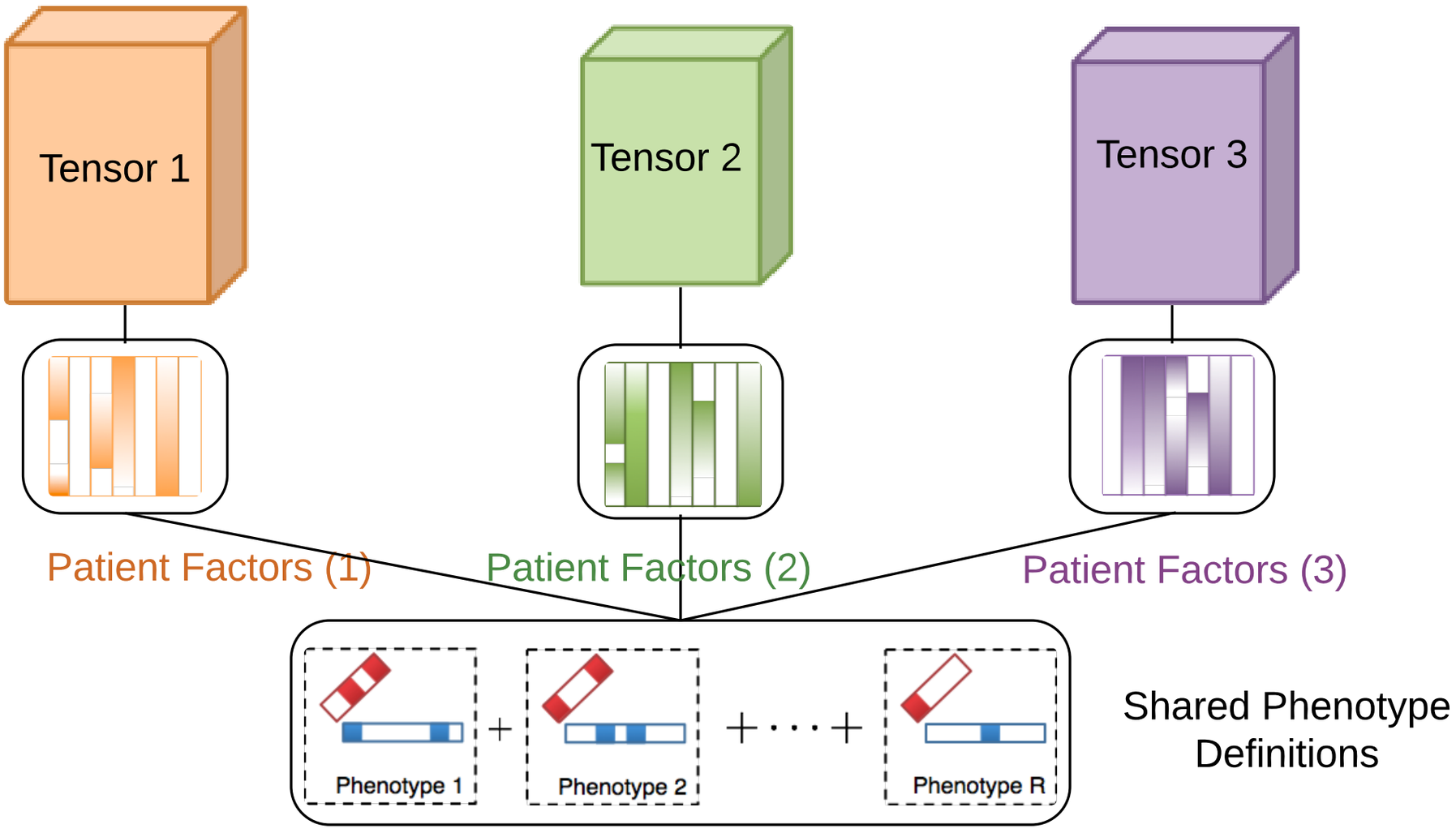




SANDSTONE: Multi-Task Tensor

- Decompose a large tensor by splitting into smaller groups
- Encode side information that would otherwise be lost (e.g., age, gender, geographic location, etc.)
- Flexible representation using partially shared latent space

SANDSTONE: Multi-Task Tensor



some phenotypes will be absent for certain tensors

SANDSTONE: Optimization Problem

$L_{2,1}$ Matrix Norm

$$\min \sum_{k=1}^{K} D(\boldsymbol{\mathcal{X}}(k), \boldsymbol{\mathcal{M}}(k)) + \beta_1 \sum_{k=1}^{K} ||A(k)||_{2,1} + \frac{\beta_2}{2} \sum_{n=2}^{N} \sum_{r \neq p} \left(\max\{0, \psi(\mathbf{b}_p^{(n)}, \mathbf{b}_r^{(n)})\} \right)^2$$

$$\text{s.t. } \boldsymbol{\mathcal{M}}(k) = [\![\hat{\mathbf{A}}(k); \hat{\mathbf{B}}^{(2)}; \ldots; \hat{\mathbf{B}}^{(N)}]\!]$$

$$\hat{\mathbf{A}}(k) \in \mathbb{R}_{+}^{P_k \times R}, \quad \forall k$$

$$\mathbf{B}_r^{(n)} \in \Delta_{I_n - 1}, \quad \forall r$$

$$\mathbf{u}^{(n)} \in \mathbb{R}_{++}^{I_n \times 1}, ||\mathbf{u}^{(n)}||_1 = 1.$$

Angular penalty term (Granite)

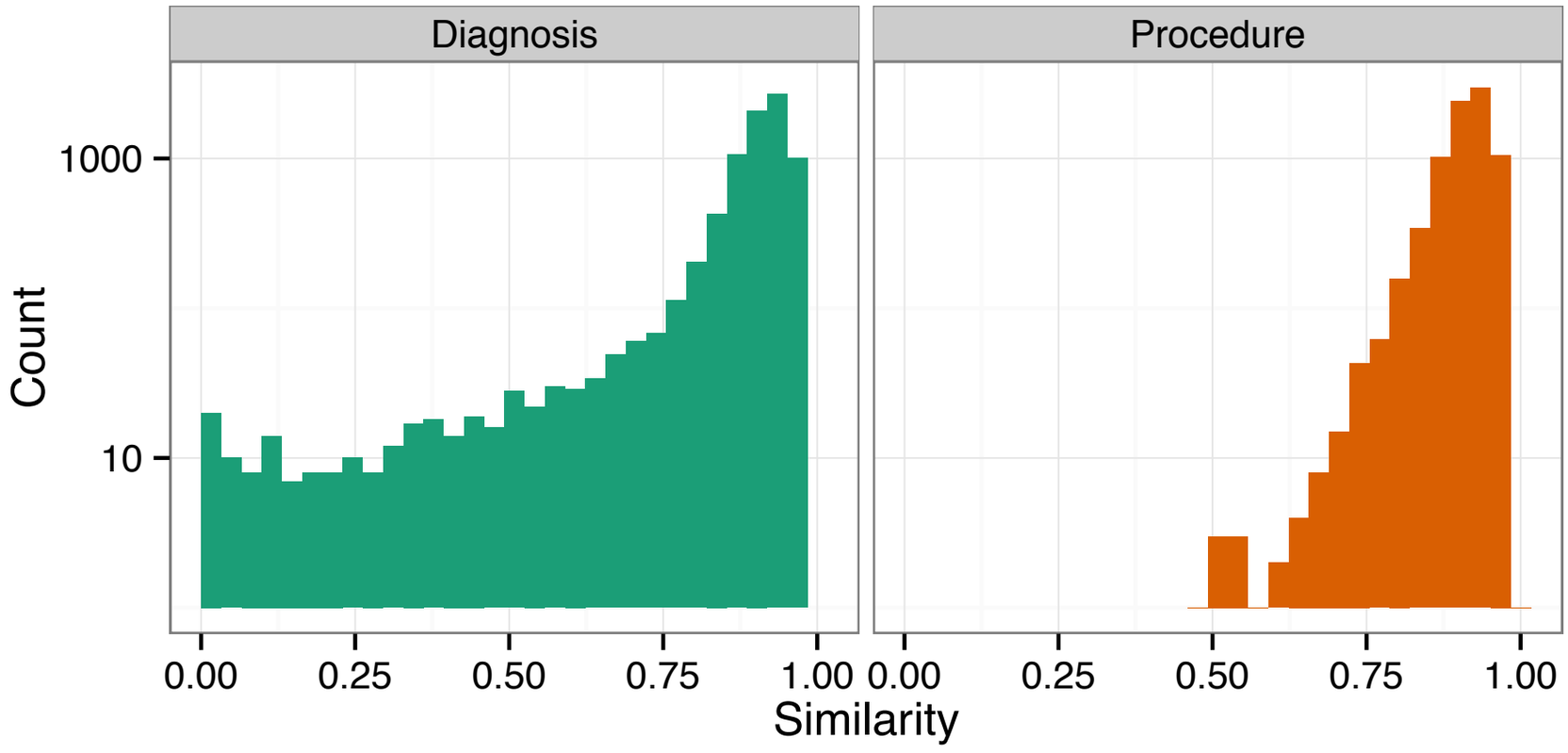
- $L_{2,1}$-norm regularization yields partially shared latent space (common in multi-task learning approaches)
- LASSO on columns such that all entries of certain columns = 0
- Encourages similar patient column sparsity patterns

SANDSTONE: Empirical Study

- CMS 2008-2010 Medicare Data Entrepreneurs' Synthetic Public Use File
- Focus on inpatient and outpatient claims for patients from three most populous states
- 4,334 patients x 240 diagnoses x 199 procedures
  - 1,582 patients from California
  - 1,444 patients from New York
  - 1,308 patients from Texas

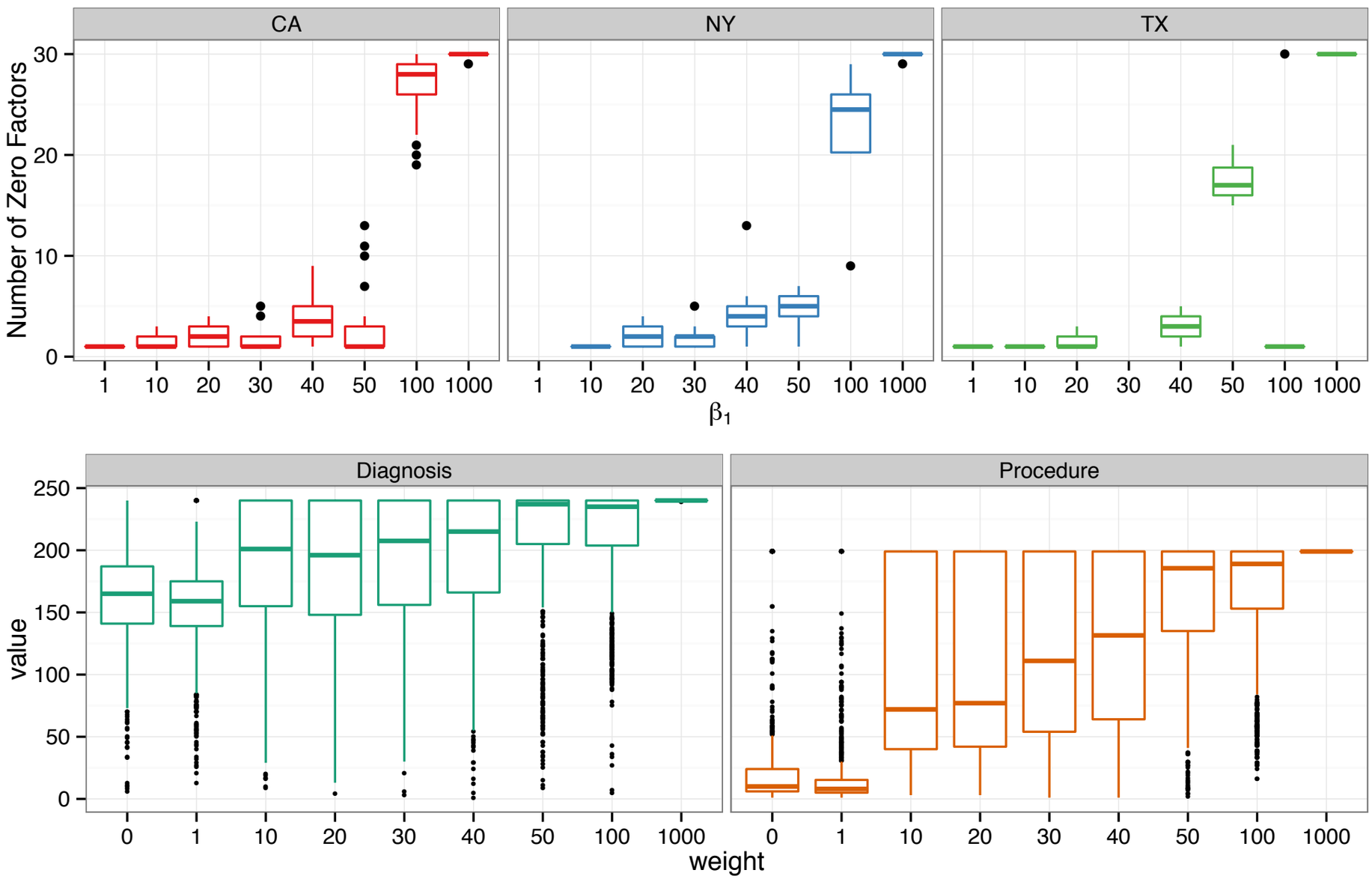
SANDSTONE: Independent Decompositions

- Separate decomposition for each state
- Phenotypes from two states are paired via the Hungarian algorithm



Wide range suggests a multi-task framework is appropriate

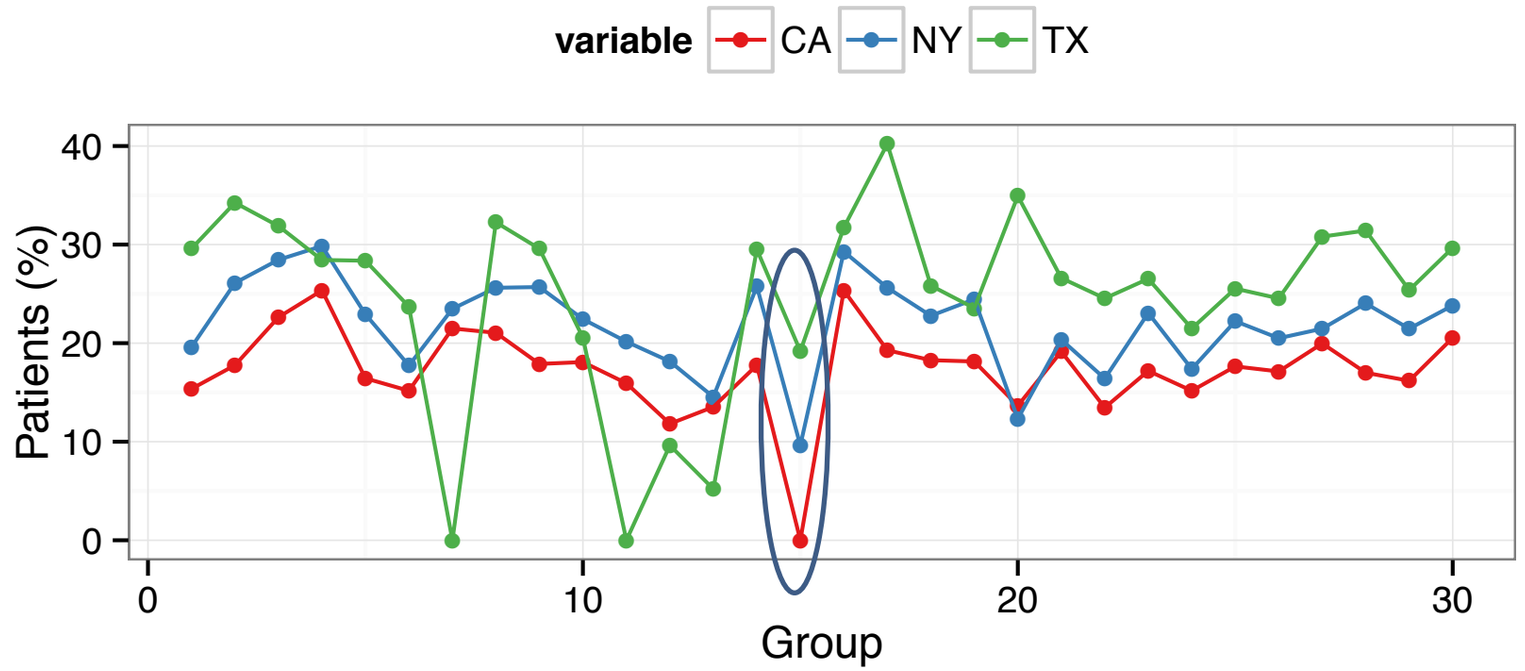
SANDSTONE: Norm Regularization



Tradeoff between unshared phenotypes and the sparsity pattern

SANDSTONE: Unique Phenotypes

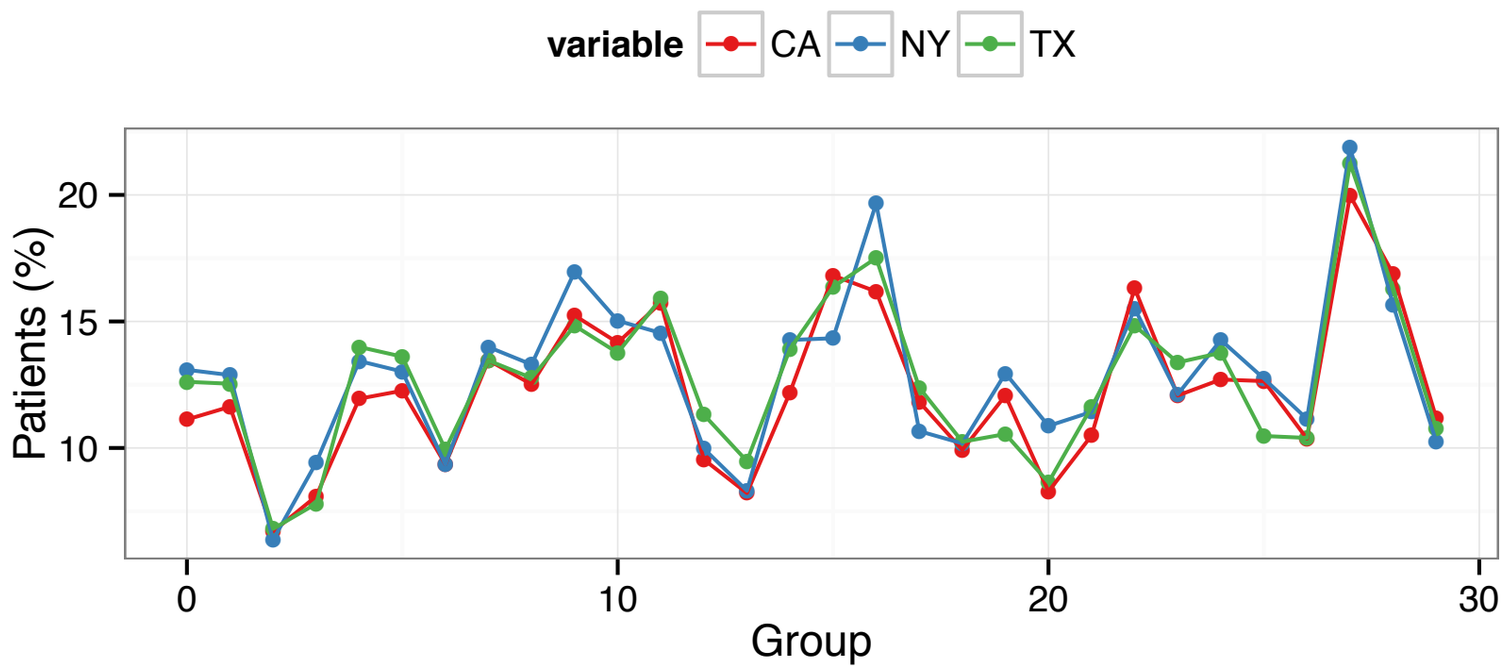
Sandstone



Group 15 NY: 9.63%; TX: 19.19%
Bacterial infection; unspecified site
Peripheral and visceral atherosclerosis
Chronic ulcer of skin
Acute cerebrovascular disease
Coronary atherosclerosis and other heart disease
Other therapeutic procedures; hemic and lymphatic system
Blood transfusion
Diagnostic ultrasound of the heart (echocardiography)
Other diagnostic cardiovascular procedures

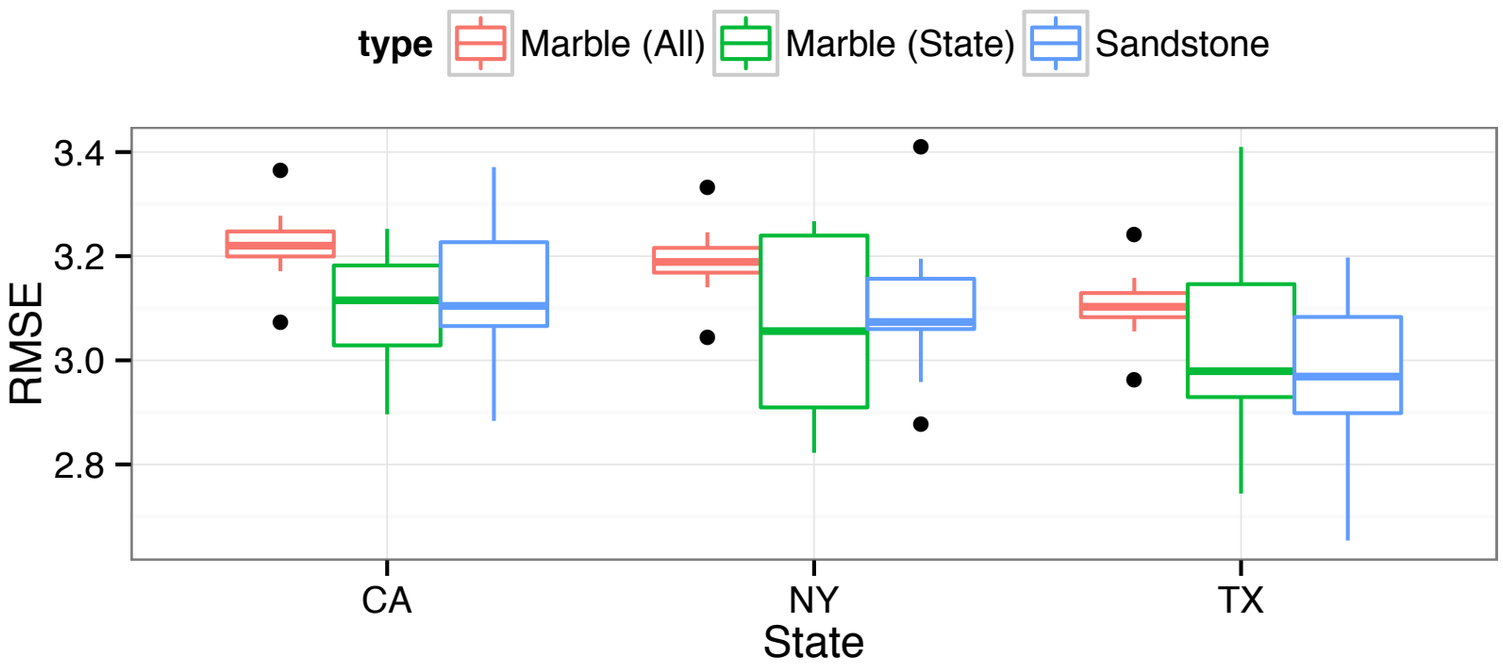
Not present in CA patients

Marble (Single)



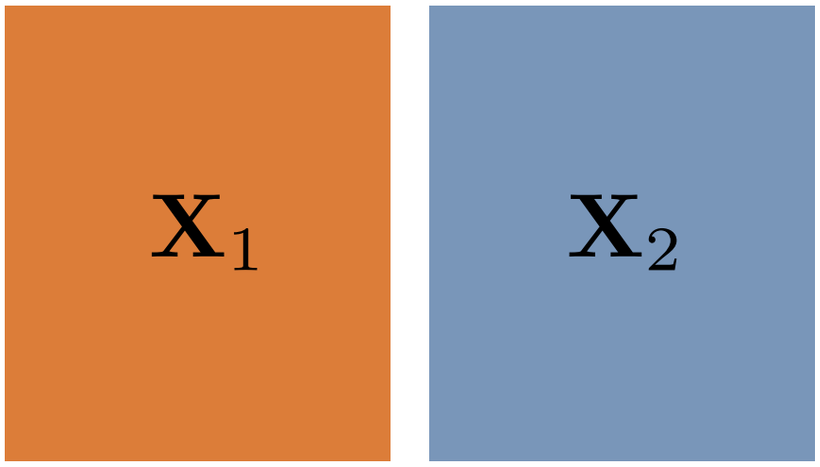
SANDSTONE: Predictive Performance

- Predict total costs (log transformed) of inpatient events in the third year using only observations in first two years
- Features are row-normalized patient matrix (patient loadings on phenotypes)
- Linear regression model is trained on each feature set



SiCNMF

Structured Collective Matrix Factorization



Motivation for Collective MF

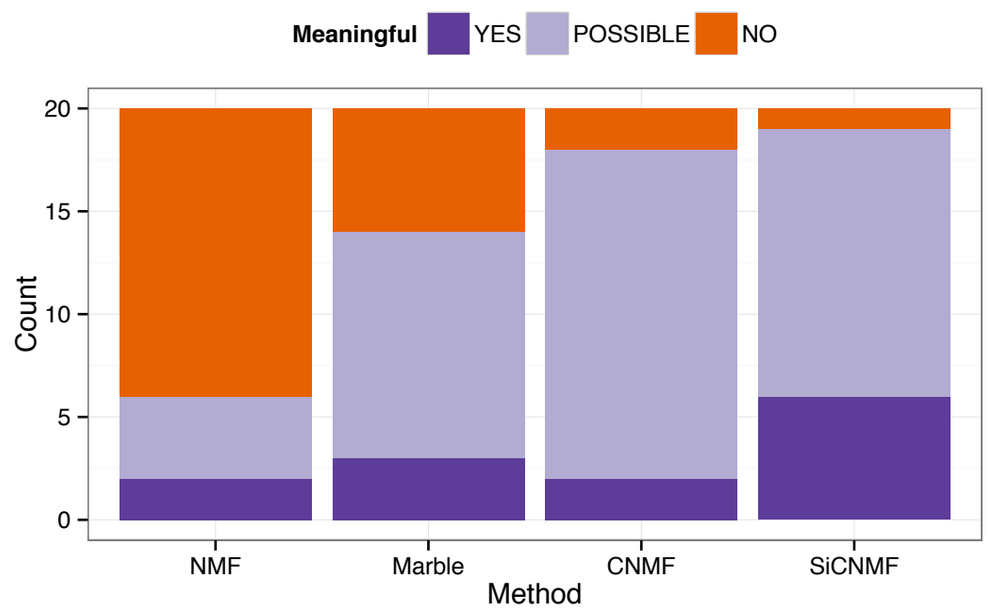
- EHR systems record data in flat formats
  - infrastructure for multi-way interactions is resource-intensive
  - legacy systems traditionally use tables
- Approximation of higher order interactions can lead to noisy correlations and biased results
- Heterogenous data (e.g., labs, demographics, diagnoses)

Collective matrix factorization offers alternative to identify shared subspace among various modes of data when data is not readily available as tensors

SiCNMF: Empirical Study

- Electronic medical records for 10,000 patients in BioVU, the Vanderbilt DNA databank
- Two count matrices:
  - Patient - Diagnosis: 2039 x 936
  - Patient - Medication: 2039 x 161
- Two medical experts annotated the results

Method	Score
NMF	0.3125
Marble	0.3750
CNMF	0.4875
SiCNMF	0.6260



Summary

- Data-driven solutions to obtain multiple concepts simultaneously with minimal human intervention
- Computational phenotypes are concise and generally clinically relevant
- Framework is flexible to incorporate side information and different data types

Thank you!

joyce.c.ho@emory.edu

http://joyceho.github.io