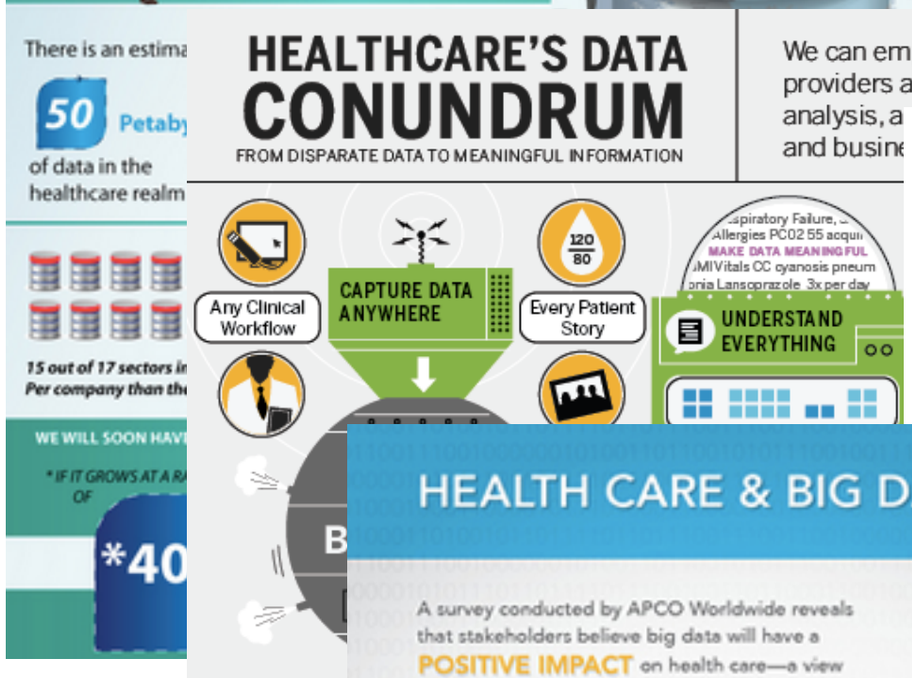
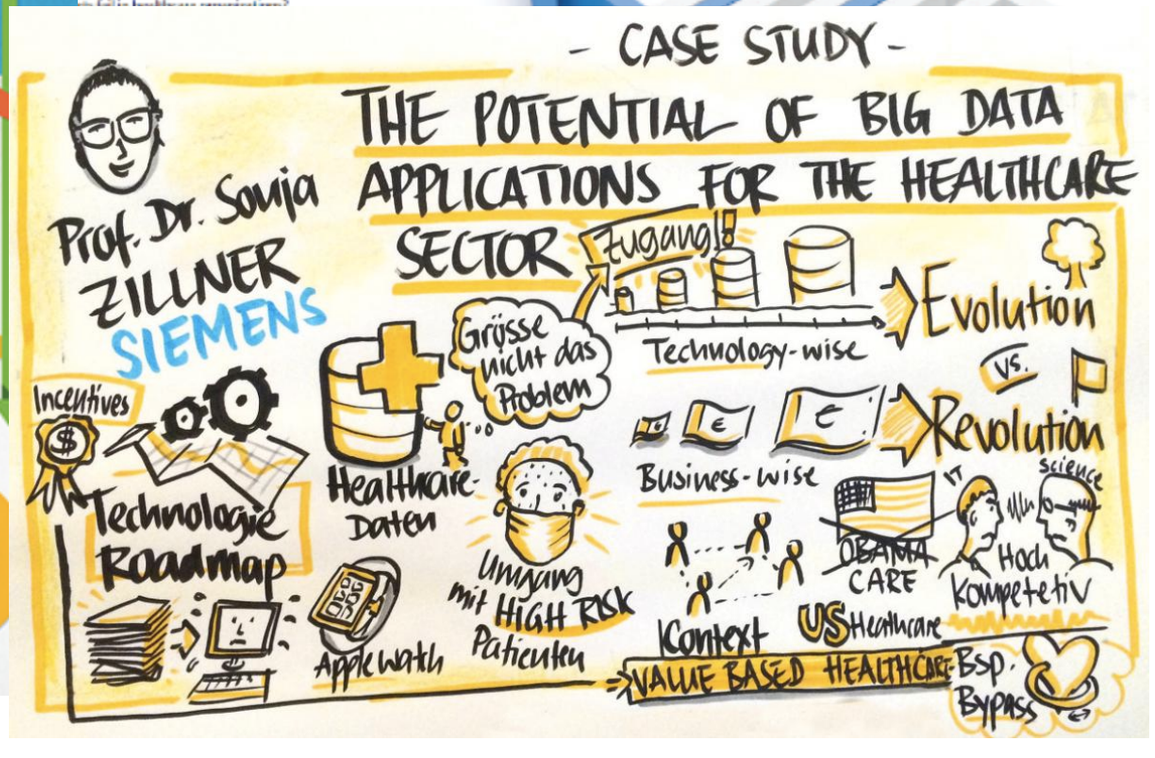
Understanding High-Dimensional Health Data via Tensor Factorization

Joyce Ho
Emory University

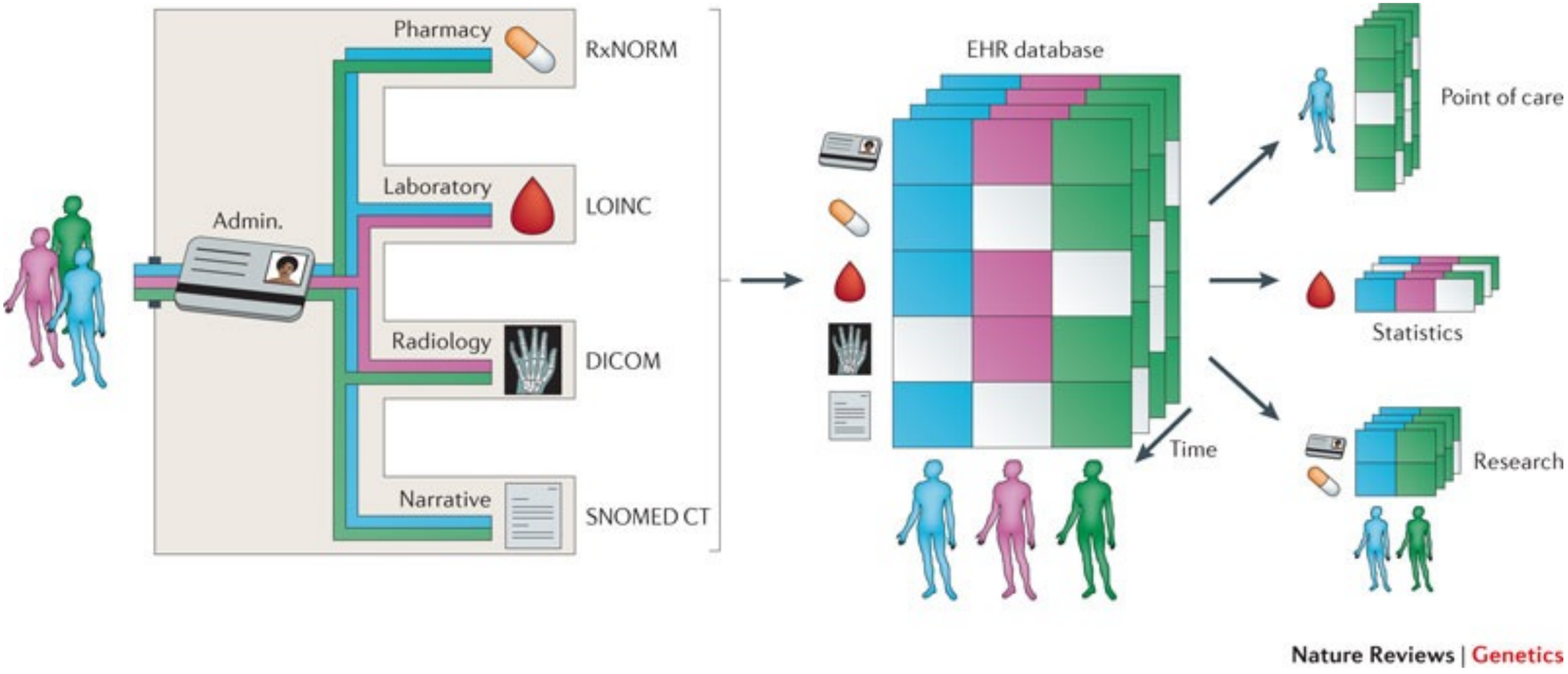
January 27, 2017

Electronic Health Data: Hype or Hope





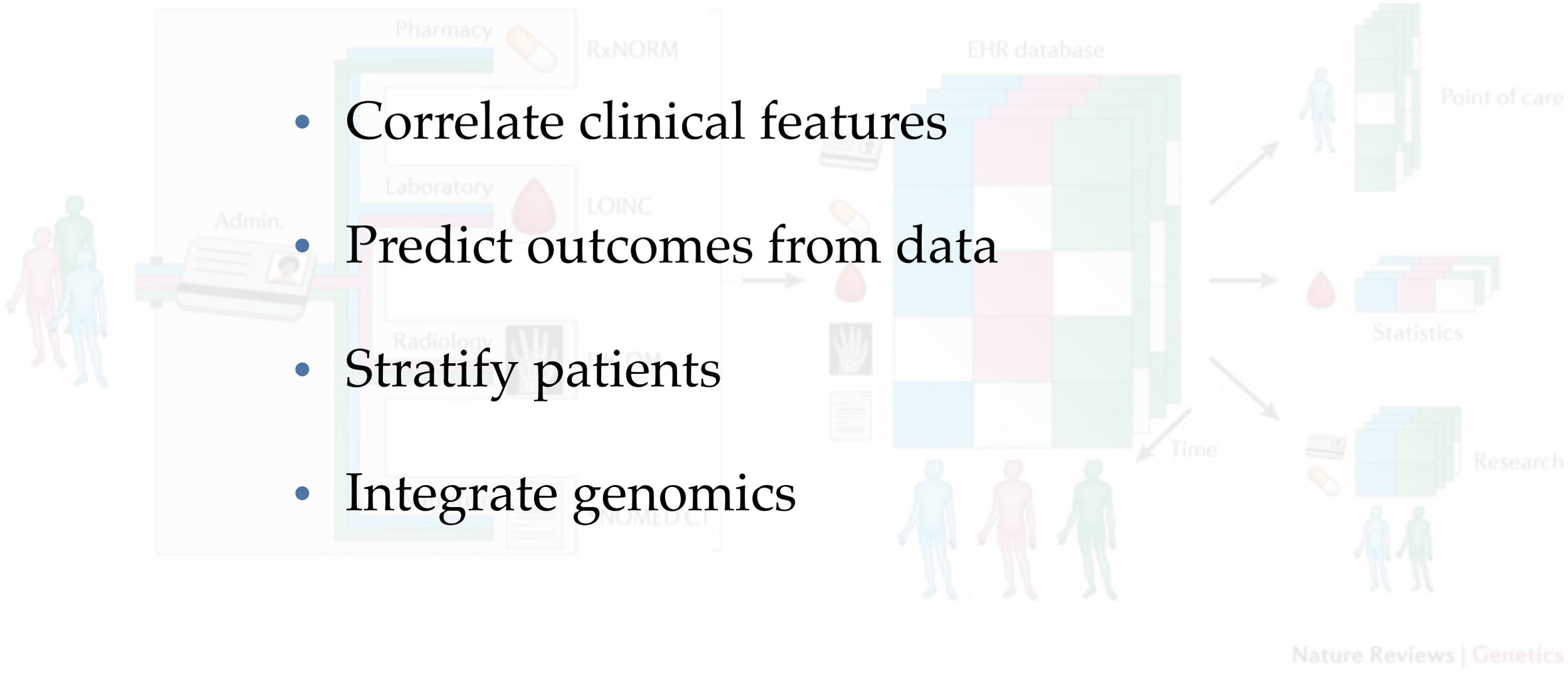
Electronic Health Records (EHRs)



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.

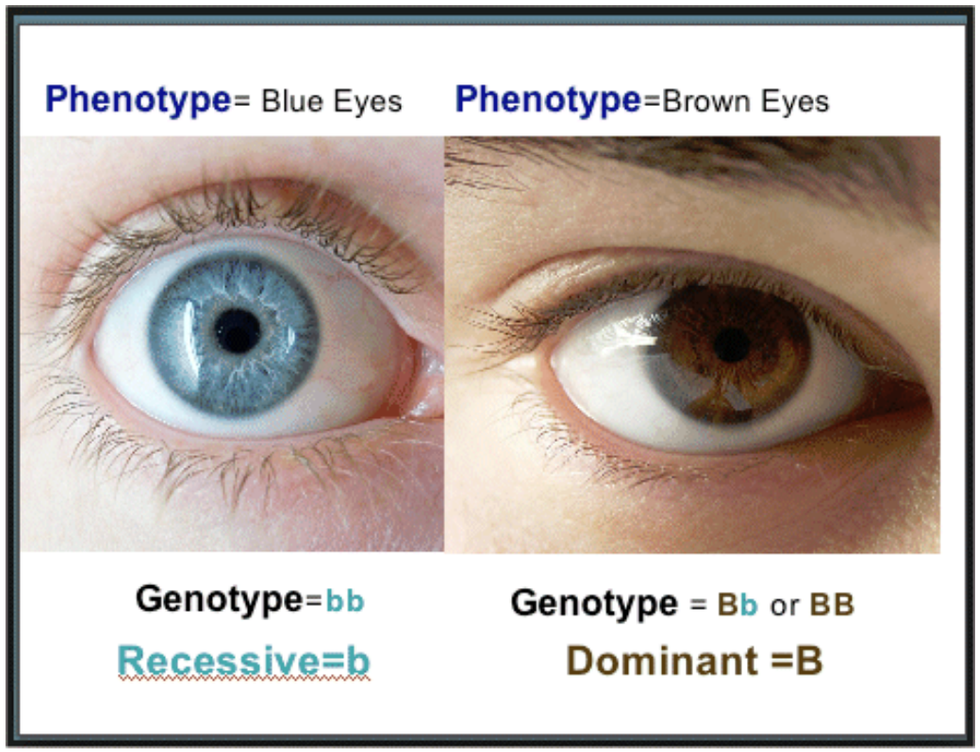
EHR Challenges

- 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

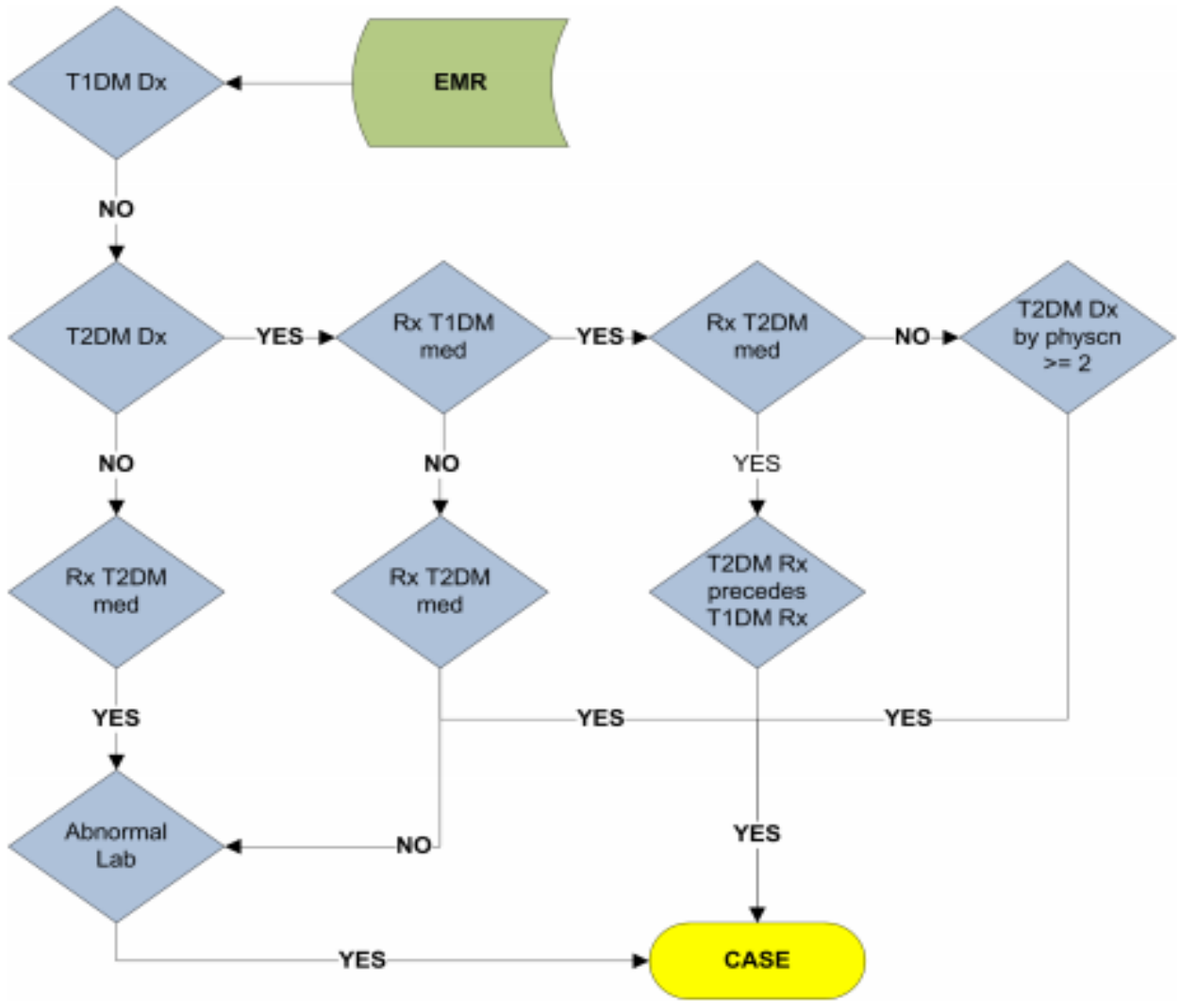


- Uses
  - Retrospective research
  - Clinical trial
  - Epidemiology / population health

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.

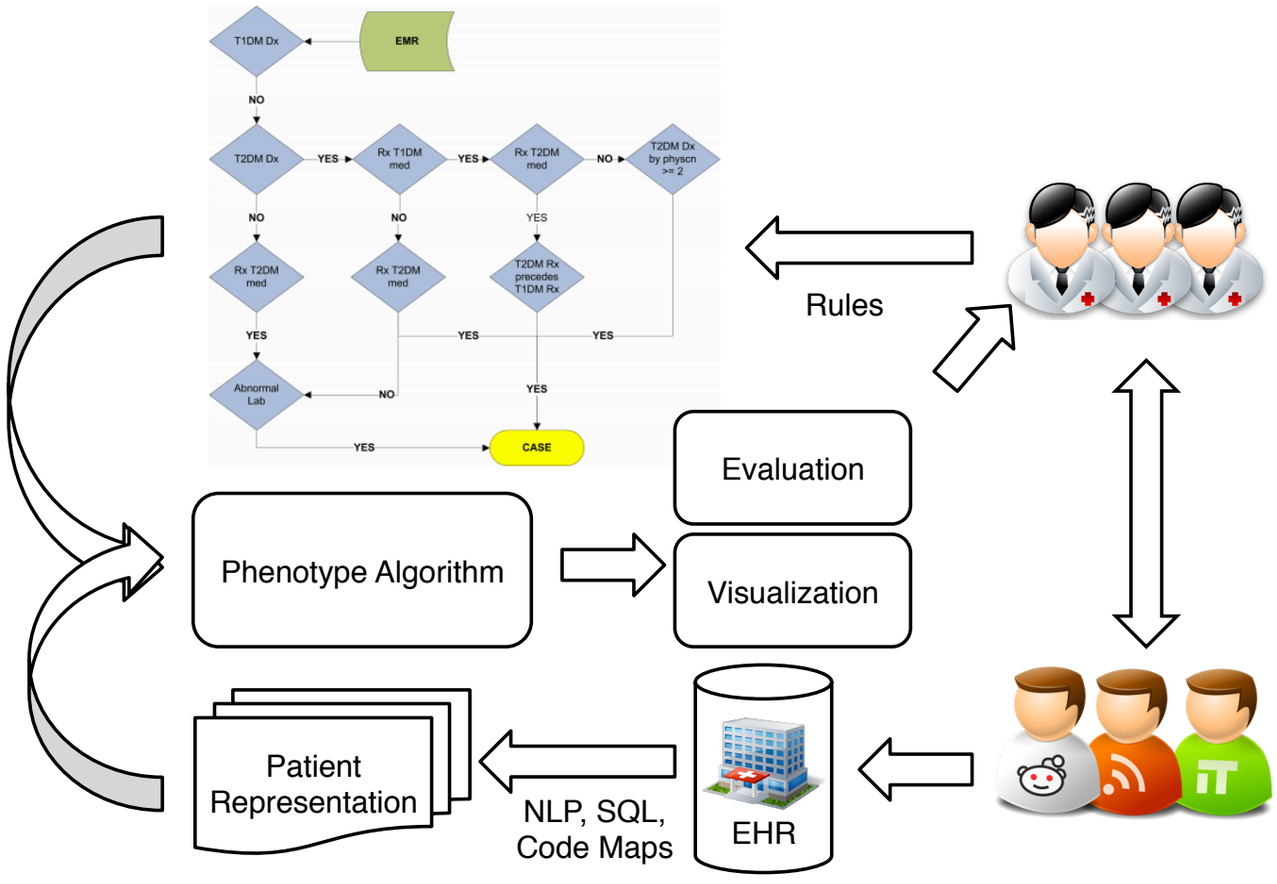
Phenotype: Modern Interpretation

- 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.

Fraiser, Brenden L. 38 years M DOB: 10/10/1970 MRN: 20378945 FIN: 050820934832 Admit: Unit LOS: 214 days Location: 7S-701-B

Dose Weight: 32kg **Actual Weight:** 33.2kg **Allergies:** Codeine, Sulfa

Active Meds

The system displays administration information charted within the past five (5) days.

Continuous Infusions

D5W 1000 ML + KCl 40 mEq + MgSO4 2 g IV ordered at 100 mL/hr (Tot Vol 1,000 ML)
Last Rate: 100 mL/hr (8/24/2009 17:52)

Dexmedetomidine IV ordered at 0.5 mcg/kg/hr
Last Dose: 0.5 mcg/kg/hr (8/24/2009 17:52)

DOBUTamine IV ordered at 10 mcg/kg/min
Last Dose: 5 mcg/kg/min (8/24/2009 17:52)

DOPamine ordered at 5 mcg/kg/min + NORepinephrine IV
Last Dose: 5 mcg/kg/min (8/24/2009 17:52)

Heparin IV ordered at 18 units/kg/hr
Last Dose: 18 units/kg/hr (8/24/2009 17:52)

Morphine IV ordered at 0.05 mg/kg/hr
Last Dose: 0.05 mg/kg/hr (8/24/2009 17:52)

Suspended Continuous Infusions Orders (1)

vecuronium IV ordered at 5 mg/hr
Last Dose: 5 mg/hr (8/24/2009 17:52)

Scheduled

ADEK 2 mL NG daily
Last Dose: 2 mL (8/24/2009 17:52)

ceftriaxone 2000 mg + D5W 50 mL IV Q24h (Tot Vol 50 mL)
Last Dose: View Details... (8/24/2009 17:52)

ciprofloxacin 400 mg IV q12h
Last Dose: 400 mg (8/24/2009 17:52)

gabapentin 300 mg NJ Q8h
Last Dose: 300 mg (8/24/2009 17:52)

methylPREDNISolone 48 mg IV daily
Last Dose: 60 mg (8/24/2009 17:52)

metoclopramide 10 mg NJ Q6h
Last Dose: 10 mg (8/24/2009 17:52)

metroNIDAZOLE 500 mg + D5W 95 mL IV Q6h (Tot Vol 100 mL)
Last Dose: View Details... (8/24/2009 17:52)

PANTOprazole 40 mg NJ daily
Last Dose: 40 mg (8/24/2009 17:52)

traZODone 200 mg NJ at bedtime
Last Dose: 200 mg (8/24/2009 17:52)

VORIconazole 300 mg IV Q12h
Last Dose: 300 mg (8/24/2009 17:52)

Zosyn 4.5 g IV Q6h

PRN

acetaminophen 650 mg PO Q6h PRN pain
Last Dose: —

acetaminophen 650 mg PR Q6h PRN fever
Last Dose: 650 mg (8/24/2009 17:52)
Given 1 times in last 24 hours

diPHENhydramine 25 mg PO Q6h PRN pruritus
Last Dose: 25 mg (8/24/2009 17:52)
Given 1 times in last 24 hours

haloperidol 2 mg IV Q6h PRN agitation
Last Dose: 2 mg (8/24/2009 17:52)
Given 1 times in last 24 hours

midazolam 5 mg IV Q2h PRN agitation
Last Dose: 5 mg (8/24/2009 17:52)
Given 6 times in last 24 hours

morphine 10 mg IV Q2h PRN pain
Last Dose: 10 mg (8/24/2009 17:52)
Given 1 times in last 24 hours

Unscheduled

clindamycin 600 mg IV on-call (Tot Vol 50 ML)
Last Dose: —

Labs

Blood Gases (Last 2 in 24 hours)

Lab	13:51	12:52
pH	7.35 ↓	
PO2	80 ↓	
PCO2	41.0	
HCO3	18 ↓	
BD	3	
pH ven		7.44
PO2 ven		78
PCO2 ven		48
HCO3 ven		24

Chemistry

Lab	Latest	Previous
iCa	1.12 7/30/2009 04:00	1.14 7/29/2009 16:00
Lactate	1.7 ↑ 7/30/2009 04:00	2.2 ! 7/29/2009 16:00
Na	138 7/30/2009 04:00	149 ↑ 7/29/2009 16:00
K	5.5 7/30/2009 04:00	2.8 ↓ 7/29/2009 16:00
K whole blood	6.1 ↑ 7/30/2009 04:00	2.9 △ 7/29/2009 16:00
Cl	101 7/30/2009 04:00	109 ! 7/29/2009 16:00
CO2	25 7/30/2009 04:00	29 ↑ 7/29/2009 16:00
AGAP	12 7/30/2009 04:00	8 7/29/2009 16:00
BUN	8 7/30/2009 04:00	9 7/29/2009 16:00
Creat	1.5 ↑ 7/30/2009 04:00	1.1 7/29/2009 16:00
Glu	83 7/30/2009 04:00	113 ↑ 7/29/2009 16:00
Ca	10.7 7/30/2009 04:00	8.5 7/29/2009 16:00
Mg	2.1 7/30/2009 04:00	1.7 7/29/2009 16:00

Data Representation: Take 1

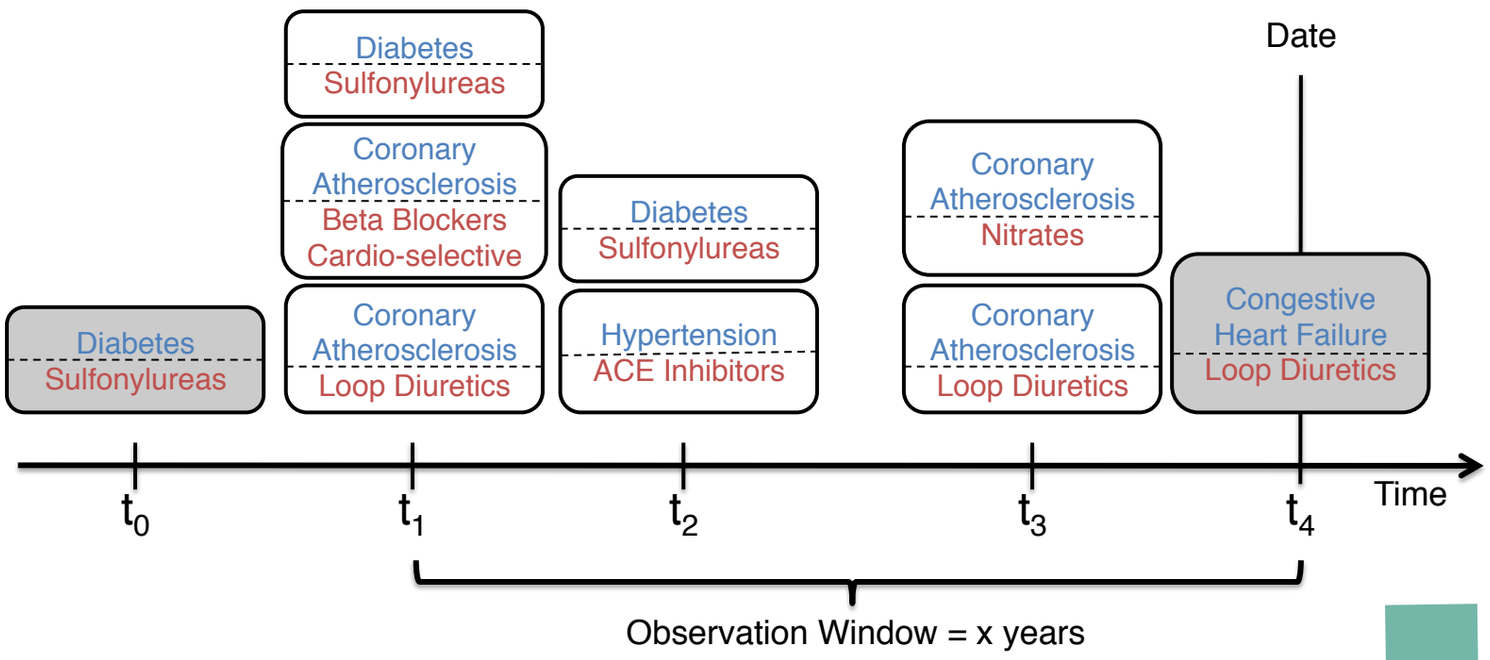
Vector Representation

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

Example: Patient Vector



Diabetes	Coronary Atherosclerosis	Hypertension	Loop Diuretics	ACE Inhibitors	Beta Blockers Cardio-selective	Sulfonylureas	Nitrates
2	2	1	2	1	1	2	1

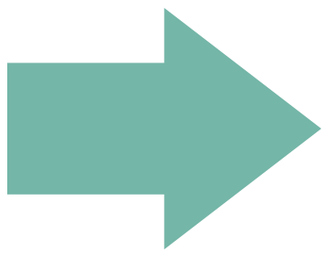
Feature Matrix Representation

Stack patient vectors to get a feature matrix

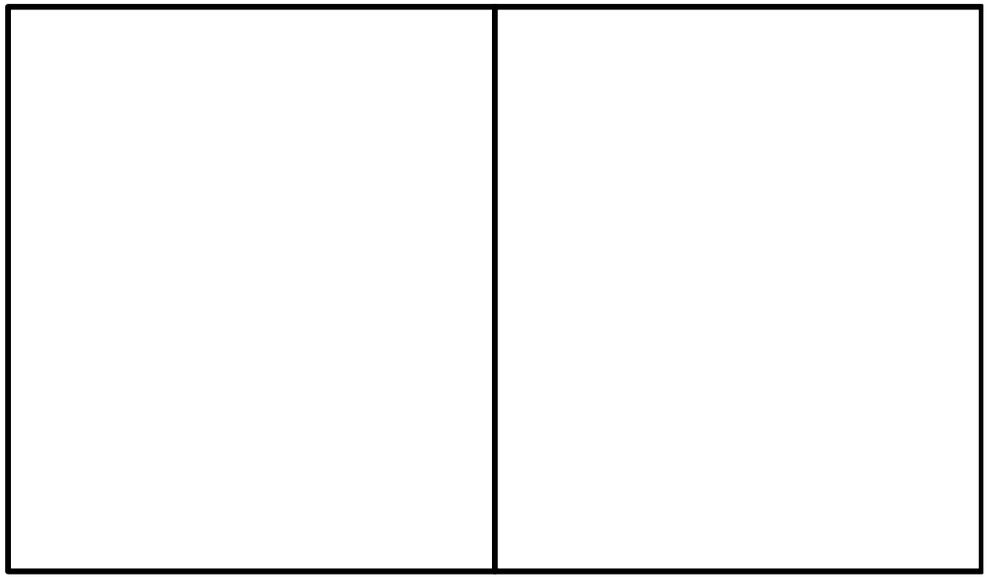
Diagnoses

Medications

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



Patients

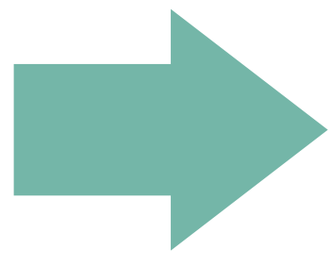


What if we want to represent same visit interactions?

Feature Interaction Matrix

Encode the same visit interactions with columns that represent the presence of that combination
(e.g. one column represents hypertension - ACE inhibitor)

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



Patients

Diagnosis

×

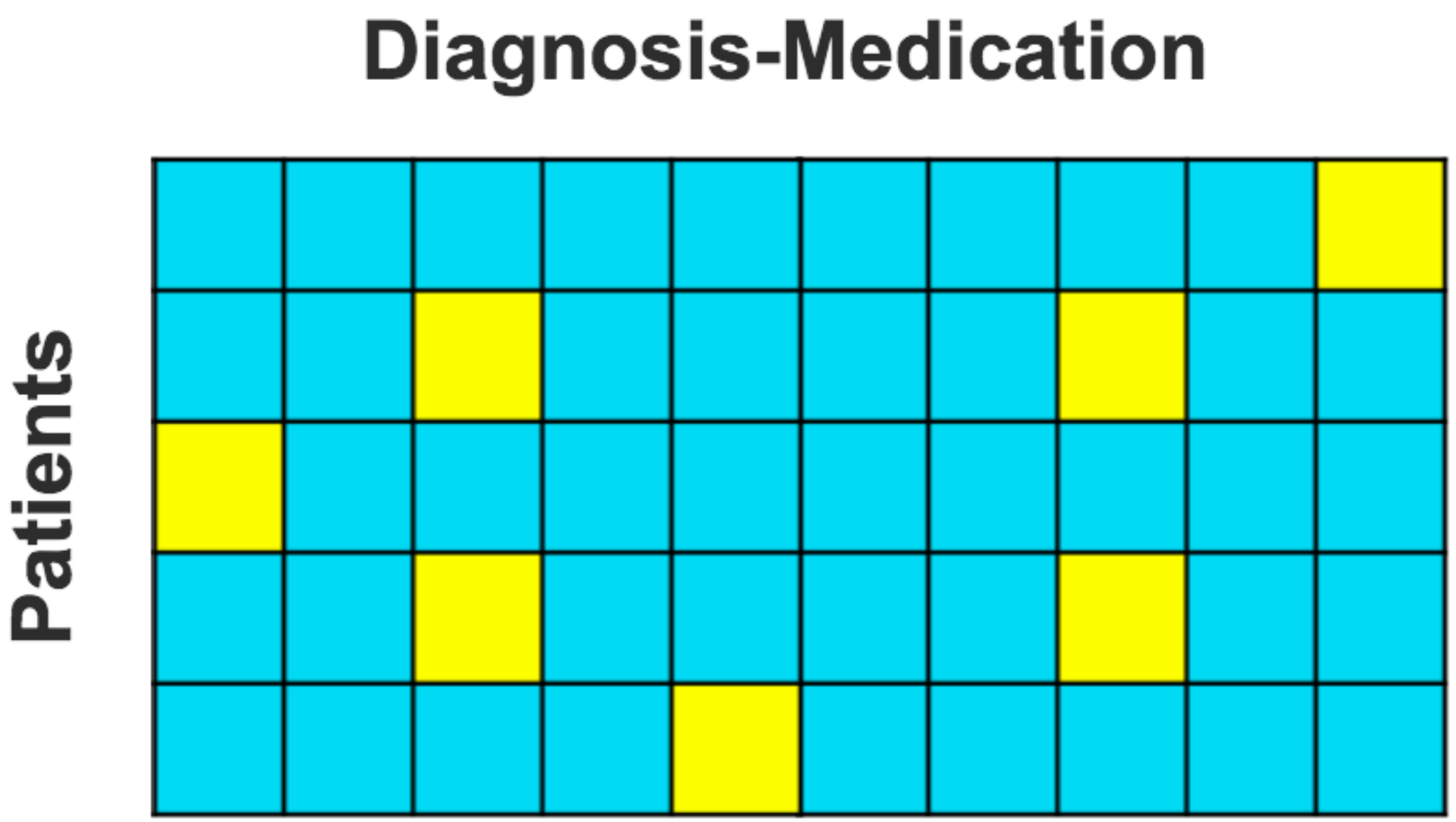
Medication

Feature Interaction Matrix

Encode the same visit interactions with columns that represent the presence of that combination
(e.g. one column represents hypertension - ACE inhibitor)

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

Matrix with many columns and many zeros (sparse)



Interaction matrix of medication
for specific disease

Dimensionality Reduction: Matrix

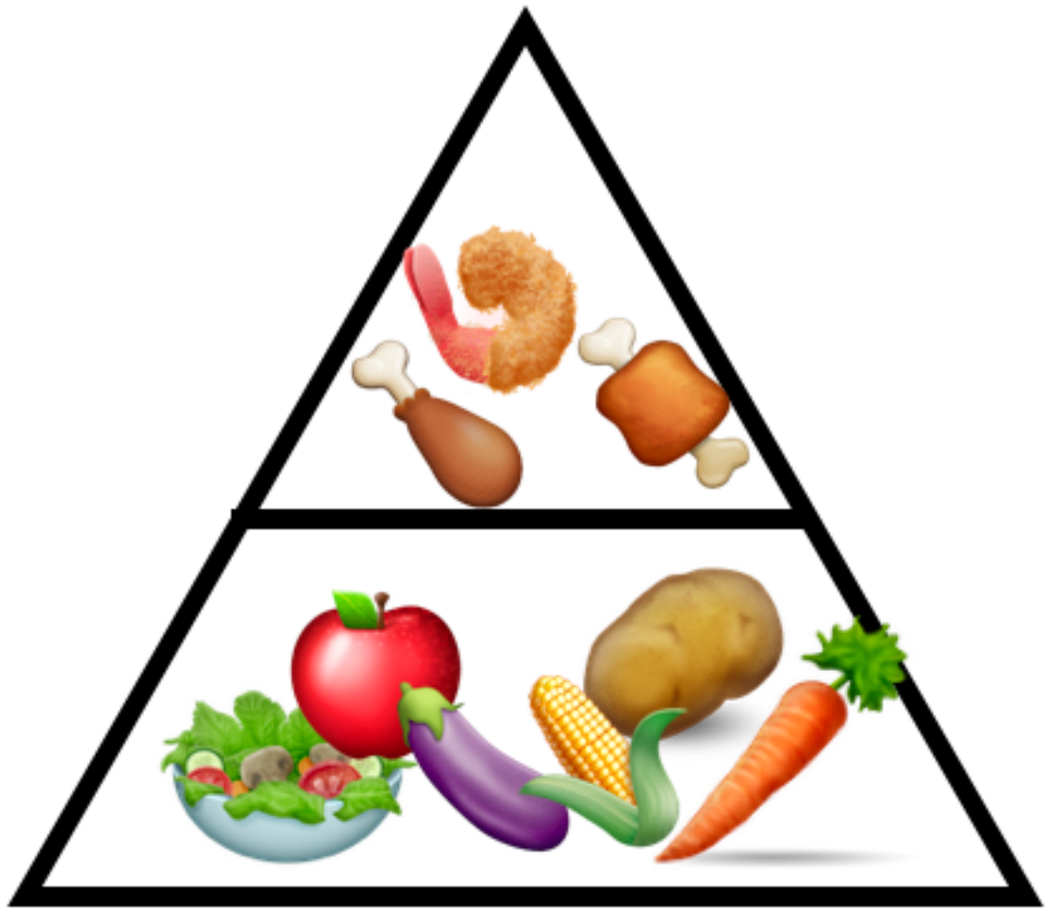
Adapted from Fei Wang's Talk @ ICHI 2016

Dimensionality Reduction

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

Example: Food Nutrition

- What is the best way to differentiate food items?
  - Vitamin content
  - Protein levels
  - Fat
  - Fiber



https://algobeans.com/2016/06/15/principal-component-analysis-tutorial/

Example: Food Nutrition Data

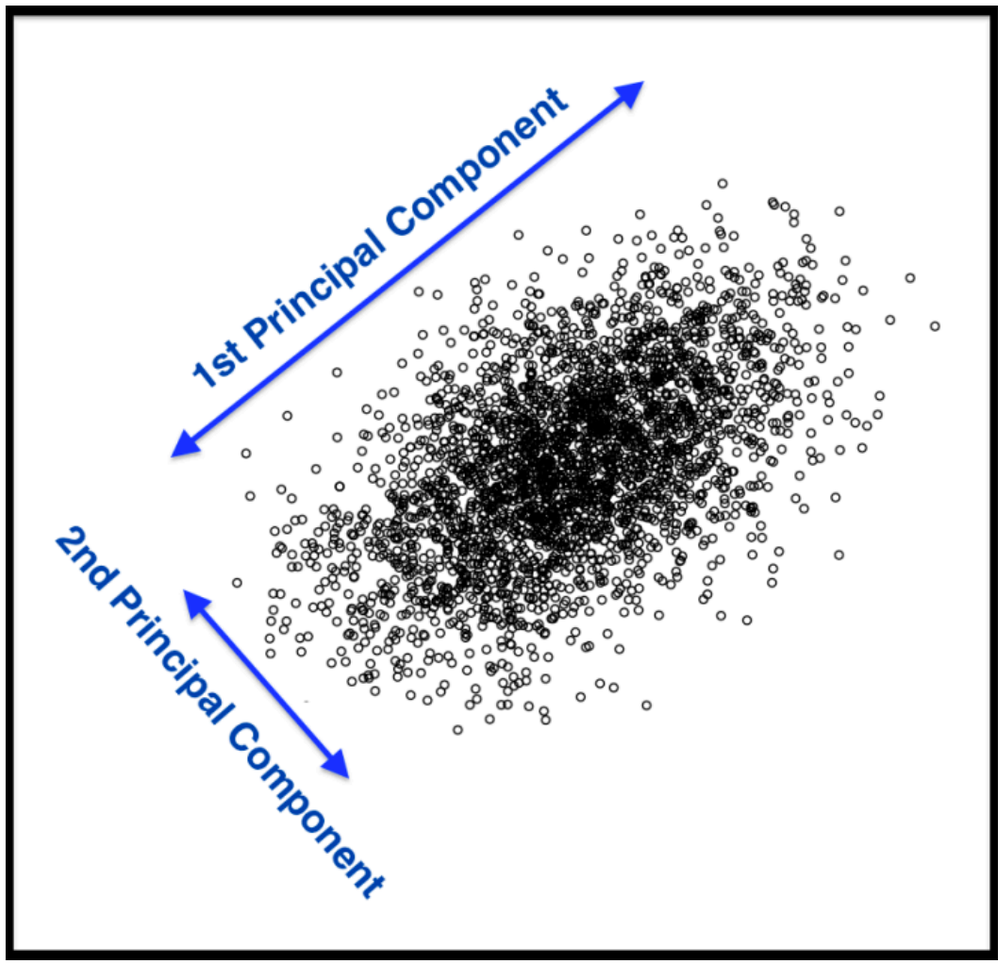
https://algobeans.com/2016/06/15/principal-component-analysis-tutorial/

Example: Linear Combinations

https://algobeans.com/2016/06/15/principal-component-analysis-tutorial/

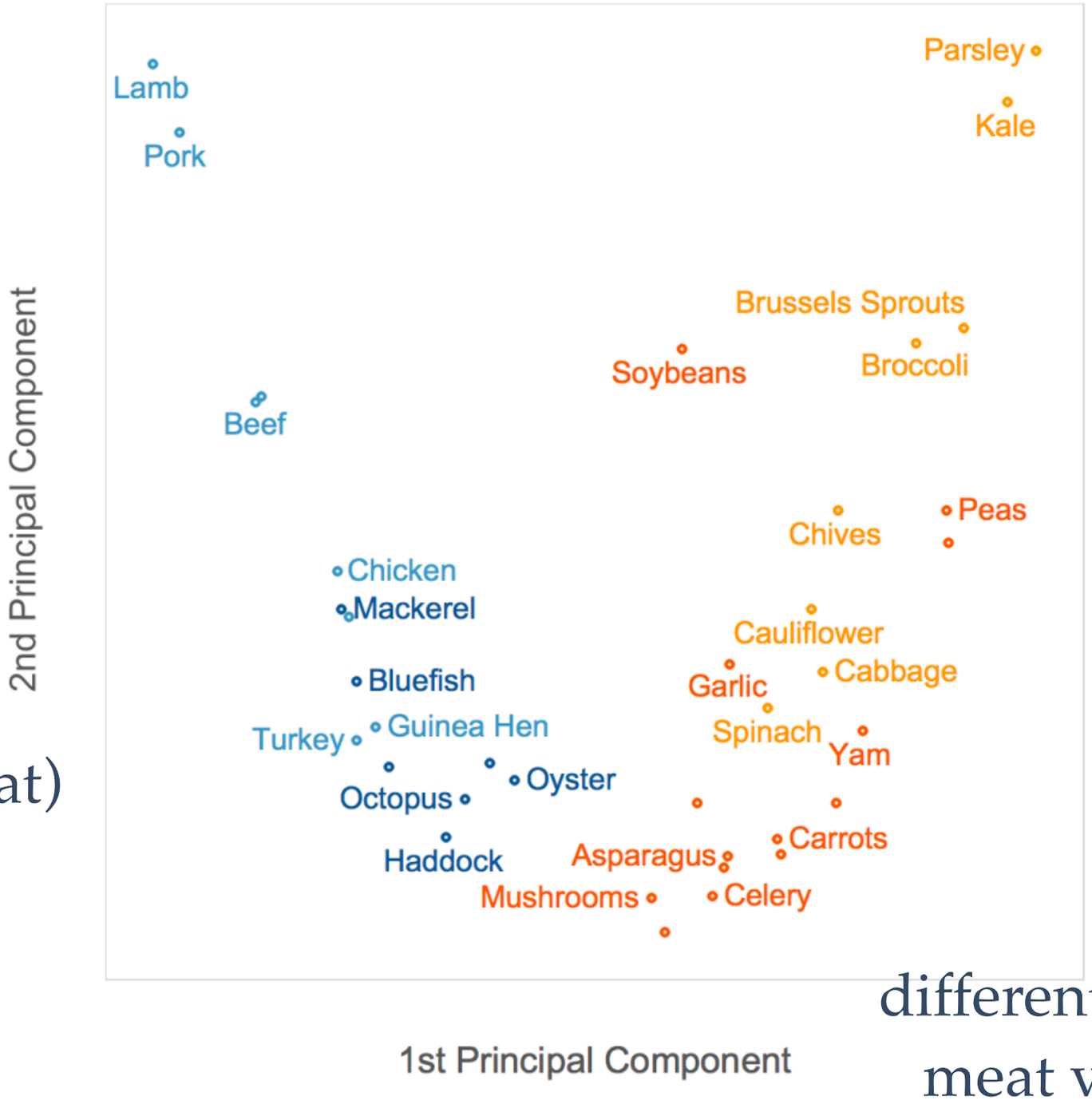
Principal Component Analysis (PCA)

- Find underlying variables (principal components) that best differentiate your data points
- PC are dimensions along which your data points are most spread out (maximizing variance)
- Examines correlation to reduce the number of dimensions



Example: PCA

differentiates between fat (meat) and vitamin c (vegetables)

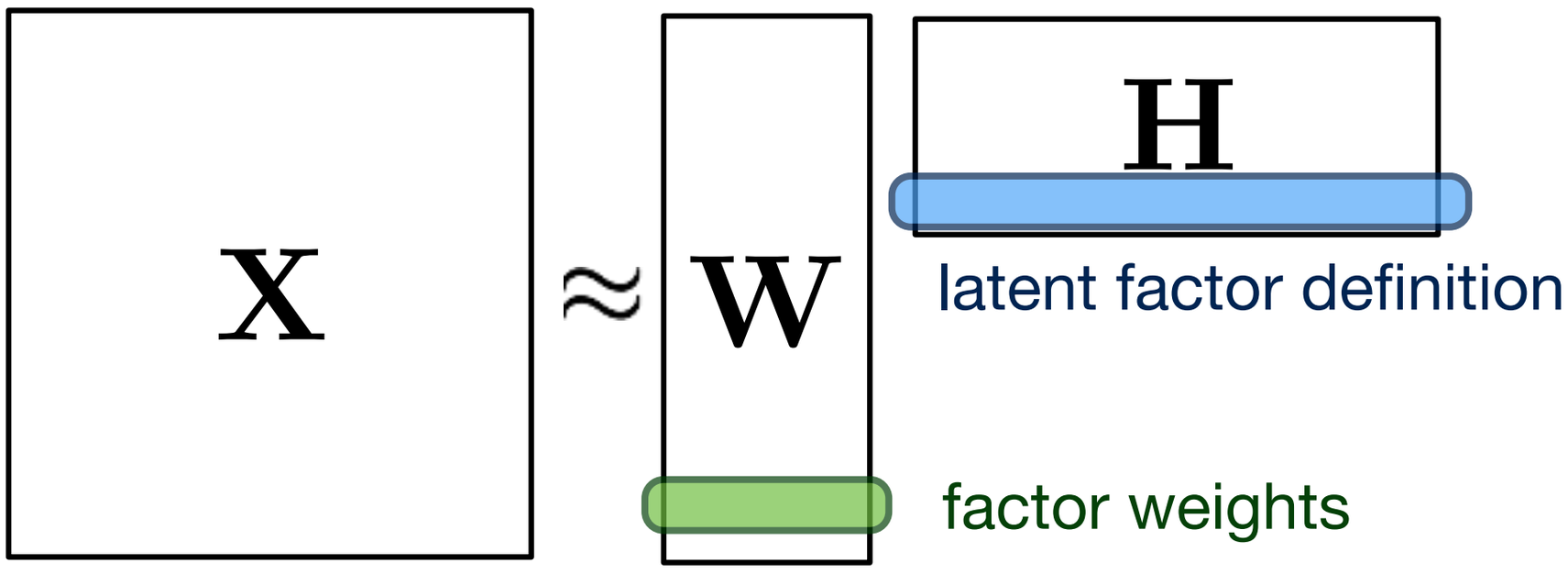


differentiates between meat vs vegetables

https://algobeans.com/2016/06/15/principal-component-analysis-tutorial/

Matrix Factorization

- Generalization of PCA
- Approximate original matrix using two low rank matrices to uncover latent relations or factors
- Common technique used in recommendation systems, clusterings, etc.



Example: PCA Loadings

	PC1	PC2	PC3	PC4
Fat	-0.45	0.66	0.58	0.18
Protein	-0.55	0.21	-0.46	-0.67
Fiber	0.55	0.19	0.43	-0.69
Vitamin C	0.44	0.70	-0.52	0.22

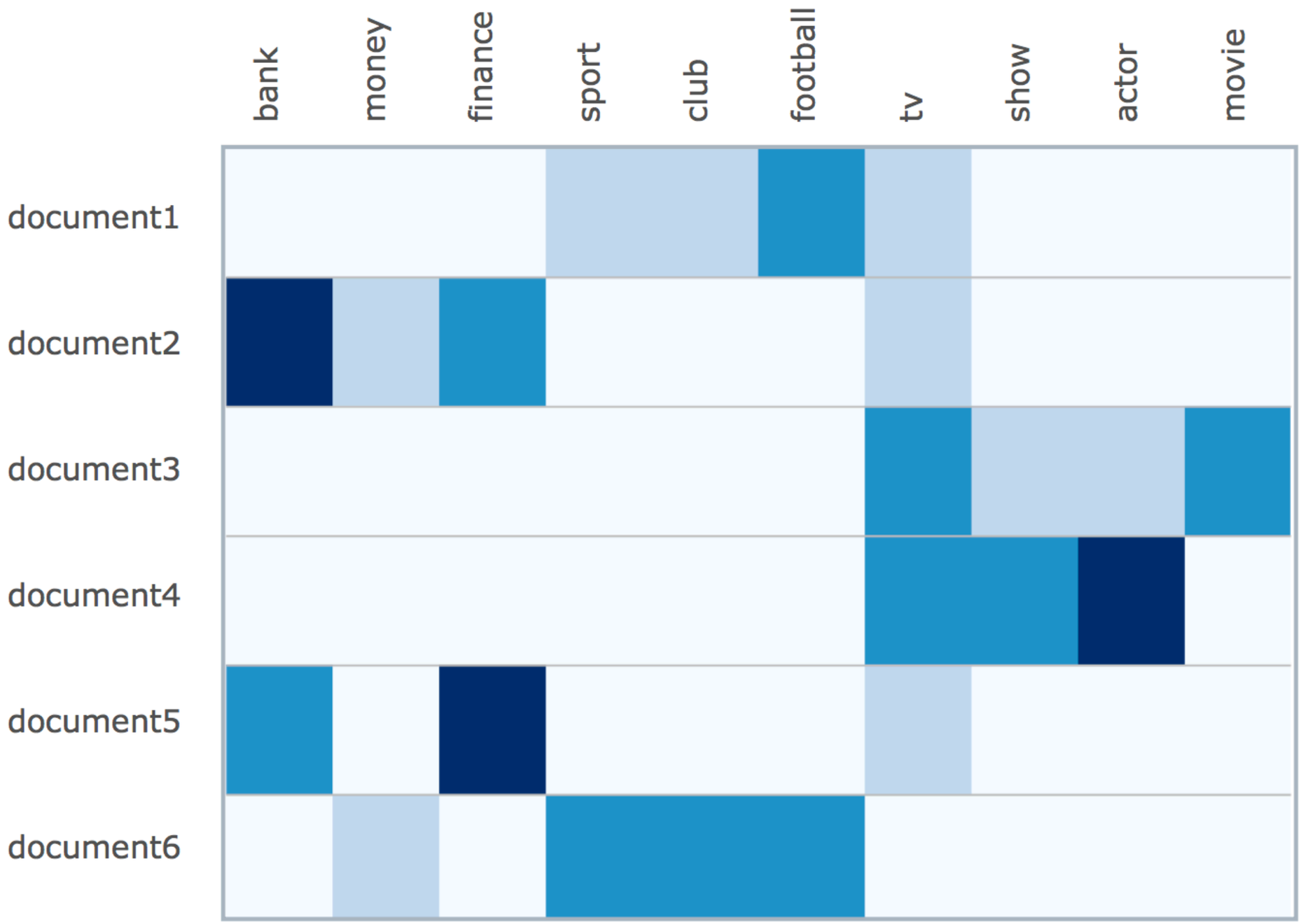
What happens if negative combinations doesn't make sense?
Example: trying to find the topics in a document

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"

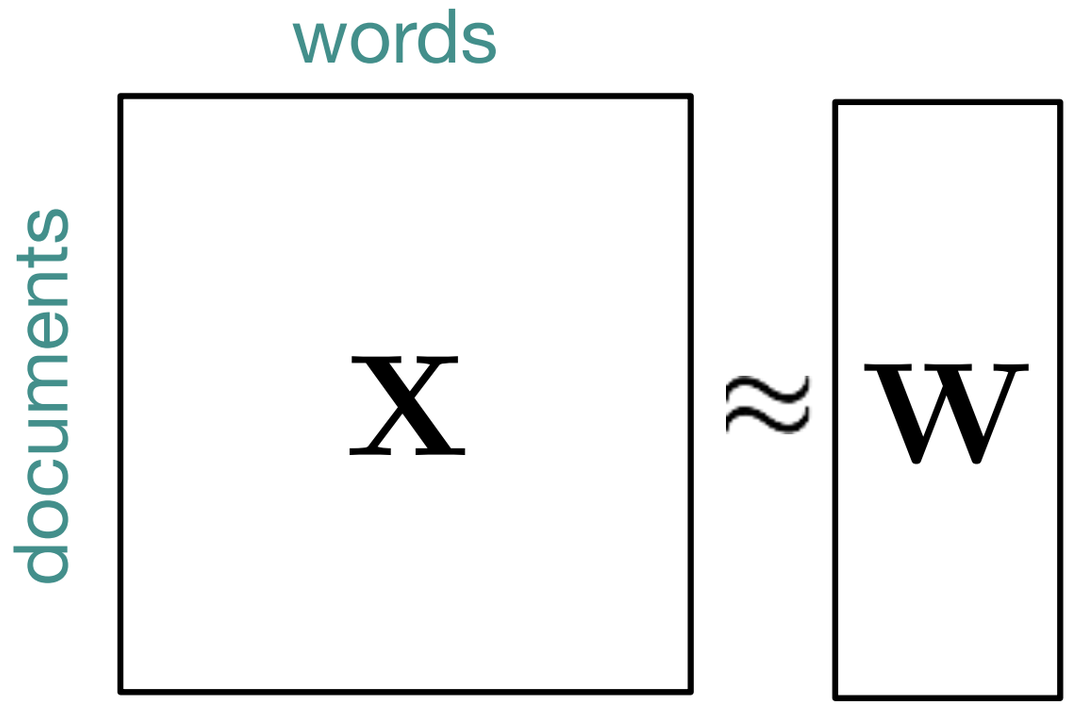
$$\mathbf{X} \approx \mathbf{W}\mathbf{H}^{\top}$$
$$\text{s.t. } \mathbf{W} \geq 0, \ \mathbf{H} \geq 0$$

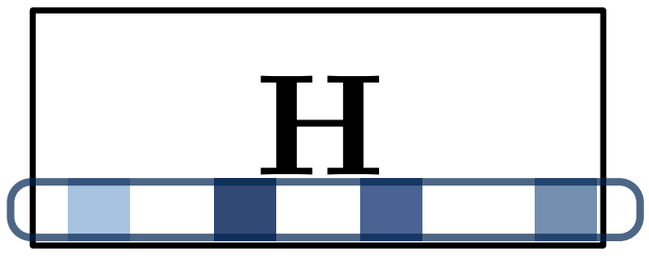
Example: Topic Modeling



http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.702.4867&rep=rep1&type=pdf

Example: NMF Words to Topics

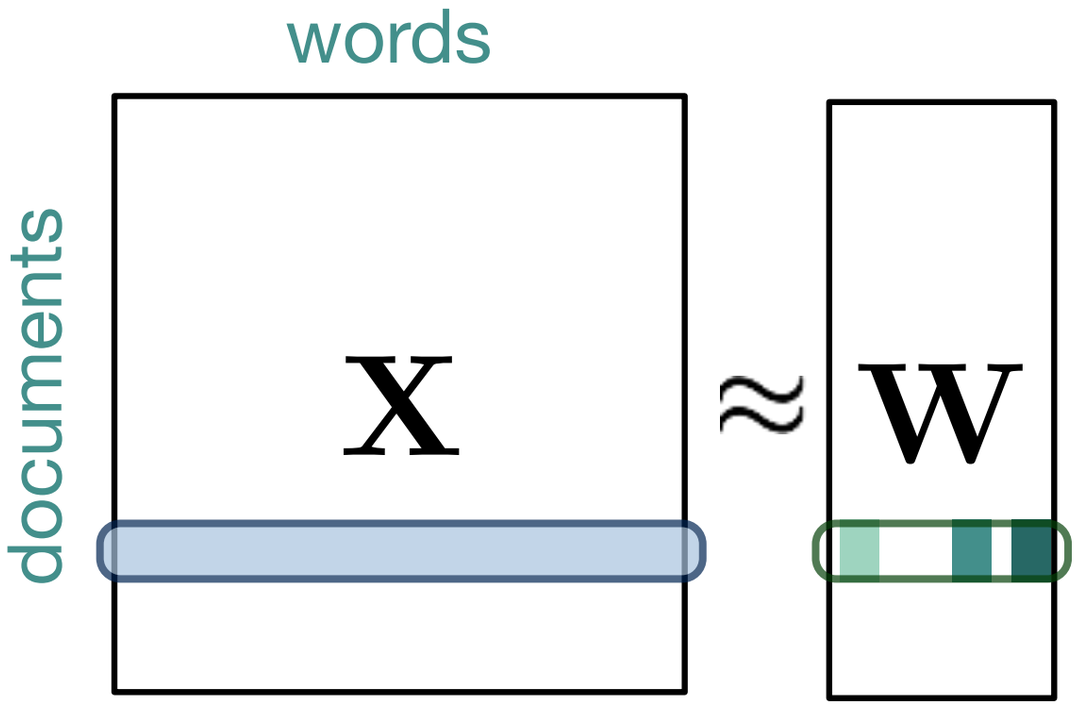




Personal finance topic

money	0.15
retire	0.10
risk	0.03
…	…

Example: NMF Documents to Topics

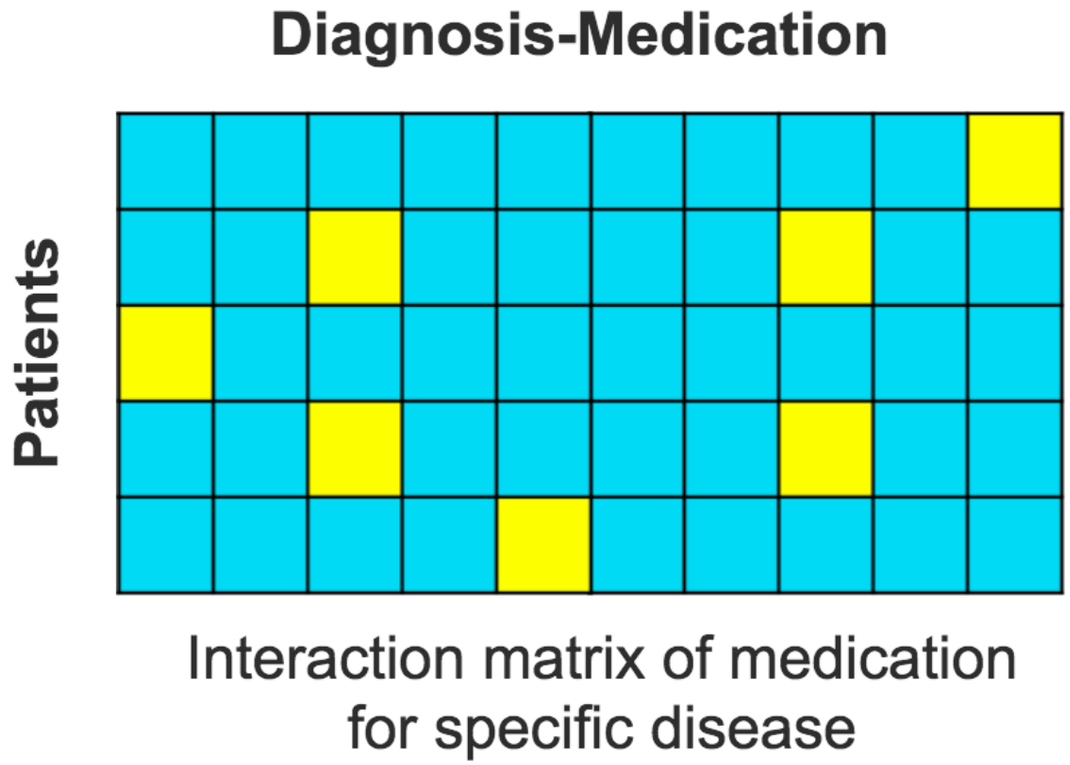


article from wall street journal

topic loadings

personal finance	0.55
bond-fund	0.30
advice	0.15

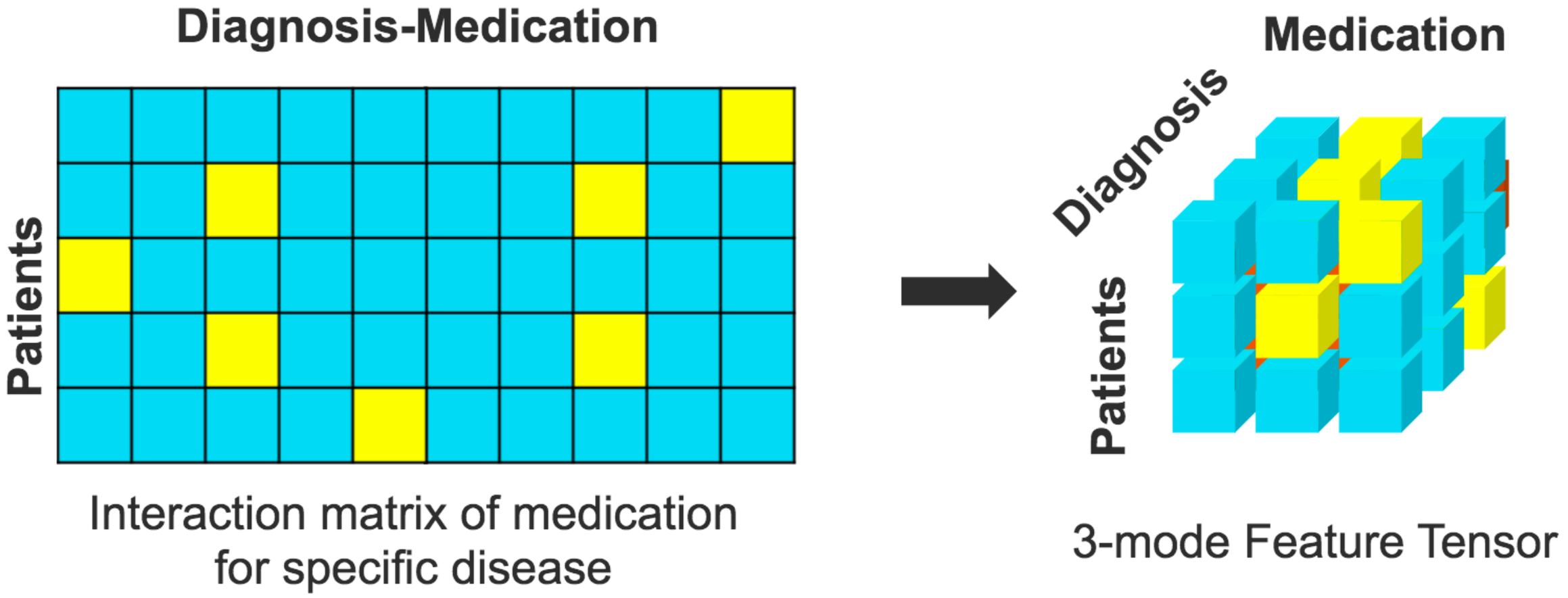
Example: Interaction Matrix



NMF

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

Very hard to interpret the results!
But note the higher level grouping



Feature Tensor to the Rescue!

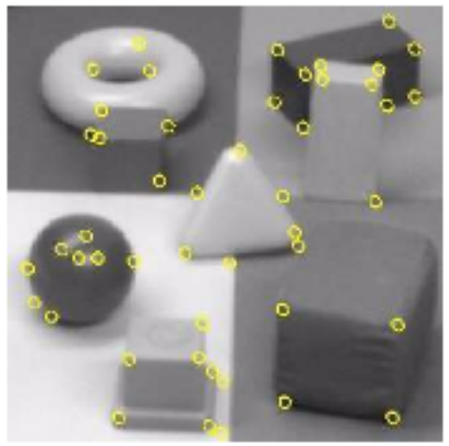
Adapted from Fei Wang's Talk @ ICHI 2016

Tensors (Multiway Arrays)

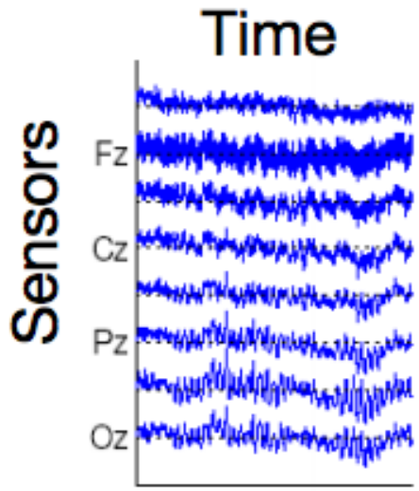
- Generalization of matrices to multidimensional array
- Representation of an n-way interaction
- Captures hierarchical information in the structure
- Used in lots of places (e.g., chemistry, neuroimaging, bioinformatics, text mining, psychology, etc.)

Tensors are Everywhere

Matrices



Black and White



Multivariate time series

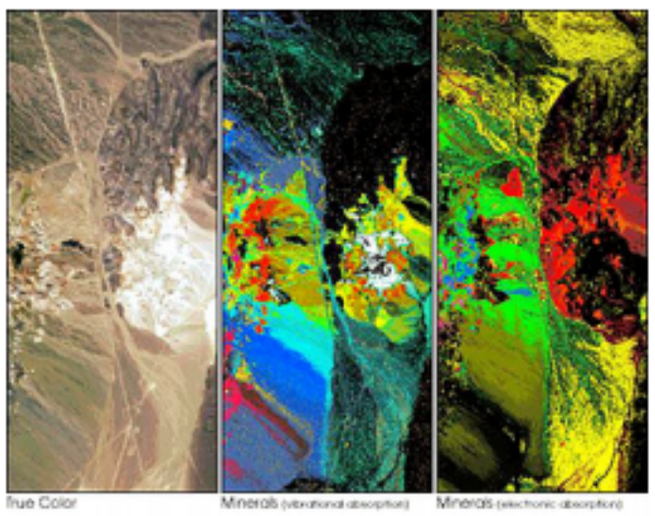
Movies

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

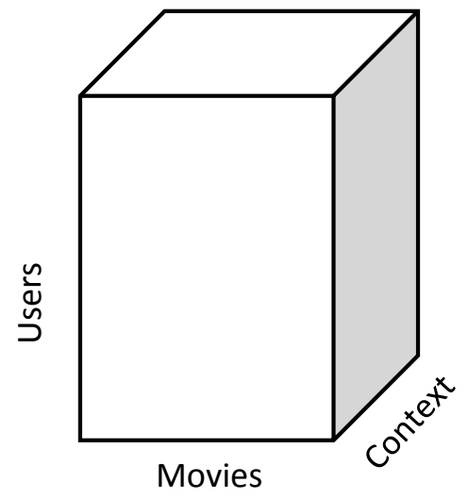
Movie recommendation

Tensors

Color



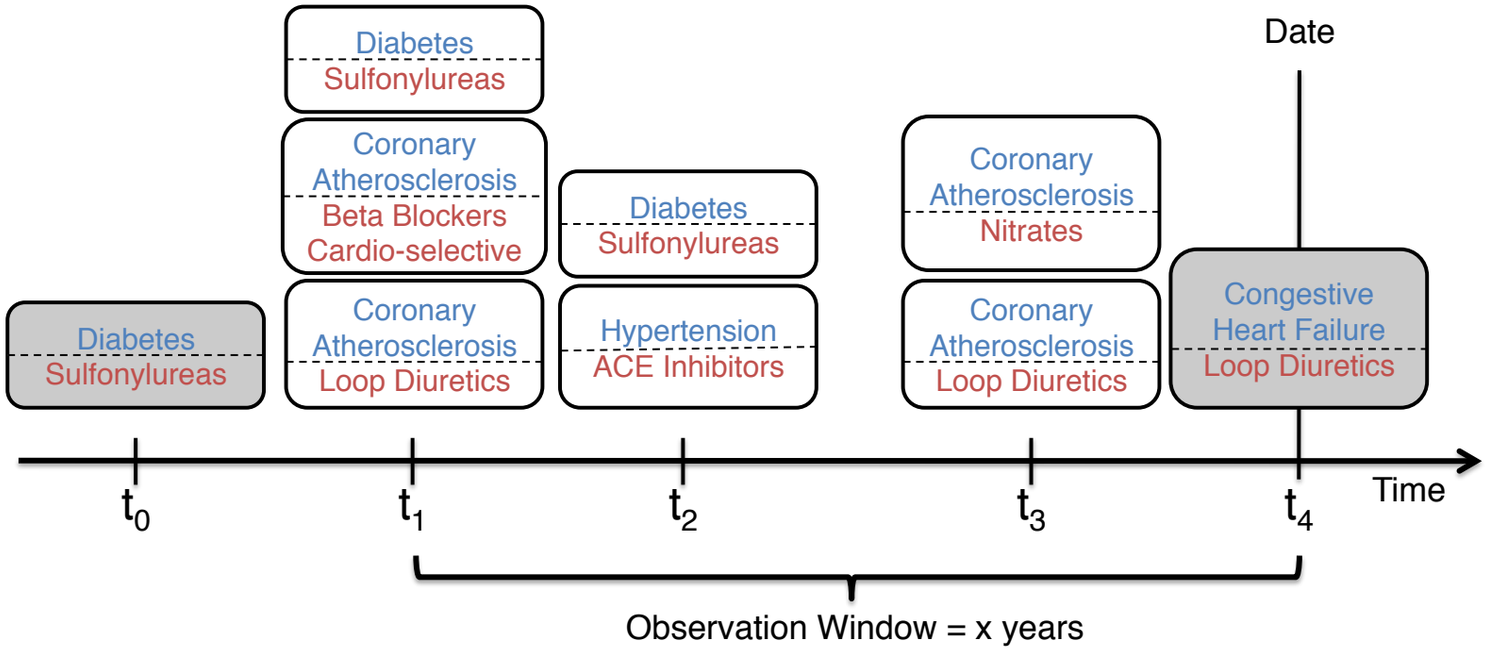
Spatio-temporal data

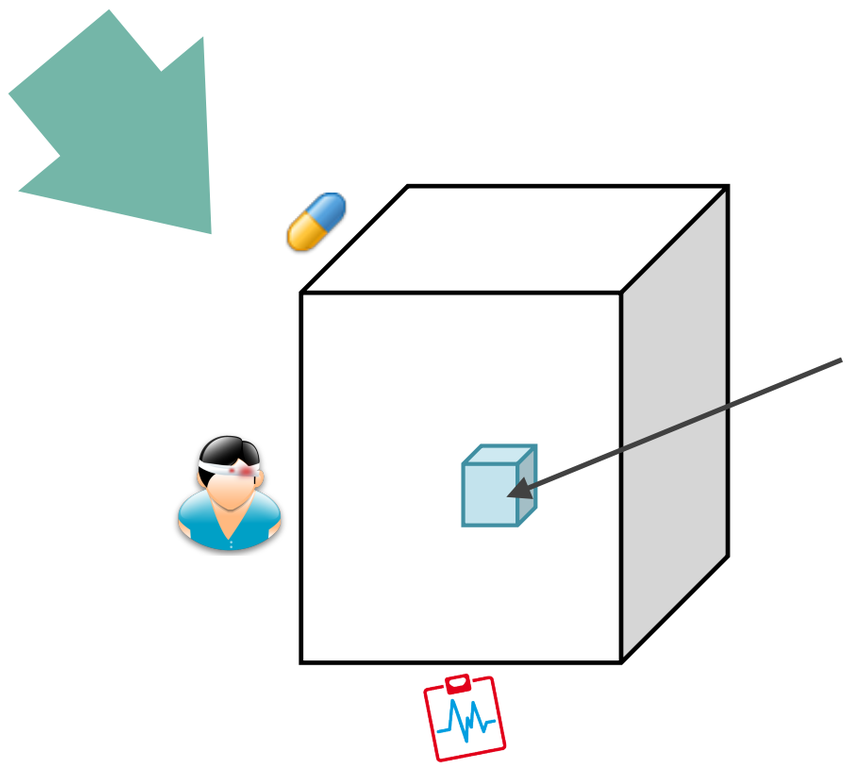


Multiple relations

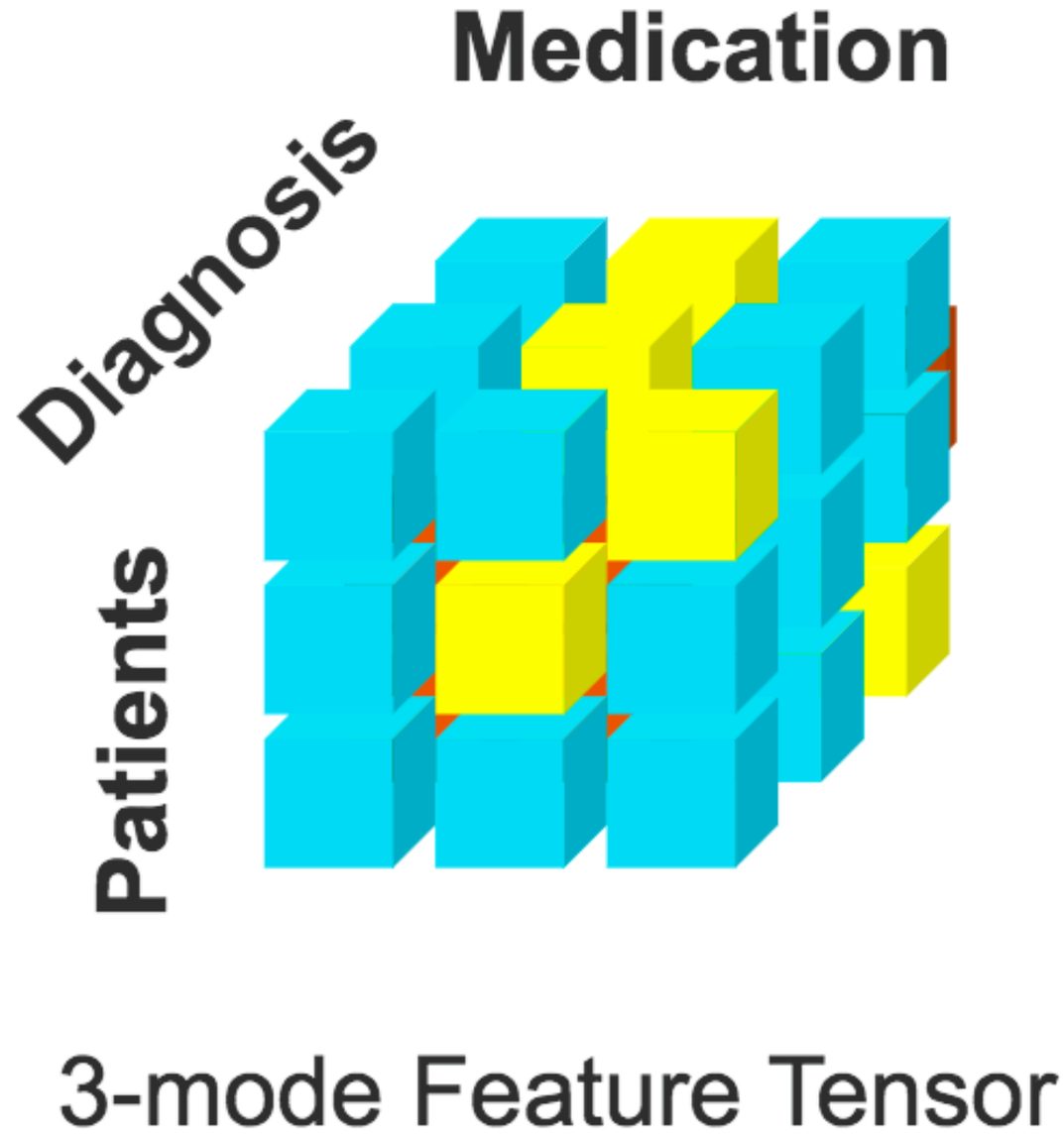
Adapted from Ryoto Tomioka's Atalk @ MLSS 2015 talk on Tensor decompositions: old, new, and beyond

EHR Tensor Generation



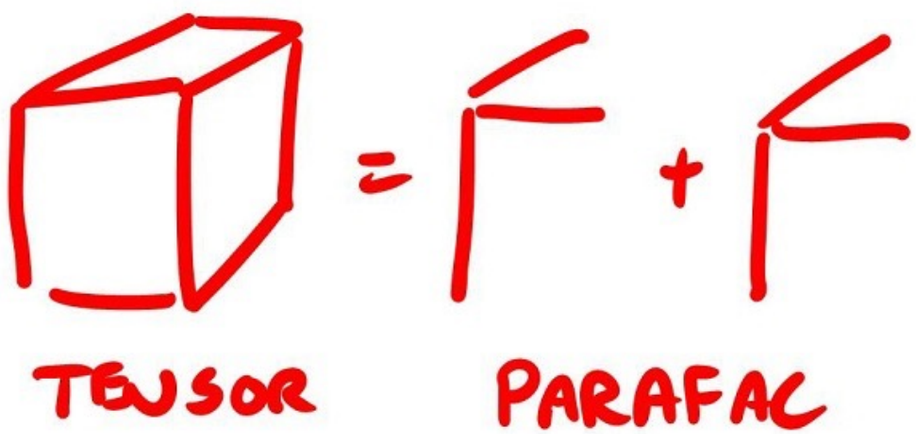


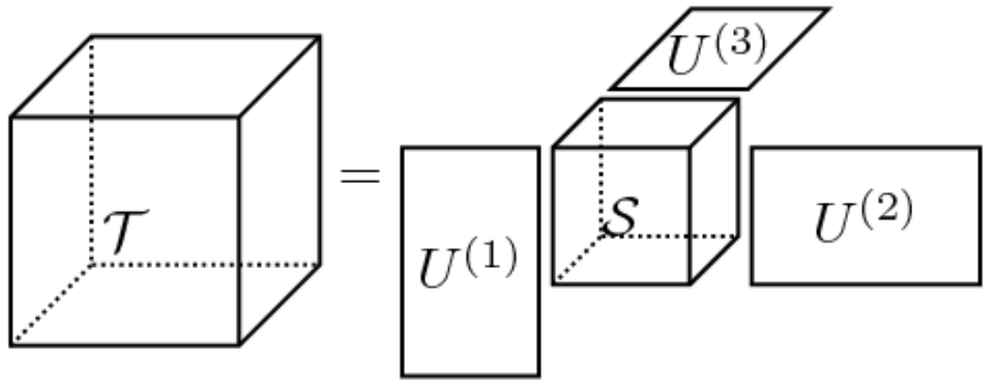
Each element represents the relationship between patient, medication, and diagnosis

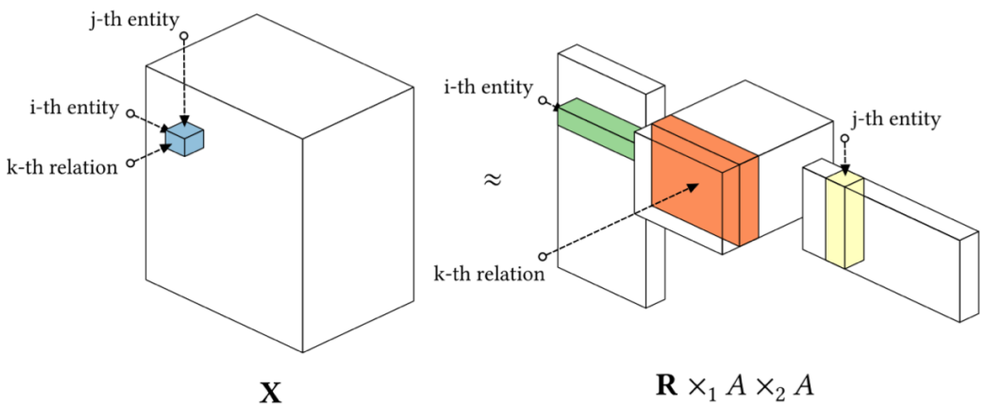


Dimensionality Reduction: Tensor

Tensor Factorization





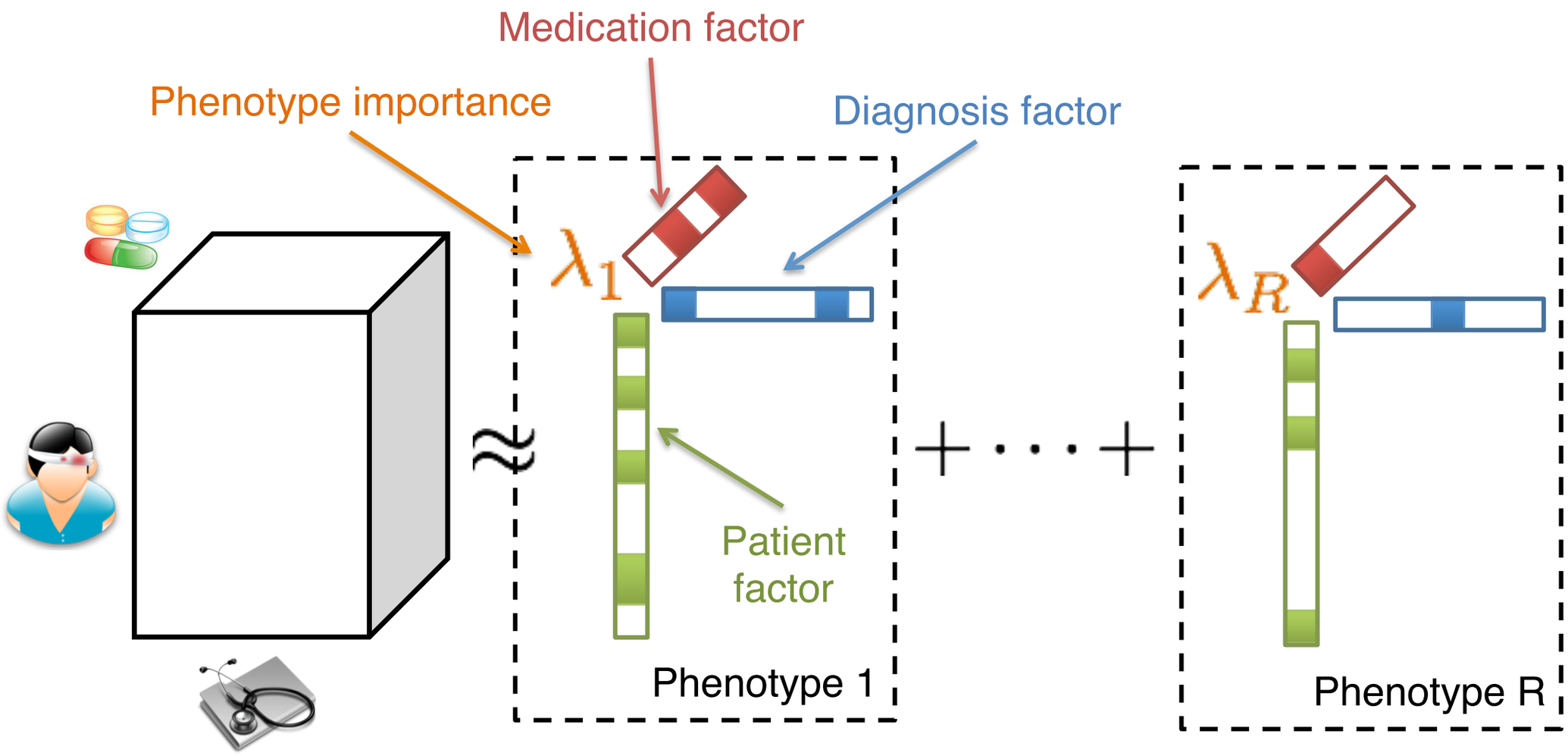


- Generalization of matrix factorization
- Multiway structure information utilized during decomposition process
- Many decomposition models: CANDECOMP / PARAFAC (CP), Tucker, Rescal

LIMESTONE

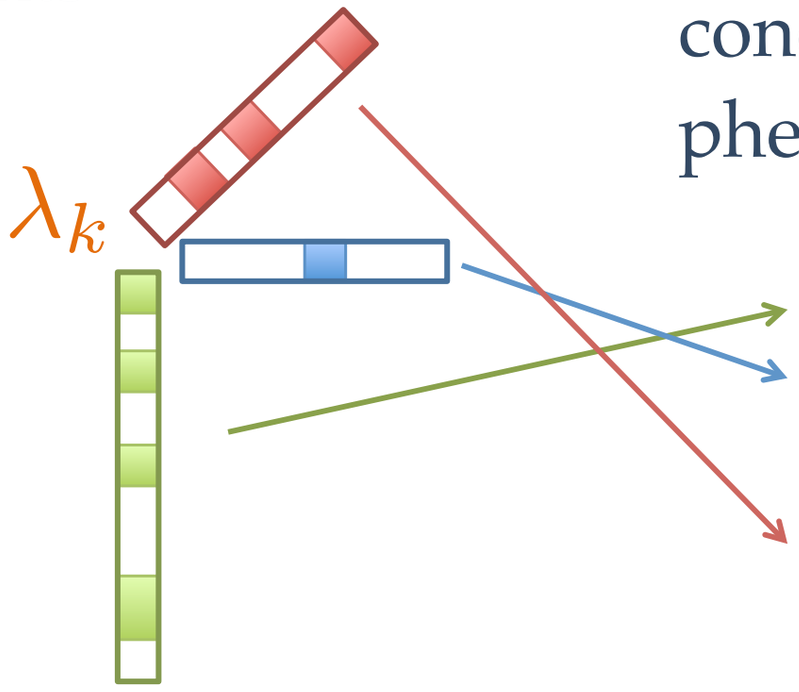
- CP decomposition of an EHR feature tensor
- Simultaneously discover multiple phenotypes
- Post-process the latent factors to remove probabilistically unlikely elements for improved interpretability

LIMESTONE: Phenotype Generation



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	0.94
Hypertensive Heart Disease	0.06
Beta Blockers Cardio-Selective	0.51
Calcium Channel Blockers	0.32
Diuretic Combinations	0.06
Nitrates	0.06
HMG CoA Reductase Inhibitors	0.06
Vasodilators	0.05

NMF Phenotype	
Hypertension – Sympathomimetics	0.0032
Hypertension – Insulin	0.0027
Hypertension – Potassium	0.0018
Hypertension – Beta Blockers Cardio-Selective	0.0004
Hypertension – HMG CoA Reductase Inhibitors	0.0003
Major Symptoms, Abnormalities – Sympathomimetics	0.0167
Major Symptoms, Abnormalities – Insulin	0.0143
Major Symptoms, Abnormalities – Sodium	0.0133
Major Symptoms, Abnormalities – Potassium	0.0097
Major Symptoms, Abnormalities – Coumarin Anticoagulants	0.0092
Vascular Disease – Sympathomimetics	0.0068
Other Gastrointestinal Disorders – Sympathomimetics	0.0065
Other Endocrine/Metabolic/Nutritional Disorders – Sympathomimetics	0.0062
History of Disease – Sympathomimetics	0.0041
Other Dermatological Disorders – Sympathomimetics	0.0040
Other Infectious Diseases – Sympathomimetics	0.0039
… 1,549 total combinations	

Limestone phenotypes are more succinct

LIMESTONE: Example Phenotypes

Hyperlipidemia

Phenotype 1 (41.6% of patients)
Other Endocrine, Metabolic, and Nutritional Disorders
HMG CoA Reductase Inhibitors
Intestinal Cholesterol Absortpion Inhibitors
Fibric Acid Derivatives
Antihyperlipidemics - Combinations
Nicotinic Acid Derivatives
Bile Acid Sequestrants
Oil Soluble Vitamins

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

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

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

LIMESTONE: Disease Subtypes

Mild Hypertension

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

Moderate Hypertension

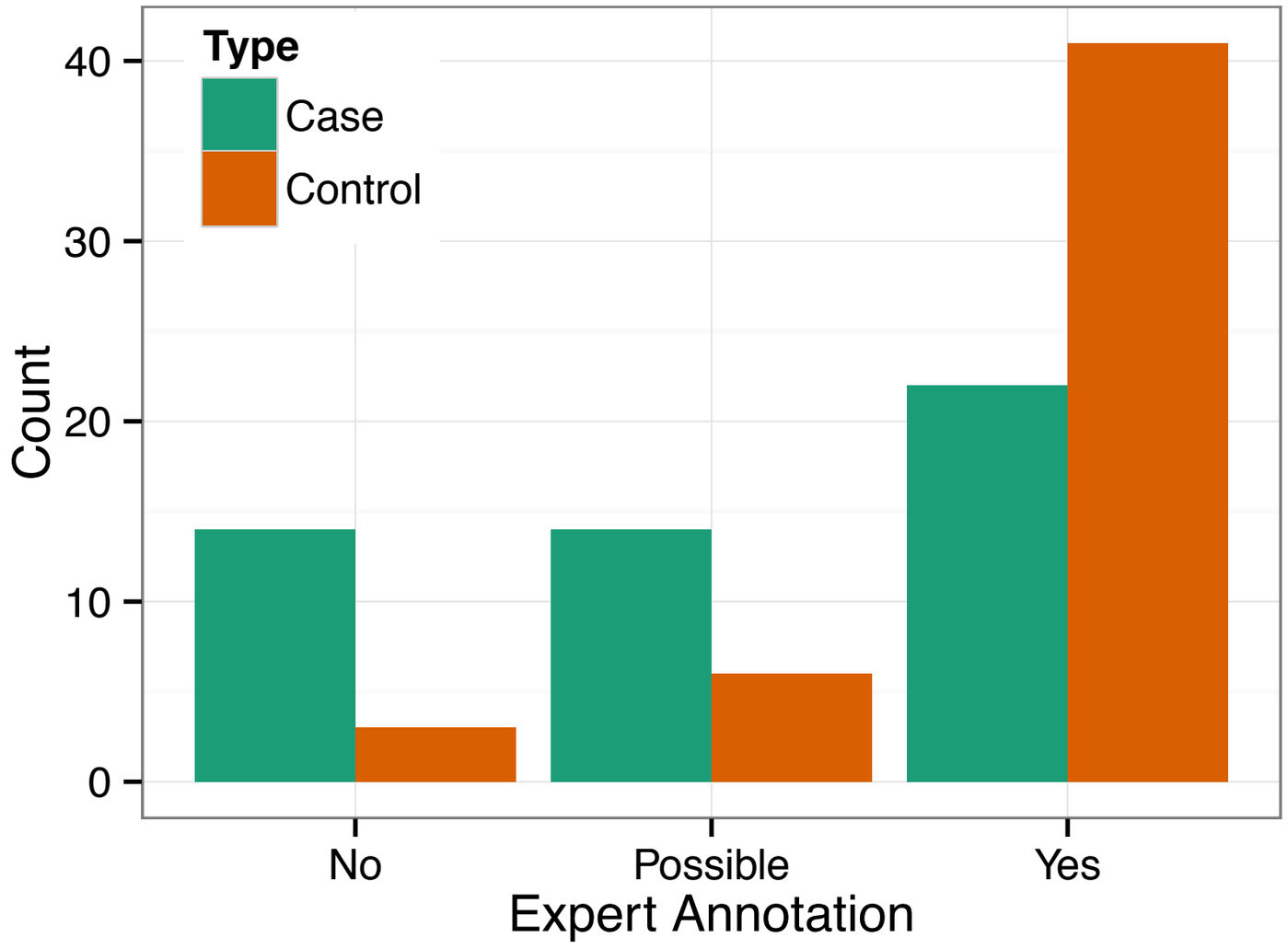
Phenotype 2 (31.5% of patients)
Hypertension
Beta Blockers Cardio-Selective
Angiotensin II Receptor Antagonists
Loop Diuretics
Potassium
Nitrates
Alpha-Beta Blockers
Vasodilators

Severe Hypertension

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

Limestone has the capability to discover disease subtype automatically from EHR!

LIMESTONE: Clinical Relevance



Domain expert confirmed 82% as clinically meaningful

LIMESTONE: Classification Task

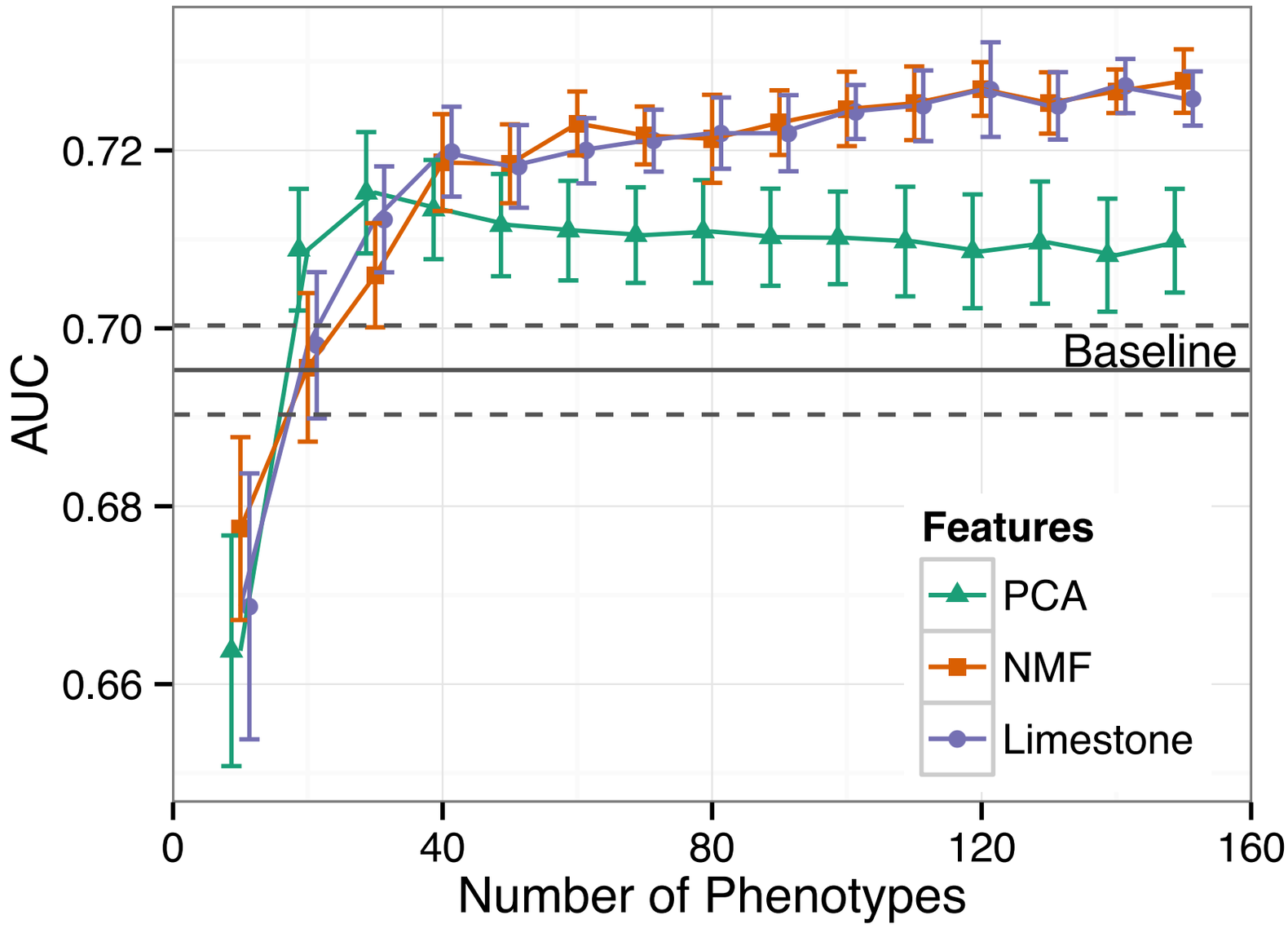
- Predict patients with heart failure
- Evaluate on four feature sets

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

Using patient factor matrix (loadings on phenotypes)

- Logistic regression model

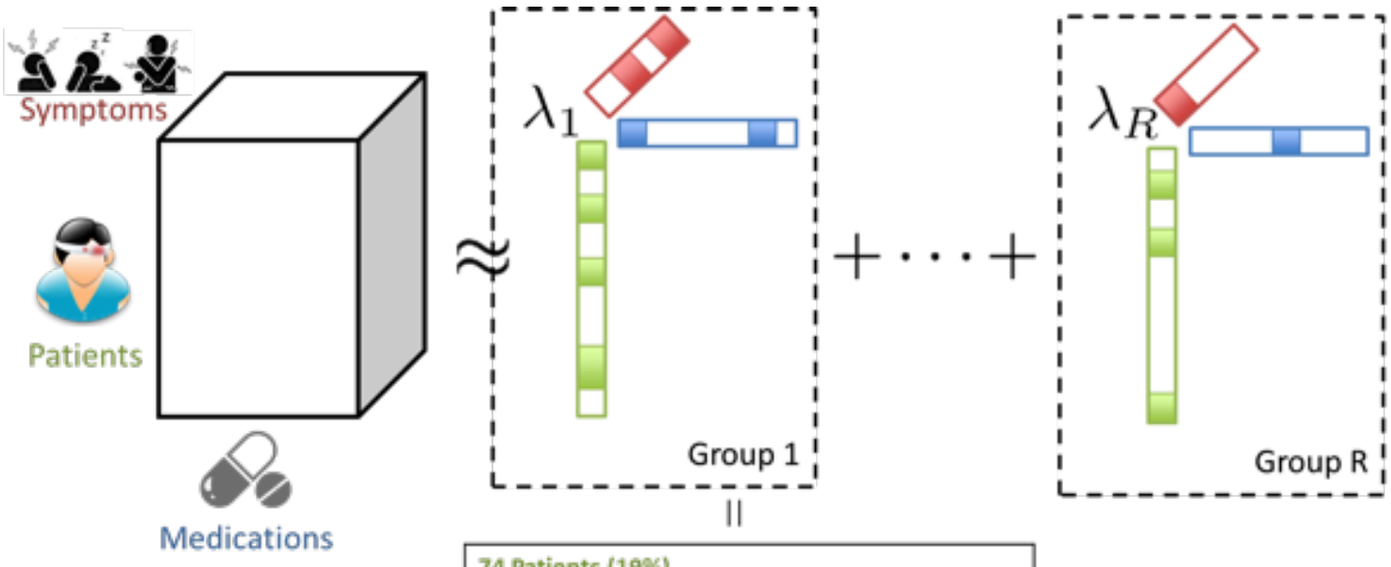
LIMESTONE: Predictive Power



40 phenotypes outperforms original 640 features!

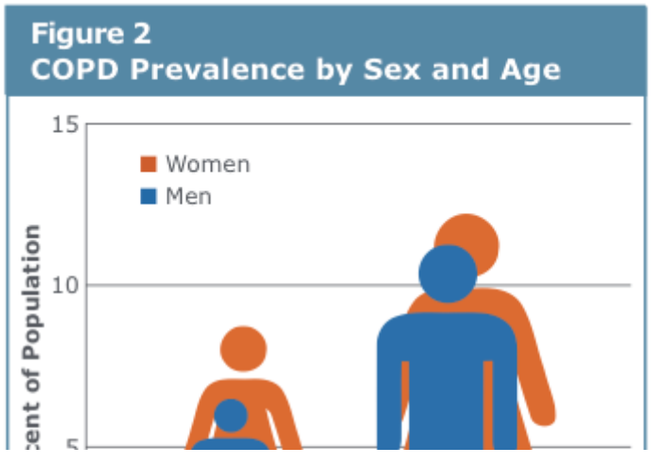
LIMESTONE: Chronic Fatigue Syndrome

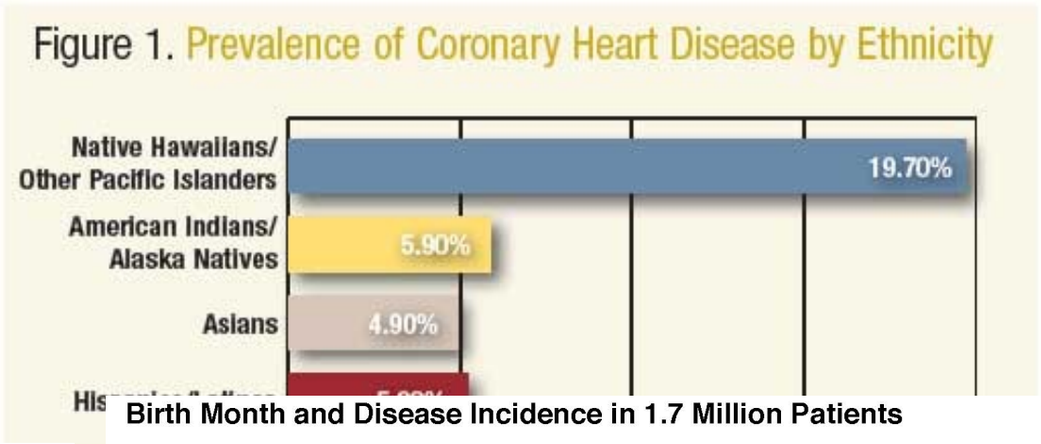
- Complex, devastating illness
- Managing symptoms is complicated
- Uncover common medication patterns with similar symptom severities

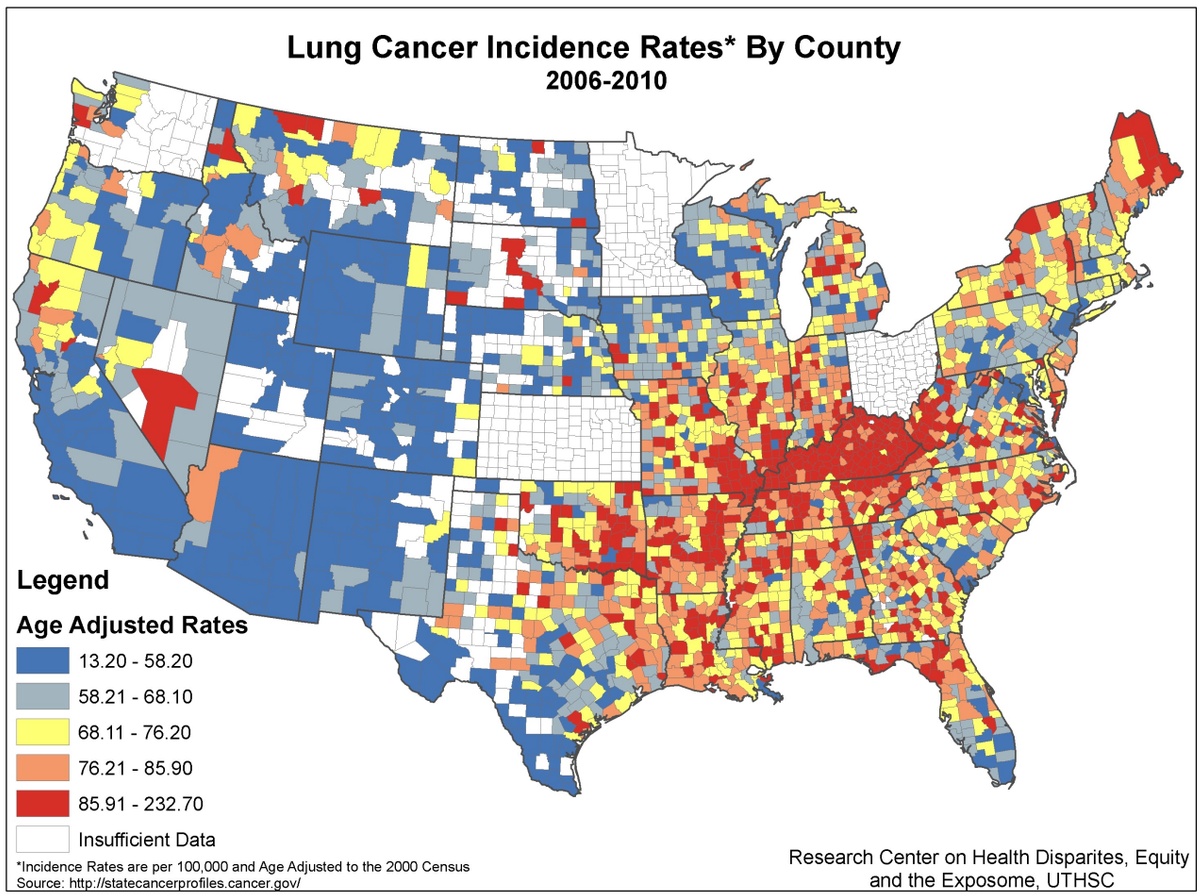


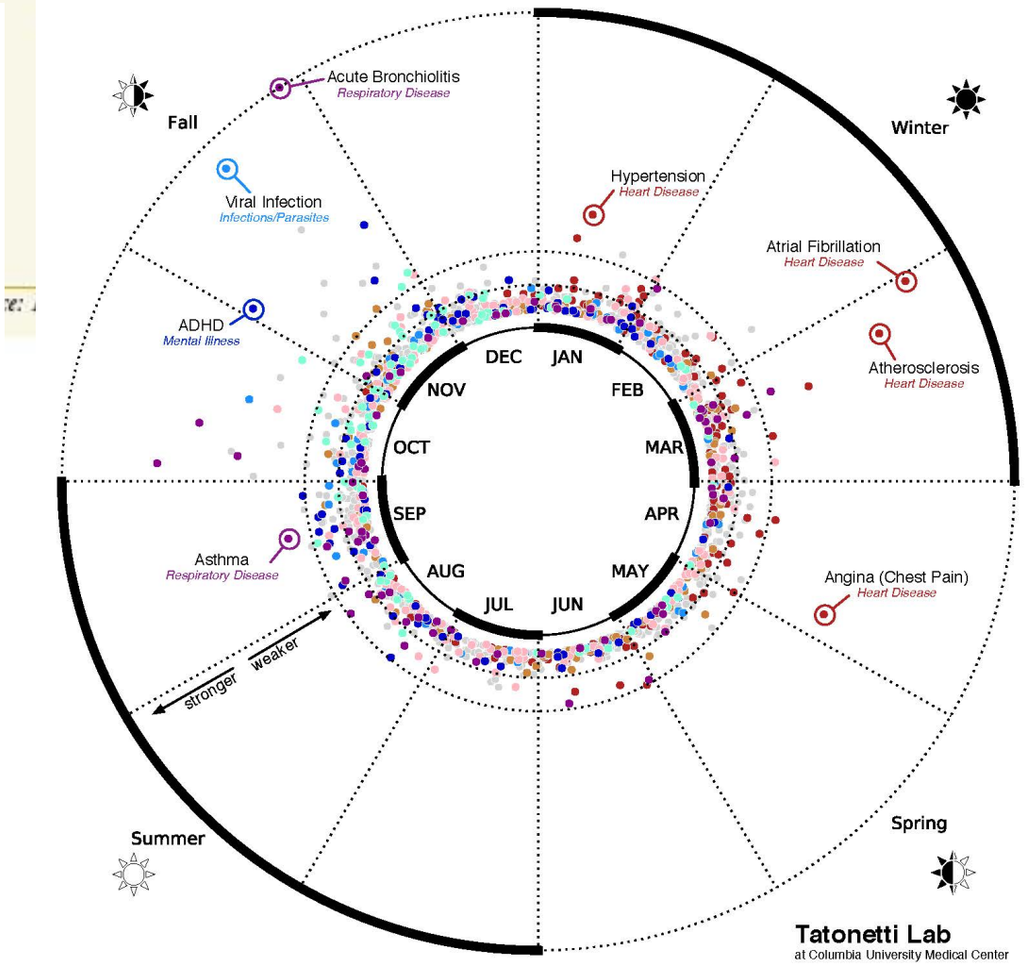
74 Patients (19%)
Analgesics
Anticonvulsants
Anxiolytics, sedatives, and hypnotics
Unusual Fatigue (13.6)
Unrefreshing Sleep (13.1)
Muscle Pain (11.5)
Joint Pain (9.5)
+ 9 medications and 8 symptoms

Splitting Patients into Subgroups





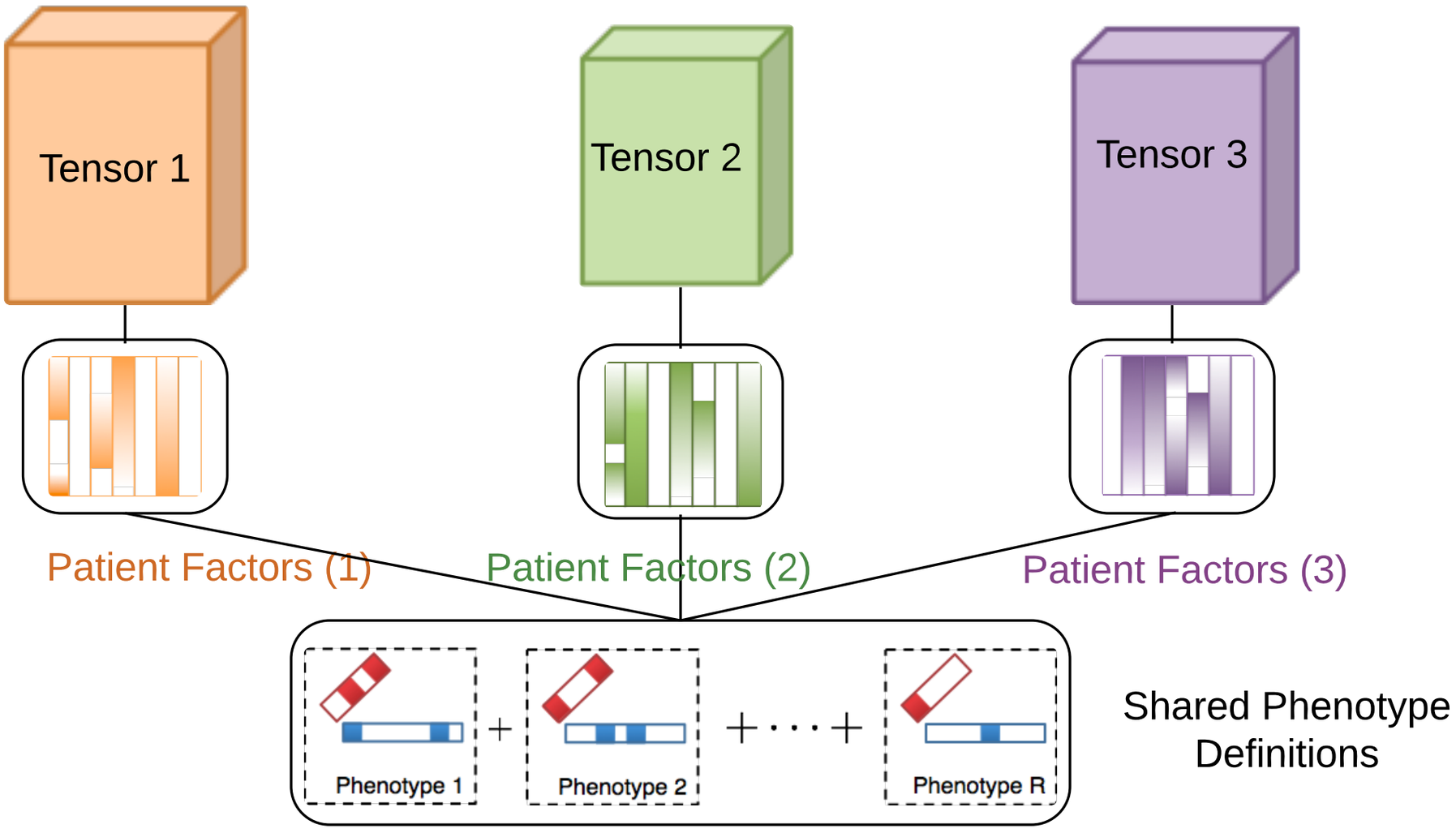




SANDSTONE: Multi-Task Tensor

- 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



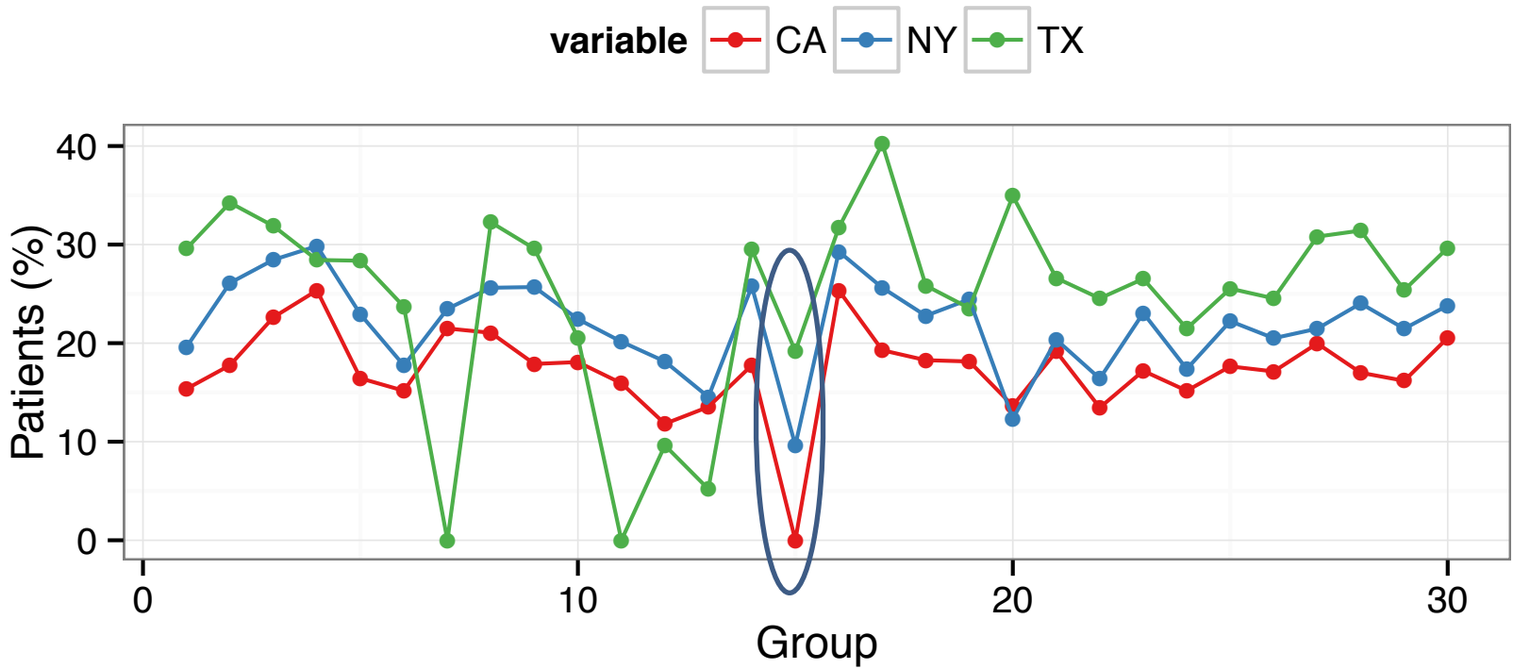
some phenotypes will be absent for certain tensors

SANDSTONE: 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: Unique Phenotypes

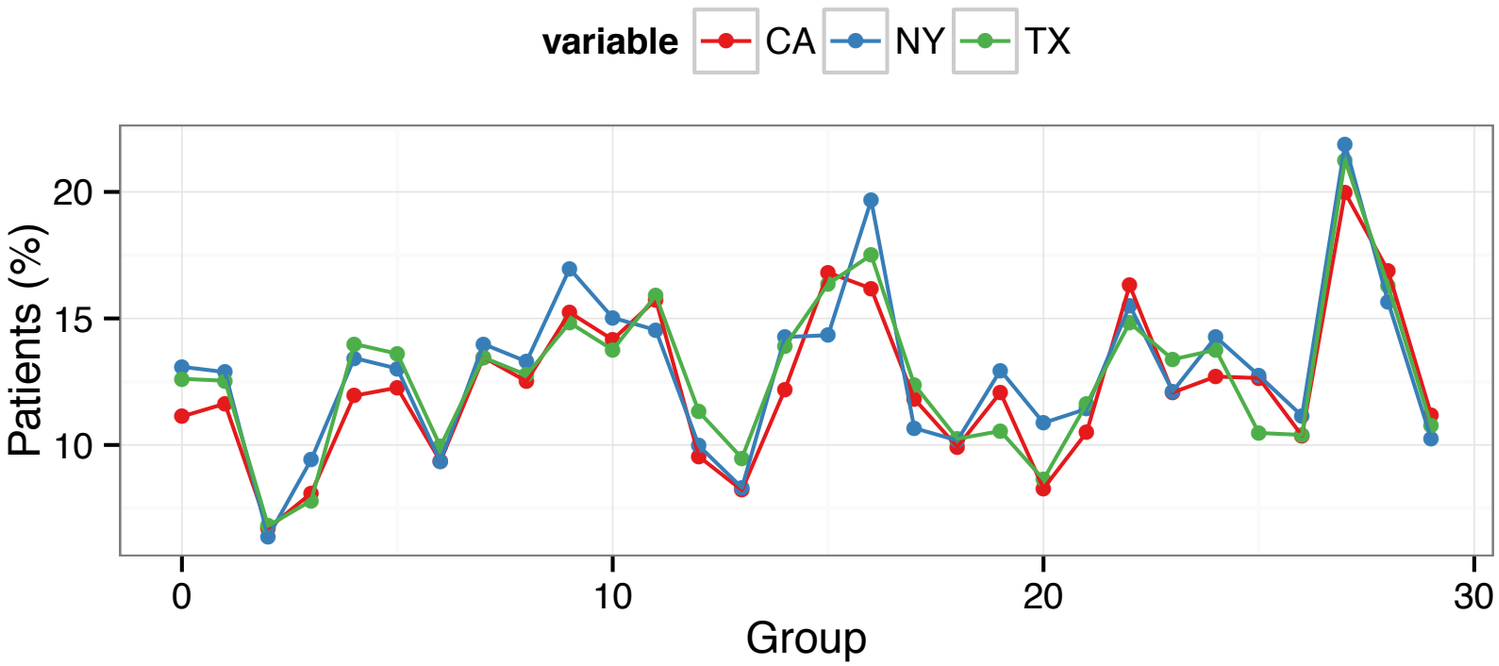
Sandstone



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

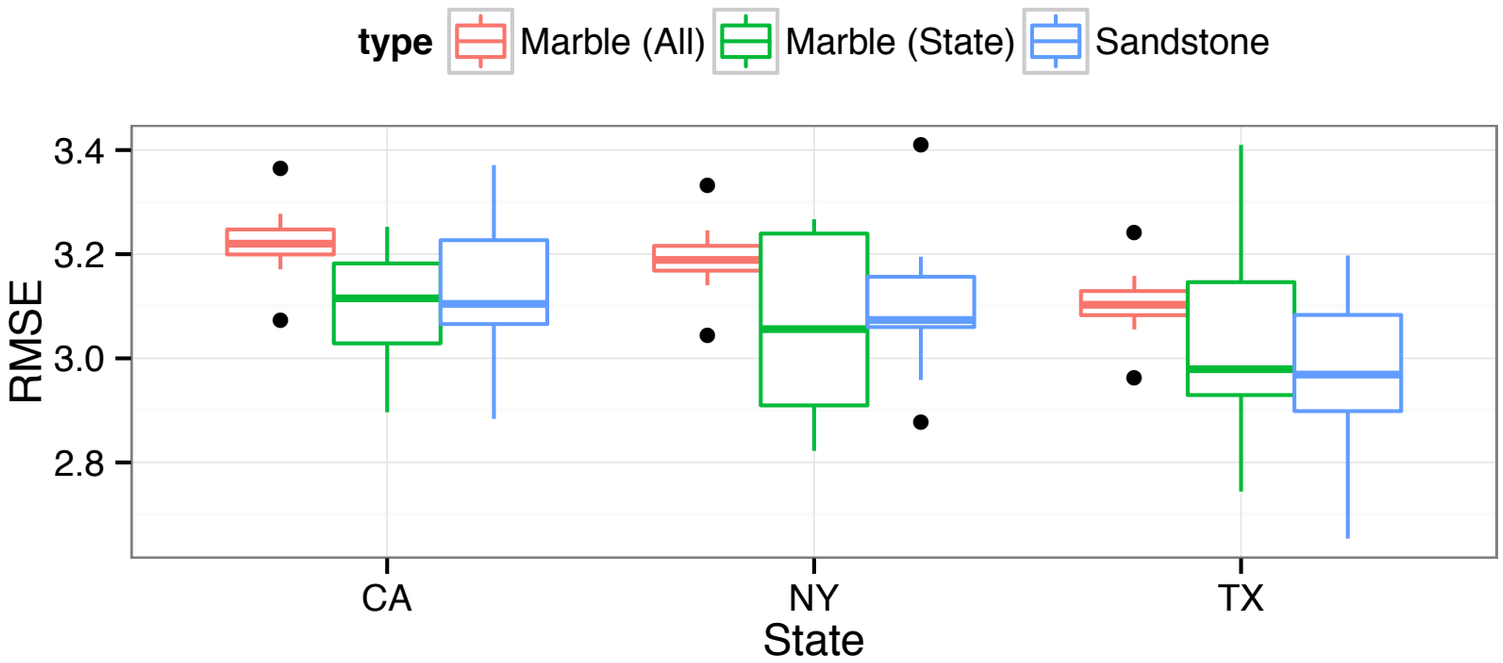
Not present in CA patients

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 patient loadings on phenotypes
- Linear regression model is trained on each feature set



Summary

- Tensor factorization can be used 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

Collaborators







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

Thank you!

joyce.c.ho@emory.edu

http://joyceho.github.io