Computational Phenotyping Methods

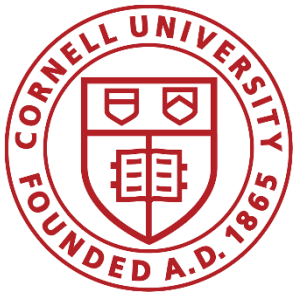
Fei Wang, PhD[1], Jimeng Sun, PhD[2], Xiaoqian Jiang, PhD[3], Yuan Luo, PhD[4]

[1]Cornell University, New York, NY;

[2]Georgia Institute of Technology, Atlanta, GA;

[3]University of California at San Diego, La Jolla, CA;

[4]Northwestern University, Chicago, IL



Tutorial Roadmap

- Supervised learning in computational phenotyping (Fei Wang)
- Unsupervised learning in computational phenotyping (Jimeng Sun)
- Computational phenotyping with unstructured data (Yuan Luo)
- Privacy in computational phenotyping (Xiaoqian Jiang)

Weill Cornell Medicine

Computational Phenotyping

Supervised Methods

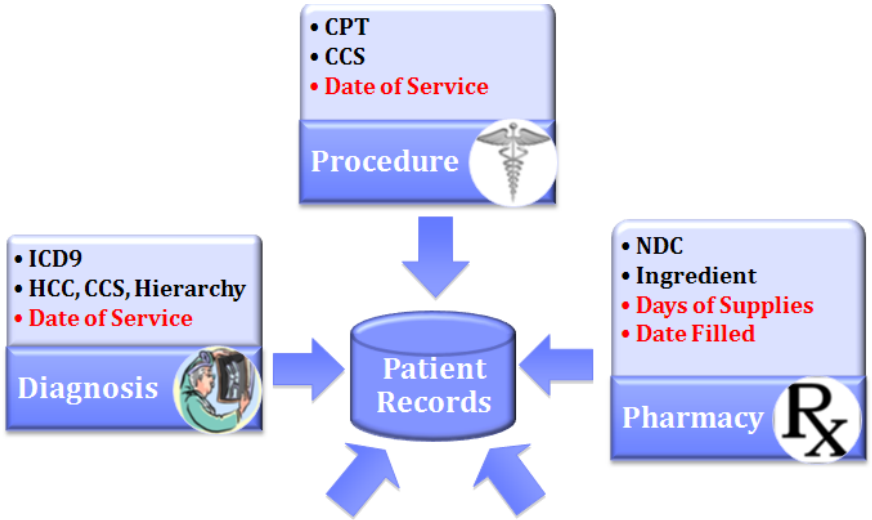
Tutorial in AMIA 2016

Fei Wang

Agenda

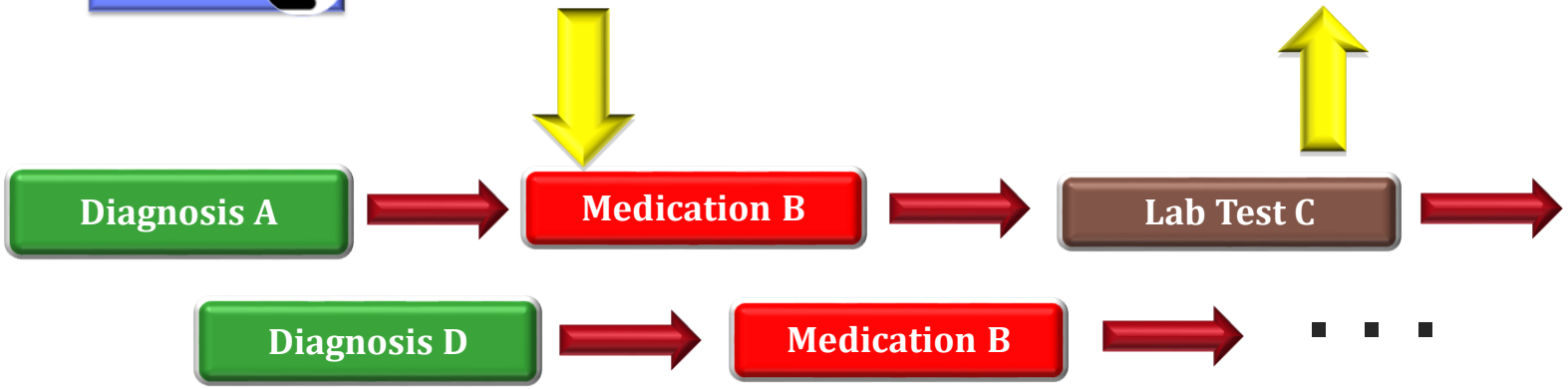
- Sequential Pattern Mining
- Deep Learning
- Anchor Based Methods

Sequentiality of Medical Events



How to interpret and make use of the sequentiality of the events?

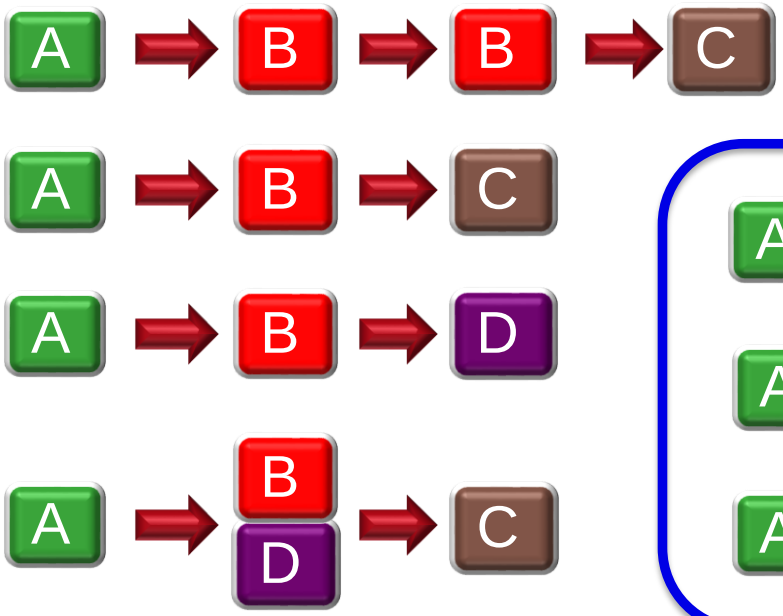
The sequentiality of those events may indicate some impending disease conditions



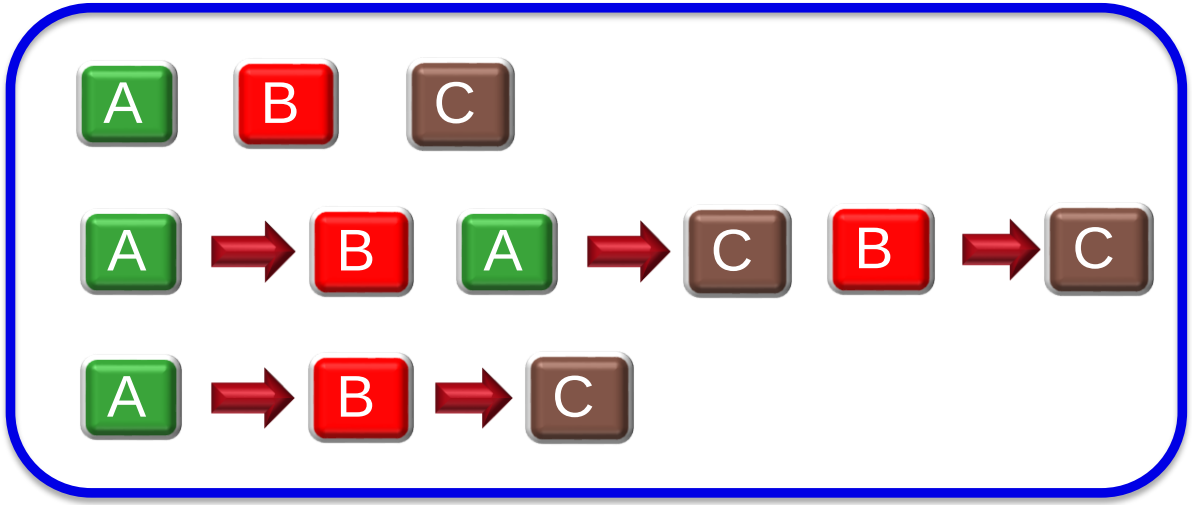
Sequential Pattern Mining

Given a set of sequences and *support* threshold, find the complete set of *frequent* subsequences

Event sequences



Frequent patterns with support 0.6



J. Han et al. Frequent pattern mining: current status and future directions. DMKD. 2007.

Pattern Explosion

Patient_ID	Day_ID	Event
2765239	74253	ALBUMIN
2765239	74253	ALKALINE PHOSPHATASE
2765239	74253	ALT
2765239	74253	AST
2765239	74253	BUN
2765239	74253	BUN/CREATININE RATIO
2765239	74253	CREAT
2765239	74253	GLUCOSE
2765239	74253	PROTEIN
2765239	74253	TSH

Study cohort: CHF patients
Medical events: Diagnosis+Lab
Size: 1032
Support threshold: 0.95
 # Patterns: >4200

Collapsing Concurrent Events

Collecting all concurrent event sets, detecting frequently co-occurred event subsets from them with pre-defined support threshold

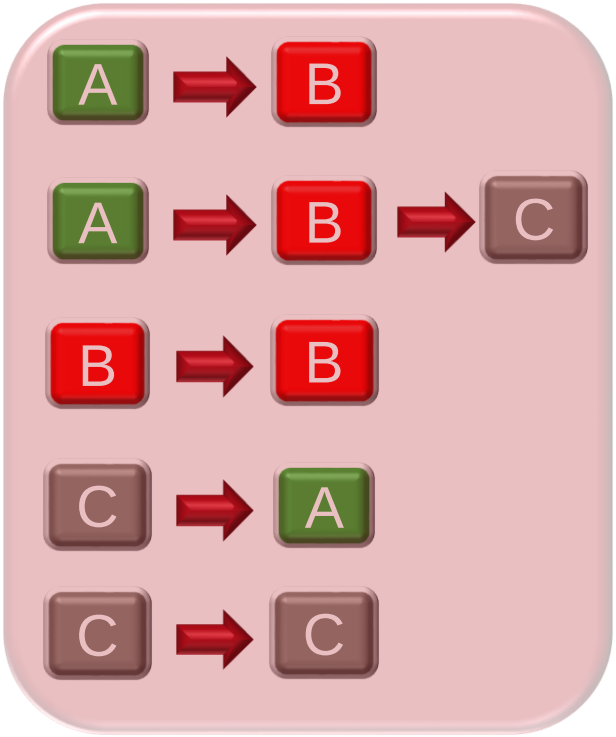
Event Package	Events
Metabolic Panel	BUN
	Creatinine
	GFR estimated
	Glucose
	Potassium
	Sodium
Hepatic Panel	HCT
	HGB

Event Package	Events
Diabetes Related Procedures	SUP-BLOOD GLUCOSE TST STRIP,50
	SPRING-PWRD DEV FOR LANCET,EACH
	SUP-LANCETS,PER BOX
Diabetes Related Diagnosis	DIABETES MELLITUS
	DISORDERS OF LIPOID METABOLISM

Geetika Lakshmanan, Szabolcs Rozsnyai, ***Fei Wang****.* Investigating Clinical Care Pathways Correlated With Outcomes. BPM 2013.

Bag-of-Pattern Representation

Pattern Dictionary

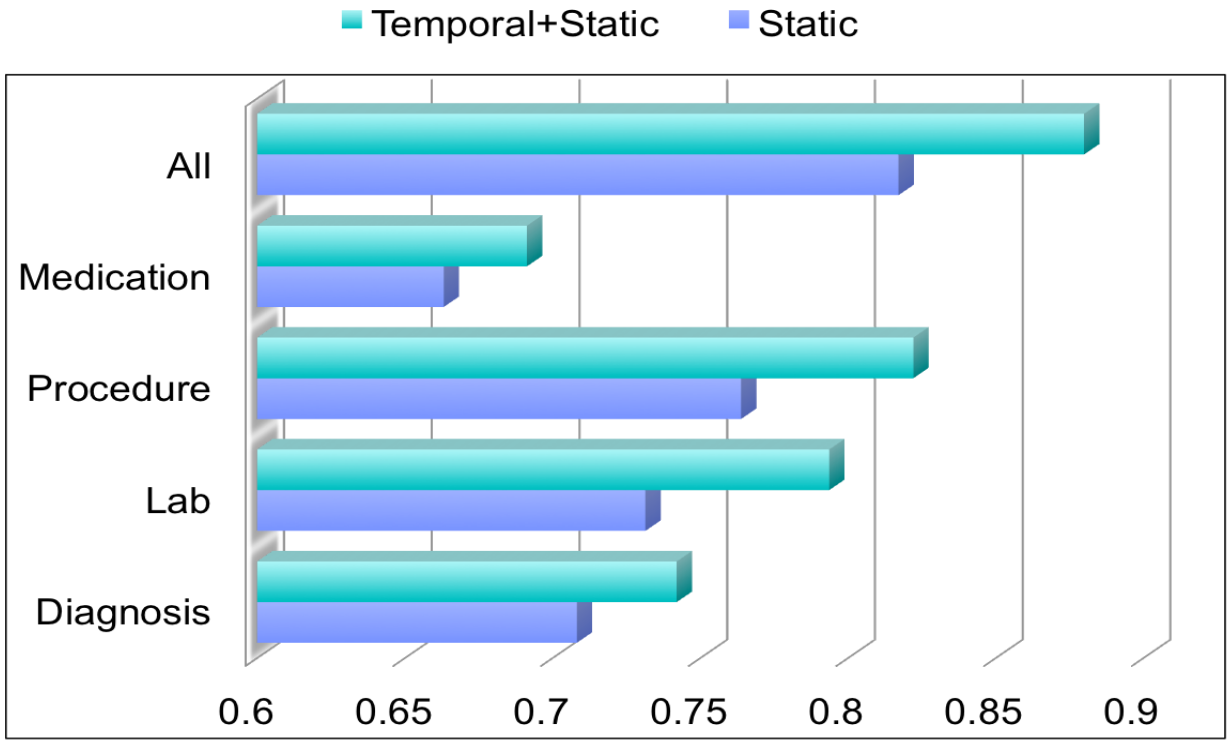


Let n be the size of the pattern dictionary, then each event sequence is represented as an n dimensional vector, with the i-th element indicating the number of times of the i-th pattern appeared in the sequence

Very similar to the bag-of-words representation of documents

Maximum interval between pairwise consecutive events: 30 days
Pattern duration: within 90 days

Prediction of the CHF Onset Risk



Case: 1,127
Control: 3,850

Prediction Window: 180 days
Observation Window: 360 days

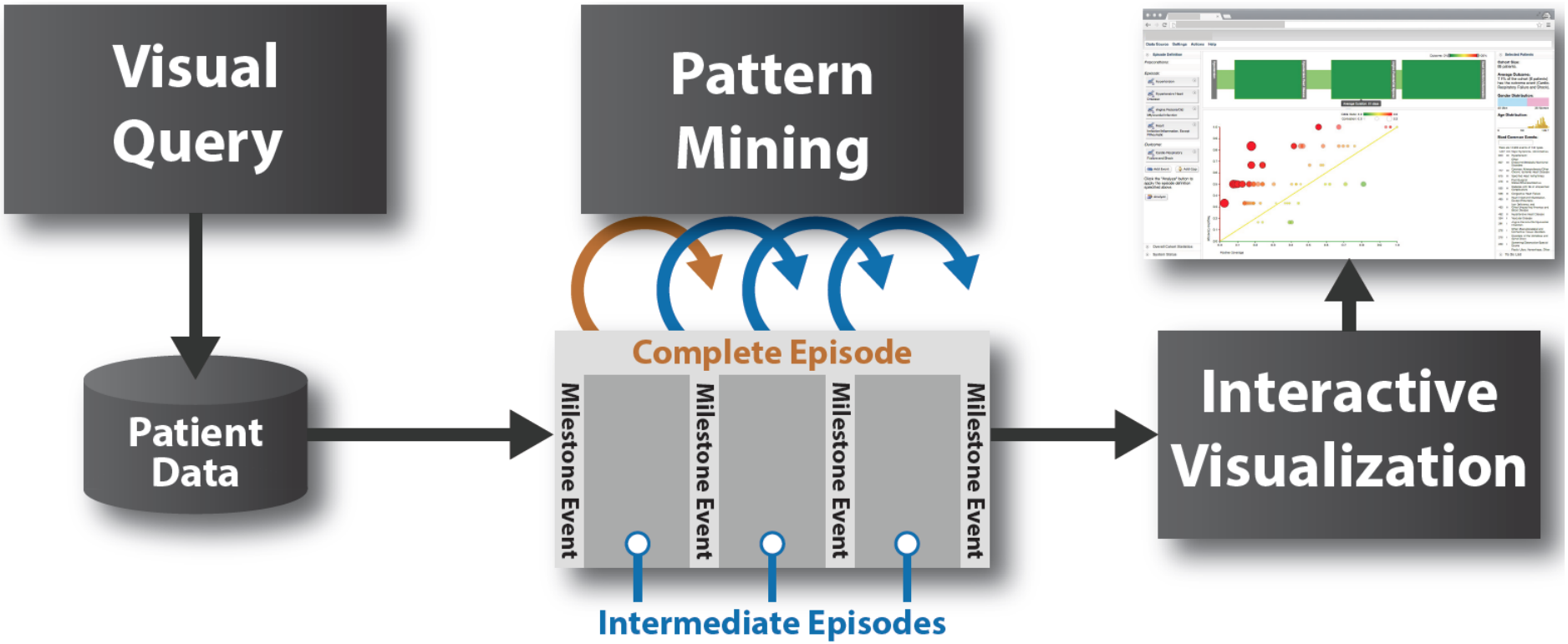
Logistic Regression

Example Patterns

COLLECTION OF VENOUS BLOOD BY VENIPUNCTURE-->Beta Blockers
DOPPLER COLOR FLOW (+)-->OV EST PT,DETAILED MODER COMPLEX
Diuretics-->DIABETES MELLITUS-->OV.EXP.PROB FOCSD,LOW COMPLX

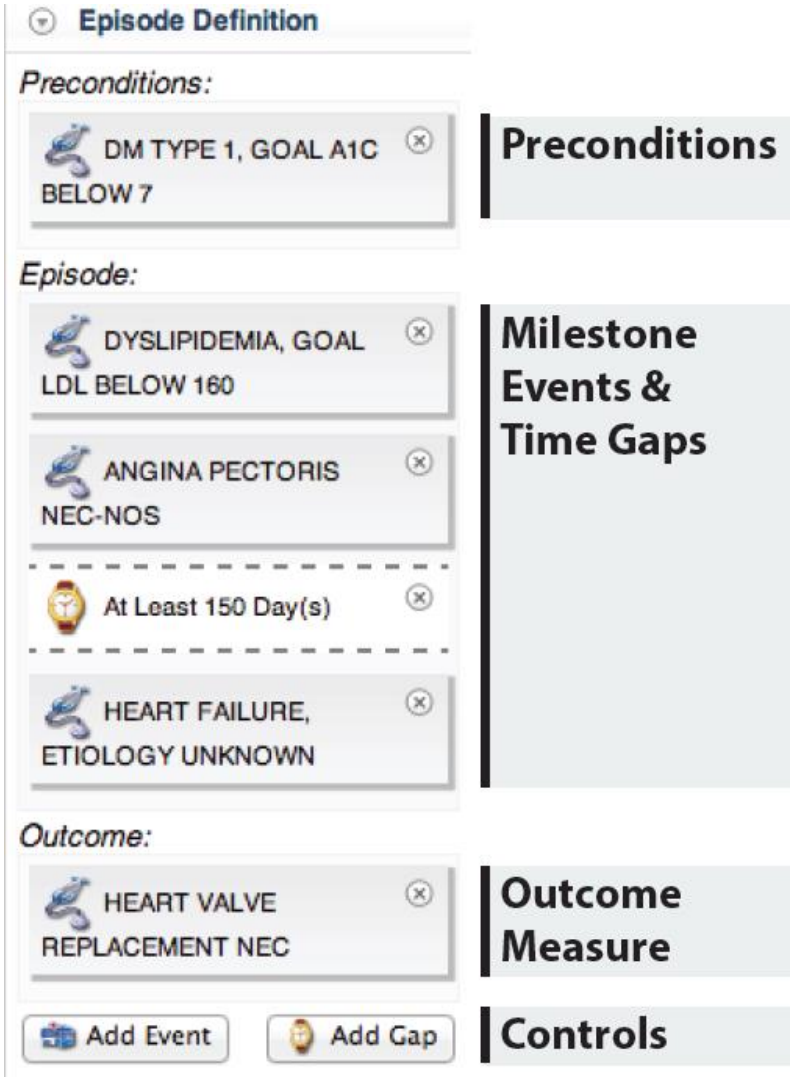
Fei Wang, Xiang Wang, Jianying Hu, Robert Sorrentino. Predictive Modeling with Data-Driven Temporal Clinical Event Patterns Discovery from Longitudinal EMR: A Case Study on Heart Failure Patients *Journal of Biomedical Informatics. Under Review. 2014.*

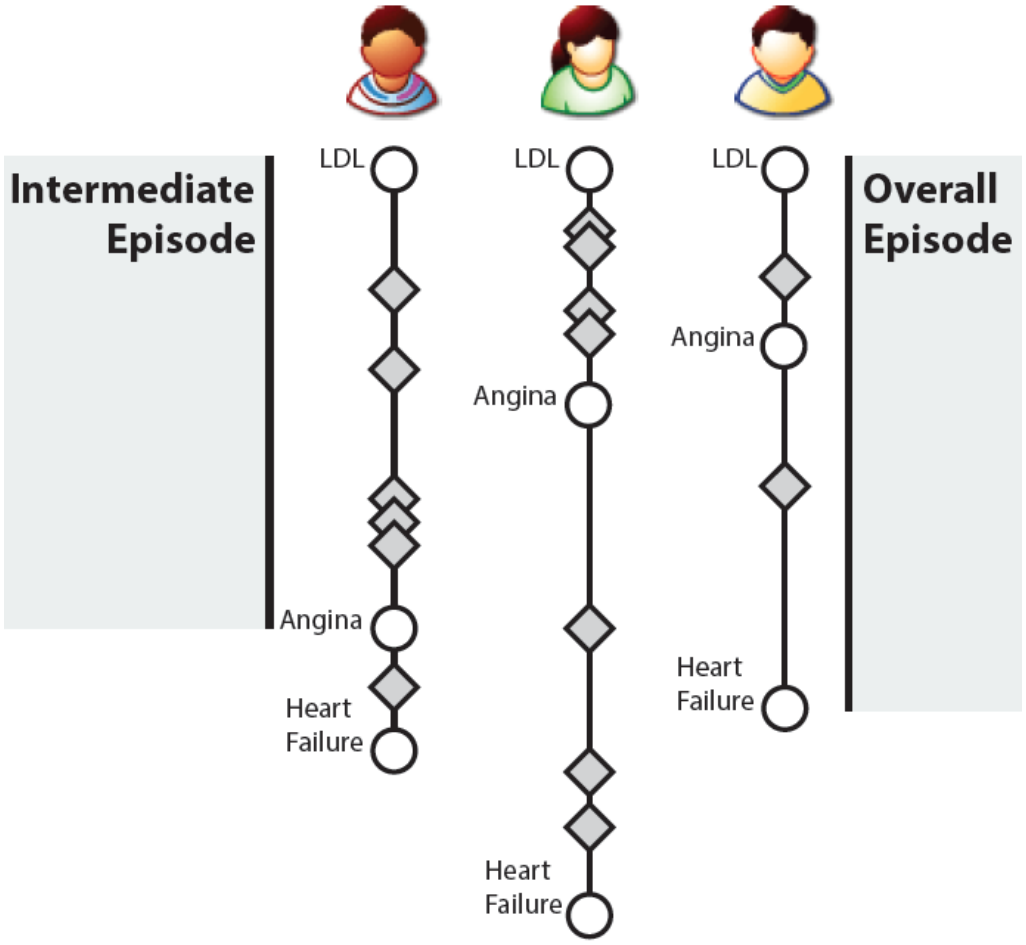
Interactive Visualization



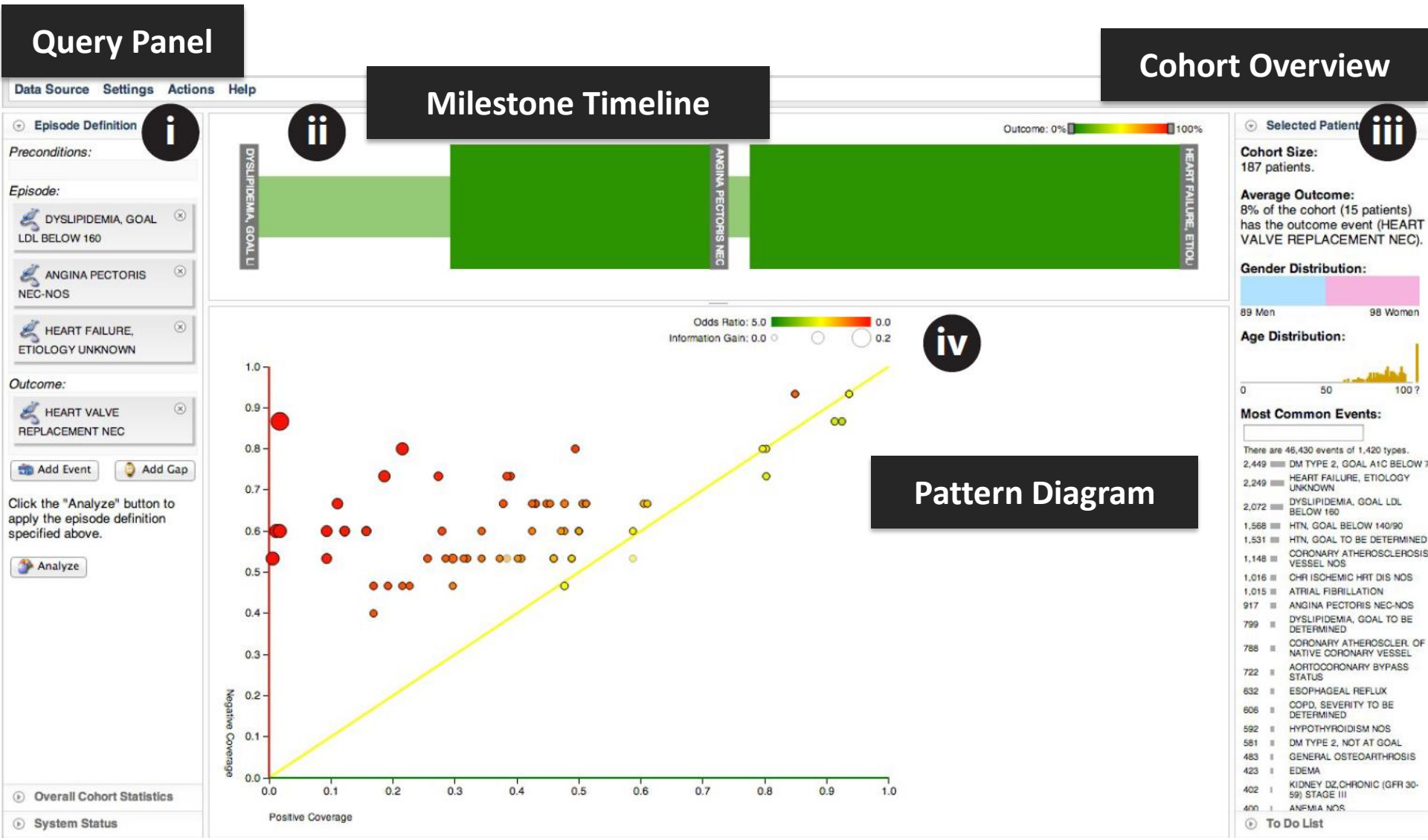
*David Gotz, **Fei Wang**, Adam Perer*. A Methodology for Interactive Mining and Visual Analysis of Clinical Event Patterns Using Electronic Health Record Data. *Journal of Biomedical Informatics. 2014.*

Visual Query

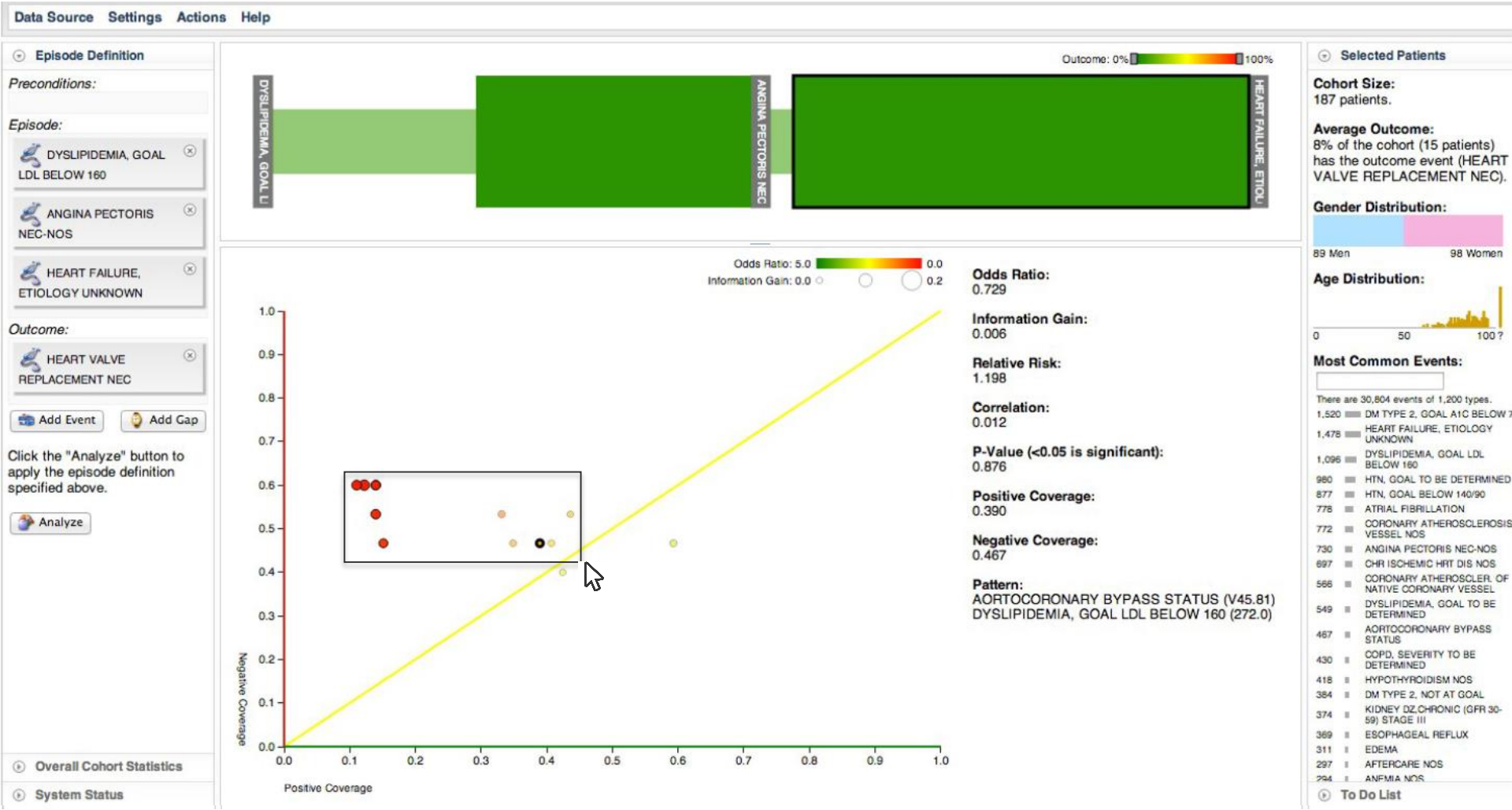




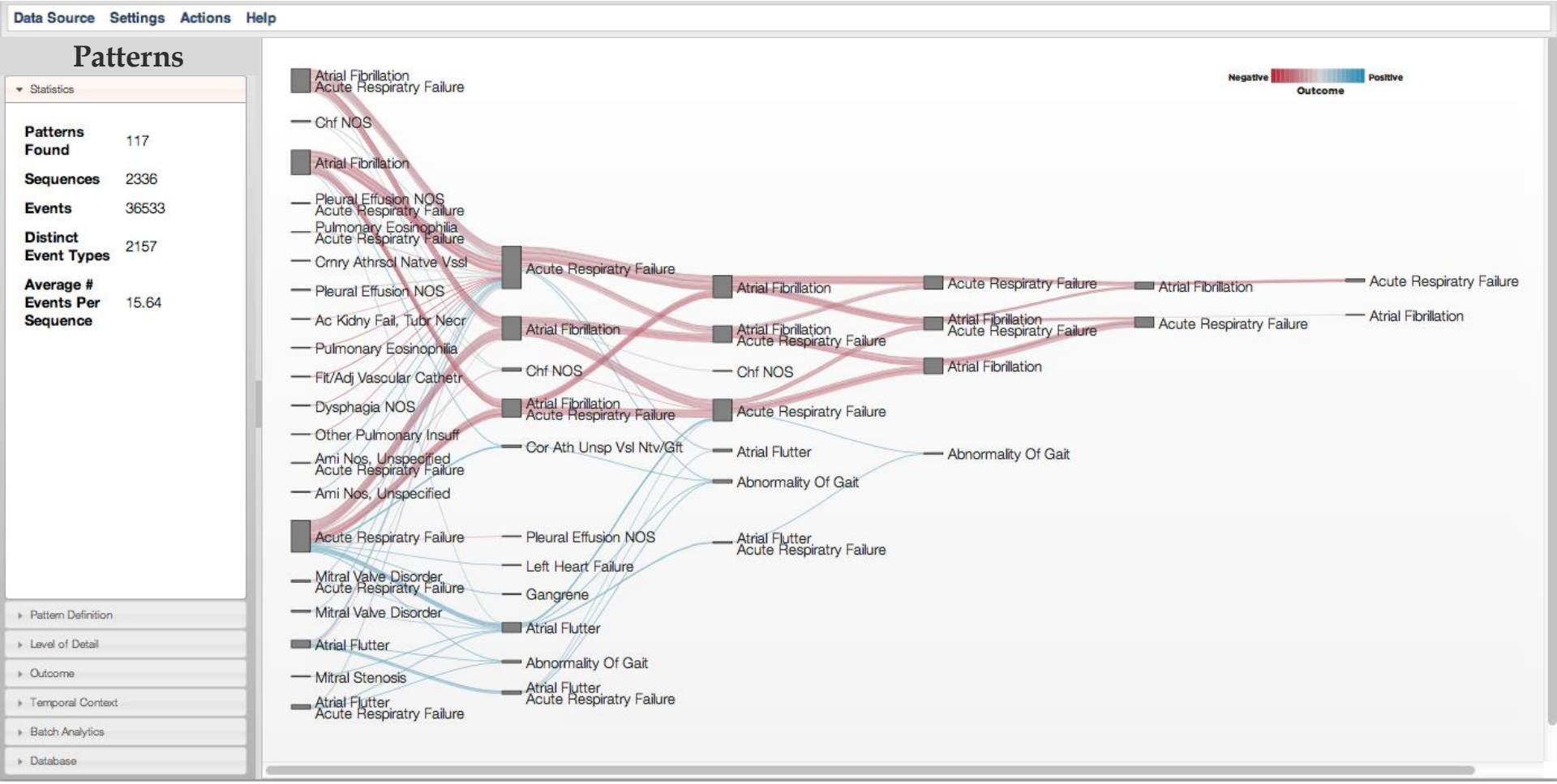
System Overview



System Overview



System Overview



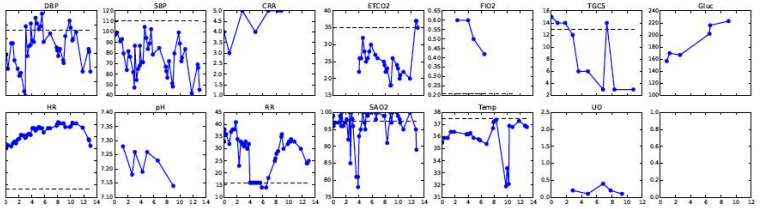
*Adam Perer, **Fei Wang**.* Frequence**:** Interactive Mining and Visualization of Temporal Frequent Event Sequences*. Proceedings of the 19th International Conference on Intelligent User Interfaces Proceedings (IUI) 2014.*

Agenda

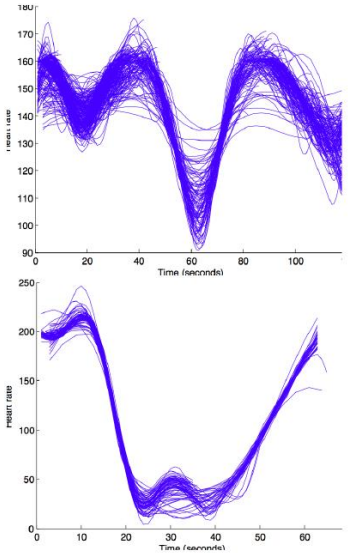
- Sequential Pattern Mining
- Deep Learning
- Anchor Based Methods

Computational phenotyping of critical illness

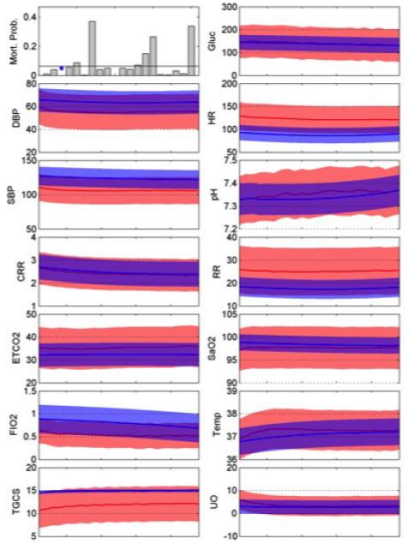
Setting: learning critical illness phenotypes from multivariate PICU time series.



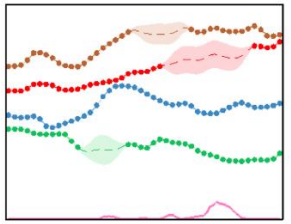
Deformable motifs [SDK11]



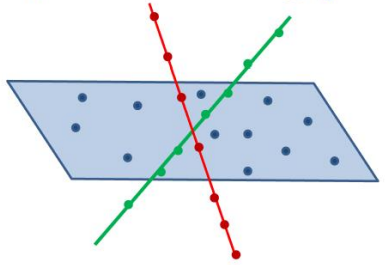
Bayesian clustering [MK12]



Multi-task GPs [GP15]



Subspace clustering [BK15]

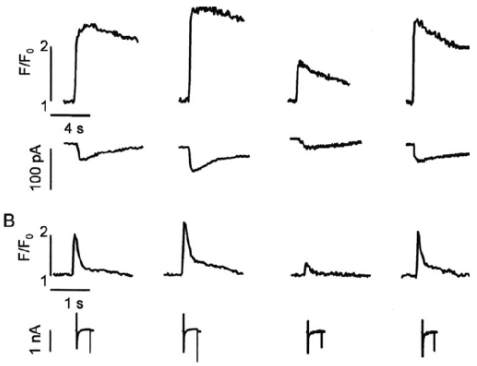


Kale, D.C., Che, Z., Bahadori, M.T., Li, W., Liu, Y. and Wetzel, R., 2015. Causal Phenotype Discovery via Deep Networks. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 677). American Medical Informatics Association.

Phenotyping as representation learning

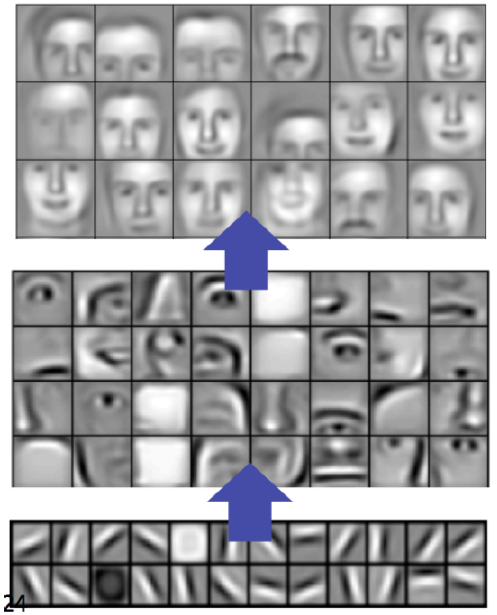
Medicine: *phenotypes, biomarkers* [BD01]

1. Measurable attributes of patient/disease.
2. Independent of other biomarkers.
3. Separate patients into meaningful groups.
4. Improve outcome prediction, risk assessment.
5. Clinically plausible, interpretable.



Machine learning: *features, representations* [BCV13]

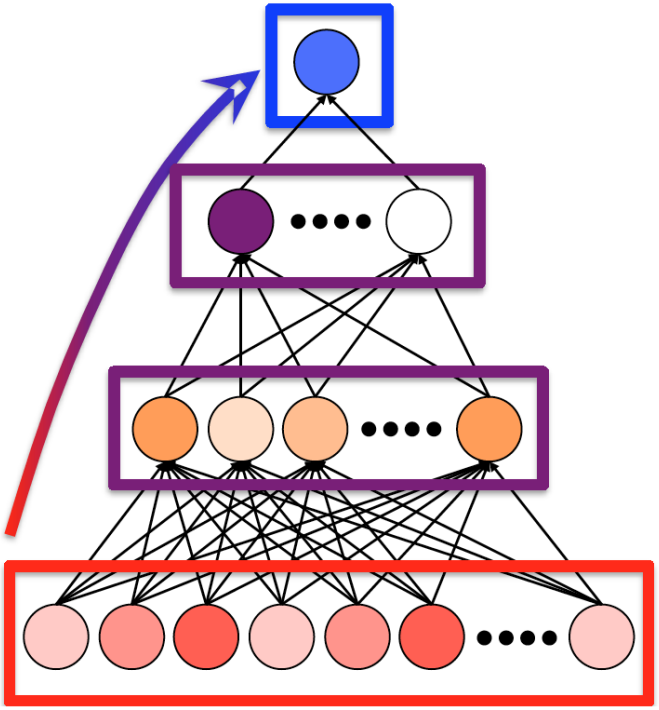
1. Measurable properties of objects.
2. Independent, disentangle factors of variation.
3. Form natural clusters.
4. Useful for discriminative, predictive tasks.
5. Interpretable, provide insight into problem.



Deep Learning of Representations

Representation learning: learn transformation of data useful for some task.

Main tool: *neural networks* (feedforward nets, ConvNets, RNNs, etc.)



Output: $\hat{y} = \mathrm{g}(\boldsymbol{h}_L \boldsymbol{W}_{\text{out}} + \boldsymbol{b}_{\text{out}})$

- sigmoid for binary classification
- softmax for multiclass classification
- identity for regression

Hidden: $\hat{h}_\ell = \mathrm{h}(\boldsymbol{h}_{\ell-1} \boldsymbol{W}_\ell + \boldsymbol{b}_\ell)$

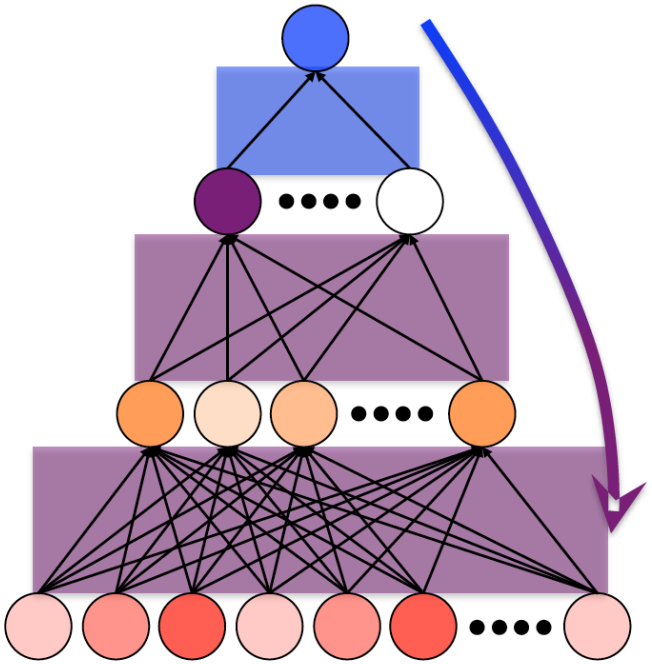
- sigmoid or $\tanh$ traditional
- rectified linear ($h(a) = \max(0, a)$) popular

Input: $\hat{h}_0 = \boldsymbol{x}$

Deep Learning of Representations

Representation learning: learn transformation of data useful for some task.

Main tool: *neural networks* (feedforward nets, ConvNets, RNNs, etc.)



Train using *gradient descent*.

Cost: $\mathcal{C}(y, \boldsymbol{x}; \{\boldsymbol{W}_\ell, \boldsymbol{b}_\ell\})$ (denote $\mathcal{C}$)

Update: $W_\ell(i,j) = W_\ell(i,j) - \alpha \frac{\partial \mathcal{C}}{\partial W_\ell(i,j)}$

Computing the gradients via backpropagation:

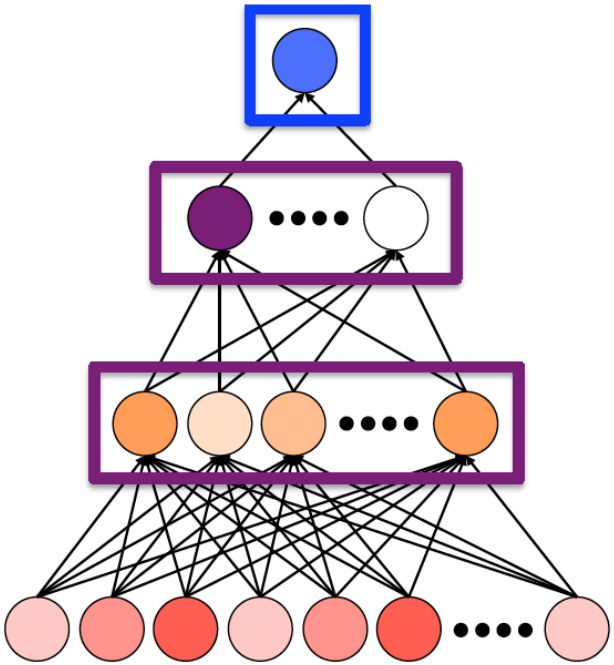
$$\frac{\partial \mathcal{C}}{\partial W_\ell(i,j)} = \frac{\partial \mathcal{C}}{\partial h_\ell(j)} \frac{\partial h_\ell(j)}{\partial a_\ell(j)} \frac{\partial a_\ell(j)}{\partial W_\ell(i,j)} \text{ where}$$

$$\frac{\partial h_\ell(j)}{\partial a_\ell(j)} = g'(a_\ell(j)) \qquad \frac{\partial a_\ell(j)}{\partial W_\ell(i,j)} = h_{\ell-1}(i)$$

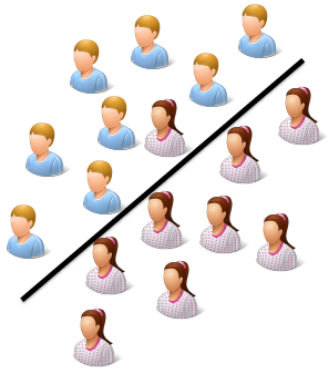
$$\frac{\partial \mathcal{C}}{\partial h_\ell(j)} = \sum_k W_{\ell+1}(j,k) \frac{\partial \mathcal{C}}{h_{\ell+1}(k)}$$

$$a_\ell(j) = \boldsymbol{h}_{\ell-1} \boldsymbol{W}_\ell(:,j) + b_j$$

Neural Networks Combine Different Views

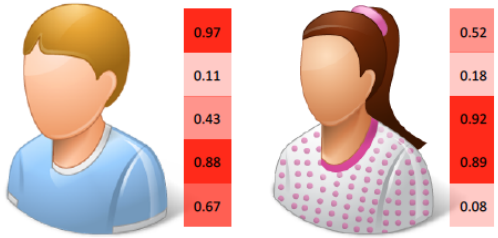


Output layer: classifier

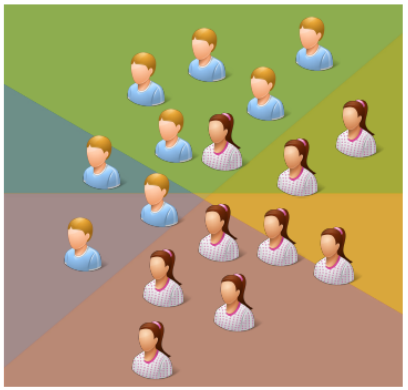


Hidden layers:

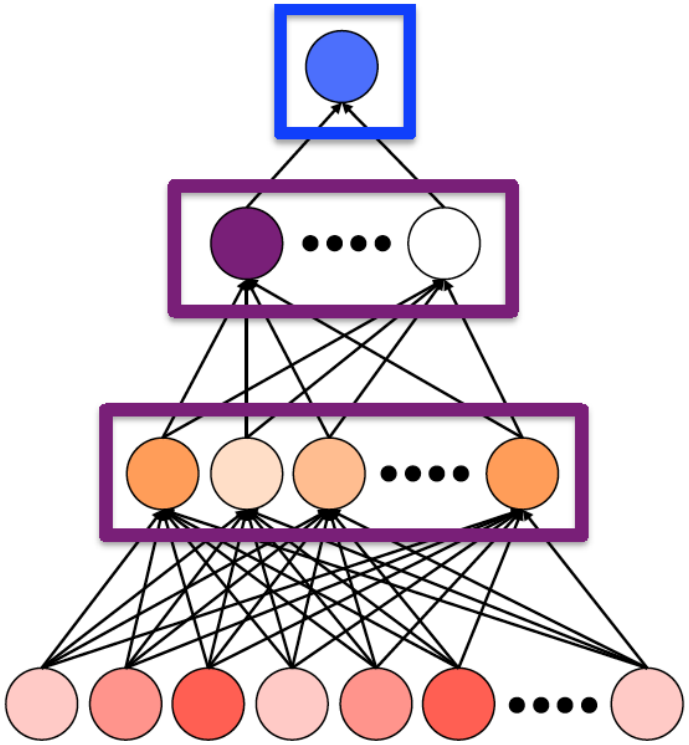
Latent factors/bases



Multiclustering [BC13]



Challenges of Interpretation

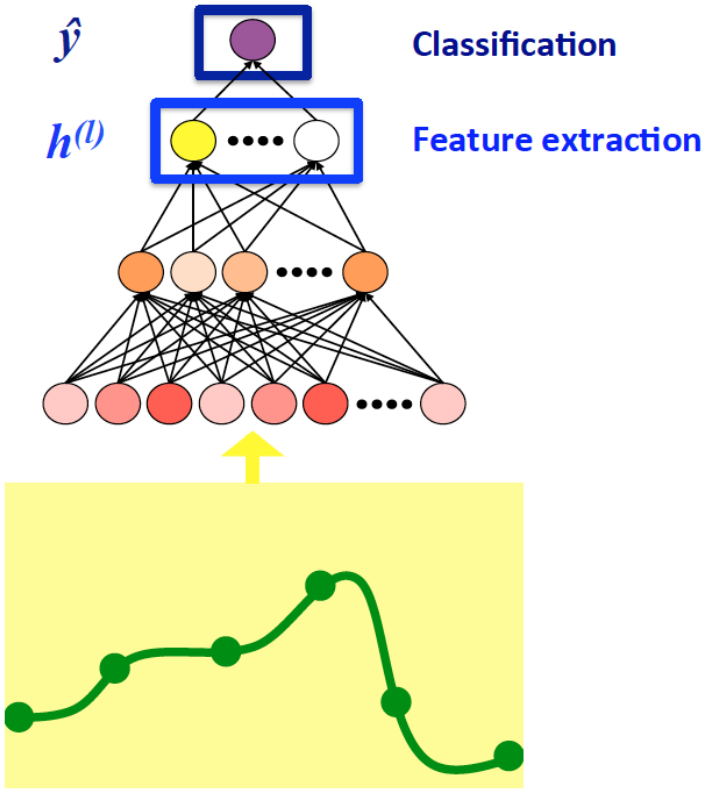


No predefined semantics
(vs. graphical model)

Learned bases not guaranteed to be uncorrelated or independent
(vs. PCA, ICA)

Information contained in distributed activations, so interpreting individual features unreliable [SZ14]

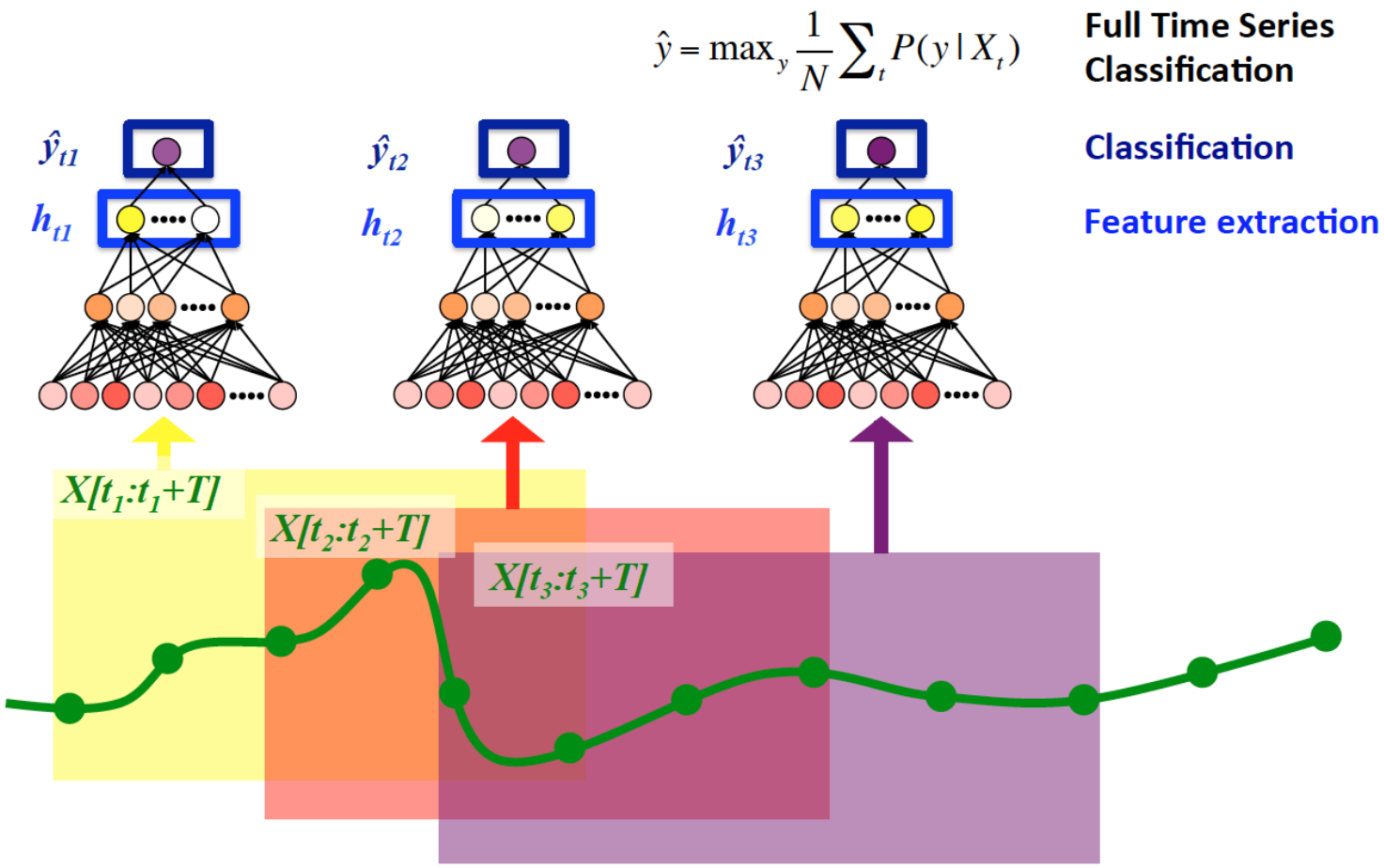
Deep Learning for Time Series: Window-based Approach



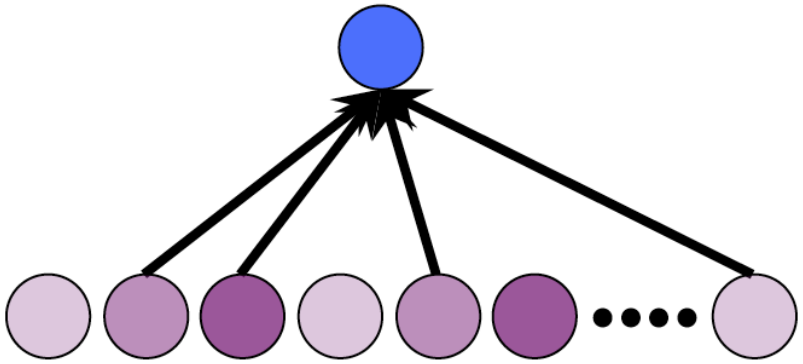
- Apply neural net (NNet) to fixed-size *windows* (*subsequences*).
- Classification, feature extraction.
- Correlations across variables, time.
- Relatively few, weak model assumptions.
- Can learn to detect smooth, trajectory-like patterns.

Deep Learning for Time Series: Window-based Approach

Can also be applied in sliding window fashion to longer time series.



Causal analysis of learned phenotypic features



- Now have set of D latent factors $\{h_i\}_{i=1}^{D}$, response y.
- Analyze of causal relationship between each factor, response.
- Choose causal direction of each edge: $h_i \rightarrow y$ or $h_i \leftarrow y$.
- Use only causal factors ($h_i \rightarrow y$) in further analysis.
- **Note:** for predictive tasks, use original network.

Causal analysis with pairwise likelihood ratios [HS13]

For two variables h and y, want to distinguish between two causal models:

$$\begin{aligned} h \to y \quad &: \quad y = \rho h + d \\ h \leftarrow y \quad &: \quad h = \rho y + e \end{aligned}$$

h, y are non-Gaussian. Noise d (e) is independent of x (y).

Model log-likelihood:

$$\log L(h \to y) = \log p_h(h) + \log p_d\left(\frac{y - \rho h}{\sqrt{1-\rho^2}}\right) - \log(1-\rho^2).$$

Sign of likelihood ratio determines direction of causal edge:

$$R = \log L(h \to y) - \log L(h \leftarrow y) \qquad \begin{cases} R > 0 & \text{if} \quad h \to y \\ R < 0 & \text{if} \quad h \leftarrow y \end{cases}$$

Important note: makes no statement about *strength* of edge. Use in combination with feature selection!

Experiment: Two Clinical Datasets

8500 multivariate time series from CHLA PICU (*PICU*) [7]:

- All > 24 hours long.
- Sampled once per hour (*after preprocessing**).
- 13 variables: vitals, labs, outputs, assessments.
- Phenotype labels: 67 groups of ICD-9 codes, 19 standard ICD-9 categories.

8000 multivariate time series from *PhysioNet Challenge 2012*† (*PC2012*):

- 48 hours long (not full episodes in all cases).
- Sampled once per hour (*after preprocessing**).
- 33 variables: vitals, labs, outputs, assessments.
- Label: in-hospital mortality

Experiment Setup

1. **Data preparation**
   - Generate 5-10 random training/validation/test splits of *episodes*.
   - Train on fixed-size windows of time series:
     - PC2012: full 48 hour time series.
     - PICU: 12 hour windows extracted in sliding window fashion.
2. **Model architecture, training details**
   - 3 hidden layers, fully connected, sigmoid activation.
   - Unsupervised pretraining with stochastic denoising autoencoders.
   - Supervised training (with early stopping) as multilayer perceptron.
3. **Evaluation**
   - Quantitative: area under ROC curve (AUROC), area under precision-recall curve (AUPRC), precision at 90% recall.
   - Qualitative: causal feature analysis + visualization.

First 48-hour mortality prediction (PC2012)

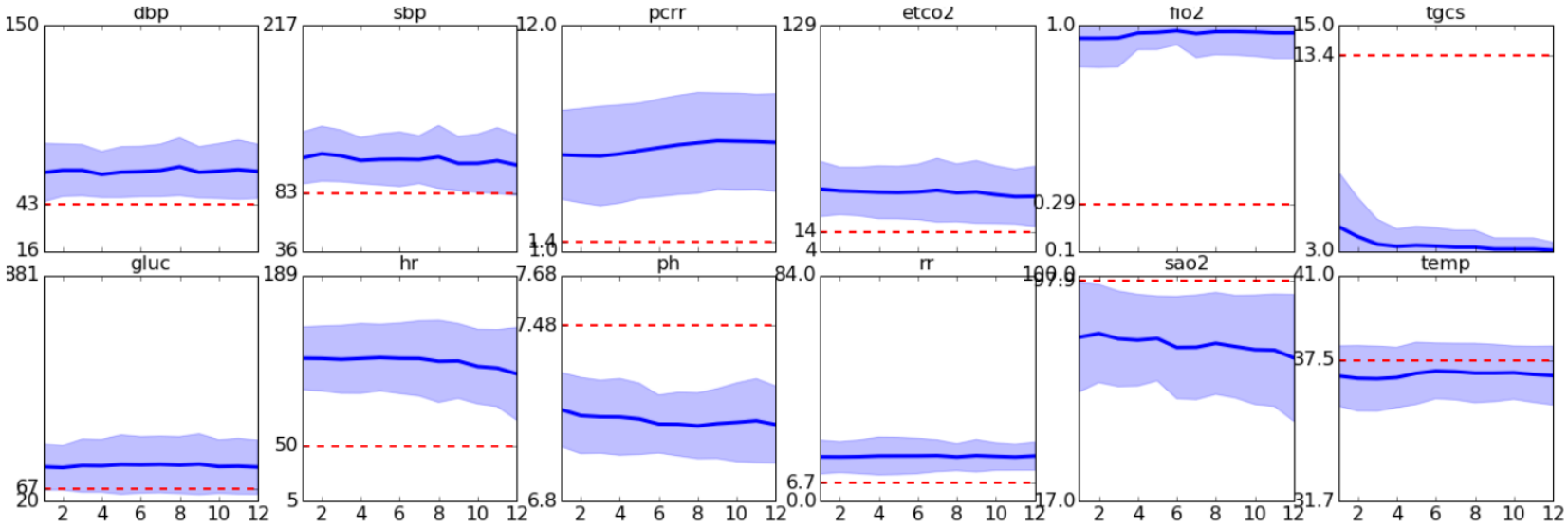
	AUROC	AUPRC	Prec@90%Rec
Raw (R)	0.787 ± 0.0290	0.407 ± 0.0429	0.221 ± 0.0171
HandDesigned (H)	0.829 ± 0.0211	**0.468 ± 0.0479**	0.259 ± 0.0494
NNet(R,3)	0.821 ± 0.0210	0.444 ± 0.0324	0.256 ± 0.0303
NNet(H,3)	**0.832 ± 0.0162**	0.462 ± 0.0480	**0.271 ± 0.0260**
H+R	0.823 ± 0.0183	0.438 ± 0.0354	0.256 ± 0.0319
H+NNet(R,3)	**0.845 ± 0.0165**	**0.487 ± 0.0473**	**0.291 ± 0.0335**

Mean performance with standard deviation (10 folds); *classifier*: linear SVM + L_1 penalty.

NeuralNet: Layer 3 hidden unit activations of neural net
(3 layer neural net, unsupervised + supervised training)
HandDesigned: extremes, central tendencies, variance, trends

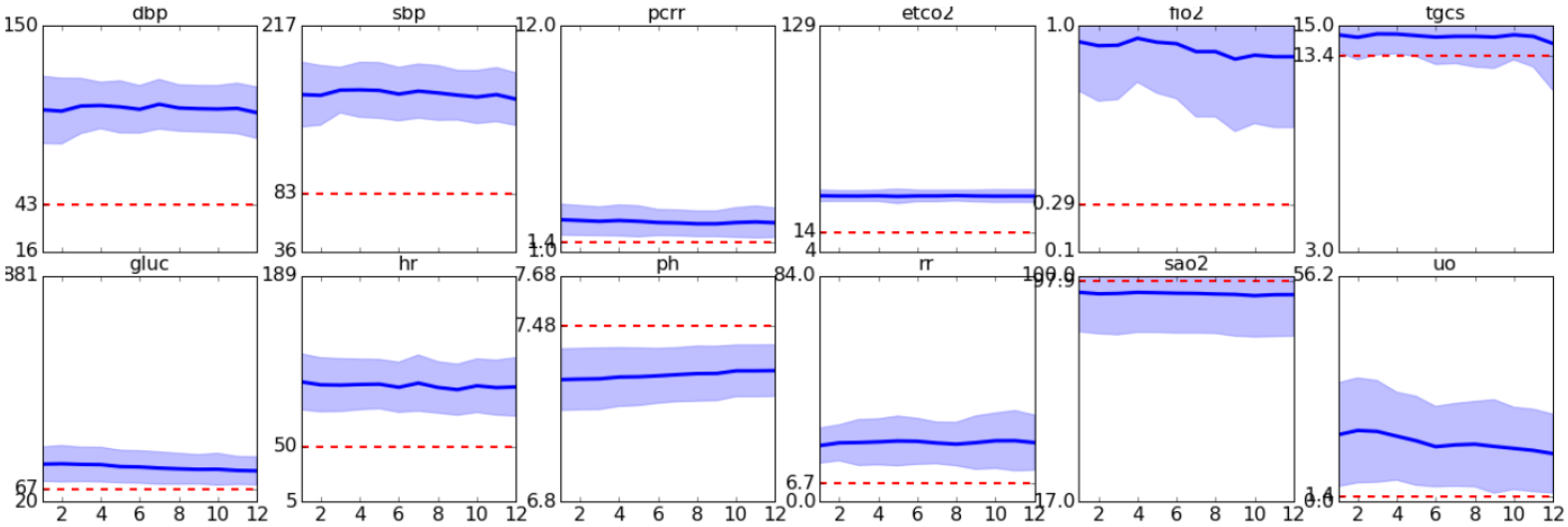
PICU classification results: Che, Kale, Li, Bahadori, and Liu, SIGKDD 2015 [CK15]

Phenotype for septic shock (ICD-9: 990-995)



- Very irregular physiology, known symptoms of sepsis.
- Low Glascow coma score indicates patient is unconscious.

Phenotype for circulatory disease (ICD-9: 390-459)



- Elevated blood pressure and heart rate, depressed pH.
- Evidence of ventilation (elevated FIO2).
- Note elevated urine output; also correlated with urinary disorders.

Contributions

- Use deep learning to discover and detect characteristic patterns in clinical time series data.
- Propose modifications to standard neural net training
  - Prior-based regularization
  - Incremental training for different length of input

Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep Computational Phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '15)

Methodology

Prior-based regularization

- Graph Laplacian-based regularization, applied to the output layer.

$$\mathcal{L} = -\sum_{i=1}^{N} \log p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\Theta}) + \lambda R(\boldsymbol{\Theta}) + \frac{\rho}{2}\mathrm{tr}(\boldsymbol{\beta}^\top \mathbf{L}\boldsymbol{\beta})$$

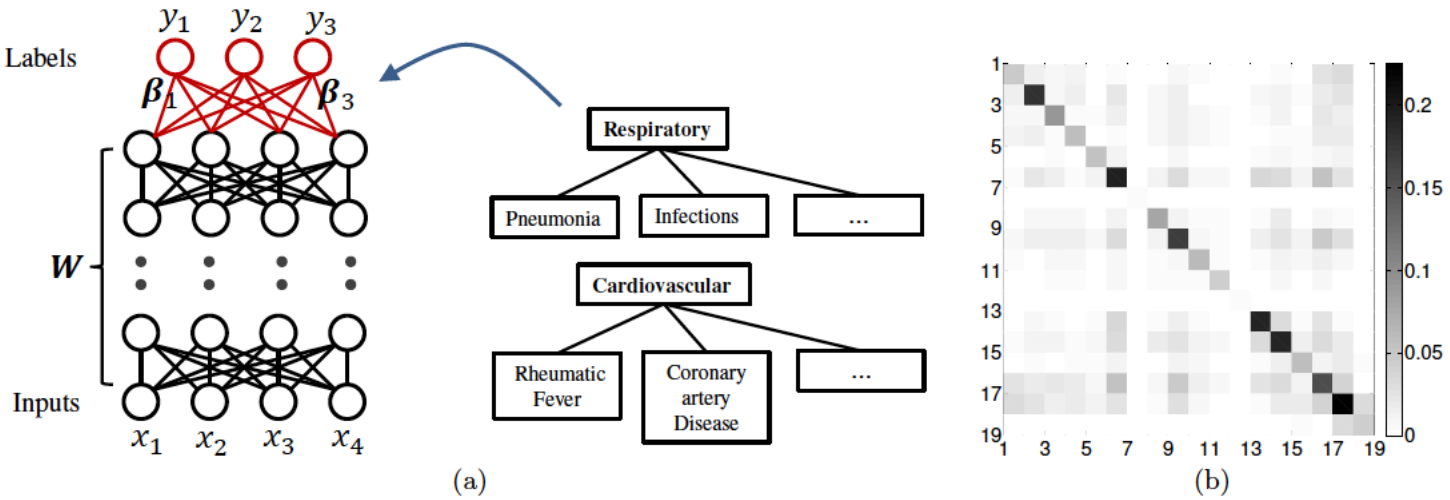


Figure 1: (a) A miniature illustration of the deep network with the regularization on categorical structure. The regularization is applied to the output layer of the network. (b) The co-occurrence matrix in the ICU dataset.

Methodology
Incremental Training

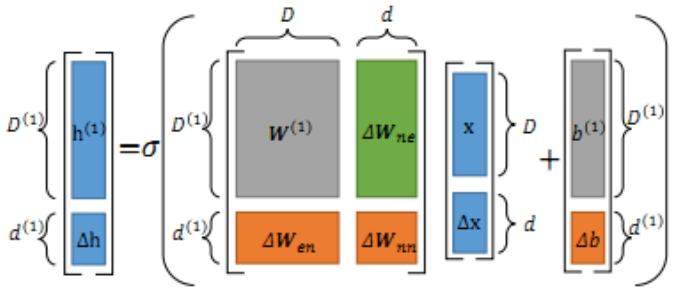


Figure 3: How adding various units changes the weights $\boldsymbol{W}$.

- $\Delta \boldsymbol{W}_{ne}$: $D^{(1)} \times d$ weights that connect new inputs to existing features.
- $\Delta \boldsymbol{W}_{en}$: $d^{(1)} \times D$ weights that connect existing inputs to new features.
- $\Delta \boldsymbol{W}_{nn}$: $d^{(1)} \times d$ weights that connect new inputs to new features.

Measurements with varying or increasing length

- Regularly re-classify a time series based on all available data.
  - waste of sources, unstable diagnose result
- Sliding window approach: Subsequences of size Ts with stride Rs.
  - Testing too many sets of Ts, Rs can be time-consuming, computationally expensive for neural network

Methodology
Incremental Training

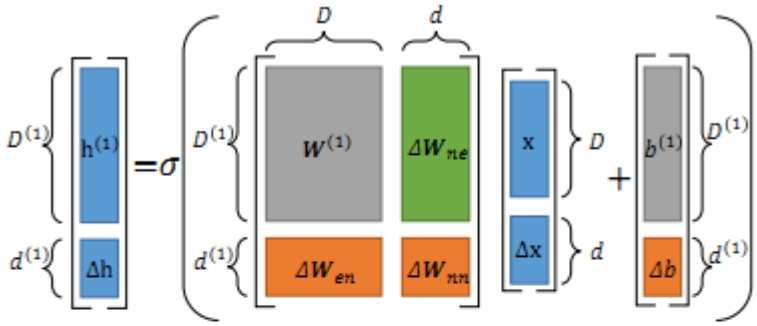


Figure 3: How adding various units changes the weights $\boldsymbol{W}$.

- $\Delta \boldsymbol{W}_{ne}$: $D^{(1)} \times d$ weights that connect new inputs to existing features.
- $\Delta \boldsymbol{W}_{en}$: $d^{(1)} \times D$ weights that connect existing inputs to new features.
- $\Delta \boldsymbol{W}_{nn}$: $d^{(1)} \times d$ weights that connect new inputs to new features.

Algorithm 1 Similarity-based Initialization

Input: Training data $\mathbf{X} \in \mathbb{R}^{N \times (D+d)}$; existing weights $\boldsymbol{W}^{(1)} \in \mathbb{R}^{D^{(1)} \times D}$; kernel function $\mathbf{k}(\cdot, \cdot)$

Output: Initialized weights $\Delta \boldsymbol{W}_{ne} \in \mathbb{R}^{D^{(1)} \times d}$

```
for each new input dimension i ∈ [1, d] do
    for each existing input dimension k ∈ [1, D] do
        Let K[D + i, k] := k(X[·, D + i], X[·, k])
    end for
    Normalize K (if necessary)
    for each existing feature j ∈ [1, D^(1)] do
        Let ΔW_ne[j, i] := Σ_{k=1}^{D} K[D + i, k] W^(1)[j, k]
    end for
end for
```

Algorithm 2 Gaussian Sampling-based Initialization

Input: Existing weights $\boldsymbol{W}^{(1)} \in \mathbb{R}^{D^{(1)} \times D}$

Output: Initialized weights $\Delta \boldsymbol{W}_{en} \in \mathbb{R}^{d^{(1)} \times D}$, $\Delta \boldsymbol{W}_{nn} \in \mathbb{R}^{d^{(1)} \times d}$

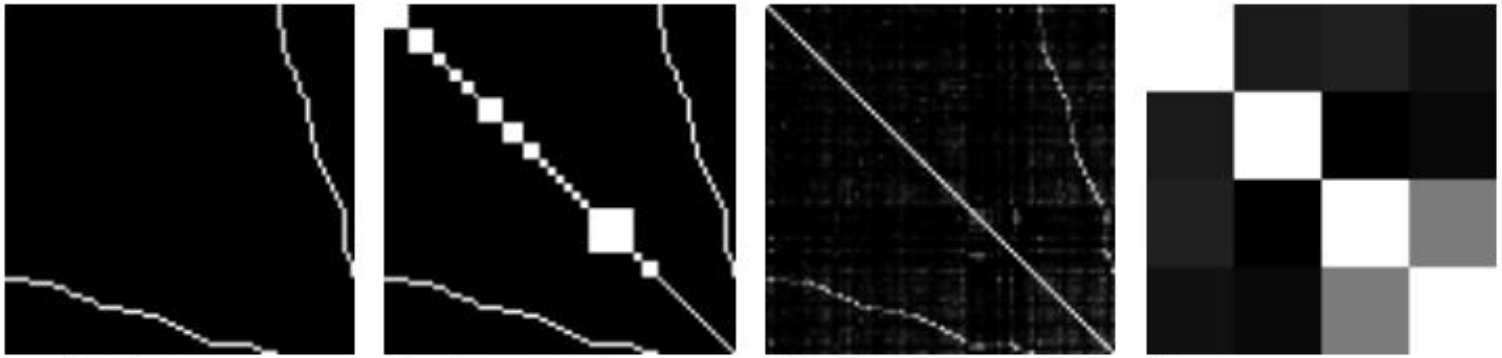
```
Let w̄ = 1/(D D^(1)) Σ_{i,j} W^(1)[i, j]
Let s̄ = 1/(D D^(1) − 1) Σ_{i,j} (W^(1)[i, j] − w̄)^2
for each new feature j ∈ [1, d^(1)] do
    for each existing input dimension i ∈ [1, D] do
        Sample ΔW_ne[j, i] ~ N(w̄, s̄)
    end for
    for each new input dimension i ∈ [1, d] do
        Sample ΔW_nn[j, i] ~ N(w̄, s̄)
    end for
end for
```

Experiments
Data

Two collections of clinical time series collected during the delivery of care in intensive care units (ICUs).

1. Physionet Challenge 2012 Data (8000 ICU units, has no natural label structure)
2. ICU Data (contains 8500 episodes varying in length from 12 to 128 hours.)

Experiments
Benefits of prior-based regularization



(a) ICD-9 Tree. (b) ICD-9 shared category. (c) ICU co-occurrence. (d) Physionet co-occur.

Figure 4: Example priors for the ICU (a-c) and Physionet (d) data sets.

Table 1: AUROC for classification.

	Tasks	*No Prior*	*Co-Occurrence*	*ICD-9 Tree*
Subsequence	*All*	0.7079 ± 0.0089	$\mathbf{0.7169 \pm 0.0087}$	0.7143 ± 0.0066
	Categories	0.6758 ± 0.0078	$\mathbf{0.6804 \pm 0.0109}$	0.6710 ± 0.0070
	Labels	0.7148 ± 0.0114	$\mathbf{0.7241 \pm 0.0093}$	0.7237 ± 0.0081
Episode	*All*	0.7245 ± 0.0077	$\mathbf{0.7348 \pm 0.0064}$	0.7316 ± 0.0062
	Categories	0.6952 ± 0.0106	$\mathbf{0.7010 \pm 0.0136}$	0.6902 ± 0.0118
	Labels	0.7308 ± 0.0099	$\mathbf{0.7414 \pm 0.0064}$	0.7407 ± 0.0070

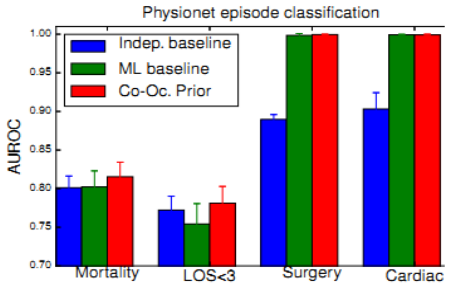


Figure 5: Physionet classification performance.

Experiments
Efficacy of incremental training

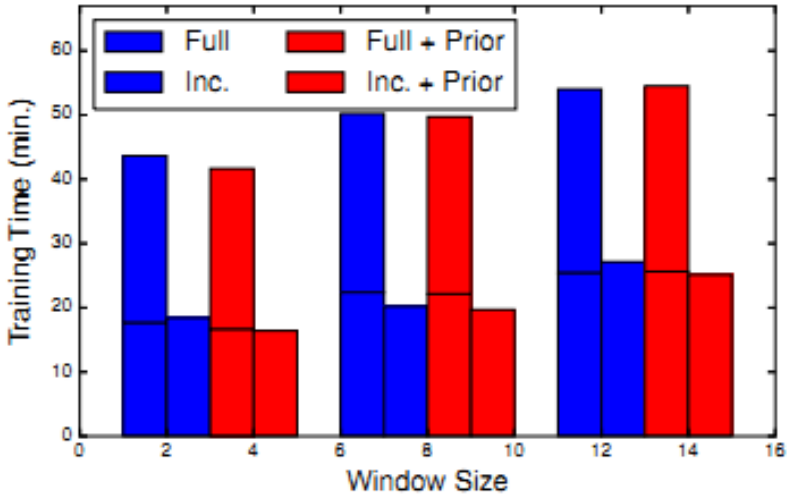


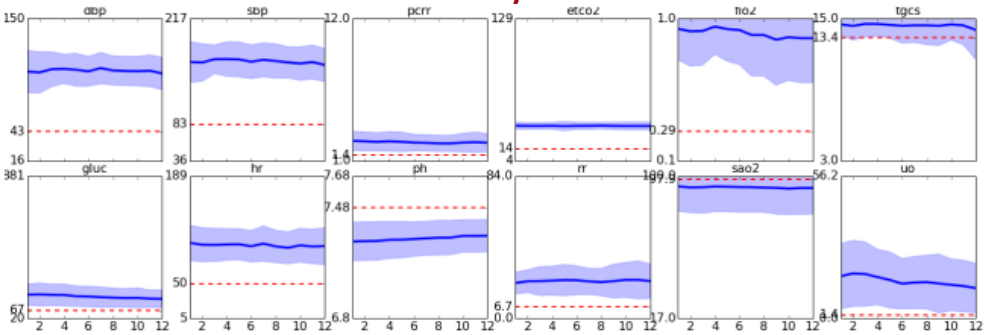
Figure 6: Training time for different neural networks for full/incremental training strategies.

Table 2: AUROC for incremental training.

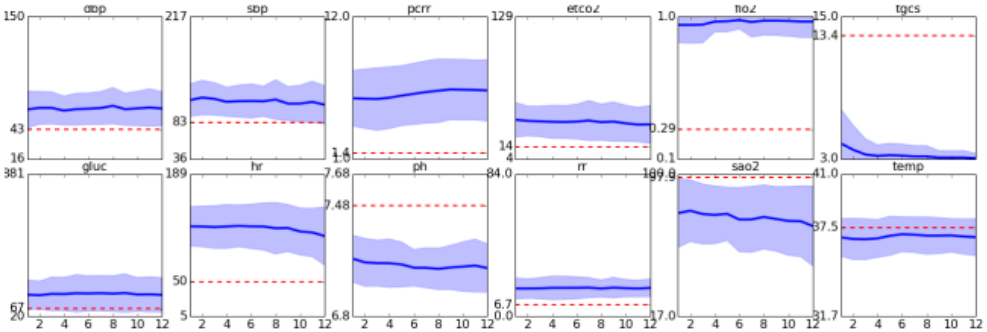
Size	Level	*Full*	*Inc*	*Prior+Full*	*Prior+Inc*
16	Subseq.	0.6928	0.6874	0.6556	0.6581
	Episode	0.7148	0.7090	0.6668	0.6744
20	Subseq.	0.6853	0.6593	0.6674	0.6746
	Episode	0.7022	0.6720	0.6794	0.6944
24	Subseq.	0.7002	0.6969	0.6946	0.7008
	Episode	0.7185	0.7156	0.7136	0.7171

Experiments

Qualitative Analysis of Features



(a) Phenotypic patterns for ICD-9 circulatory disease category (390-459).



(b) Phenotypic patterns for conditions related to septic shock (ICD-9 codes 990-995).

Figure 7: Example features learned from the ICU data.

Causality Analysis

Table 3: Magnitude of the causal relationships identified by using the representations learned by the deep learning algorithm.

No Prior	*Co-occurrence*	*ICD-9 Tree*
252.61	**270.28**	242.50

Agenda

- Sequential Pattern Mining
- Deep Learning
- Anchor Based Methods

Challenges Anchor Method Tries to Solve

- How to combine domain expertise (simple rules) and vast amounts of data (machine learning)?
- How to learn without manually created labels?
- How the learned models could generalize across institutions?

What is Anchor?

Rather than provide gold-standard labels, construct a simple rule that can catch some positive cases.

Phenotype	Possible Anchor
Diabetic	gsn:016313 (insulin) in Medications
Cardiac	ICD9:428.X (heart failure) in Diagnoses
Nursing home	“from nursing home” in text
Social work	“social work consulted” in text

Theoretical Basis for Anchor

Unobserved variable: Y, Observation: A

Definition: A is an anchor for Y if conditioning on A=1 gives uniform samples from the set of positive cases.

Alternative formulation: two necessary conditions

$P(Y = 1|A = 1) = 1$ AND $A \perp \chi| Y$

Positive condition

e.g. If patient is taking insulin, the patient is surely diabetic.

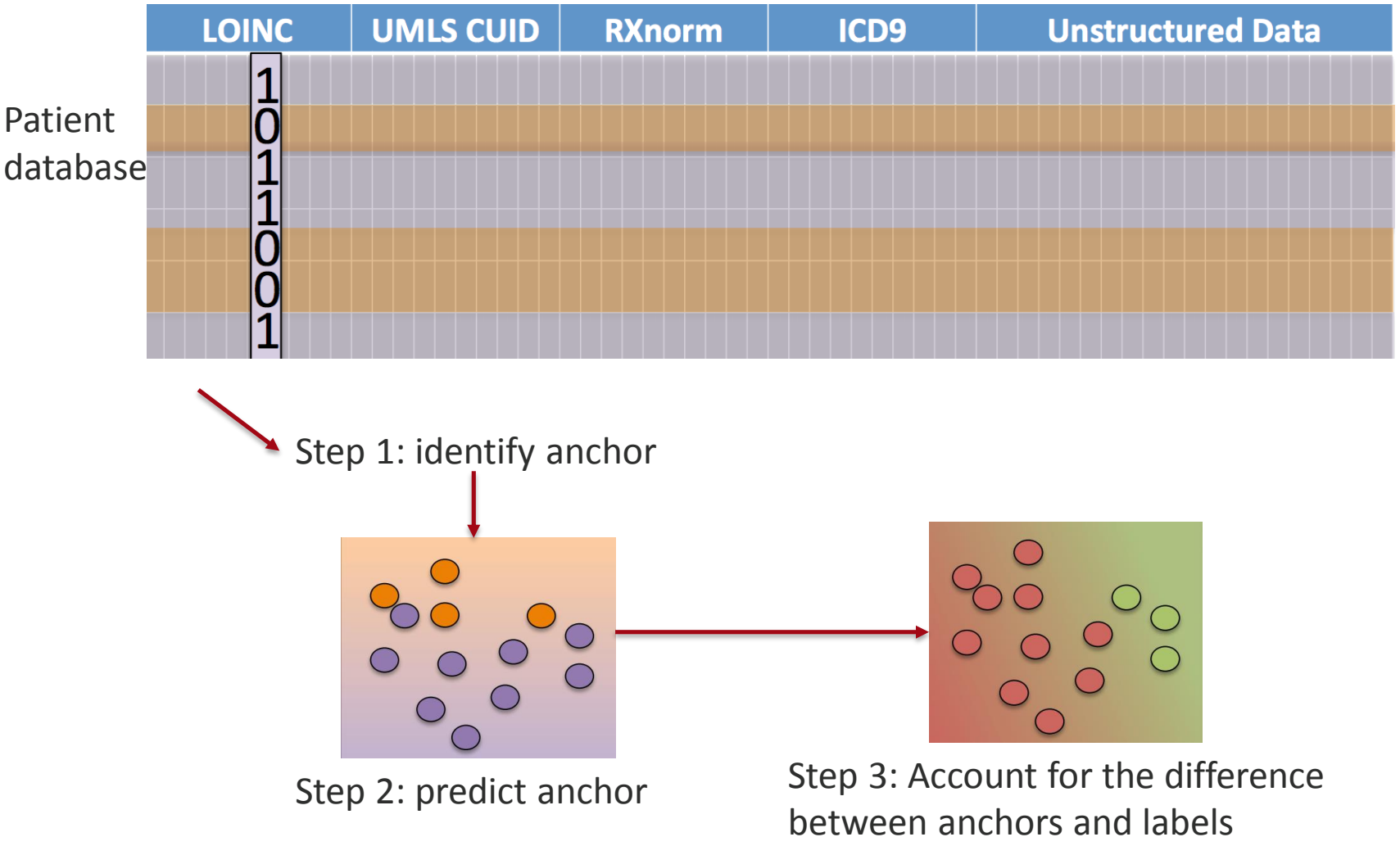
Conditional independence

e.g. If we know the patient had heart failure, knowing whether the diagnosis code appears does inform us about the rest of the record.

(χ all other observations)

Theorem [Elkan & Noto 2008]: In the above setting, a function to predict A can be transformed to predict Y

Learning with Anchor (Illustration)



Learning with Anchor (Algorithm)

Input: anchor A

unlabeled patients

Output: prediction rule

1. Learn a calibrated classifier (e.g. logistic regression) to predict:

$$\Pr(A = 1 \mid \mathcal{X})$$

2. Using a validate set, let *P* be the patients with A=1. Compute:

$$C = \frac{1}{|\mathcal{P}|} \sum_{k \in \mathcal{P}} \Pr(A = 1 \mid \mathcal{X}^{(k)})$$

3. For a previously unseen patient *t*, predict:

$$\begin{cases} \frac{1}{C} \Pr(A = 1 | \mathcal{X}^{(t)}) & \text{if } A^{(t)} = 0 \\ 1 & \text{if } A^{(t)} = 1 \end{cases}$$

Learning

Learn to predict A from the other variables.

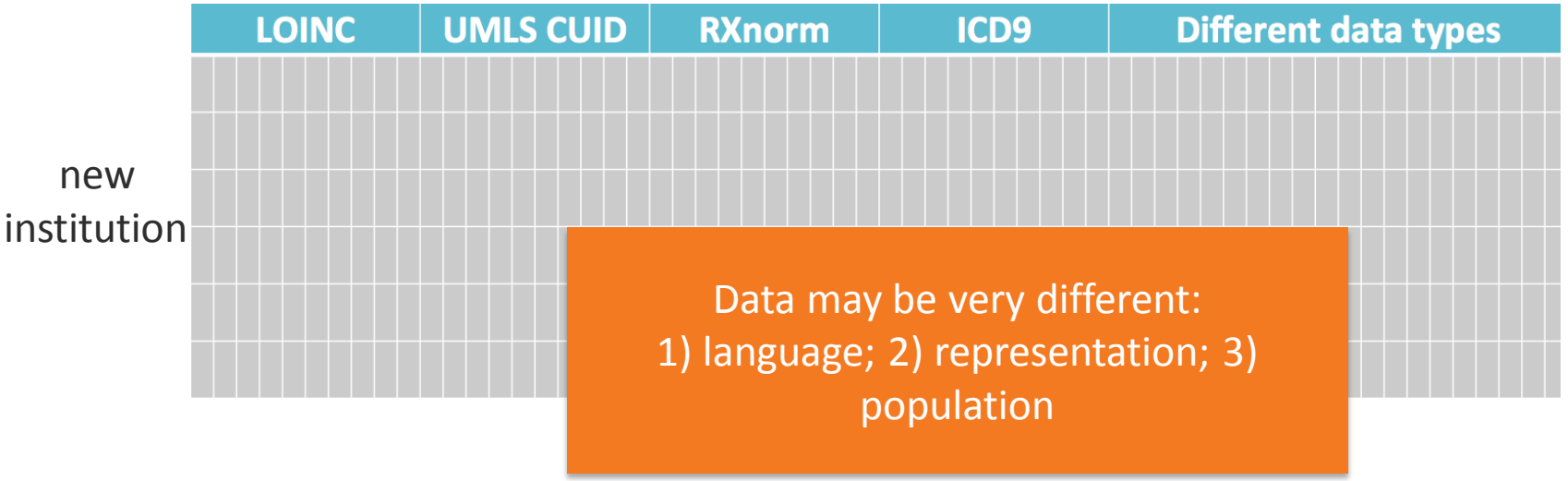
Calibration

C is the average model prediction for patients with anchors.

Transformation

If no anchor present, according to a scaled version of the anchor-prediction model

Generalizability/Portability



As long as our anchors appear in the new data as well… Can learn a new model, specific to the new institution.

Only need to share anchor definitions.
Each site trains models on its own data.

Experiment

Table 1: Features used to build binary patient description vectors

	Representation	Dimension
Age	Binned by decade	11
Sex	M/F	2
Medication History	Indicators by medication generic sequence number	1947
Medication Dispensing Record		279
Triage Vitals	Binned by decision tree	77
Lab Results		2805
Triage Assessment	Binary bag-of-words	7073
MD Comments		8909

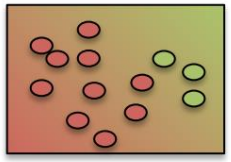
Table 2: Phenotype variables used for evaluation

Phenotype	Disposition Question	N	Pos
Cardiac – acute	In the workup of this patient, was a cardiac etiology suspected?	17 258	0.068
Infection – acute	Do you think this patient has an infection? (Suspected or proven viral, fungal, protozoal, or bacterial infection)	62 589	0.213
Pneumonia – acute	Do you think this patient has pneumonia?	9934	0.073
Septic shock – acute	Is the patient in septic shock?	6867	0.020
Nursing home – history	Is the patient from a nursing home or similar facility? (Interpret as if you would be giving broad-spectrum antibiotics)	36 256	0.045
Anticoagulated – history	Prior to this visit, was the patient on anticoagulation? (Excluding antiplatelet agents like aspirin or Plavix)	1082	0.047
Cancer – history	Does the patient have an active malignancy? (Malignancy not in remission, and recent enough to change clinical thinking)	4091	0.042
Immunosuppressed – history	Is the patient currently immunocompromised?	12 857	0.040

Learned Models: Nursing Home

Anchors

nursing facility
nursing home
nsg facility
nsg home
nsg. home



Highly weighted features

Ages	Medications	Pyxis
age=90+	senna mirtazapine	vancomycin
age=80-90	colace maalox	levofloxacin
age=70-80	trazodone tums	

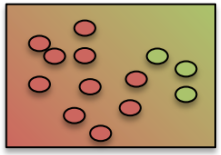
baseline	from	dnr
changes	staff	full code
nonverbal	at	g tube
ams	resident	foley
unwitnessed_fall	sent	nh
confusion	reported	

Unstructured text

Conditional independence assumption?

Learned Models: Cardiac Etiology

Anchors



Highly weighted features

ICD9 codes
410.* acute MI
411.* other acute …
413.* angina pectoris
785.51 card. shock

Pyxis
coron. vasodilators

cardiac medicine
BIDMC shortform

Ages
age=80-90
age=70-80
age=90+

Medications
lasix
furosemide

Sex=M

Pyxis
aspirin
clopidogrel
Heparin Sodium
Metoprolol
Tartrate
Morphine Sulfate
Integrilin
Labetalol

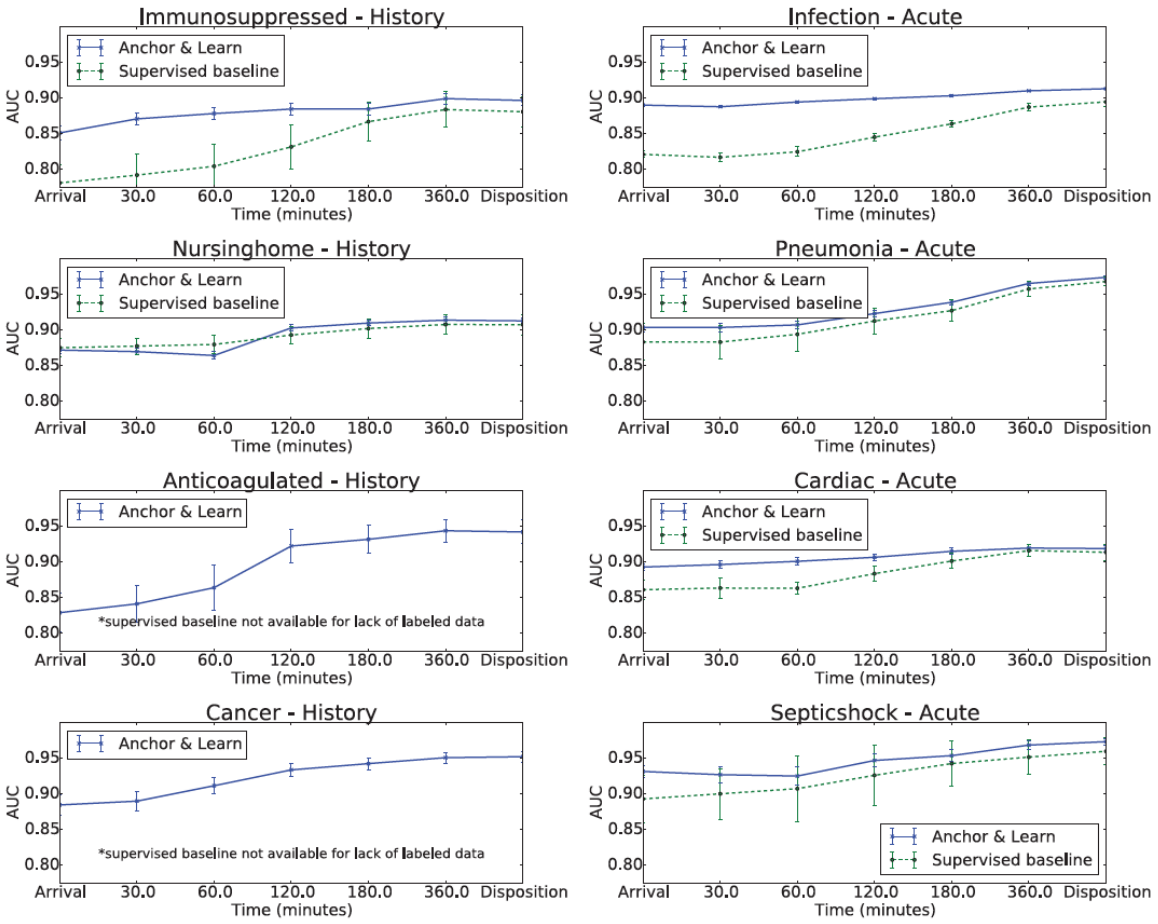
nstemi
stemi
ntg
lasix
nitro

cp
chest pain
edema
cmed
chf exacerbation
sob
pedal edema

Unstructured text

Phenotype predictions

Figure 1: Comparison of performance of phenotypes learned with 200 000 unlabeled patients using the semi-supervised anchor based method, and phenotypes learned with supervised classification using 5000 gold-standard labels. Error bars indicate 2 * standard error. For anticoagulated and cancer, there were not a sufficient number of gold-standard labels to learn with 5000 patients, so the fully supervised baseline is omitted.



LR

Weill Cornell Medicine

Computational Phenotyping using Tensor Factorization and Tensor Network

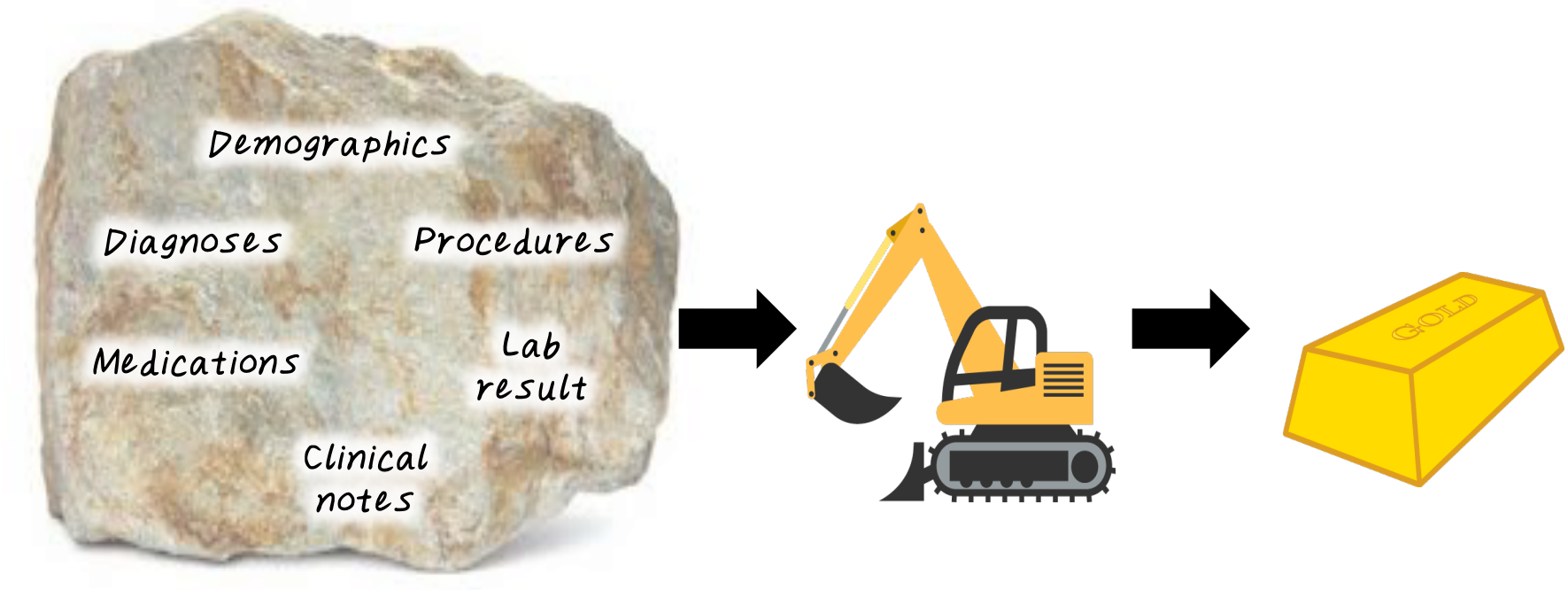
Jimeng Sun

jsun@cc.gatech.edu

Georgia Tech College of Computing
Computational Science and Engineering

SunLab
Enlighten the World

COMPUTATIONAL PHENOTYPING



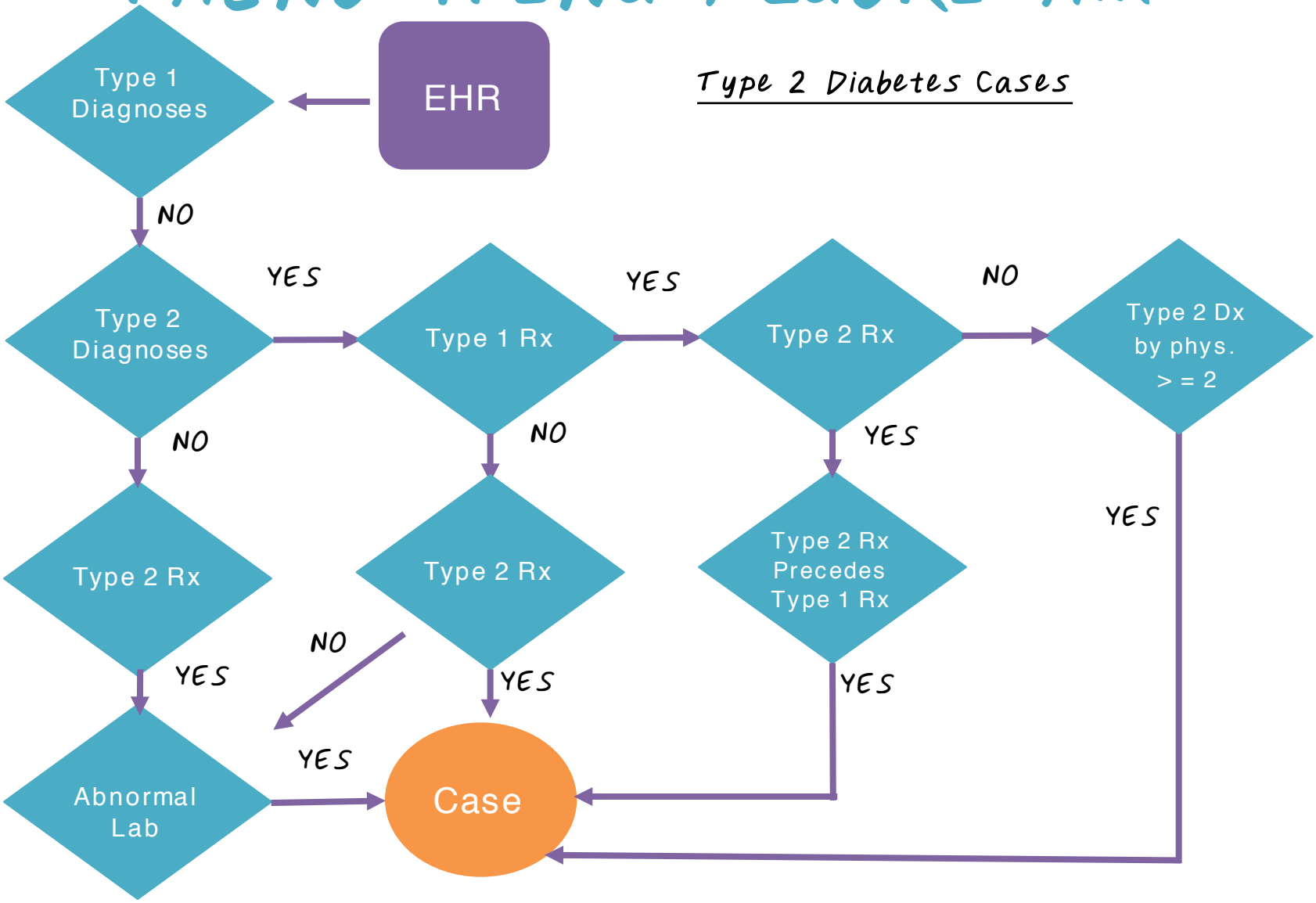
Raw Data

Phenotyping

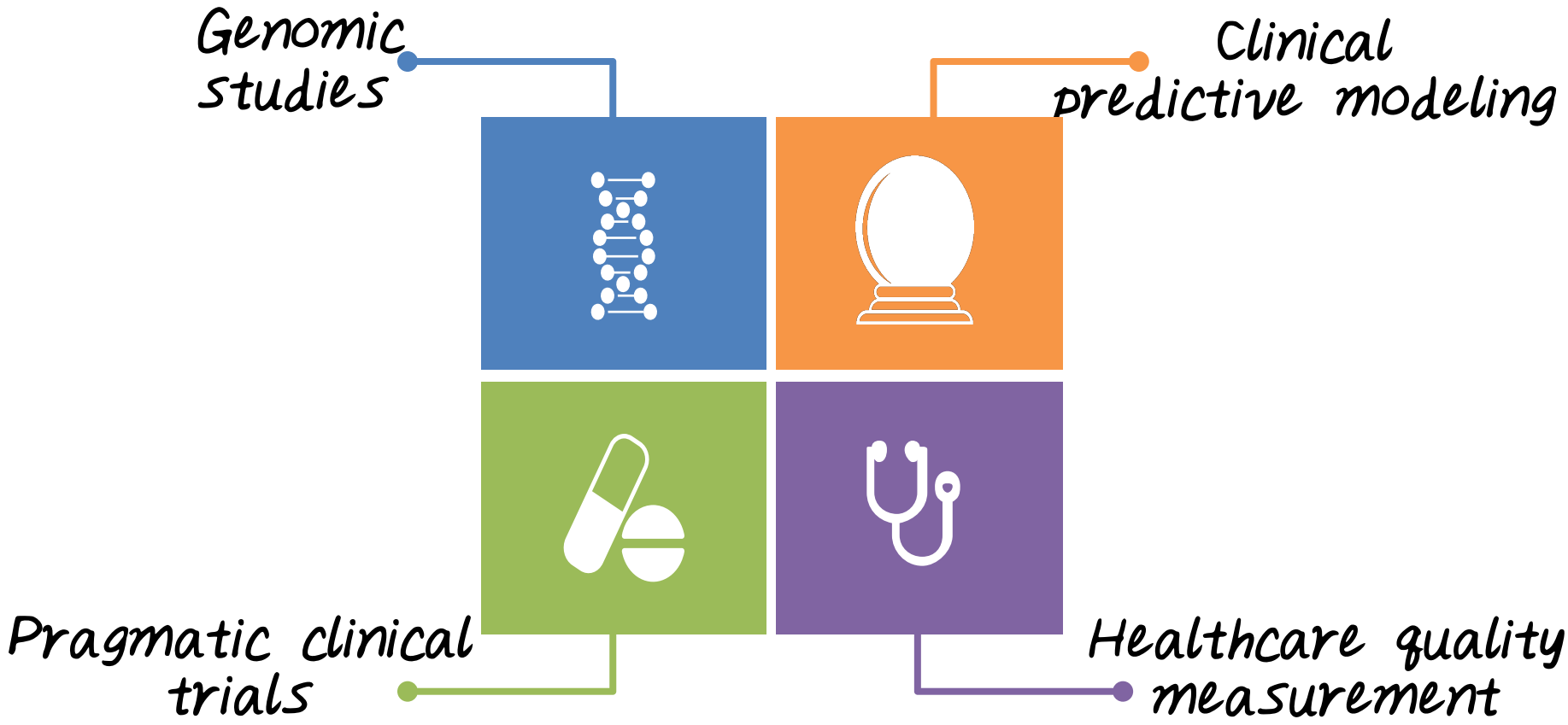
Medical Concepts (Phenotypes)

A phenotype = a cluster of patients that share similar characteristics

PHENOTYPING ALGORITHM



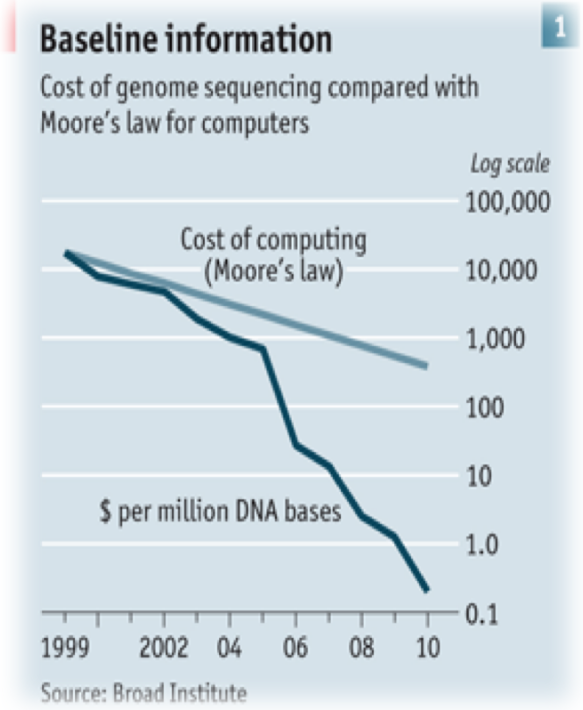
APPLICATIONS OF PHENOTYPING



WHY DO WE CARE ABOUT PHENOTYPING?

Precision Medicine

We need rich and deep phenotypic data in order to analyze genomic data.



Cost of phenotypic data

Cost of genomic data

CLINICAL PREDICTIVE MODELING

Raw Data

Predictive Modeling

Model

Phenotyping

Medical Concepts (Phenotypes)

PRAGMATIC CLINICAL TRIALS

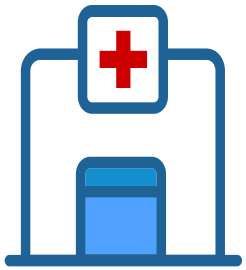
TRADITIONAL

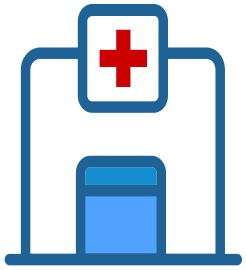
- One condition
- One drug
- Must randomize
- Careful selection
- Carefully controlled

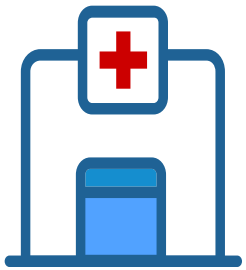
PRAGMATIC

- Multiple conditions
- Potentially multiple drugs
- No randomization
- Any patient
- Real-world environment

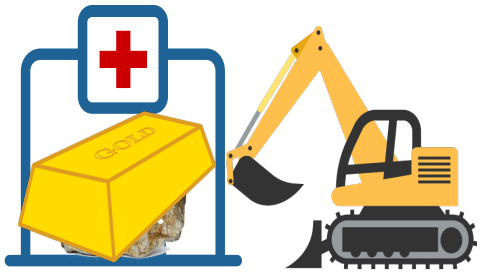
HEALTHCARE QUALITY MEASUREMENT

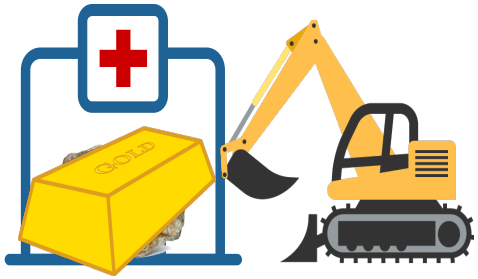






HEALTHCARE QUALITY MEASUREMENT





PHENOTYPING METHODS

SUPERVISED LEARNING

- Expert-defined rules
- Classification
- Natural language processing

UNSUPERVISED LEARNING

- Dimensionality Reduction
  - Tensor factorization

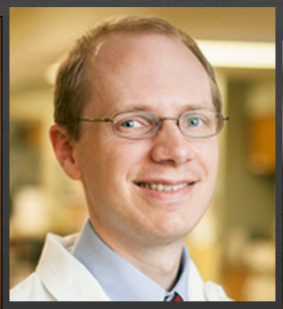
PHENOTYPING FROM EHR WITH TENSOR FACTORIZATION

Project: SCH: INT: Collaborative Research: High-throughput Phenotyping on Electronic Health Records using Multi-Tensor Factorization

Principal Investigators

Jimeng Sun
Associate Professor
College of Computing
Georgia Tech

Bradley Malin
Associate Professor
Biomedical Informatics
and Computer Science
Vanderbilt University

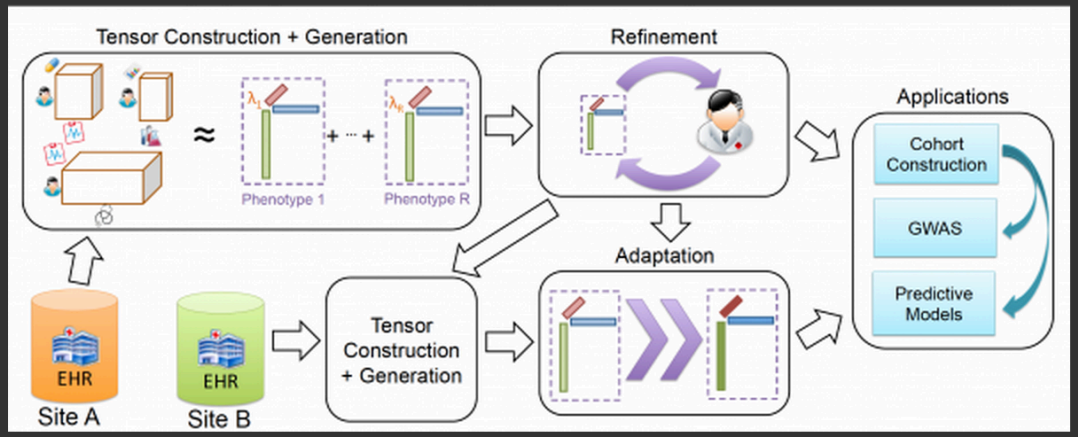


Joshua Denny
Associate Professor
Biomedical Informatics
and Medicine
Vanderbilt Utnvierty

Joydeep Ghosh
Professor
Electrical & Computer
Engineering
Univ of Texas, Austin

Abel Kho
Assistant Professor
Medicine - Biomedical
Informatics
Northwestern Univ

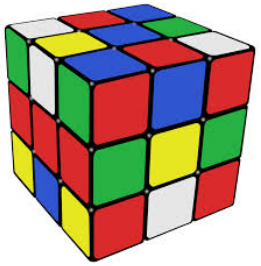
Funding Source: NSF Smart Connect Health Integrated Grant: Award Number 1418511



Tensor Factorization

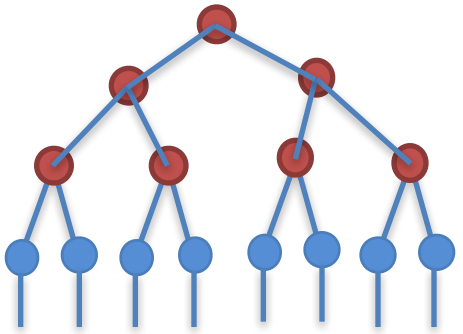
Limestone JBI'14

Marble KDD'14



Rubik KDD'15

Tensor Network



Sparse Hierarchical Tucker

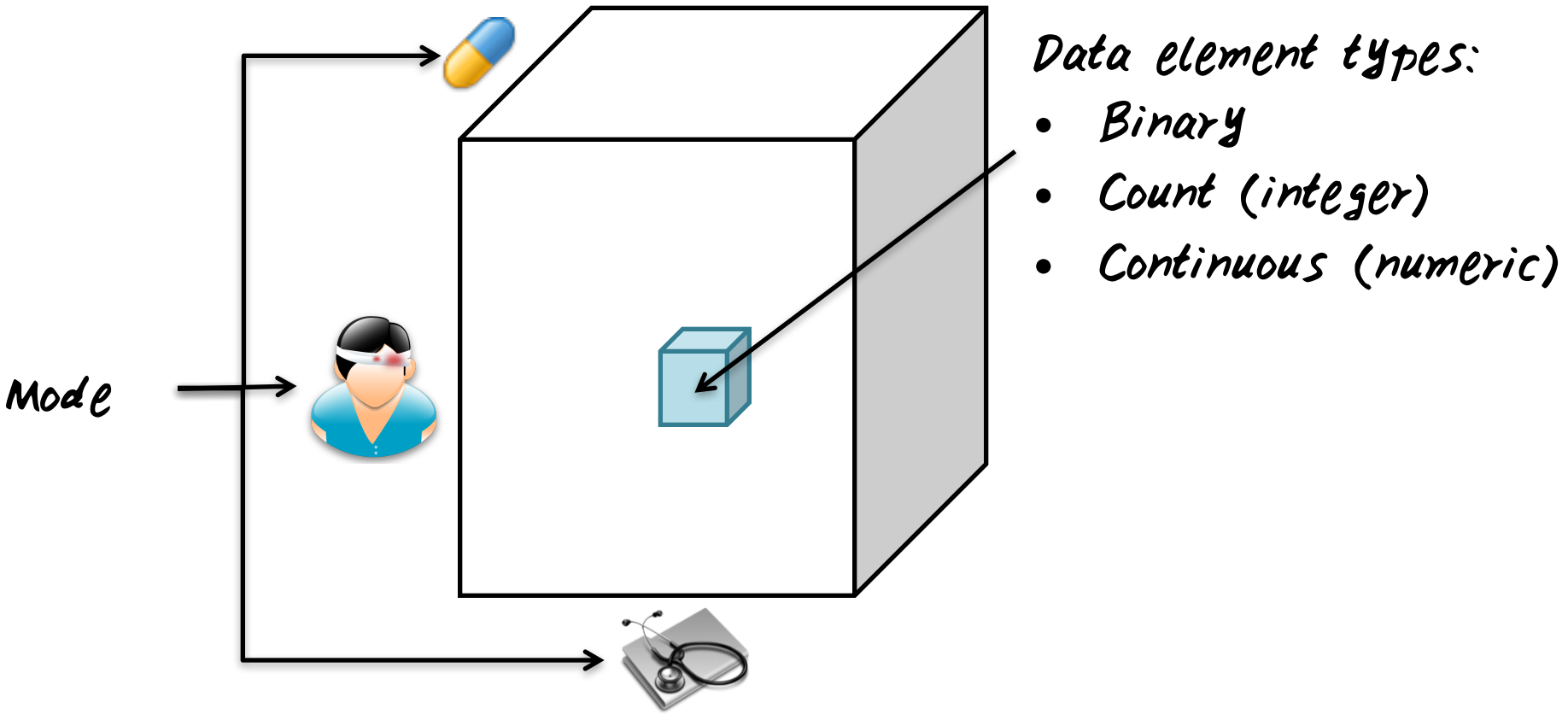
ICDM'15

Limestone

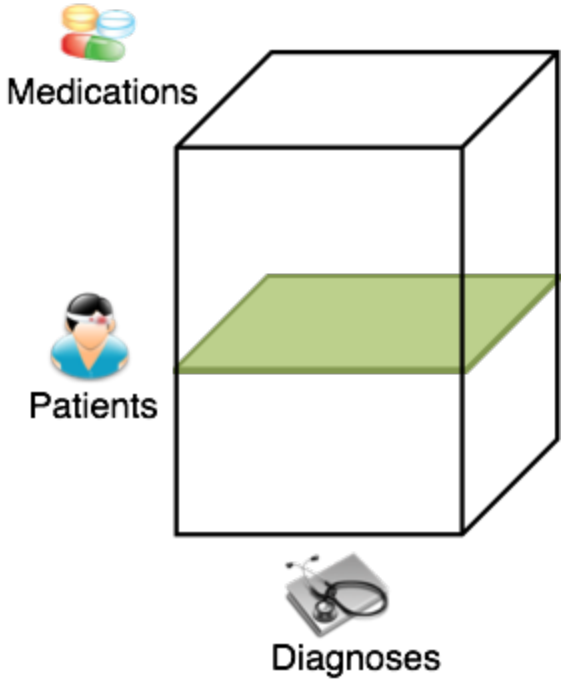
Ho, Joyce C., Joydeep Ghosh, Steve R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. “Limestone: High-Throughput Candidate Phenotype Generation via Tensor Factorization.” Journal of Biomedical Informatics. 2014

Use Tensor to model EHR data

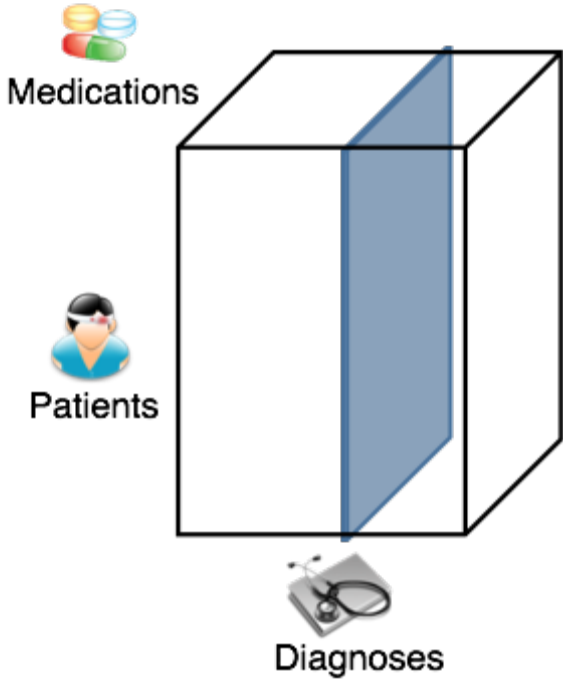
- Tensor is a generalization of matrix



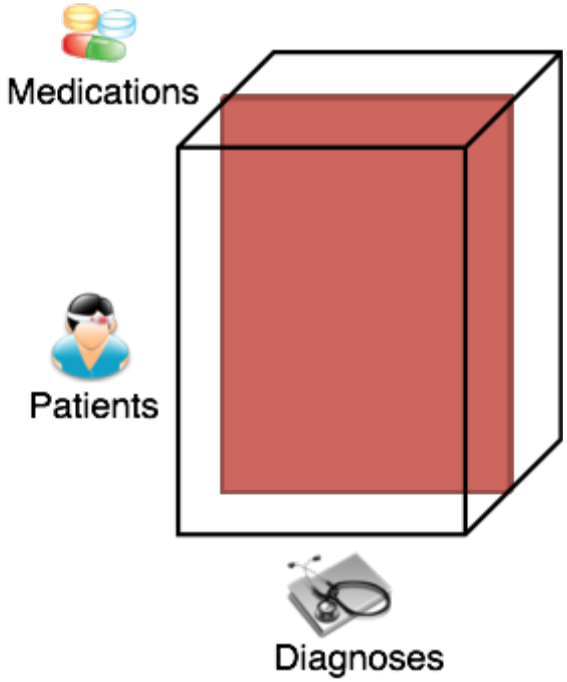
Tensor Operations: Slicing



View of Bob's diagnosis and associated medications

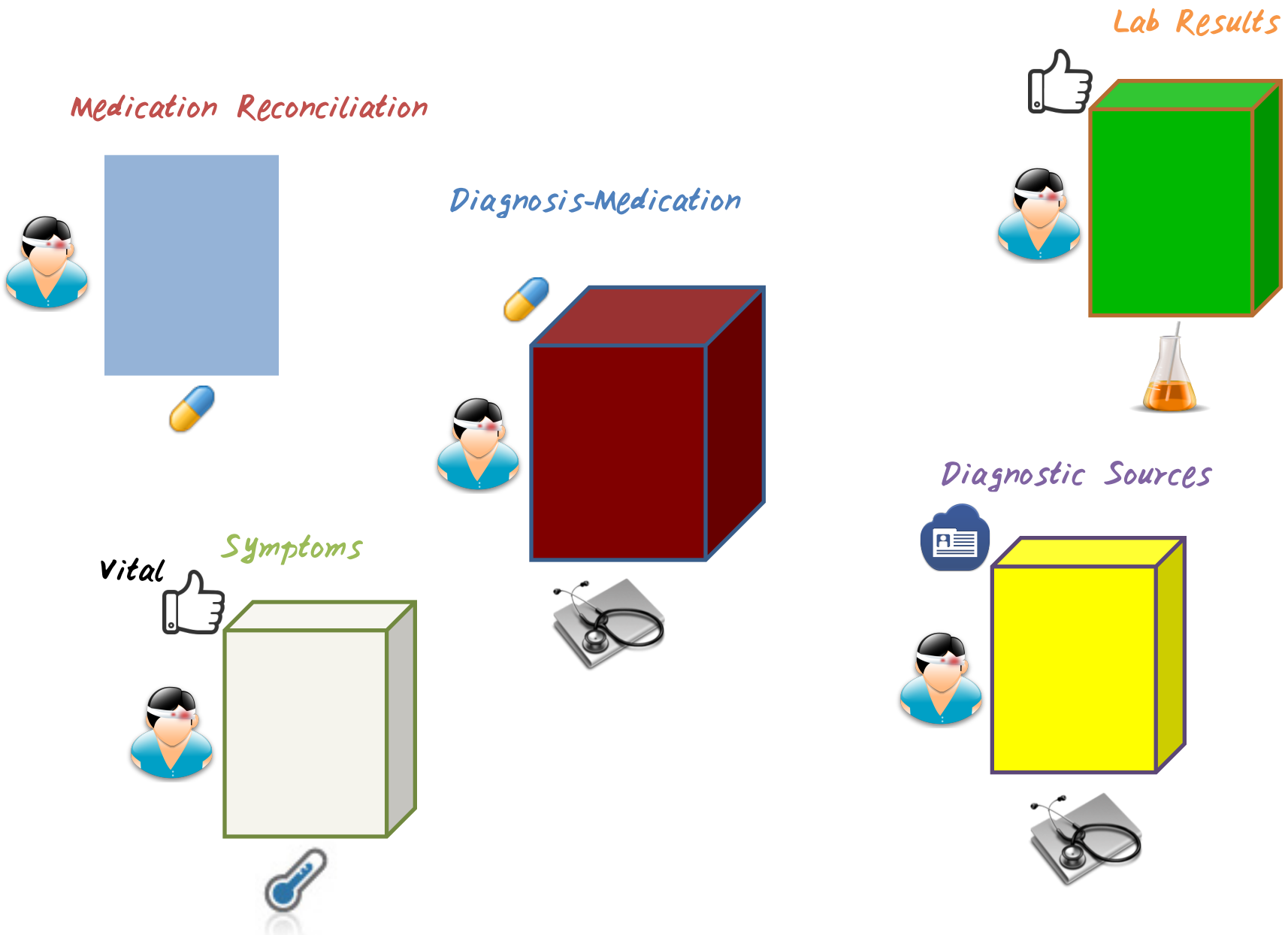


View of prescribed medications to treat asthma for all patients



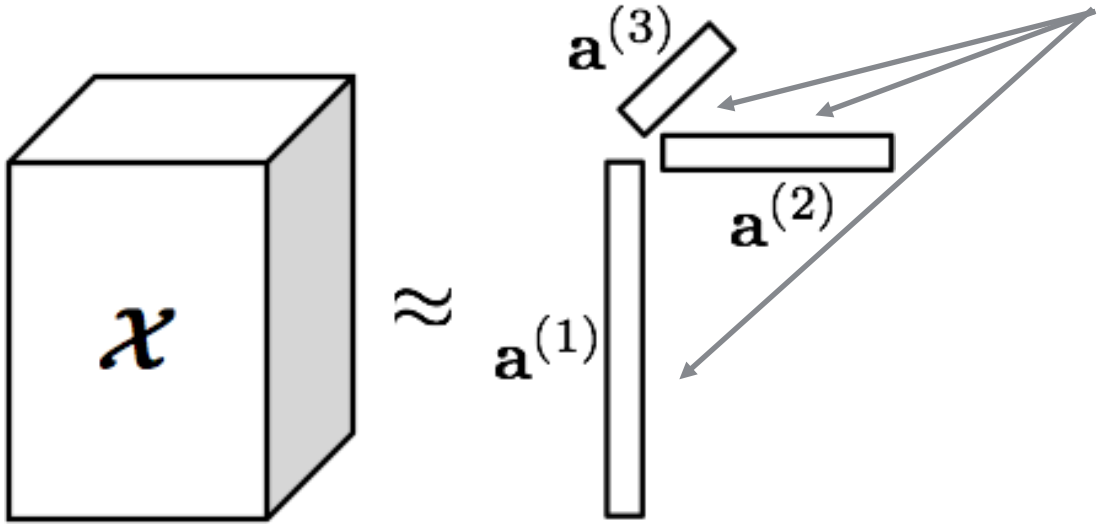
View of patients and diagnosis treated with beta-blocker

Multiple Tensors



Rank-One Tensor

basis vector along each mode

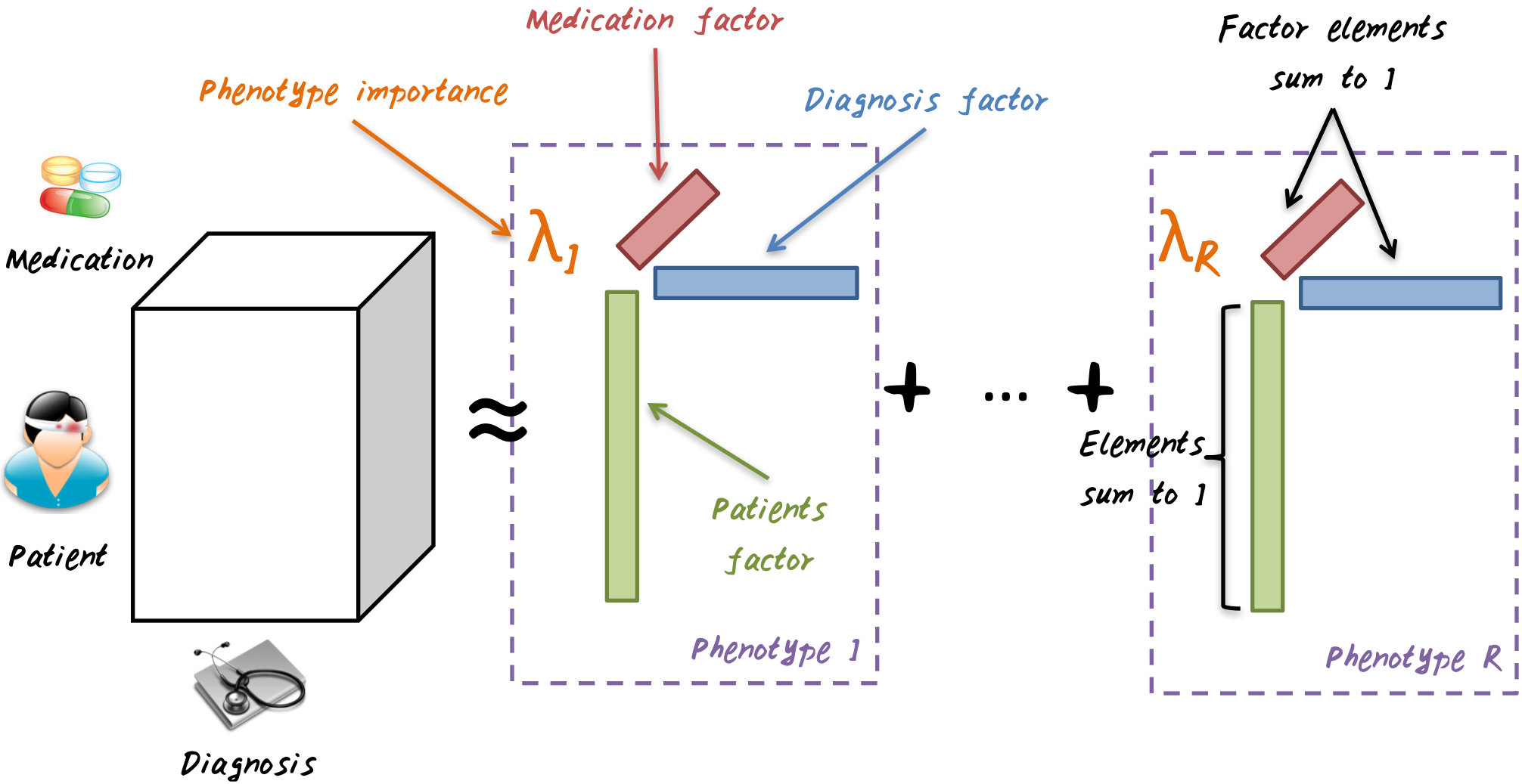


$$\mathcal{X} = \mathbf{a}^{(1)} \circ \mathbf{a}^{(2)} \circ \cdots \circ \mathbf{a}^{(N)}$$

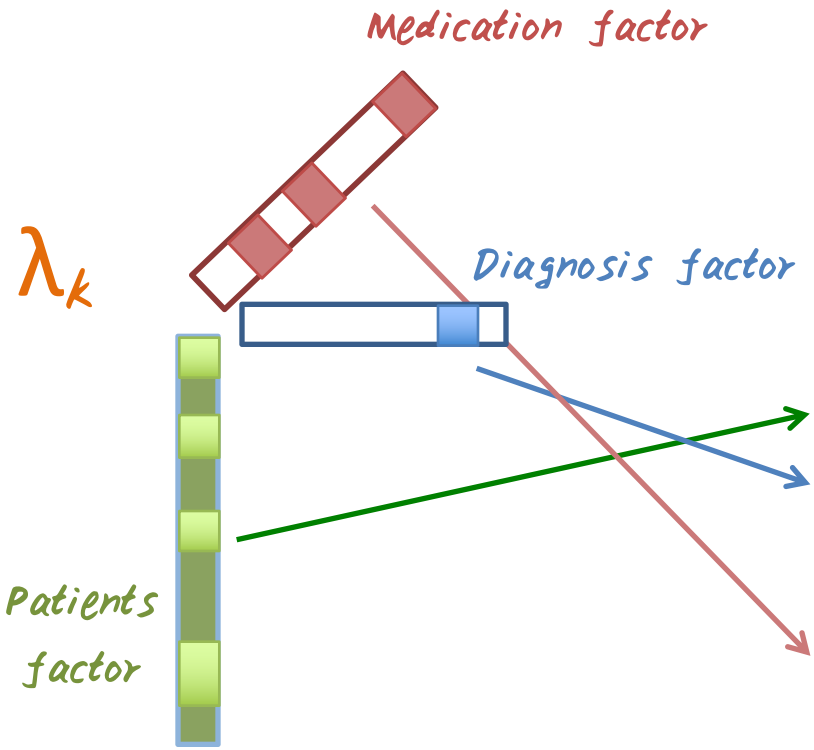
$$\Vert\Vert$$

$$x_{i_1 i_2 \cdots i_N} = a^{(1)}_{i_1} a^{(2)}_{i_2} \cdots a^{(N)}_{i_N}$$

Phenotyping through Tensor Factorization



Example Phenotype

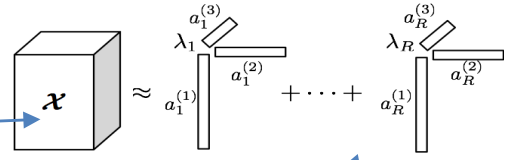


Candidate Phenotype k (40% of patients)
Hypertension
Beta Blockers Cardio-Selective
Thiazides and Thiazide-Like Diuretics
HMG CoA Reductase Inhibitors

a phenotype = a group of patients that share common characteristics

CP Alternating Poisson Regression (CP-APR)

- Nonnegative input tensor
- Nonnegative constraints
- Stochastic column constraints on factor matrices



KL divergence

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

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

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

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

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

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

- Nonnegative input tensor
- Nonnegative constraints
- Stochastic column constraints on factor matrices
- Hard thresholding on elements in factor matrices

$$\min f(\mathcal{M}) \equiv \min \sum_{\mathbf{i}} [m_{\mathbf{i}} - x_{\mathbf{i}} \log m_{\mathbf{i}}] + \gamma \sum_{j,n,r} \mathbb{1}_{\{a_{jr}^{(n)} > 0\}}$$

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

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

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

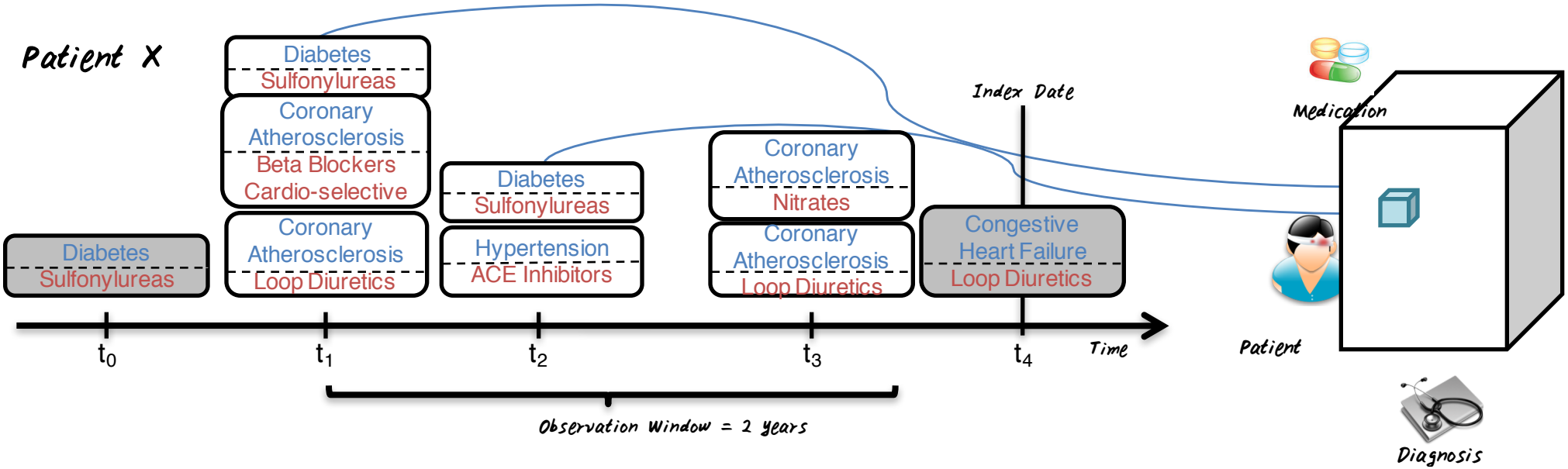
$$\Omega_n = \{\mathbf{A} \in [0,1]^{I_n \times R} \mid ||\mathbf{a}_r||_1 = 1 \ \forall r\}$$

Ho, Joyce C., Joydeep Ghosh, Steve R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. "Limestone: High-Throughput Candidate Phenotype Generation via Tensor Factorization." *Journal of Biomedical Informatics*. 2014

EXPERIMENTS OF LIMESTONE

Tensor Construction

- Medication orders from Geisinger dataset
- 31,816 patients x 169 diagnoses x 471 medications



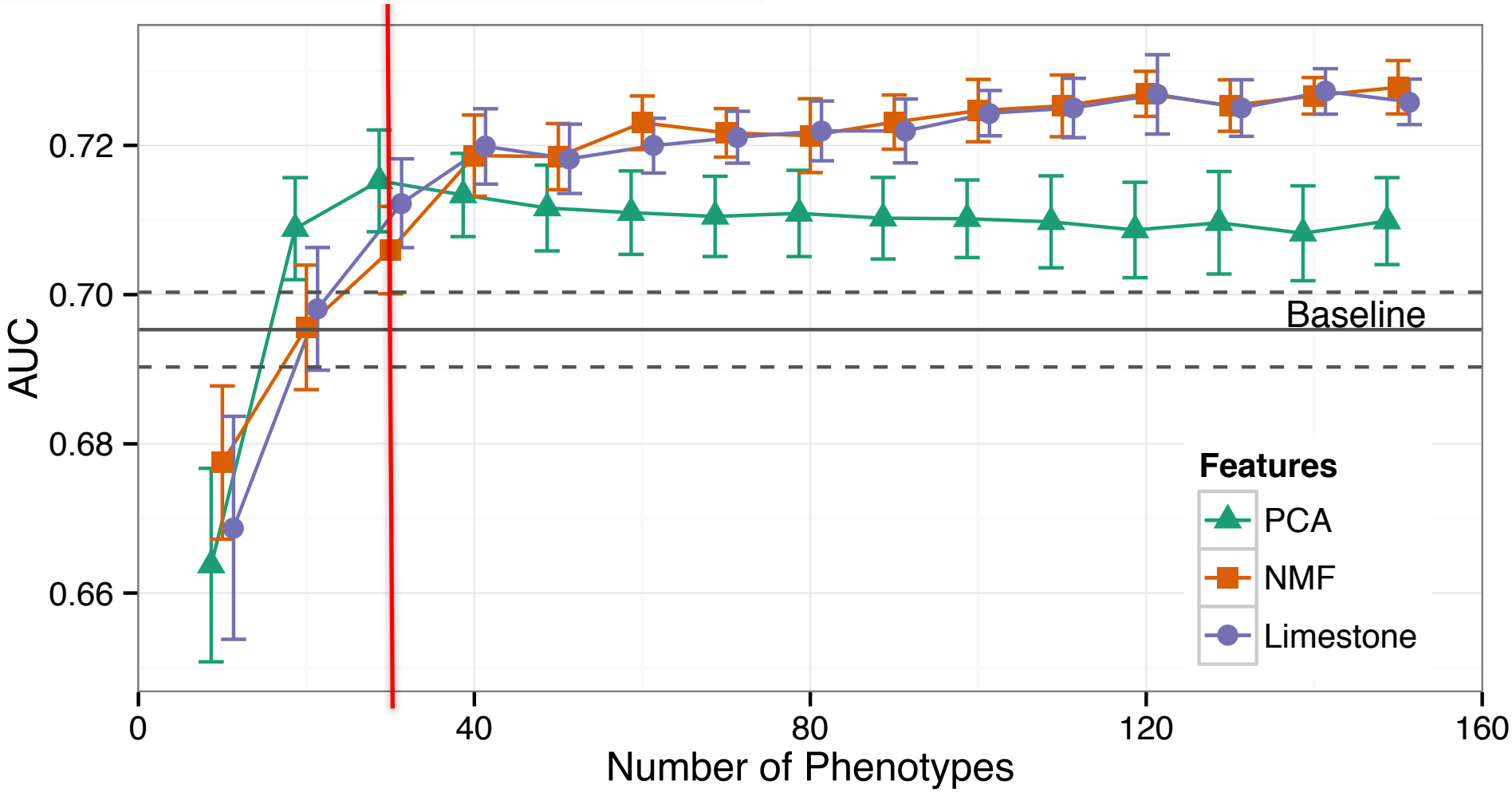
Evaluation: Heart Failure (HF) Prediction

- Task: predict patients with HF
- Model: logistic regression with ℓ_1 regularization
- Evaluation : Cross-validation

- Methods for feature construction:
  1. Baseline using source independence matrix
  2. Principal Component Analysis (PCA)
  3. Nonnegative Matrix Factorization (NMF)
  4. Limestone

Predictive Performance

Small # of features outperforms 640 features



Major disease phenotypes can be identified

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

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

Disease subtypes can be identified

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

Over 80% phenotype factors are clinically meaningful

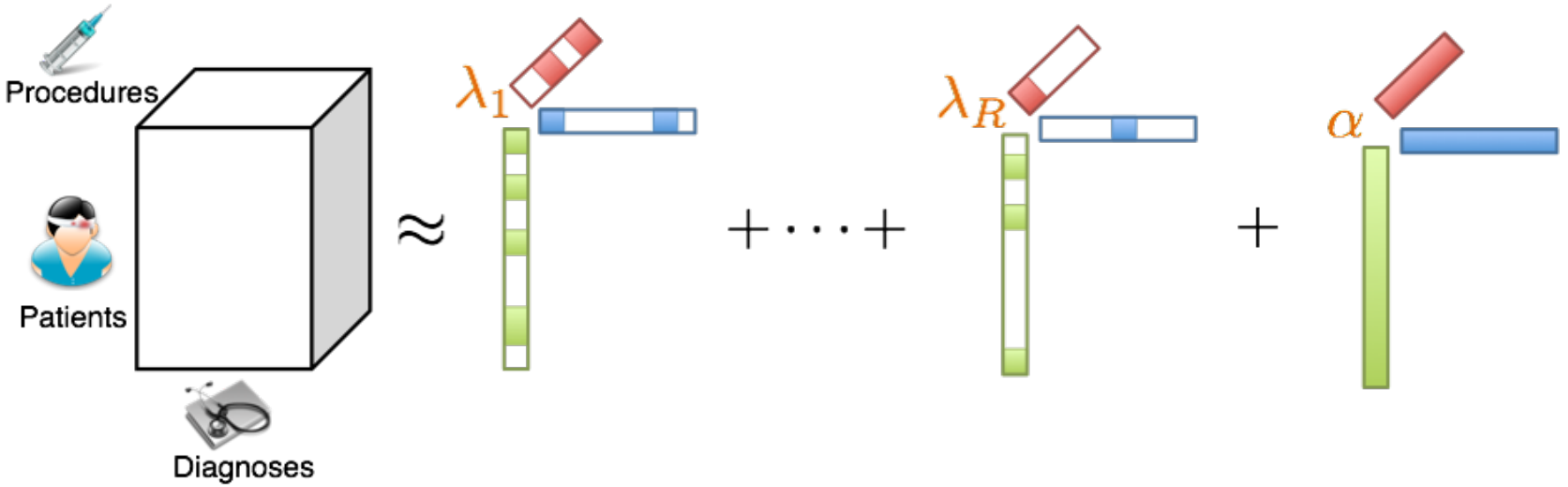
Limestone vs. NMF

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 provides more concise phenotype representation than NMF

Limestone



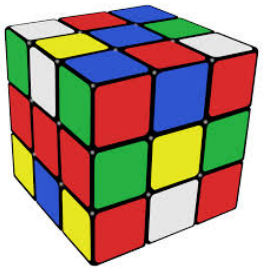
Advantages

- Unsupervised
- Intuitive phenotypes
- Predictive

Limitations

- Unable to leverage knowledge
- Overlapping phenotypes
- Missing data

RUBIK



Yichen Wang, Robert Chen, Joydeep Ghosh, Joshua C,Denny, Abel,Kho, You,Chen, Bradley,A. Malin, Jimeng; Sun. Rubik: Knowledge Guided Tensor Factorization and Completion for Health Data Analytics. KDD'15

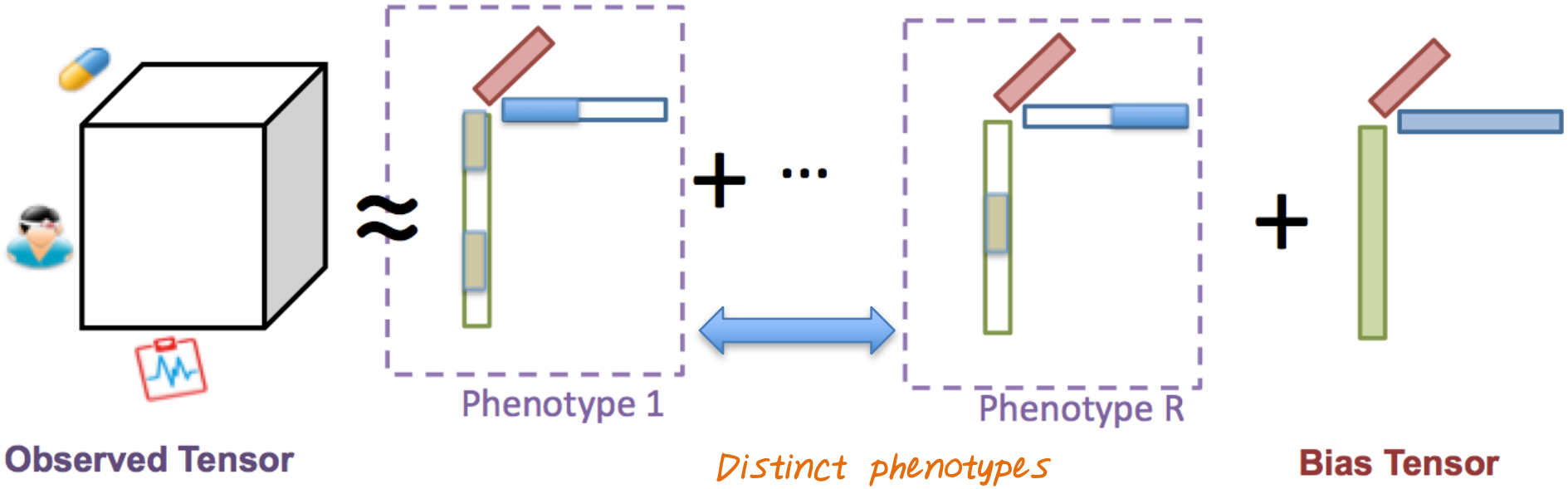
Ideal Phenotyping Algorithms

Guidance: incorporate medical knowledge

Non-overlap: discover distinct phenotypes

Robust: handle noisy and missing data

Rubik



Observed Tensor

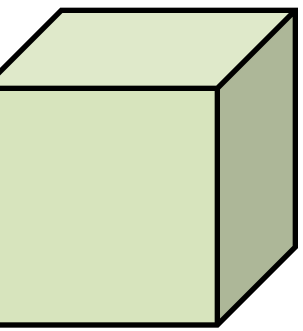
Distinct phenotypes

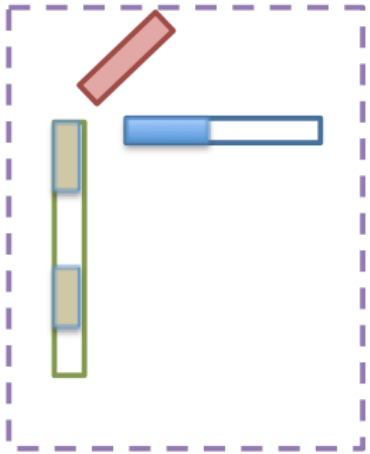
Bias Tensor

Complete
Missing data

Align with medical knowledge

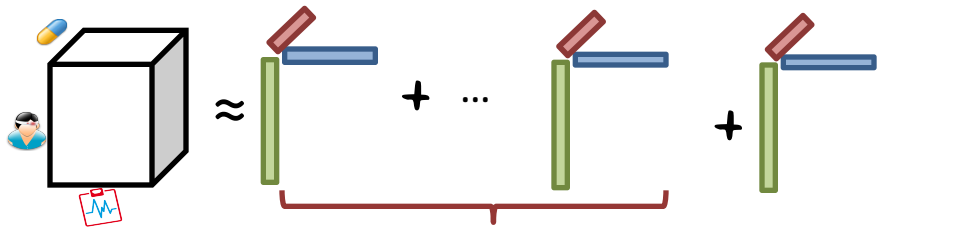
Full Tensor





Scalable

Rubik Model Formulation



X: Tensor T: Interaction tensor C: Bias tensor

$$\min_{\mathcal{X},\mathcal{T},\mathcal{C}} \quad \Psi = \|\mathcal{X} - \mathcal{C} - \mathcal{T}\|_F^2$$

Factorization error

$$\text{s.t.} \quad \mathcal{P}_\Omega(\mathcal{X}) = \mathcal{P}_\Omega(\mathcal{O})$$

O: Observed tensor

$$\mathcal{T} = [\![\mathbf{A}^{(1)}; \cdots ; \mathbf{A}^{(N)}]\!] \in \Omega_\mathcal{T}, \quad \mathcal{C} = [\![\mathbf{u}^{(1)}; \cdots ; \mathbf{u}^{(N)}]\!] \in \Omega_\mathcal{C}$$

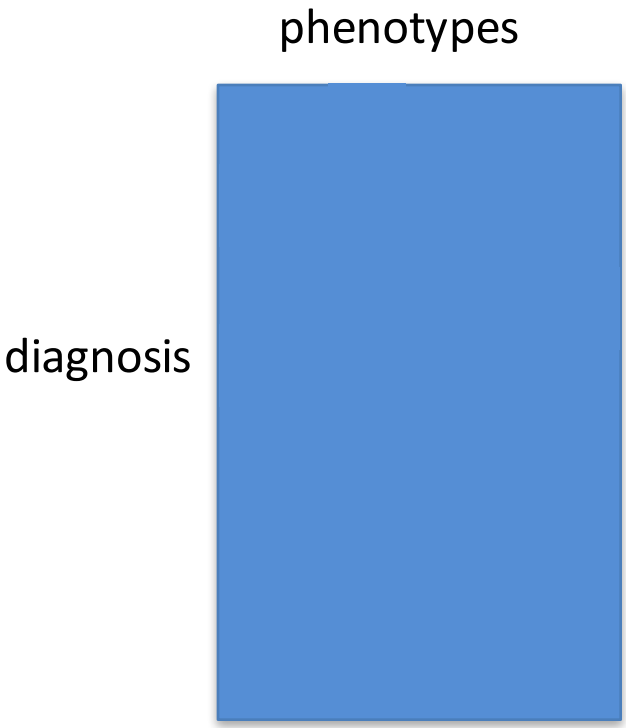
$$\Omega_\mathcal{T} = \Omega_{A_1} \times \cdots \times \Omega_{A_N}, \ \Omega_{A_n} = \{\mathbf{A} \in \{0\} \cup [\gamma_n, +\infty)^{I_n \times R}\}$$

Sparsity

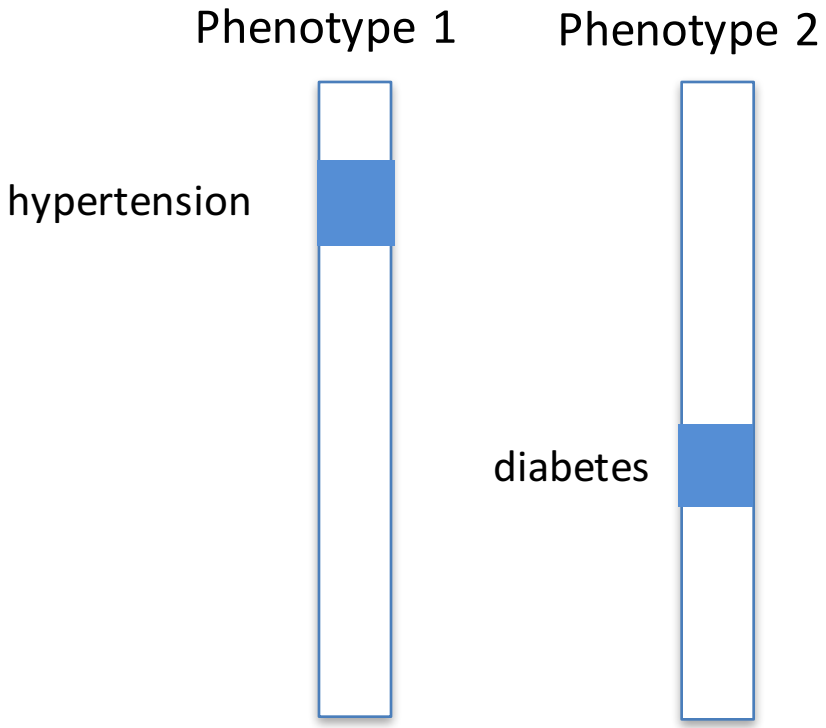
$$\Omega_\mathcal{C} = \Omega_{u_1} \times \cdots \times \Omega_{u_N}, \ \Omega_{u_n} = \{\mathbf{u} \in [0, +\infty)^{I_n \times 1}\}$$

Nonnegativity

Guidance Information



Factor matrix $\mathbf{A}$



Guidance knowledge $\hat{\mathbf{A}}$

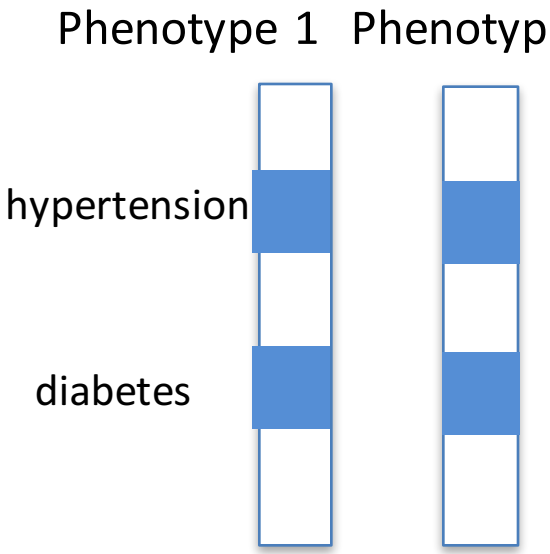
$$\mathbf{W} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$$

$$\min_{\mathbf{A}} \|(\mathbf{A} - \hat{\mathbf{A}})\mathbf{W}\|_F^2$$

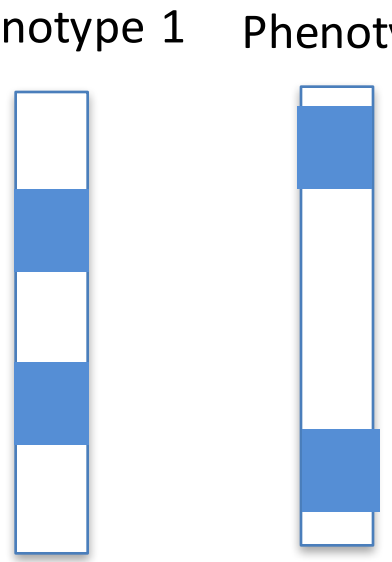
Guidance is limited

Pairwise Constraints

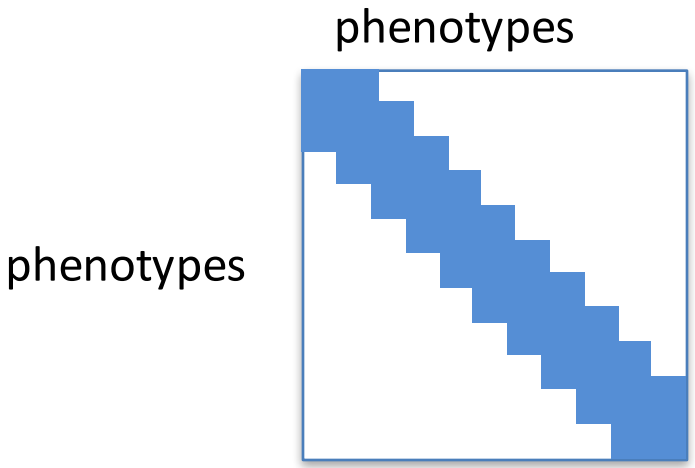
- We can penalize the cases where phenotypes have overlapping dimensions.



Overlapping case



Non-overlap case



Q: similarity matrix

$$\min_{\mathbf{A}} \|\mathbf{Q} - \mathbf{A}^T\mathbf{A}\|_F^2$$

Rubik Overview

Objective

$$\Psi = \|\mathcal{X} - \mathcal{C} - \mathcal{T}\|_F^2 + \frac{\lambda_a}{2}\|(\mathbf{A}^{(p)} - \hat{\mathbf{A}}^{(p)})\mathbf{W}\|_F^2 + \frac{\lambda_q}{2}\|\mathbf{Q} - \mathbf{B}^{(k)T}\mathbf{A}^{(k)}\|_F^2$$

$$\mathbf{A}^{(k)} = \mathbf{B}^{(k)} \qquad \mathbf{v}^{(n)} = \mathbf{u}^{(n)}$$

Lagrange Function (ADMM)

$$\mathcal{L} = \Psi + \sum_{n=1}^{N}(\langle \mathbf{p}^{(n)}, \mathbf{v}^{(n)} - \mathbf{u}^{(n)}\rangle + \frac{\eta}{2}\|\mathbf{v}^{(n)} - \mathbf{u}^{(n)}\|_F^2) + \sum_{n=1}^{N}(\langle \mathbf{Y}^{(n)}, \mathbf{B}^{(n)} - \mathbf{A}^{(n)}\rangle + \frac{\mu}{2}\|\mathbf{B}^{(n)} - \mathbf{A}^{(n)}\|_F^2)$$

Block coordinate scheme

p and Y are Lagrange multipliers

Block coordinate scheme

Update interaction tensor (T)

$$\min_{\mathbf{A}^{(n)}} \|\mathbf{A}^{(n)}\mathbf{\Pi}^{(n)T} - \mathbf{R}_{(n)}\|_F^2$$
$$+ \frac{\lambda_q}{2}\|\mathbf{Q} - \mathbf{B}_t^{(n)T}\mathbf{A}^{(n)}\|_F^2 + \frac{\lambda_a}{2}\|(\mathbf{A}^{(n)} - \hat{\mathbf{A}}^{(n)})\mathbf{W}\|_F^2$$
$$+ \frac{\mu_t}{2}\|\mathbf{A}^{(n)} - \mathbf{B}_t^{(n)} - \mathbf{Y}_t^{(n)}/\mu_t\|_F^2$$
$$\mathbf{B}_{t+1}^{(n)} = \begin{cases} \mathbf{A}_{t+1}^{(n)} + \frac{1}{\mu_t}\mathbf{Y}_t^{(n)} & \text{if } \gamma_n \leq \mathbf{A}_{t+1}^{(n)} + \frac{1}{\mu_t}\mathbf{Y}_t^{(n)} \\ 0 & \text{otherwise} \end{cases}$$

Update bias tensor (C)

$$\min_{\mathbf{u}^{(n)}} \|\mathbf{u}^{(n)}(\mathbf{\Lambda}^{(n)})^T - \mathbf{E}_{(n)}\|_F^2 + \frac{\eta_t}{2}\|\mathbf{u}^{(n)} - \mathbf{v}^{(n)} - \mathbf{p}^{(n)}/\eta_t\|_F^2$$
$$\mathbf{v}_{t+1}^{(n)} = \max(0, \mathbf{u}_{t+1}^{(n)} + \frac{1}{\eta_t}\mathbf{p}_t^{(n)})$$

Update Lagrange multipliers (p, Y)

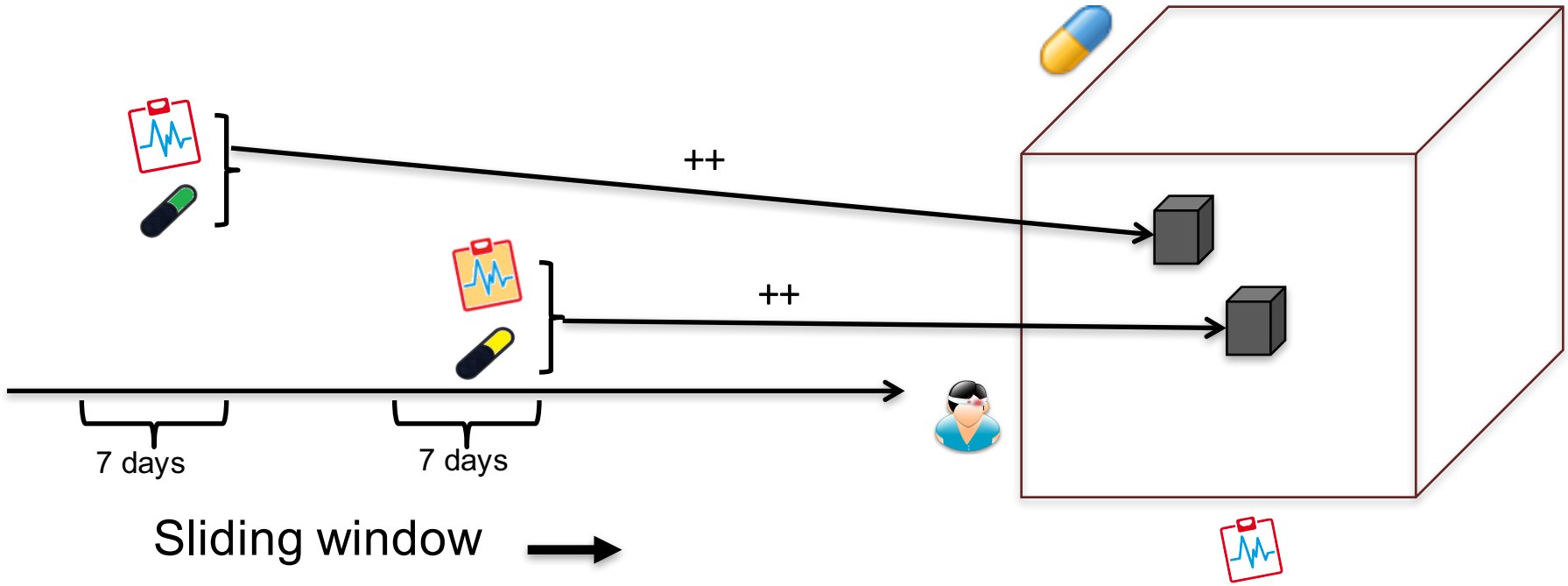
$$\mathbf{Y}_{t+1}^{(n)} = \mathbf{Y}_t^{(n)} + \mu_t(\mathbf{B}_{t+1}^{(n)} - \mathbf{A}_{t+1}^{(n)})$$
$$\mathbf{p}_{t+1}^{(n)} = \mathbf{p}_t^{(n)} + \eta_t(\mathbf{v}_{t+1}^{(n)} - \mathbf{u}_{t+1}^{(n)})$$

Update full tensor (X)

$$\mathcal{X}_{t+1} = \mathcal{P}_{\Omega^c}(\mathcal{T}_{t+1} + \mathcal{C}_{t+1}) + \mathcal{P}_{\Omega}(\mathcal{O})$$

EXPERIMENTS

Constructing Tensor



Time series data → Binary Tensor

Co-occurrences of events → tensor elements

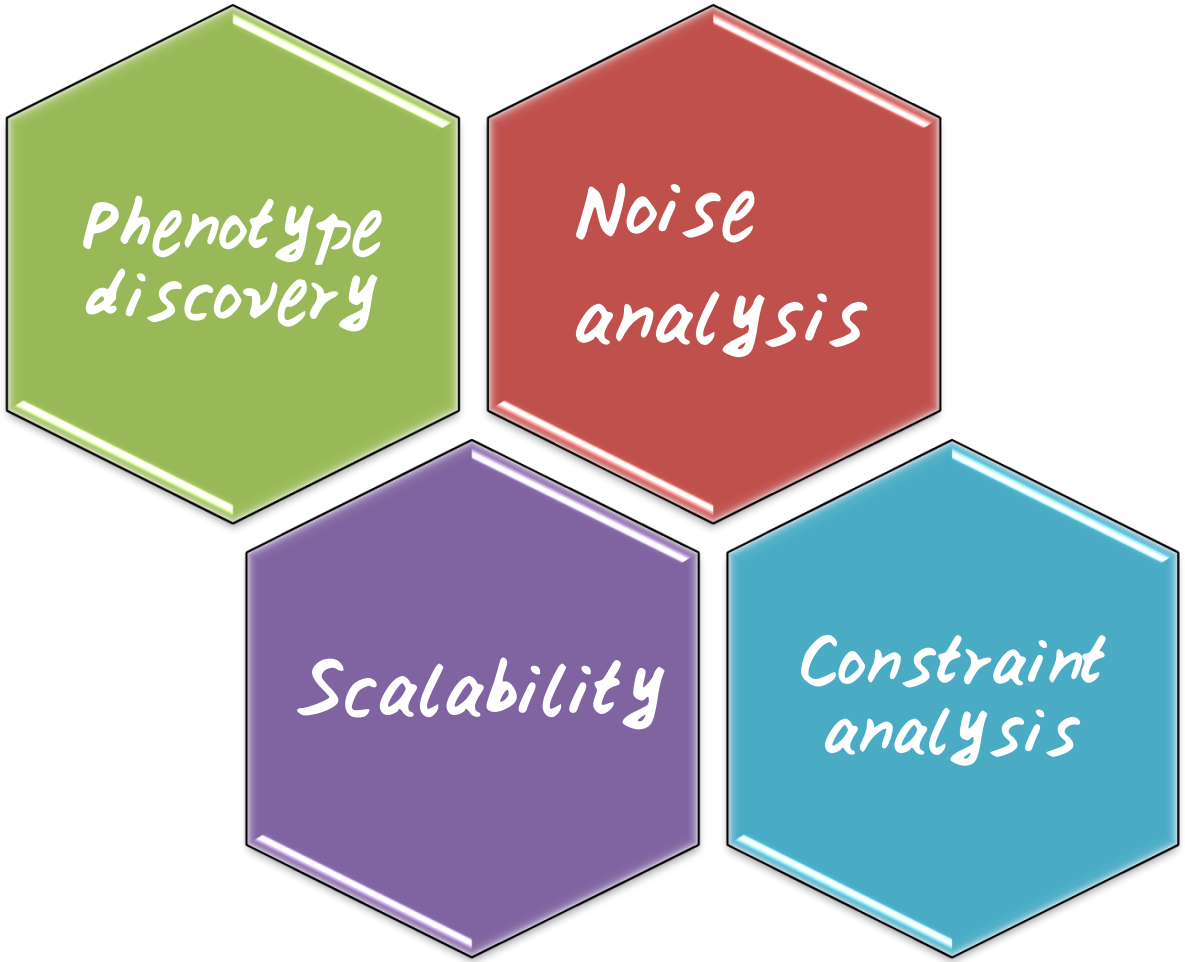
Datasets

- Vanderbilt: 7,744 patients by 1,059 diagnoses by 501 medications

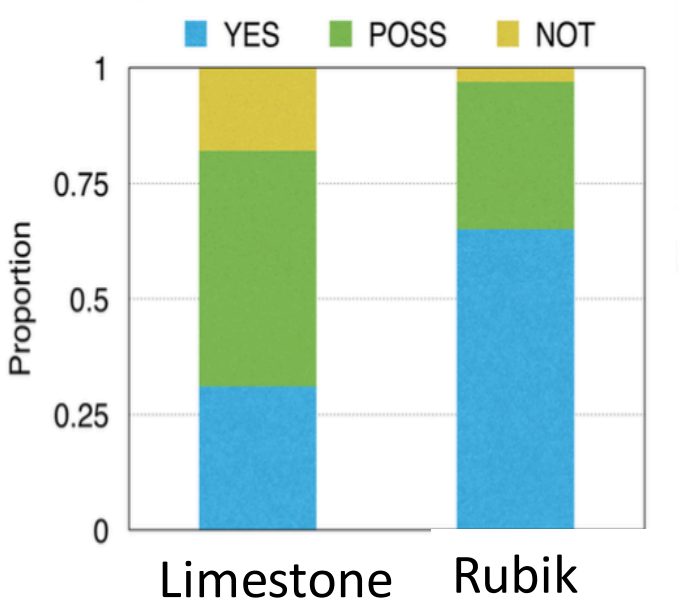
- CMS: 472,645 patients by 11,424 diagnoses by 262,312 medication events.

1,416,352,720,661,760 elements

Experiment Goals



Phenotype Discovery: Meaningful phenotypes



Survey medical experts on 30 phenotypes

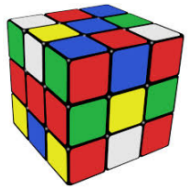
- YES = clinically meaningful
- POSS = possibly meaningful
- NOT = not meaningful

Figure 1: A comparison of the meaningfulness of the phenotypes discovered by Marble and Rubik.

Rubik generates more meaningful phenotypes

Phenotype Discovery: Knowledge-guided Sub-phenotypes

Limestone 

Rubik 

Diagnoses	Medications
Chronic kidney disease	Central sympatholytics
Hypertension	Angiotensin receptor blockers
Unspecified anemias	ACE inhibitors
Fluid electrolyte imbalance	Immunosuppressants
Type 2 diabetes mellitus	Loop diuretics
Other kidney disorders	Gabapentin

Table 6: An example of a Marble-derived phenotype.

A. Metabolic syndrome phenotype

Diagnoses	Medications
Hypertension	Calcineurin inhibitors
Chronic kidney disease	Insulin
Ischemic heart disease	Immunosuppressants
Disorders of lipoid metabolism	ACE inhibitors
Anemia of chronic disease	Calcium channel blockers
	Antibiotics
	Statins
	Calcium
	Cox-2 inhibitors

B. Secondary hypertension phenotype

Diagnoses	Medications
Secondary hypertension	Class V antiarrhythmics
Fluid & electrolyte imbalance	Salicylates
Unspecified anemias	Antianginal agents
Hypertension	ACE inhibitors
	Calcium channel blockers
	Immunosuppressants

Phenotype Discovery: More Distinct Phenotypes

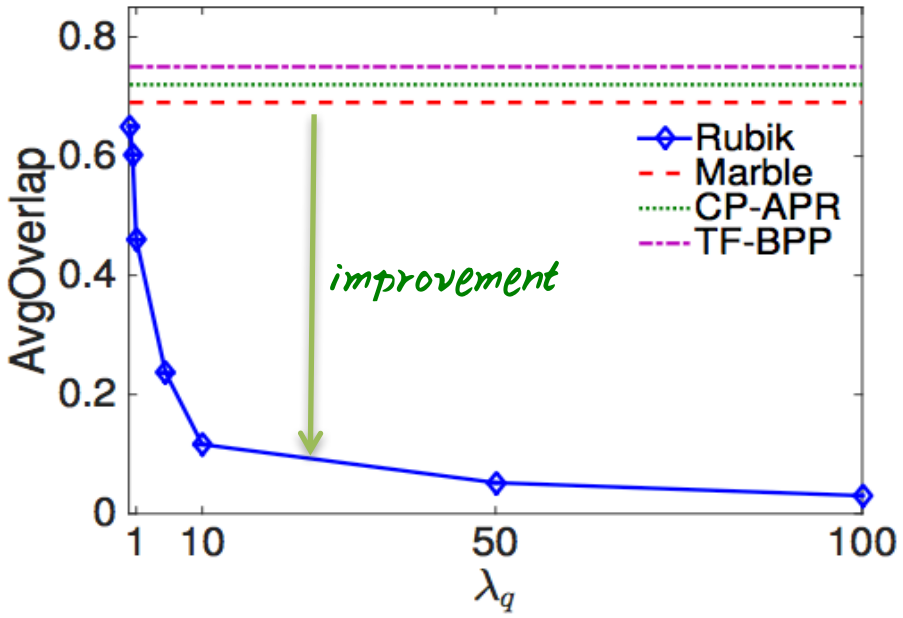


Figure 2: The average level of overlap in the phenotypes as a function of the pairwise constraint coefficient λ_q.

Pairwise constraints lead to more distinct phenotypes

Noise analysis: Missing Data Analysis

Generate missing data:

randomly set the observed values to be 0

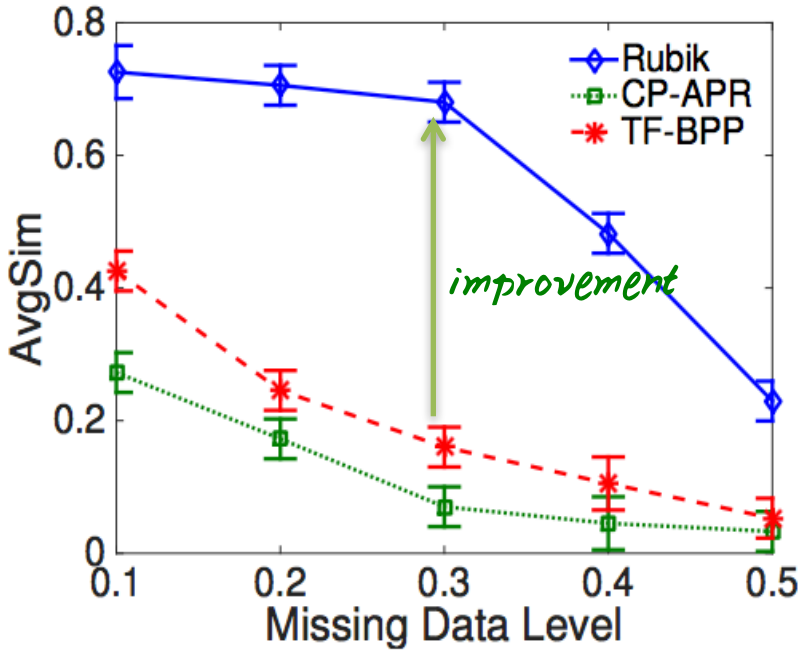


Figure 3: An average similarity comparison of different methods as a function of the missing data level.

Noise analysis: Noise Analysis

Generate noise:

randomly set the unobserved entries to be 1

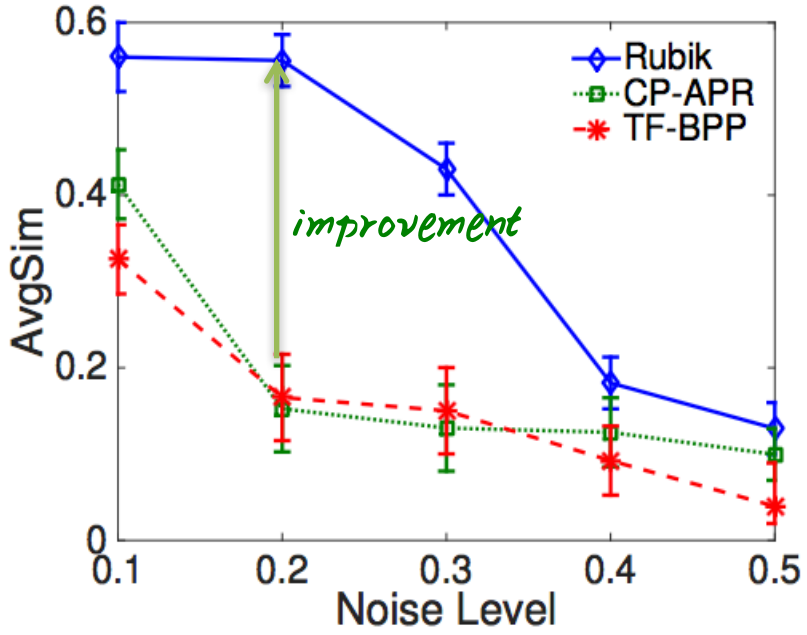


Figure 4: An average similarity comparison of different methods as a function of the noise level.

Noise analysis: Incorrect Guidance

Generate incorrect guidance:

randomly set entries in guidance matrix to be 1

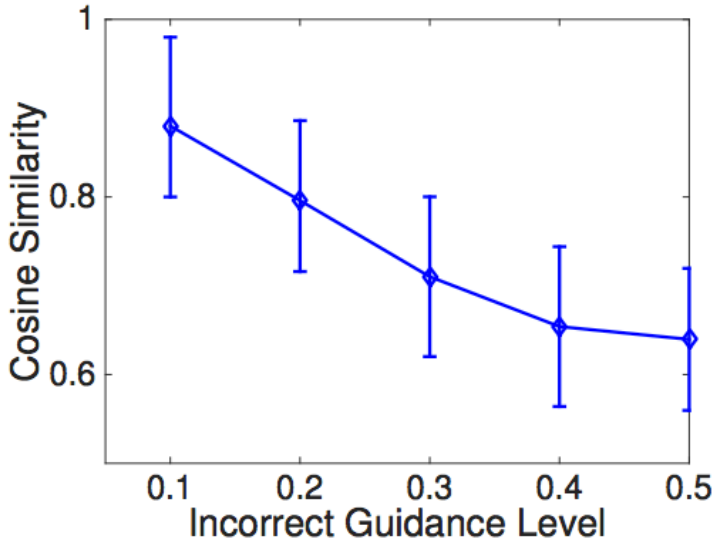


Figure 5: The similarity between the true solution and the solution under incorrect guidance as a function of the incorrect guidance level.

Scalability

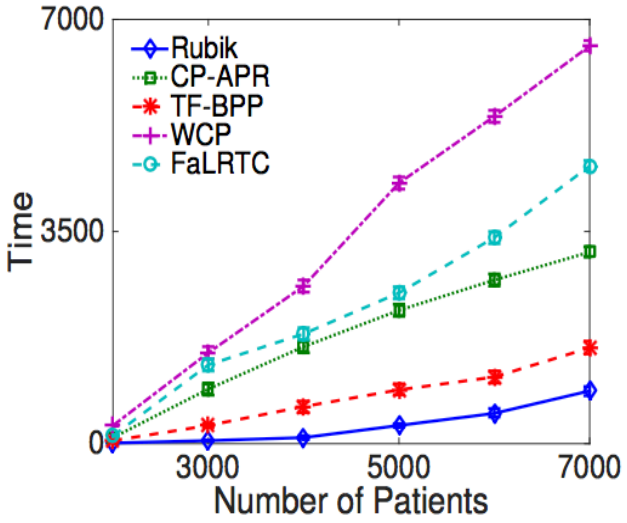


Figure 7: A runtime comparison of different methods on the *Vanderbilt* dataset as a function of the number of patients.

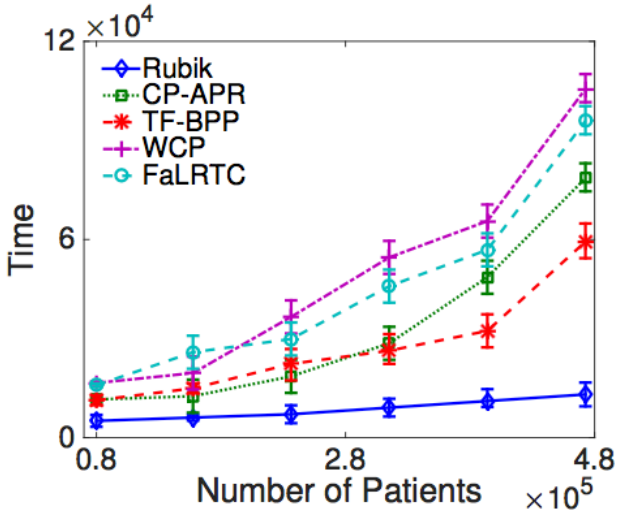


Figure 8: A runtime comparison of different methods on the *CMS* dataset as a function of the number of patients.

Rubik is at least six times faster than competitors

Constraints Analysis

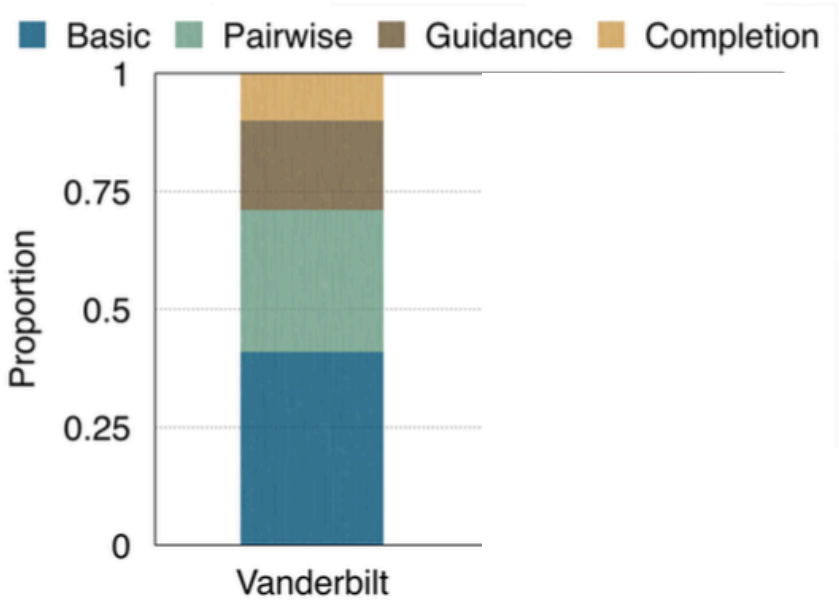


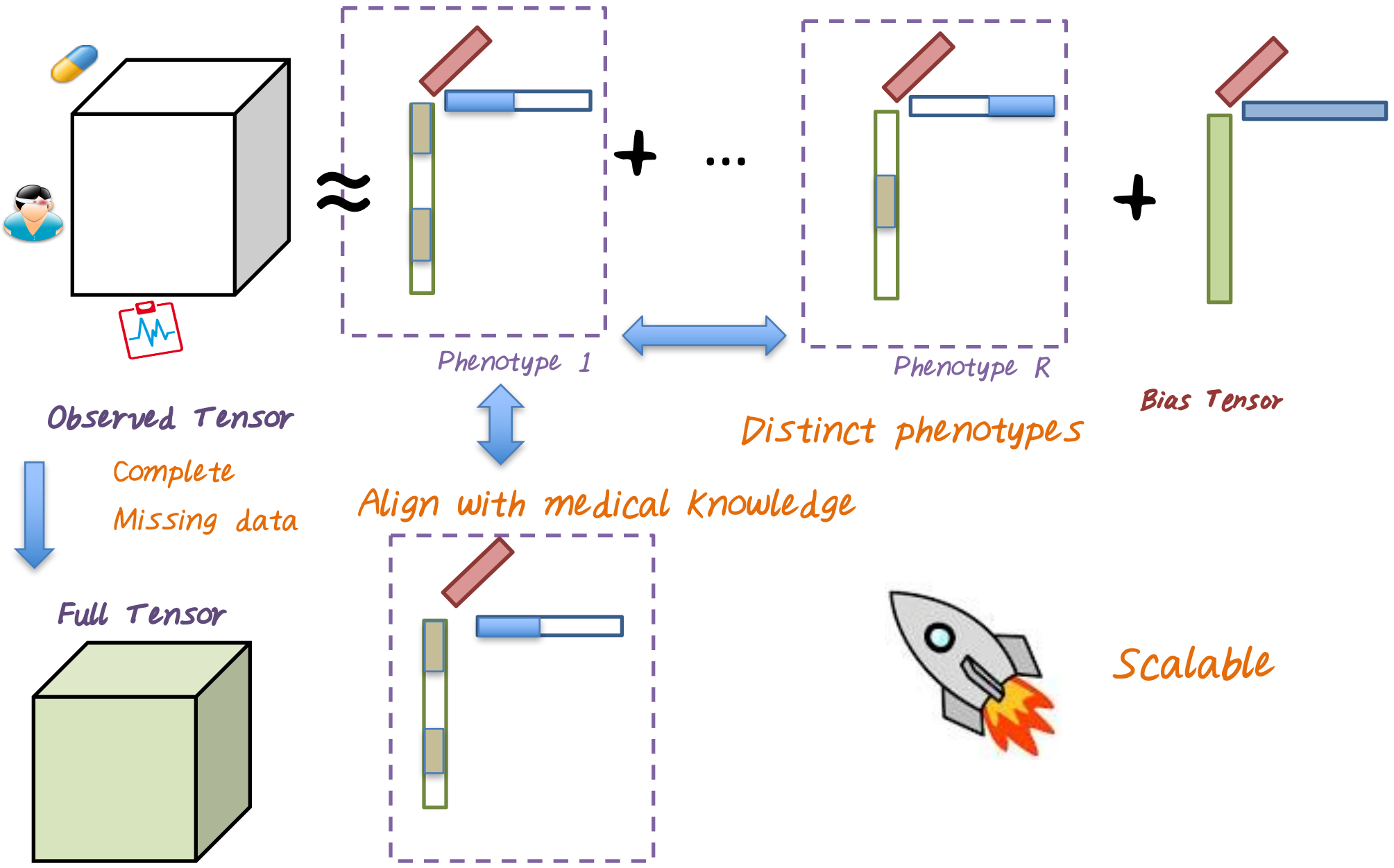
Figure 9: Proportion of contribution of each constraint.

- All constraints are important!
- Pairwise constraint provides the largest boost

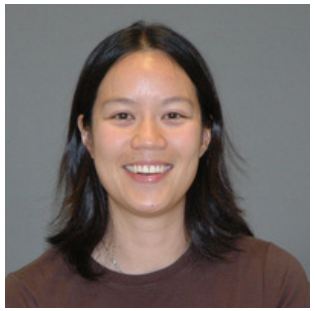
Conclusions

- The resulting phenotypes are concise, distinct, and interpretable
- RubiK can also incorporate guidance from medications and patients
- RubiK is robust to noisy and missing data
- RubiK is scalable

Summary: Rubik



Collaborators



Joyce Ho

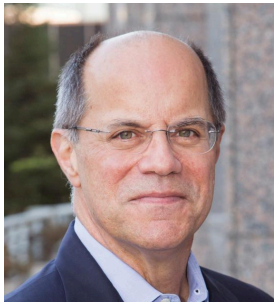
EMORY UNIVERSITY

Joydeep Ghosh

THE UNIVERSITY OF TEXAS AT AUSTIN

Steve Steinhubl

THE SCRIPPS RESEARCH INSTITUTE

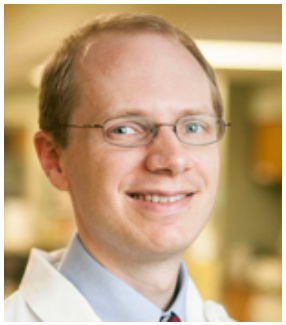


Buzz Stewart

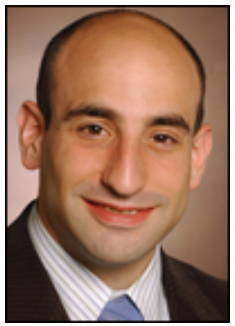
Sutter Health Plus
Your Health Plan

Abel Kho

NORTHWESTERN UNIVERSITY
QUAECUMQUE SUNT VERA
1851



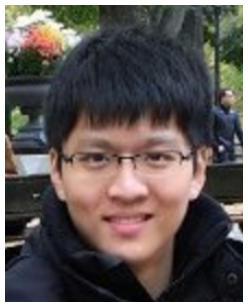
Josh Denny



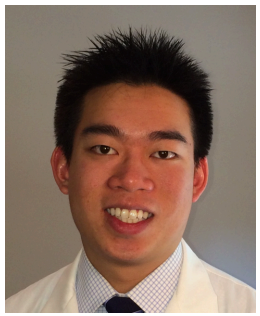
Brad Malin

You Chen

VANDERBILT UNIVERSITY



Yichen Wang



Robert Chen

Kimis Perros

Rich Vuduc

Georgia Institute of Technology

Unsupervised Computational Phenotyping using Tensor Factorization

Jimeng Sun

jsun@cc.gatech.edu

Georgia Tech College of Computing
Computational Science and Engineering

SunLab
Enlighten the World

Computational Phenotyping with Unstructured Data

Yuan Luo

Assistant Professor

Department of Preventive Medicine

Departments of IEMS and EECS (Courtesy)

Northwestern University

yuan.luo@northwestern.edu

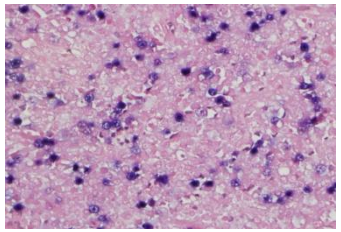
10/26/2016

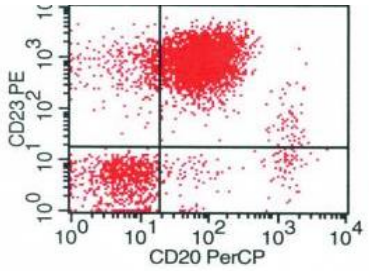
Why Using Unstructured Data

Raw

Present

Raw Observations





Atomic data

127	101	19	103
3.8	21	1.86	

Interpretive comments

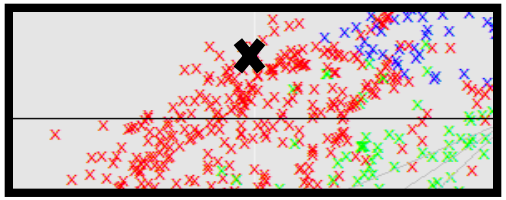
"Patient creatinine values are increasing…"

Diagnoses

"EBV-positive immunoblastic reaction, consistent with infectious mononucleosis"

Computational Pathology

Integrative Information



Predictive Information

"Given this patient's test results and clinical data, administration of vancomycin will improve odds of survival from 37% to 87%."

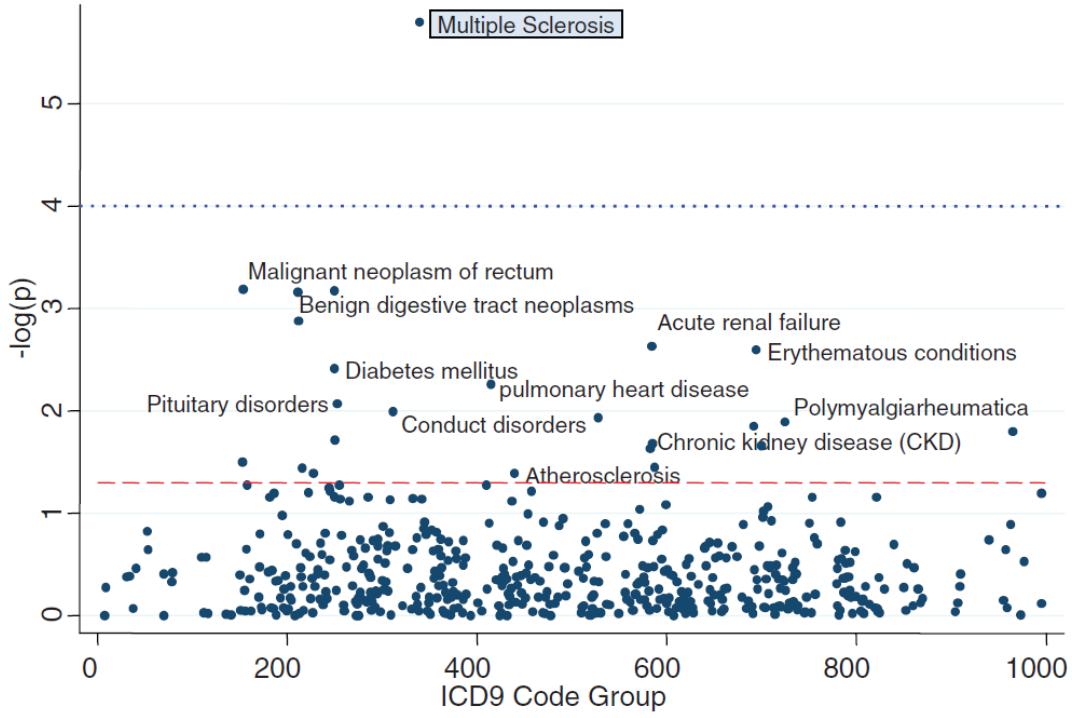
Processed

Why Using Unstructured Data

- Much of the EHR content is locked up in the narrative text
  - Physicians' and nurses' notes
  - Referring letters
  - Specialists' reports
  - Discharge summaries
  - Communications between the physician and patient
- Extracting a phenotype from complex and heterogeneous data sources needs clinical narratives to augment structured data
  - Laboratory values
  - Medication prescriptions
  - Vital sign records.

Clinical NLP for Computational Phenotyping

- Improve the selection algorithms for identifying cohorts of past patients appropriate for clinical studies (Liao et al. 2010)
- NLP that extracts medical concepts and relations can support applications such as PheWAS (Denny et al. 2010)



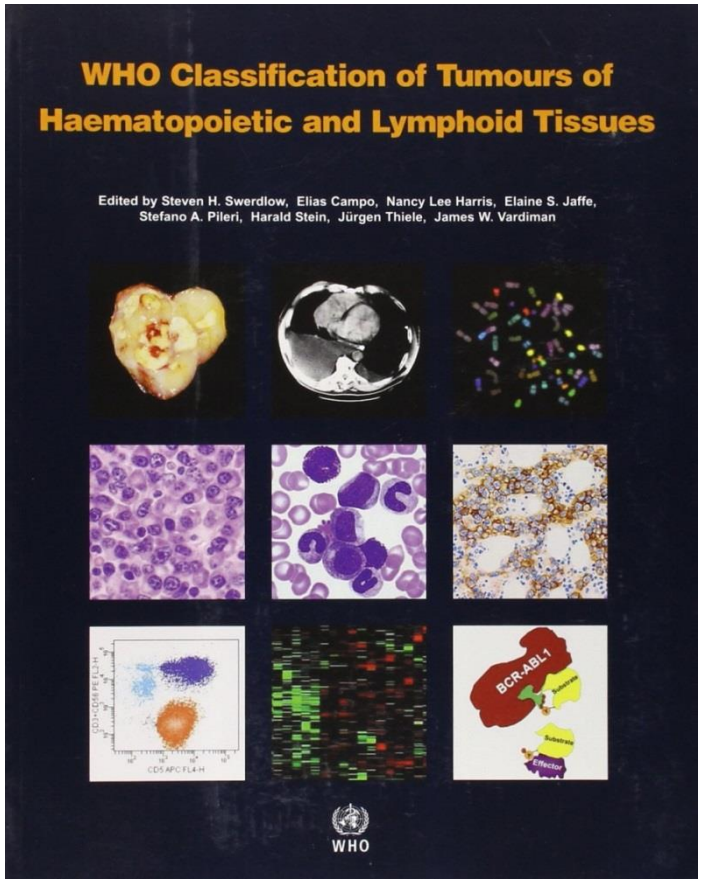
Clinical NLP for Computational Phenotyping

- Algorithms have a wide range of complexity and may use heterogeneous data sources
- Keywords search or customized rules to extract phenotypes
- named entity or concept recognition systems to extract information (e.g., symptoms, drugs) determining patient phenotypes (Collier et al. 2015)
- Statistical machine learning to train phenotype classification model (Zeng et al. 2006, Bejan et al. 2012, Lehman et al. 2012, Luo et al. 2015)
- Graph mining to learn relation features that characterize patient phenotypes (Herskovic et al. 2012, Luo et al. 2014)

Clinical NLP for Computational Phenotyping

- Learning from multiple data modalities
- Combining unstructured and structured data (Peissig et al. 2012)
- Combining Pubmed knowledge (Zhao et al. 2011)
- combining heterogeneous EHR database (Kho et al. 2012)
- Combining EHR with billing data (Xu et al. 2011)

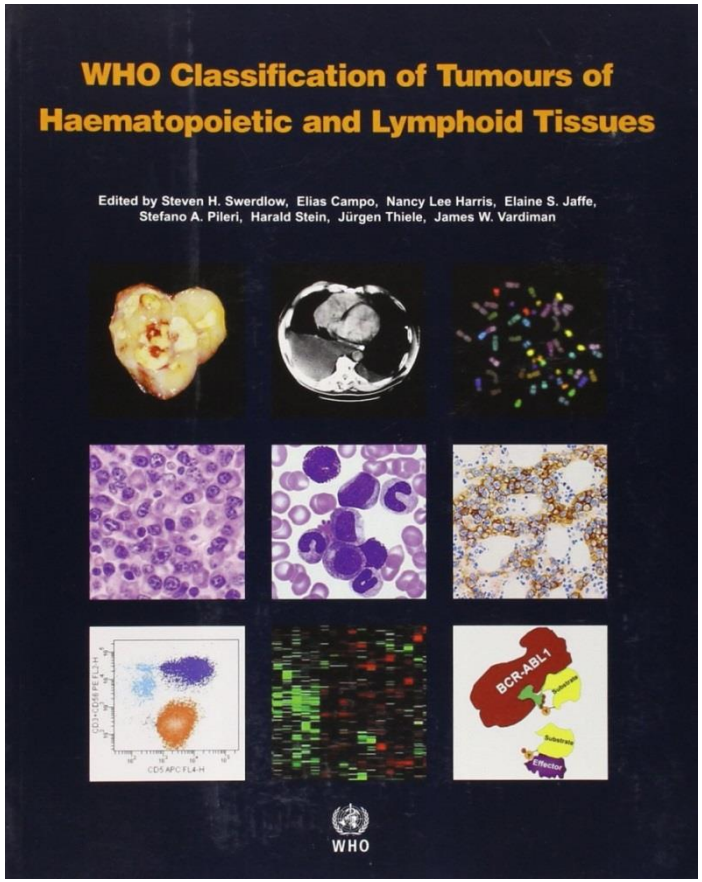
Automatically Identify Lymphoma Subtypes



Y Luo, A Sohani, E Hochberg and P Szolovits. **Automatic Lymphoma Classification with Sentence Subgraph Mining from Pathology Reports**. *JAMIA 2014 21(5):824-832.*

Y Luo, Y Xin, E Hochberg, R Joshi, O Uzuner, P Szolovits. **Subgraph Augmented Non-Negative Tensor Factorization (SANTF) for Modeling Clinical Text**. *JAMIA 2015 doi: 10.1093/jamia/ocv016.*

Automatically Identify Lymphoma Subtypes



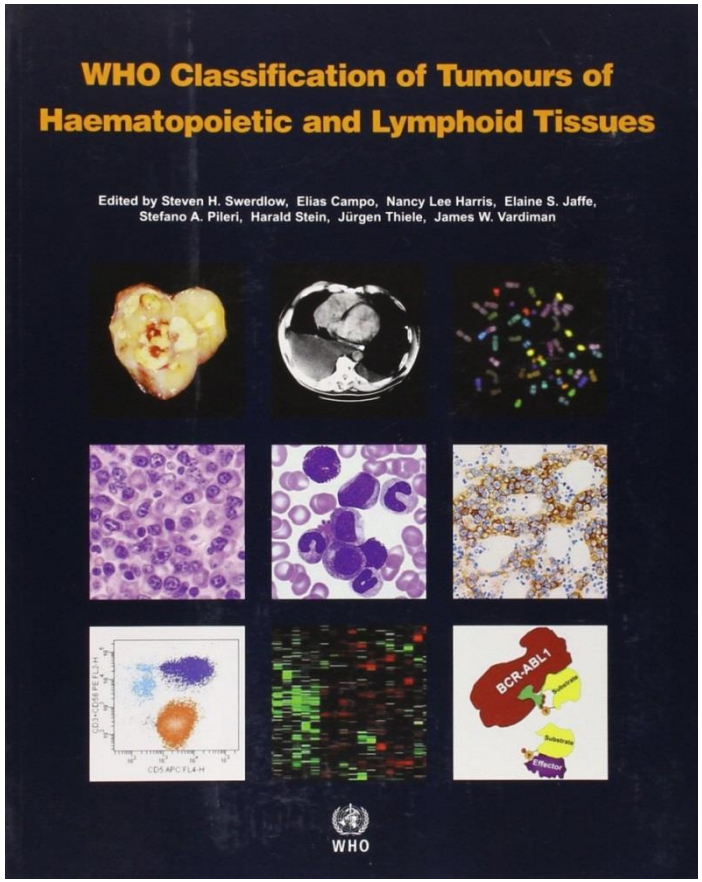
Guideline Construction and revision is "expert intensive"

- 1400 patient cases, Caucasian focus
- >50 subtypes
- 8 member steering committee
- 130 pathologists & hematologists

Y Luo, A Sohani, E Hochberg and P Szolovits. **Automatic Lymphoma Classification with Sentence Subgraph Mining from Pathology Reports**. *JAMIA 2014 21(5):824-832.*

Y Luo, Y Xin, E Hochberg, R Joshi, O Uzuner, P Szolovits. **Subgraph Augmented Non-Negative Tensor Factorization (SANTF) for Modeling Clinical Text**. *JAMIA 2015 doi: 10.1093/jamia/ocv016.*

Automatically Identify Lymphoma Subtypes



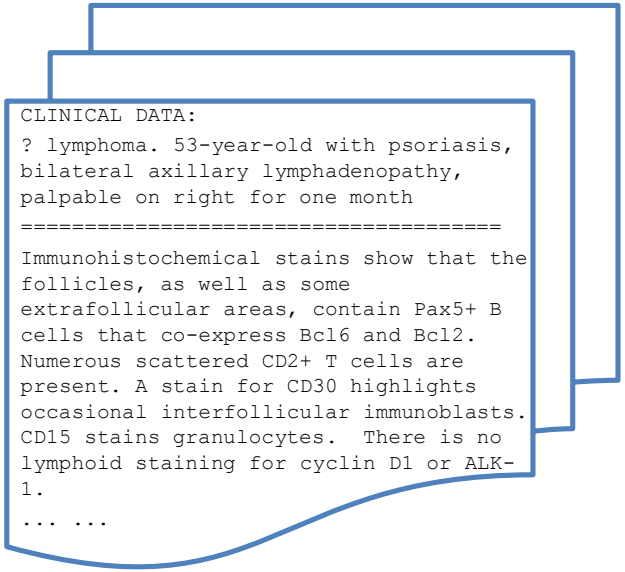
Guideline Construction and revision is "expert intensive"

- 1400 patient cases, Caucasian focus
- >50 subtypes
- 8 member steering committee
- 130 pathologists & hematologists

Y Luo, A Sohani, E Hochberg and P Szolovits. **Automatic Lymphoma Classification with Sentence Subgraph Mining from Pathology Reports**. *JAMIA 2014 21(5):824-832.*

Y Luo, Y Xin, E Hochberg, R Joshi, O Uzuner, P Szolovits. **Subgraph Augmented Non-Negative Tensor Factorization (SANTF) for Modeling Clinical Text**. *JAMIA 2015 doi: 10.1093/jamia/ocv016.*

Traditional Medical Classification/Clustering Task



Distinguishing Lymphoma Subtypes

DLBCL OR **Follicular Lymphoma** OR **Hodgkin's Lymphoma**

ML algorithm

Preprocessing

BCL2	BCL6	CD10	large cells	positive	negative	...	Class label
1	1	0	1	1	0		y
0	0	1	1	0	1		n
1	1	1	1	1	1		n

DLBCL: Diffuse Large B-cell Lymphoma

Traditional Medical Classification/Clustering Task

2d modeling

Not intuitive to model interactions

CD10 positive on large cells

or

CD10 negative on large cells?

Columns: features

Rows:	BCL2	BCL6	CD10	large cells	positive	negative	…	**Class label**
Patient 1	1	1	0	1	1	0		y
Patient 2	0	0	1	1	0	1		n
Patient 3	1	1	1	1	1	1		n

Traditional Medical Classification/Clustering Task

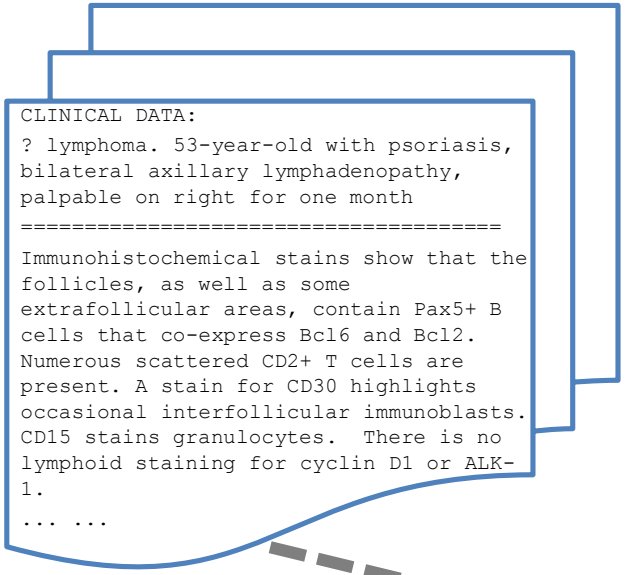
WHO guideline: The neoplastic cells express pan B-cell markers such as CD19, CD20, CD22 and CD79a, but may lack one or more of these.

Interpretability

Panel of immunophenotypic test result
Neoplastic cells express CD19
Neoplastic cells express CD20
Neoplastic cells express CD22
Neoplastic cells express CD79a
… …

BCL2	BCL6	CD10	large cells	positive	negative	…	Class label
1	1	0	1	1	0		y
0	0	1	1	0	1		n
1	1	1	1	1	1		n

Traditional Medical Classification/Clustering Task



Distinguishing Lymphoma Subtypes

DLBCL OR **Follicular Lymphoma** OR **Hodgkin's Lymphoma**

Training data

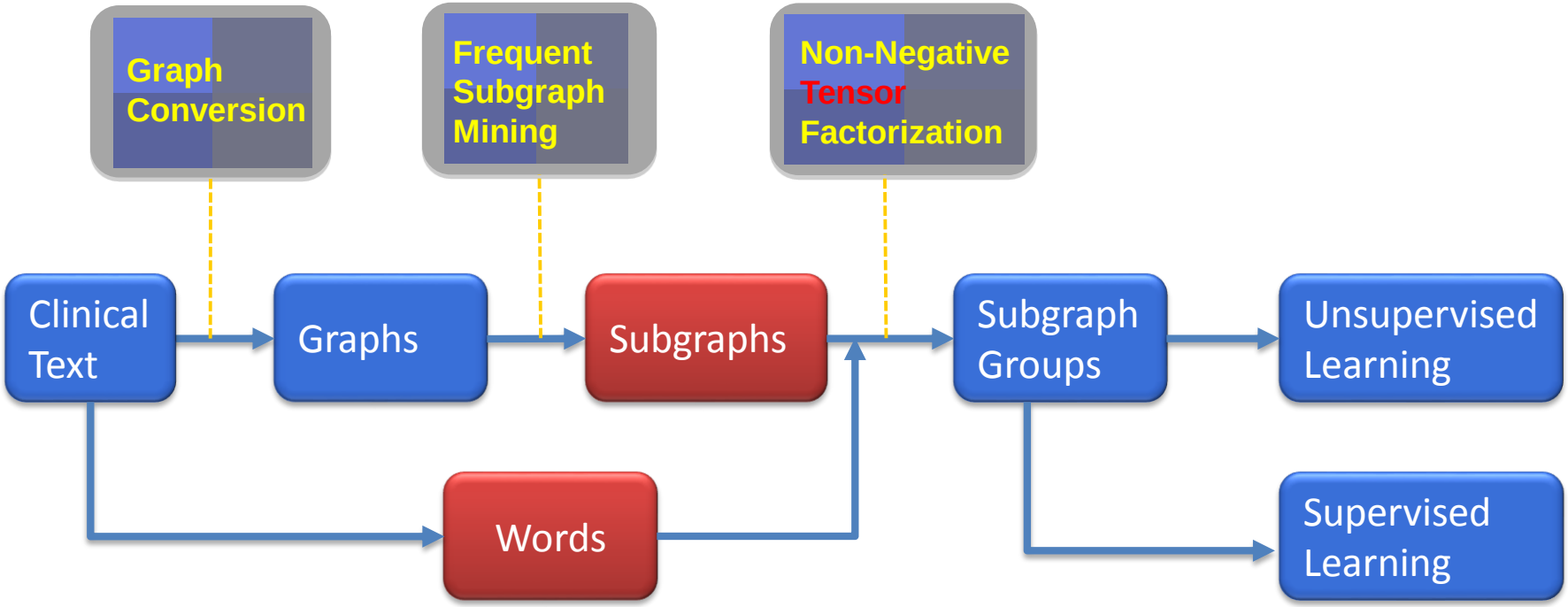
Expert annotation expensive

Evolving guideline

DLBCL: Diffuse Large B-cell Lymphoma

SANTF Overview

Subgraph Augmented Non-Negative Tensor Factorization



Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."

Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."
- The sentence tells relationships among procedures, cells, and immunologic factors

Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."
- The sentence tells relationships among procedures, cells, and immunologic factors
- Feature choices
  - Words
  - UMLS (Unified Medical Language System) concepts, e.g. LCA and CD45

Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."
- The sentence tells relationships among procedures, cells, and immunologic factors
- Feature choices
  - Words
  - UMLS (Unified Medical Language System) concepts, e.g. LCA and CD45
- Can we do better? Relations?

Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."
- The sentence tells relationships among procedures, cells, and immunologic factors
- Feature choices
  - Words
  - UMLS (Unified Medical Language System) concepts, e.g. LCA and CD45
- Can we do better? Relations?

Graph representation is the universal language for modeling relationships among flexible number of concepts

Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."

Two Phase Parsing

Two Phase Parsing Brief Outline

- Match token subsequences to UMLS (Unified Medical Language System) concepts
  - E.g., In situ hybridization

Two Phase Parsing Brief Outline

- Match token subsequences to UMLS (Unified Medical Language System) concepts
  - E.g., In situ hybridization

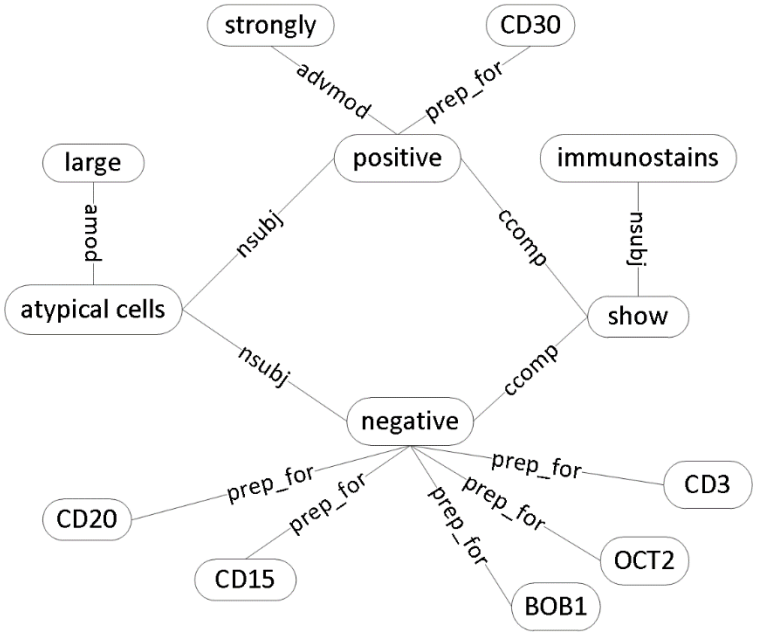
- Augment the Stanford Parser with UMLS specialist lexicon and Part-of-speech dictionary
- Apply the augmented Stanford Parser with grouped tokens as one token
  - E.g., “In situ hybridization” as noun

- Translate the dependency linkage into a graph representation

Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."

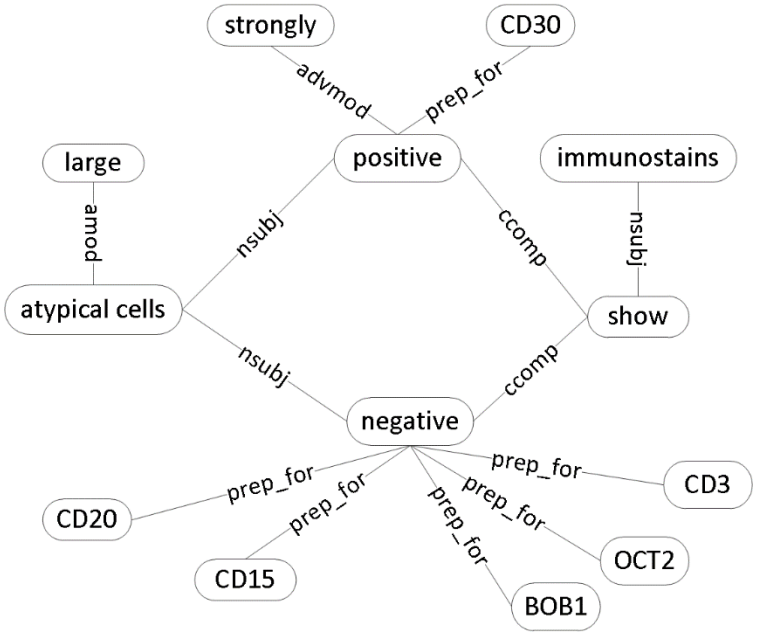
Two Phase Parsing



Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."

Two Phase Parsing

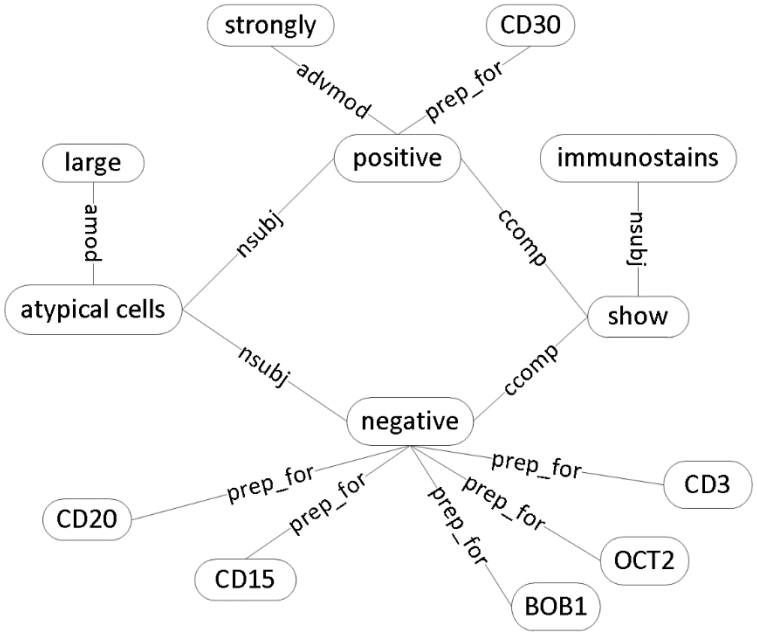


Important relations are likely to be repeated in pathology daily practice:
large atypical cells are positive for CD30 ⇒ sign of Hodgkin lymphoma etc. ⇒ frequently ordered test

Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."

Two Phase Parsing

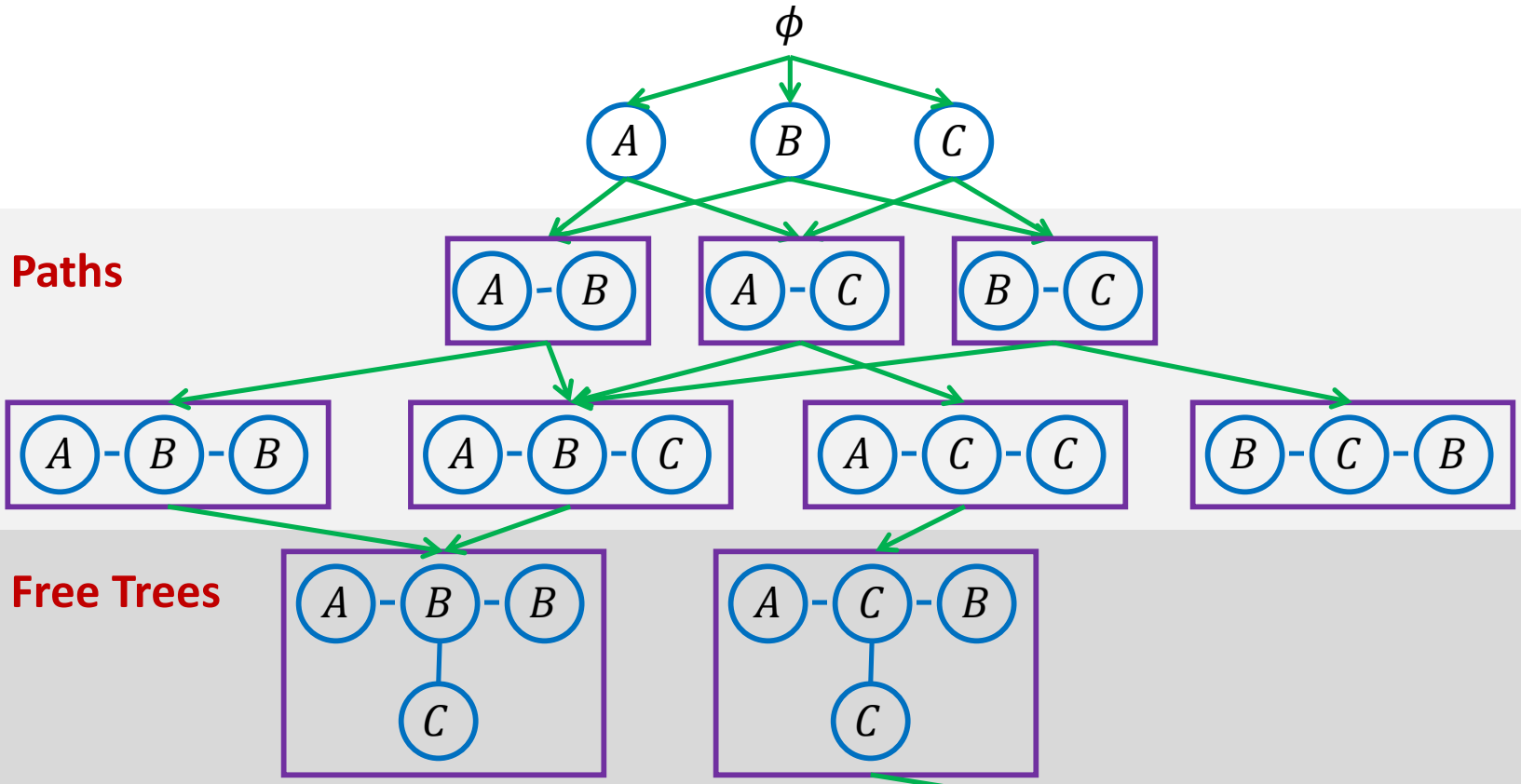


FSM

Subisomorphism filtering

FSM: frequent subgraph mining

Enumeration Order is Key to FSM Efficiency

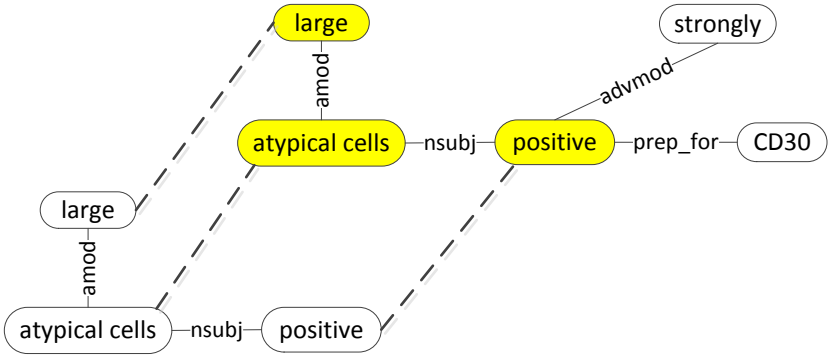


GASTON (Nijssen et al, 2005)

Larger subgraph cannot be frequent if part of it not

Subisomorphism Between Frequent Subgraphs

- Smaller frequent subgraphs isomorphic to larger ones and overwhelm their signals

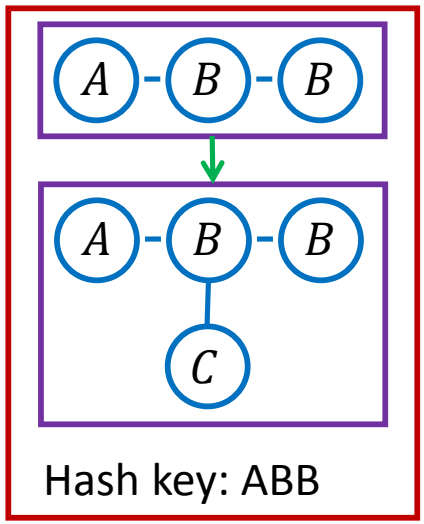


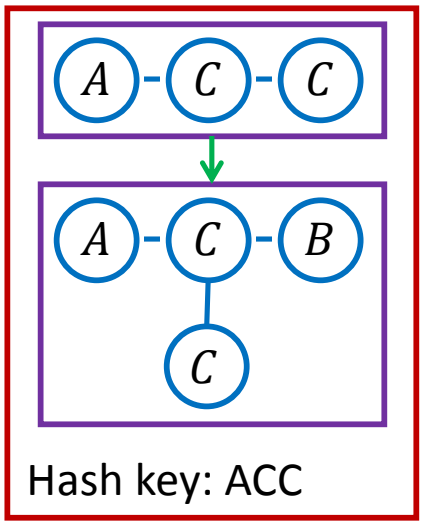
Subisomorphism Between Frequent Subgraphs

- Smaller frequent subgraphs isomorphic to larger ones and overwhelm their signals
- Only need to check subgraph pairs whose sizes differ by one node

Subisomorphism Between Frequent Subgraphs

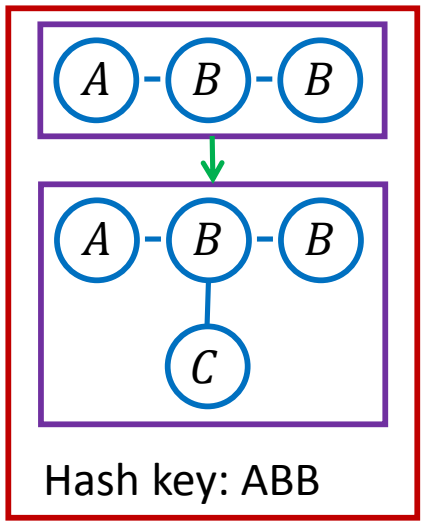
- Smaller frequent subgraphs isomorphic to larger ones and overwhelm their signals
- Only need to check subgraph pairs whose sizes differ by one node
- Hash the n-node subgraphs n times using n-1 node label subsets as keys
- Hash the (n-1)-node subgraphs using n-1 node label subsets as keys

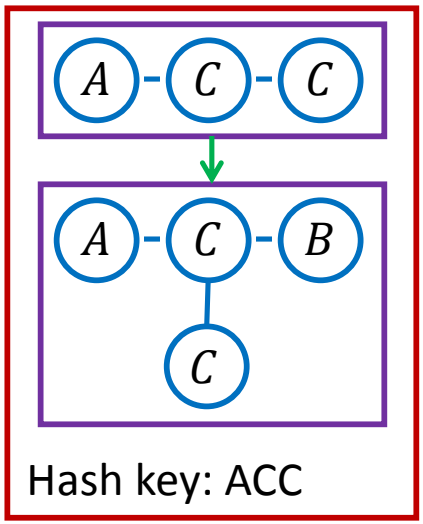




Subisomorphism Between Frequent Subgraphs

- Smaller frequent subgraphs isomorphic to larger ones and overwhelm their signals
- Only need to check subgraph pairs whose sizes differ by one node
- Hash the n-node subgraphs n times using n-1 node label subsets as keys
- Hash the (n-1)-node subgraphs using n-1 node label subsets as keys



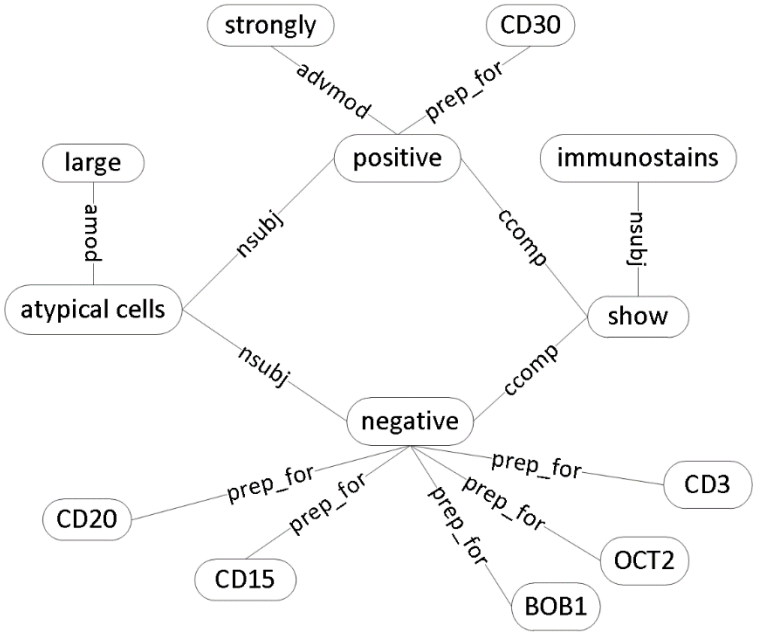


Over 100 X reduction in # subisomorphism

Representation of Narrative Sentences

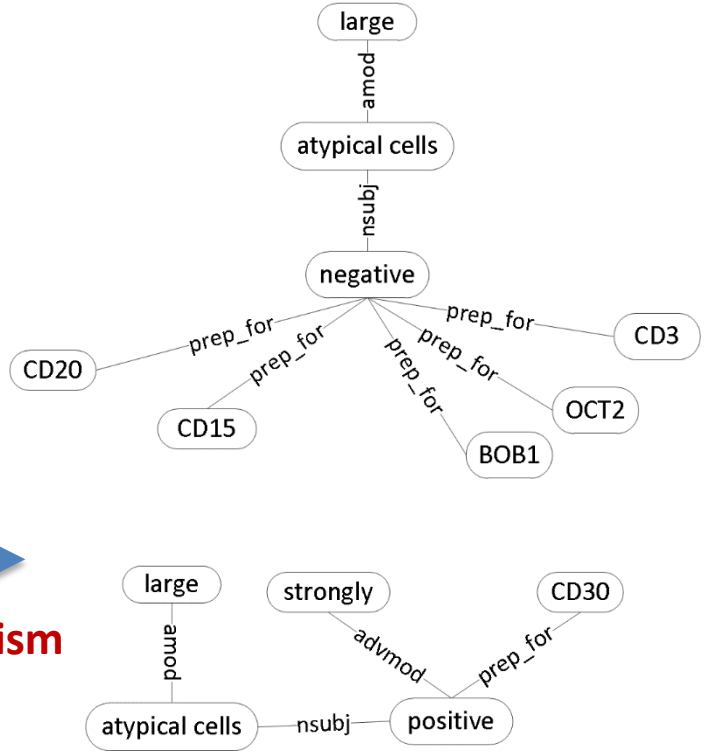
- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."

Two Phase Parsing



FSM

Subisomorphism filtering

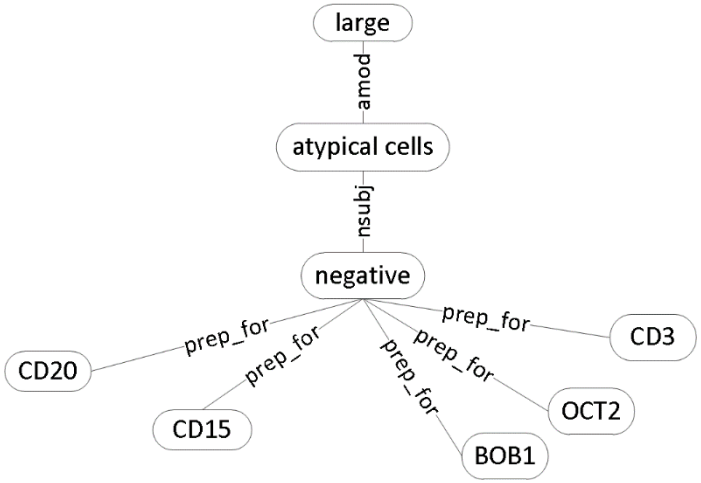


FSM: frequent subgraph mining

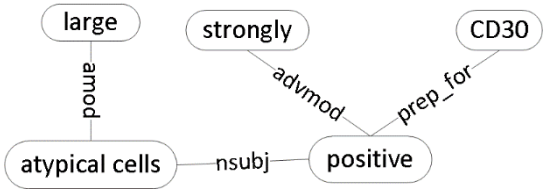
Challenges to Unsupervised Learning

- Distance metric
  - Correlation between subgraphs is lost

Subgraph 1



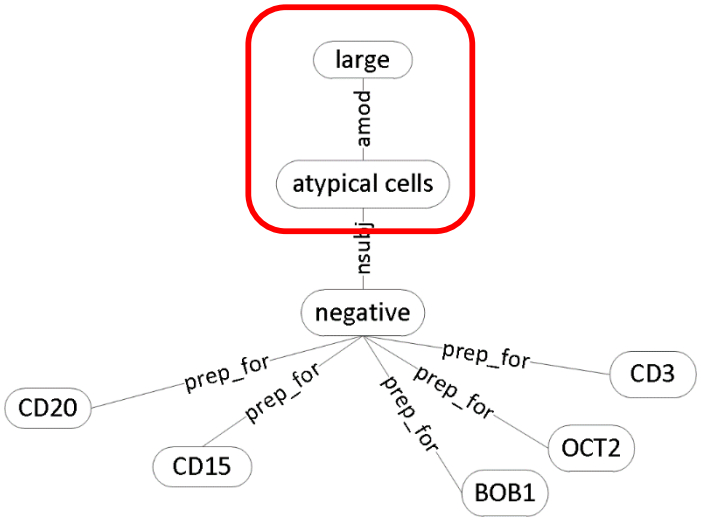
Subgraph 2



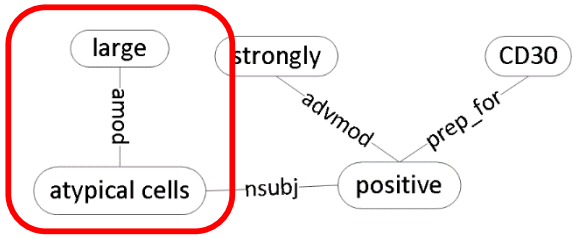
Challenges to Unsupervised Learning

- Distance metric
  - Correlation between subgraphs is lost

Subgraph 1



Subgraph 2



Challenges to Unsupervised Learning

- Correlation between subgraphs is lost
  - Add words covered by subgraph nodes
  - Add words close to subgraph nodes

Challenges to Unsupervised Learning

- Correlation between subgraphs is lost
  - Add words covered by subgraph nodes
  - Add words close to subgraph nodes
  - Distributional representation, success in general NLP tasks

Challenges to Unsupervised Learning

- Correlation between subgraphs is lost
  - Add words covered by subgraph nodes
  - Add words close to subgraph nodes
  - Distributional representation, success in general NLP tasks
- High dimensionality of multiple feature modes

Challenges to Unsupervised Learning

- Correlation between subgraphs is lost
  - Add words covered by subgraph nodes
  - Add words close to subgraph nodes
  - Distributional representation, success in general NLP tasks
- High dimensionality of multiple feature modes
  - Dimensionality reduction

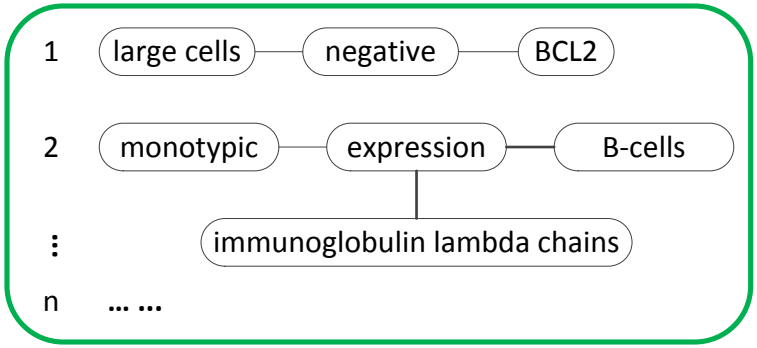
Challenges to Unsupervised Learning

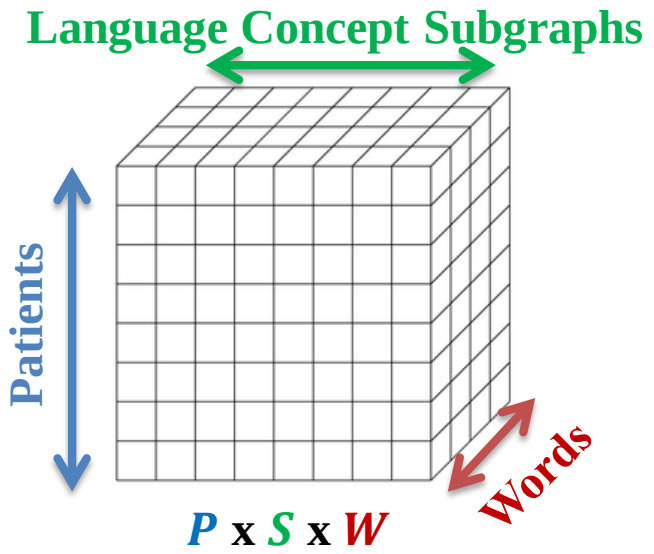
- Correlation between subgraphs is lost
  - Add words covered by subgraph nodes
  - Add words close to subgraph nodes
  - Distributional representation, success in general NLP tasks
- High dimensionality of multiple feature modes
  - Dimensionality reduction

We need to capture correlation at two modes and reduce dimensionality simultaneously

Subgraph Augmented Tensor Modeling

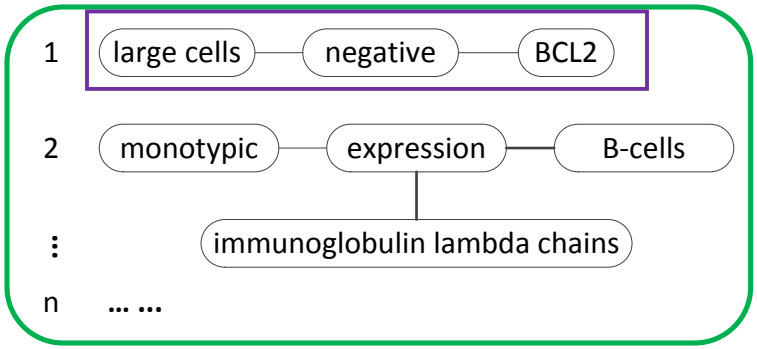
- Subgraph augmented tensor modeling, a three-mode tensor in our setting
  - One mode represents the patients
  - A second the sentence subgraphs
  - The third a set of covered and contextual words

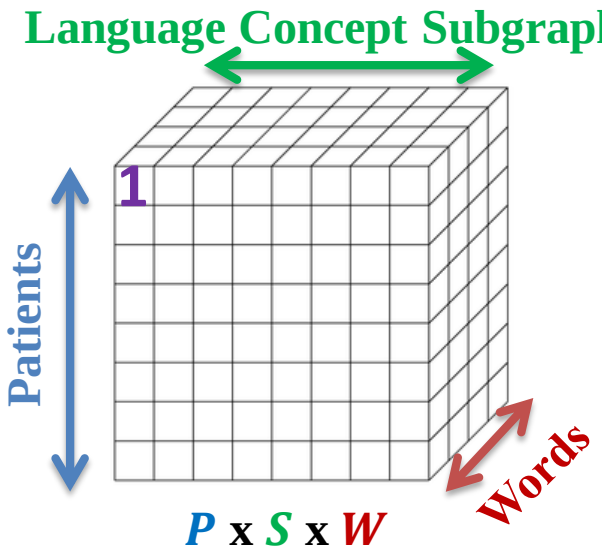


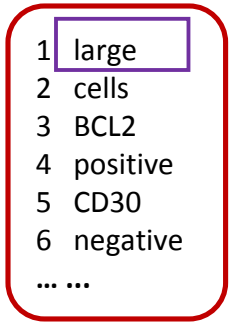


1 large
2 cells
3 BCL2
4 positive
5 CD30
6 negative
… …

Subgraph Augmented Tensor Modeling

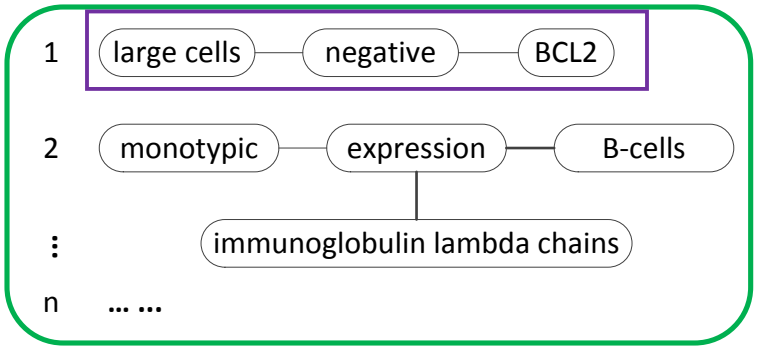


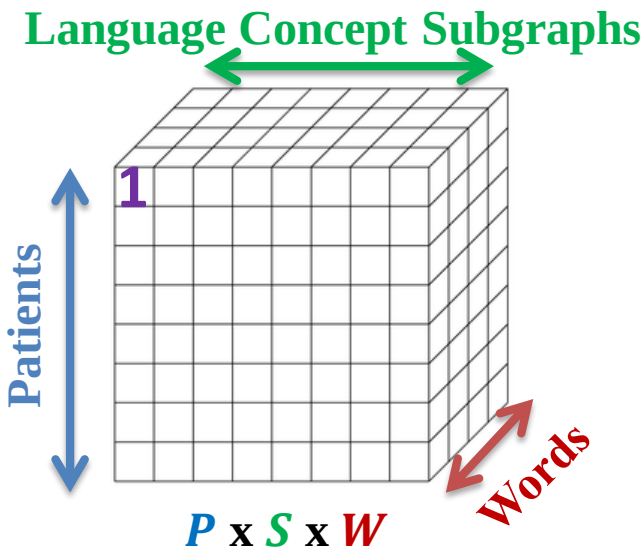


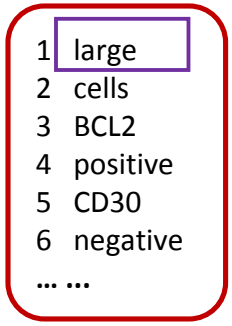


Subgraph Augmented Tensor Modeling

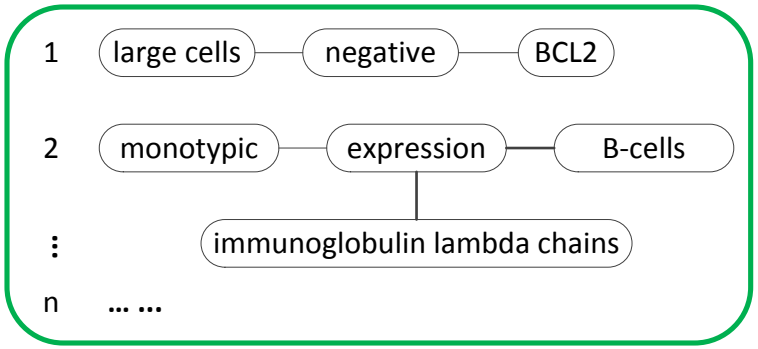
This captures correlations, what about dimensionality reduction?



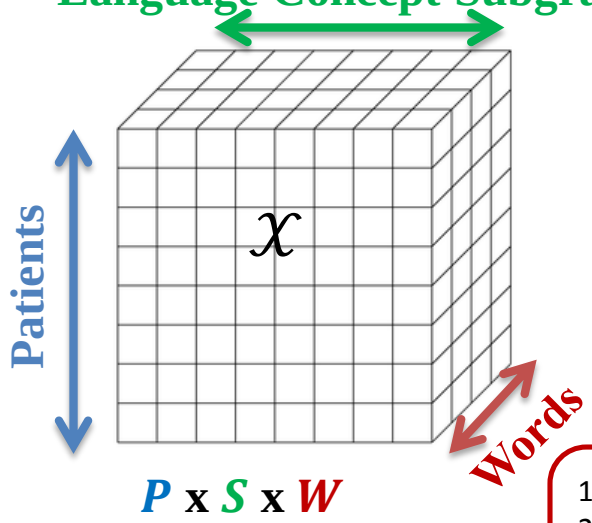


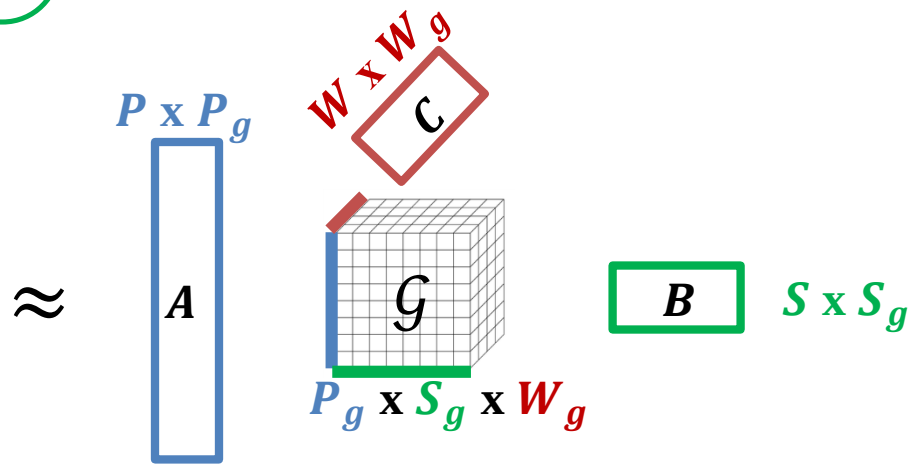


Subgraph Augmented Nonnegative Tensor Factorization (SANTF)



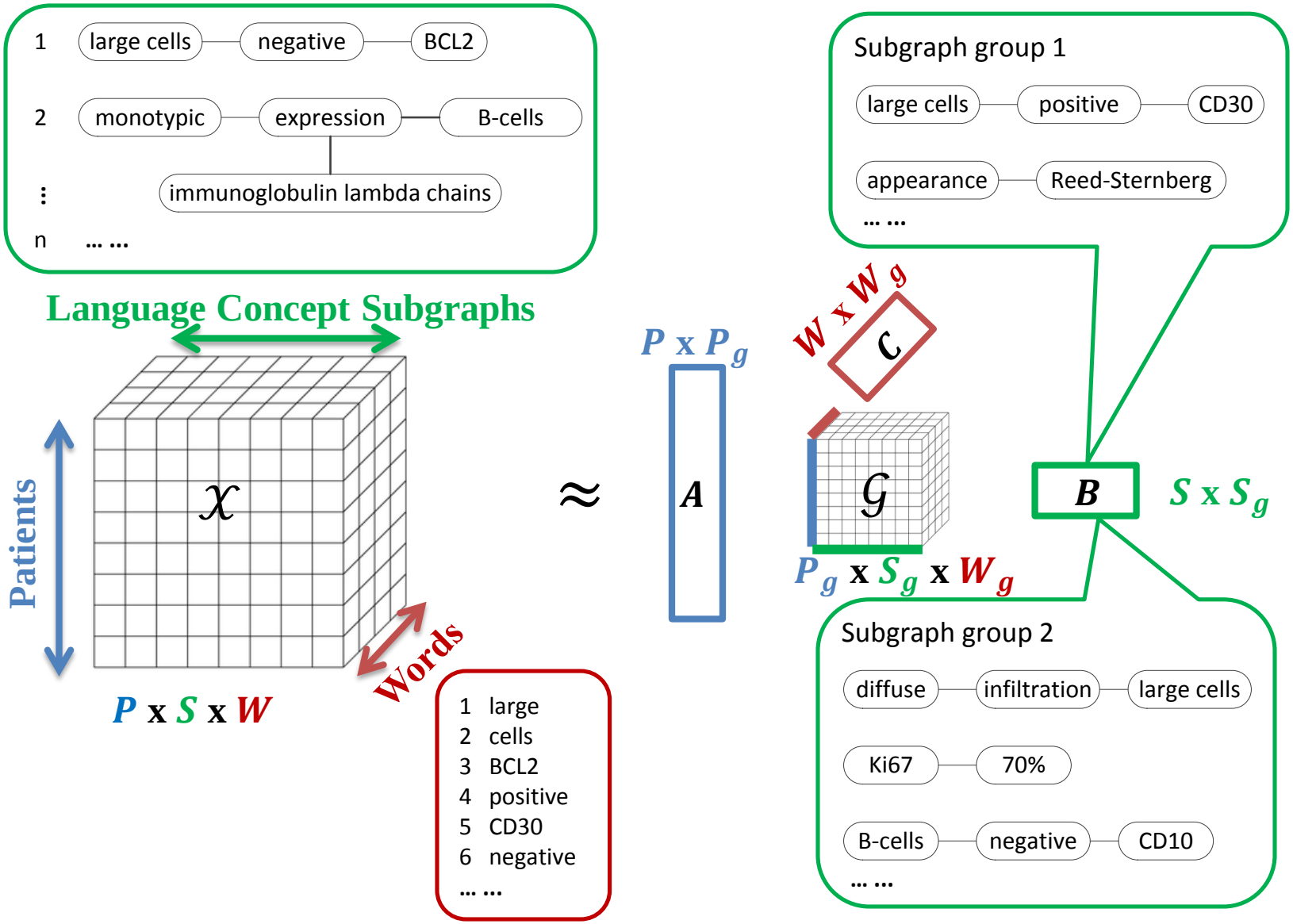
Language Concept Subgraphs





1 large
2 cells
3 BCL2
4 positive
5 CD30
6 negative
… …

Subgraph Augmented Nonnegative Tensor Factorization (SANTF)



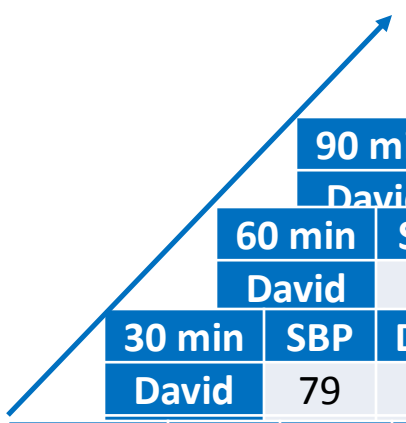
Matrix

- Two-dimension data structure
- Example: patient and physiologic measurements

	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.1	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

Generalizing Matrix to Tensor

- N-dimension data structure ($N \geq 3$)
- Example: patient and timed physiologic measurements



90 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	80	52	146	6	113	167	6.0	3.8

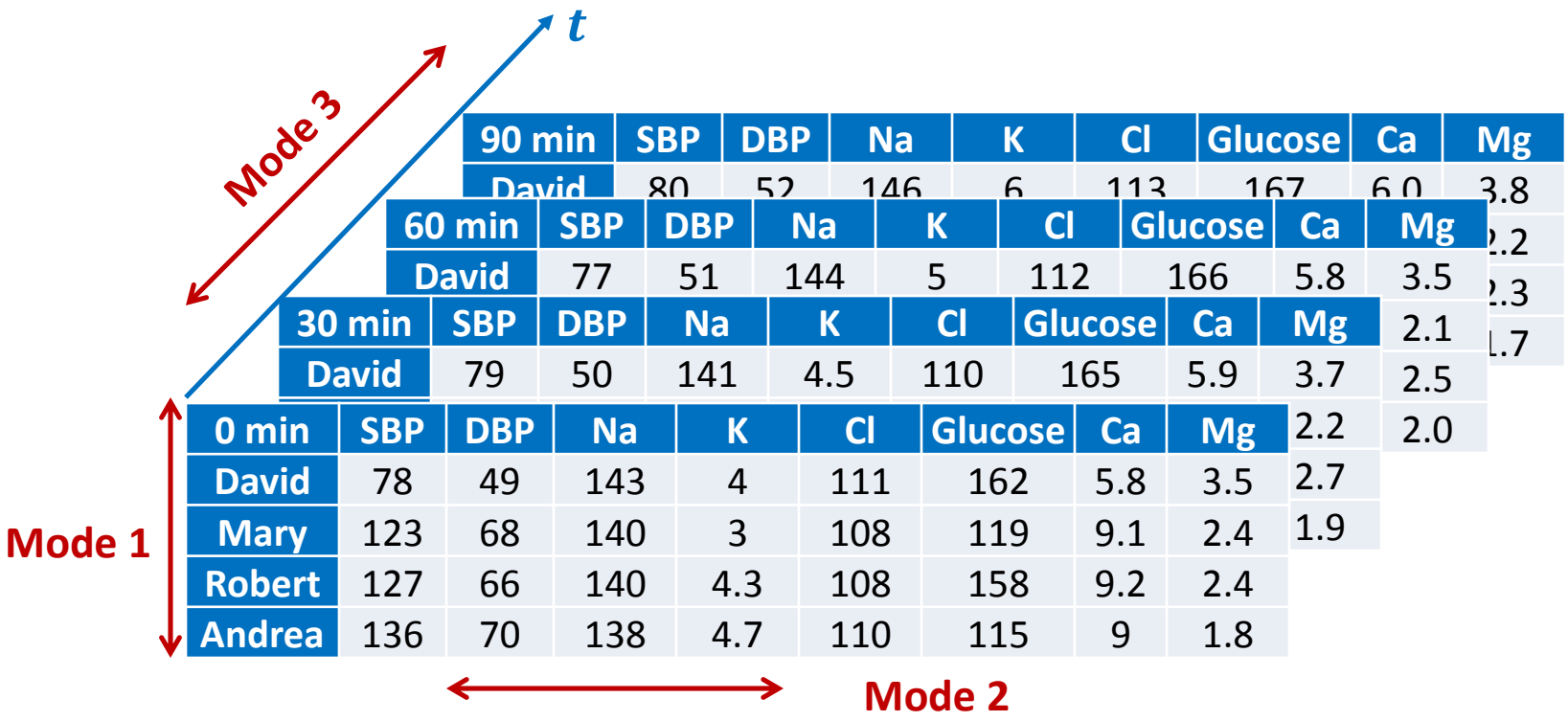
60 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	77	51	144	5	112	166	5.8	3.5

30 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	79	50	141	4.5	110	165	5.9	3.7

0 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.2	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

Tensor Terminology

- Mode
- Fiber
- Slice



90 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	80	52	146	6	113	167	6.0	3.8

60 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	77	51	144	5	112	166	5.8	3.5

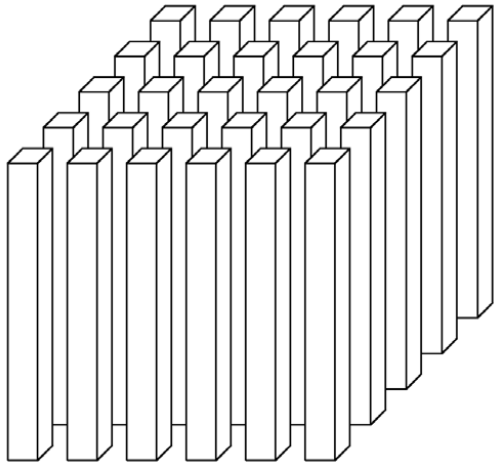
30 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	79	50	141	4.5	110	165	5.9	3.7

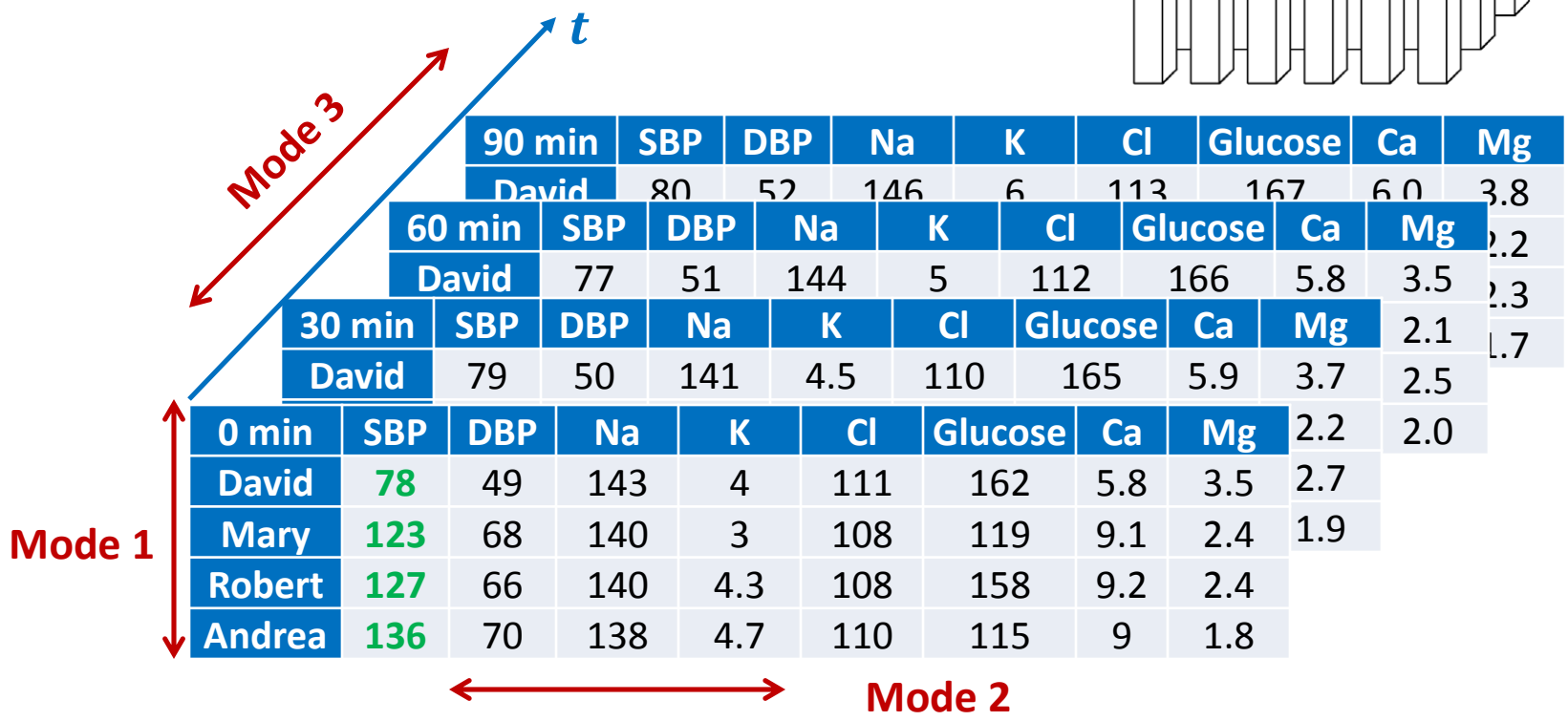
0 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.2	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

Tensor Terminology

- Mode
- Fiber
- Slice

Mode 1 fibers





90 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	80	52	146	6	113	167	6.0	3.8

60 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	77	51	144	5	112	166	5.8	3.5

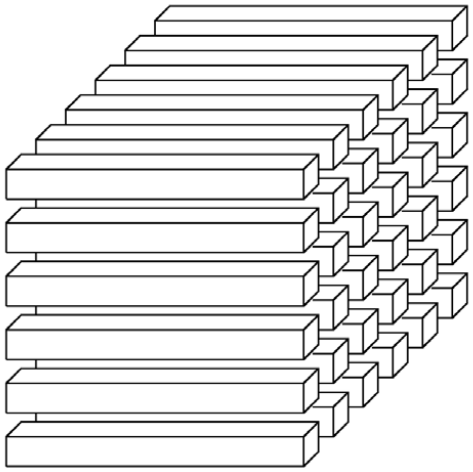
30 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	79	50	141	4.5	110	165	5.9	3.7

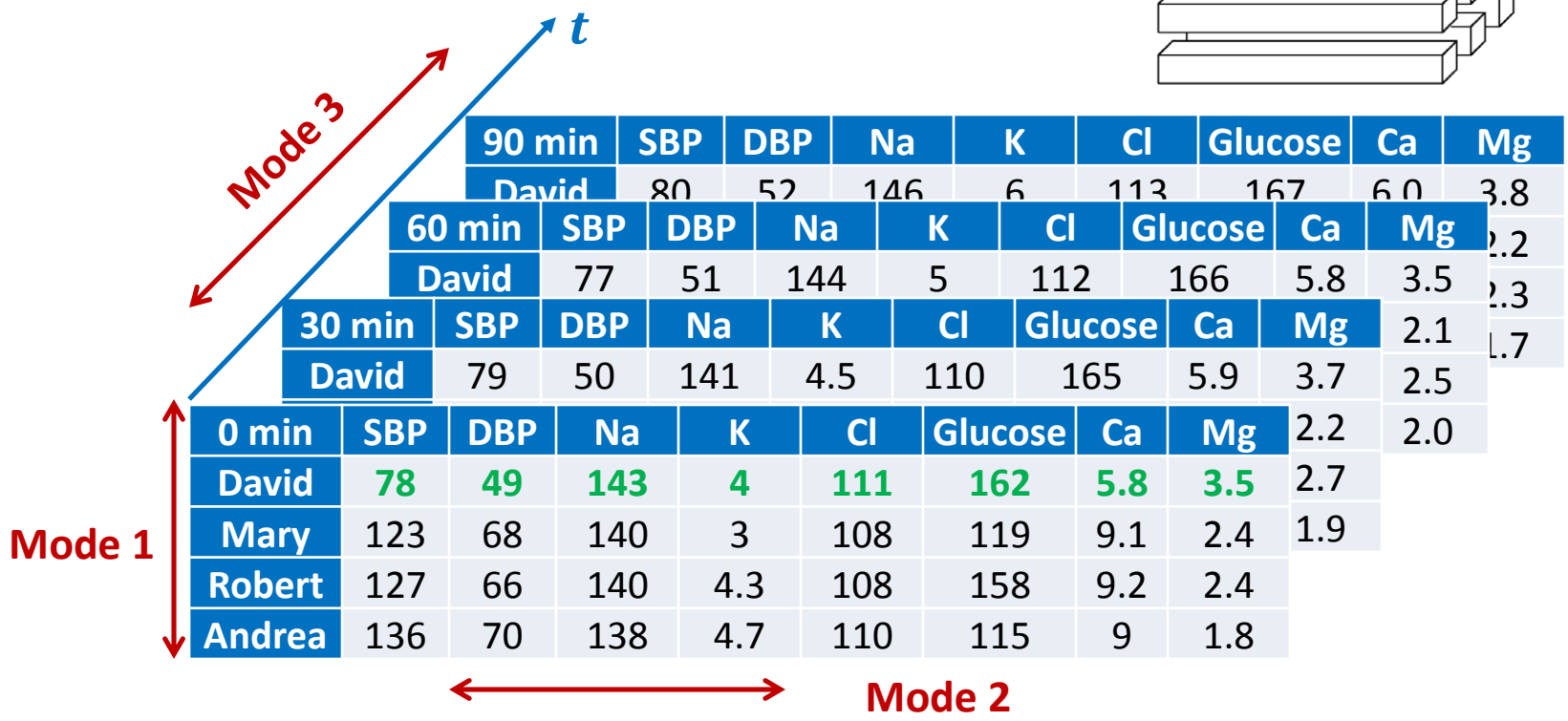
0 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.2	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

Tensor Terminology

- Mode
- Fiber
- Slice

Mode 2 fibers





90 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	80	52	146	6	113	167	6.0	3.8

60 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	77	51	144	5	112	166	5.8	3.5

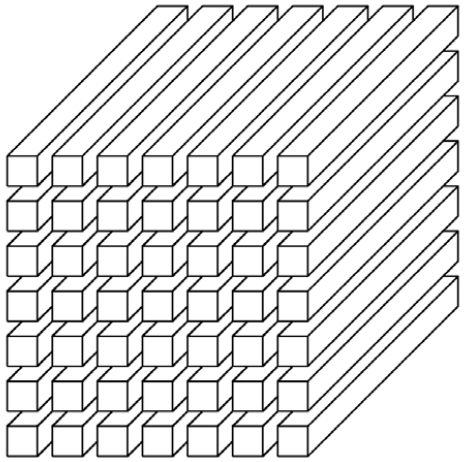
30 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	79	50	141	4.5	110	165	5.9	3.7

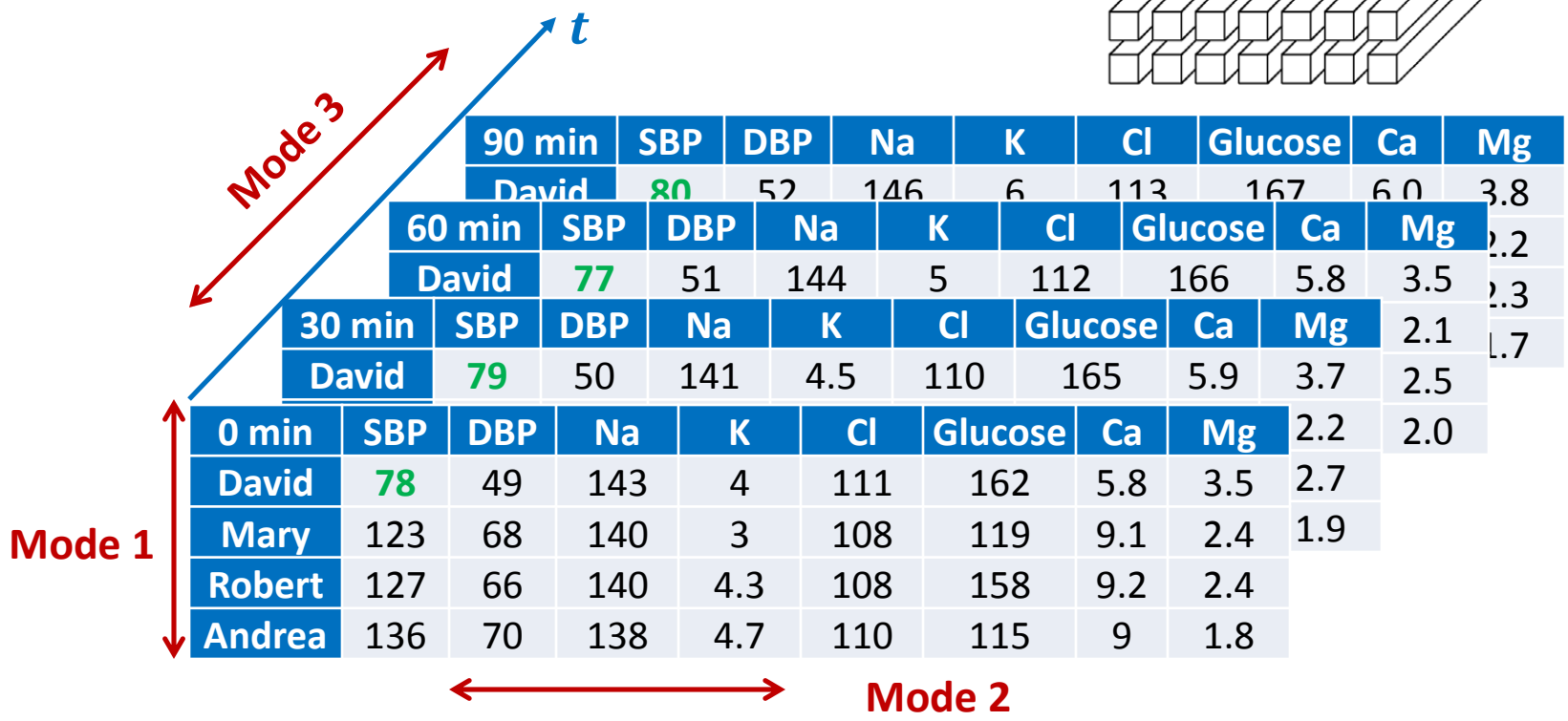
0 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.2	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

Tensor Terminology

- Mode
- Fiber
- Slice

Mode 3 fibers





90 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	80	52	146	6	113	167	6.0	3.8

60 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	77	51	144	5	112	166	5.8	3.5

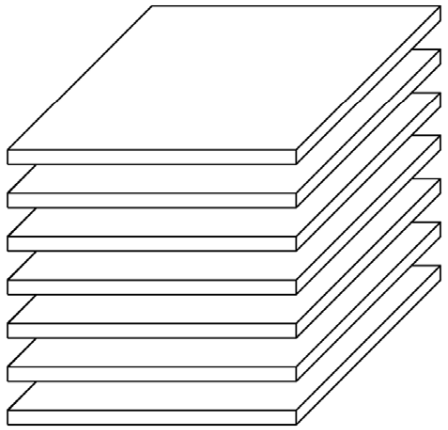
30 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	79	50	141	4.5	110	165	5.9	3.7

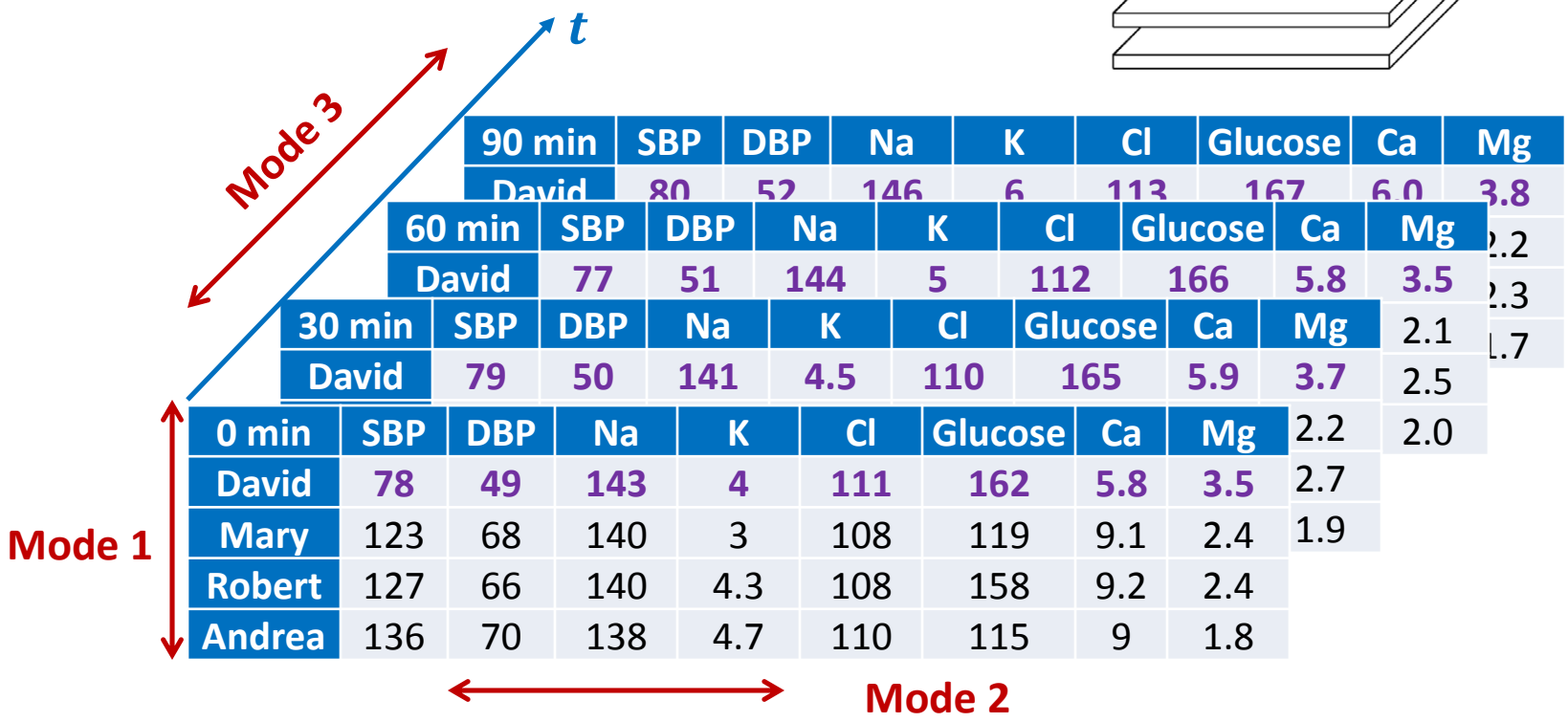
0 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.2	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

Tensor Terminology

- Mode
- Fiber
- Slice

Horizontal Slices $X_{i::}$





90 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	80	52	146	6	113	167	6.0	3.8

60 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	77	51	144	5	112	166	5.8	3.5

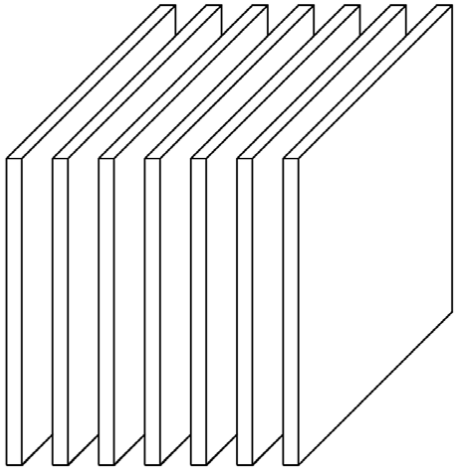
30 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	79	50	141	4.5	110	165	5.9	3.7

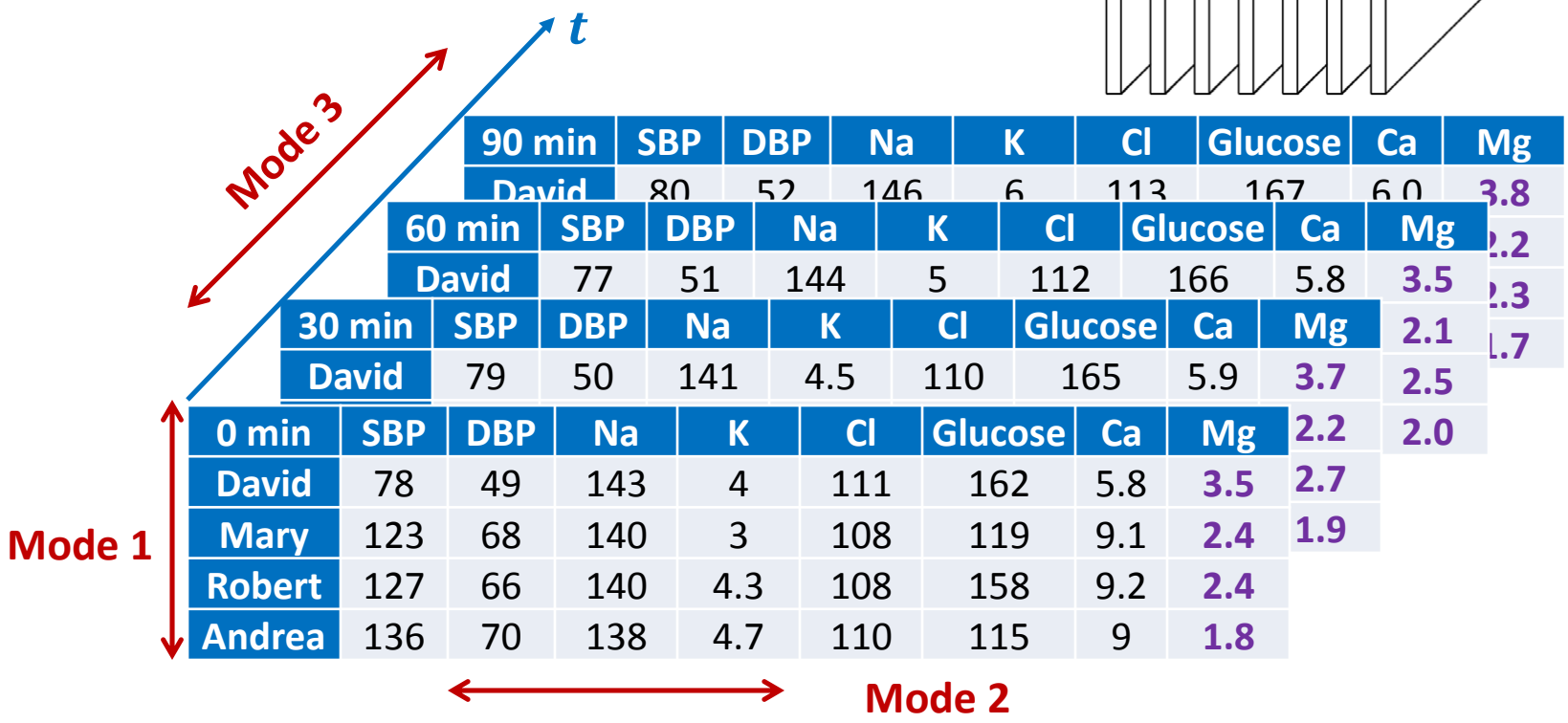
0 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.2	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

Tensor Terminology

- Mode
- Fiber
- Slice

Lateral Slices $X_{:j:}$





90 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	80	52	146	6	113	167	6.0	3.8

60 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	77	51	144	5	112	166	5.8	3.5

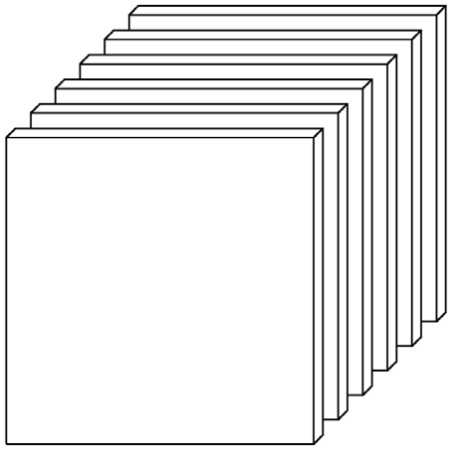
30 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	79	50	141	4.5	110	165	5.9	3.7

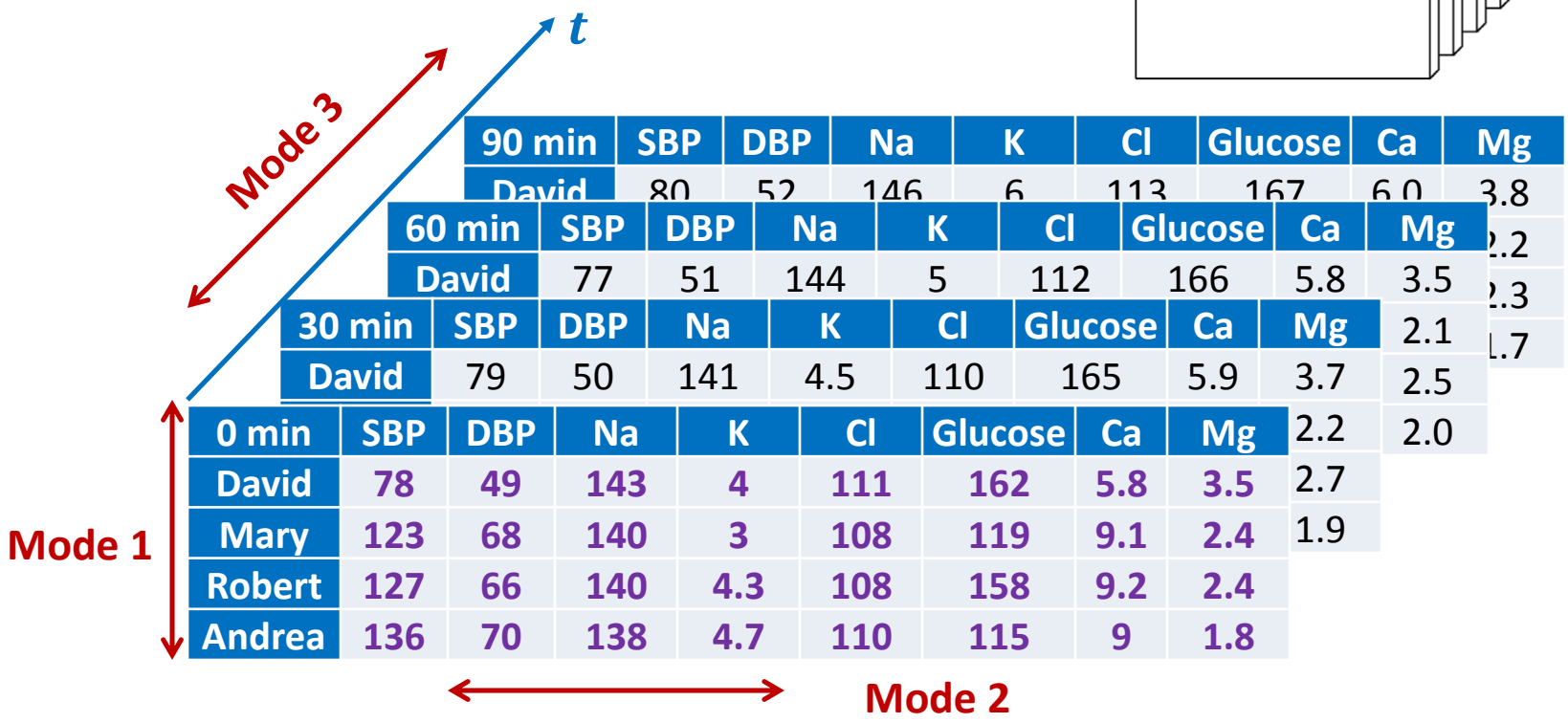
0 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.2	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

Tensor Terminology

- Mode
- Fiber
- Slice

Frontal Slices $X_{::k}$





90 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	80	52	146	6	113	167	6.0	3.8

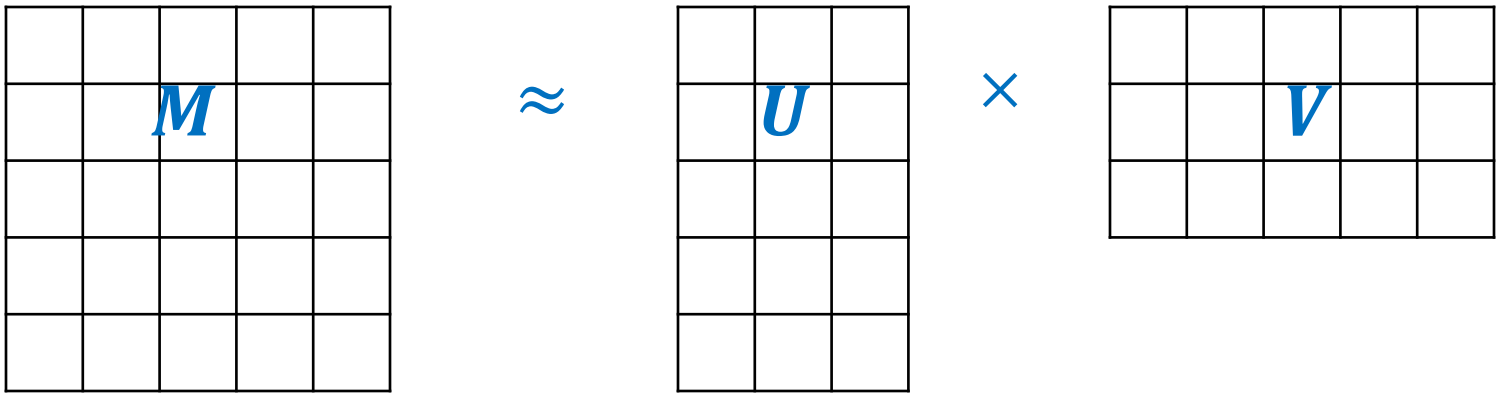
60 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	77	51	144	5	112	166	5.8	3.5

30 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	79	50	141	4.5	110	165	5.9	3.7

0 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.2	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

Non-Negative Matrix Factorization

- Non-negative tensor factorization is higher-order generalization to non-negative matrix factorization (NMF)
- Dimensionality reduction
- Group wise interaction
- $M \approx UV$, where $M \in R^{m \times n}, U \in R^{m \times p}, V \in R^{p \times n}$, all matrices are entry-wise non-negative



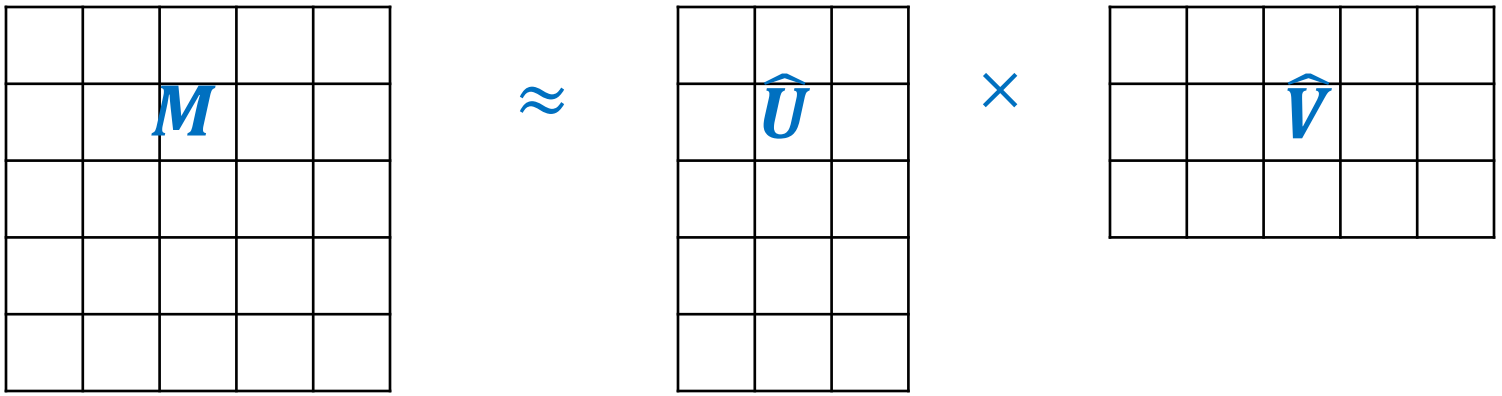
Why Non-Negativity?

- Many real-world data are non-negative
  - Text analysis use relative frequencies of words as basic building blocks
  - Household expenditures in different commodity/service groups are recorded as a relative proportion
  - Gene expression requires non-negativity to give physical and physiological meaning

- Direct interpretation can be obtained through additive combinations of non-negative basis vectors

Non-Negative Matrix Factorization

- Assume columns of M each sum to one, document-by-term
- D is diagonal with ith entry being the reciprocal of the sum of the entries in the ith columns of U
- $M = \hat{U}\hat{V}$, where $\hat{U} = UD, \hat{V} = D^{-1}V$, and columns of $\hat{U}, \hat{V}$ each sum to one
- $\hat{U}$ document-by-topic, $\hat{V}$ topic-by-word



Non-Negative Matrix Factorization

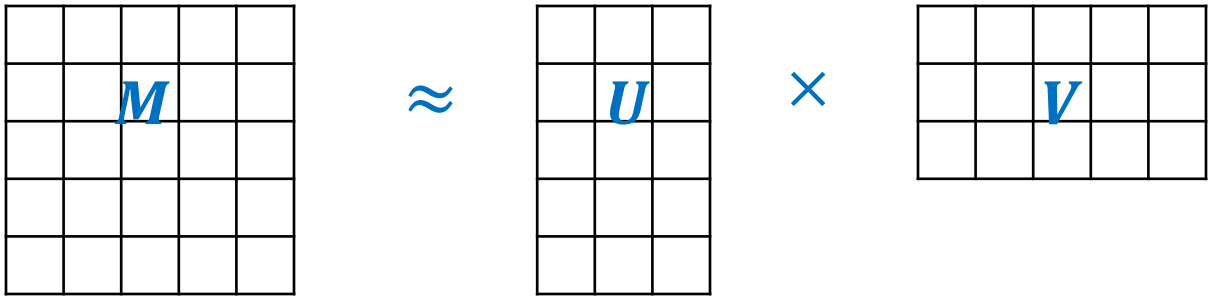
- NMF applied in natural language processing, image segmentation, collaborative filtering
- Optimization $min_{U,V} \|M - UV\|_F$, where $\|.\|_F$ is the Frobenius Norm, i.e. $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$
- Non-convex, NP-hard

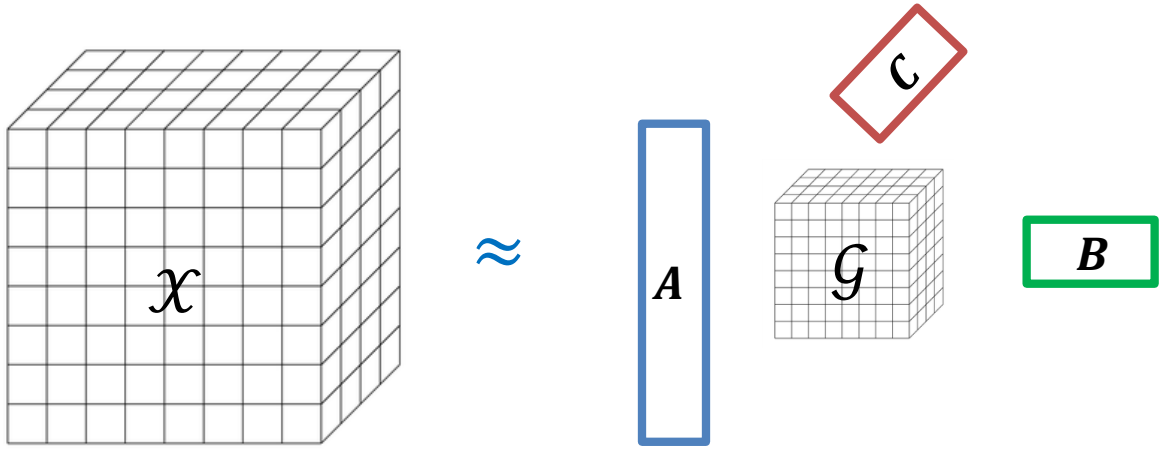
Non-Negative Matrix Factorization

- Alternating minimization
  - First guess U, compute the best V minimizing $\|M - UV\|_F$
  - Fix V, compute the best U
  - Fix U, compute the best V
  - ……
- Local minima
- Expectation-Maximization
- Gradient Descent
- Relaxed form of K-means clustering

Non-Negative Tensor Factorization

- NMF extension to tensors of arbitrary order
- Tucker model





More Tensor-Matrix Operations

- Matricization, or unfolding or flattening
- The mode-n matricization of $\mathcal{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$, denoted by $X_{(n)}$, arranges the mode-n fibers to be the columns of the resulting matrix
- $X_1 = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}, X_2 = \begin{bmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{bmatrix}$
- $X_{(1)} = \begin{bmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \end{bmatrix}$
- $X_{(2)} = \begin{bmatrix} 1 & 2 & 3 & 13 & 14 & 15 \\ 4 & 5 & 6 & 16 & 17 & 18 \\ 7 & 8 & 9 & 19 & 20 & 21 \\ 10 & 11 & 12 & 22 & 23 & 24 \end{bmatrix}$
- $X_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \end{bmatrix}$

More Tensor-Matrix Operations

- The n-mode matrix product of a tensor $\mathcal{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ with a matrix $M \in R^{J \times I_n}$ is denoted as $\mathcal{X} \times_n M$ of size $I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N$
- $(\mathcal{X} \times_n M)_{i_1 \ldots i_{n-1} j i_{n+1} \ldots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \ldots i_N} m_{j i_n}$
- Each mode-n fiber is left multiplied by M
- $X_1 = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}, X_2 = \begin{bmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{bmatrix}, M = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$
- $(\mathcal{X} \times_1 M)_1 = \begin{bmatrix} 22 & 49 & 76 & 103 \\ 28 & 64 & 100 & 136 \end{bmatrix}$
- $(\mathcal{X} \times_1 M)_2 = \begin{bmatrix} 130 & 157 & 184 & 211 \\ 172 & 208 & 244 & 280 \end{bmatrix}$

More Tensor-Matrix Operations

- The n-mode vector product of a tensor $\mathcal{X}$ with a vector $v \in R^{I_n}$ is denoted as $\mathcal{X} \bar{\times}_n v$ of order $N-1$ and size $I_1 \times \cdots \times I_{n-1} \times I_{n+1} \times \cdots \times I_N$
- $(\mathcal{X} \bar{\times}_n v)_{i_1 \ldots i_{n-1} i_{n+1} \ldots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \ldots i_N} v_{i_n}$
- Inner product of each mode-n fiber with v
- $X_1 = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}, X_2 = \begin{bmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{bmatrix}, v = [1\ 2\ 3\ 4]^T$
- $\mathcal{X} \bar{\times}_2 v = \begin{bmatrix} 70 & 190 \\ 80 & 200 \\ 90 & 210 \end{bmatrix}$

Solving NTF

- NTF mathematic formulation

$$f(A,B,C,\mathcal{G}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}\left(\mathcal{X}_{i,j,k} - \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R}\mathcal{G}_{p,q,r}A_{i,p}B_{j,q}C_{k,r}\right)^2$$

$$\min_{\mathcal{G},A,B,C} f(A,B,C,\mathcal{G}) \text{ subject to } \mathcal{G} \geq 0, A \geq 0, B \geq 0, C \geq 0$$

Solving NTF

- NTF mathematic formulation

$$f(A,B,C,\mathcal{G}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}\left(\mathcal{X}_{i,j,k} - \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R}\mathcal{G}_{p,q,r}A_{i,p}B_{j,q}C_{k,r}\right)^2$$

$$\min_{\mathcal{G},A,B,C} f(A,B,C,\mathcal{G}) \text{ subject to } \mathcal{G} \geq 0, A \geq 0, B \geq 0, C \geq 0$$

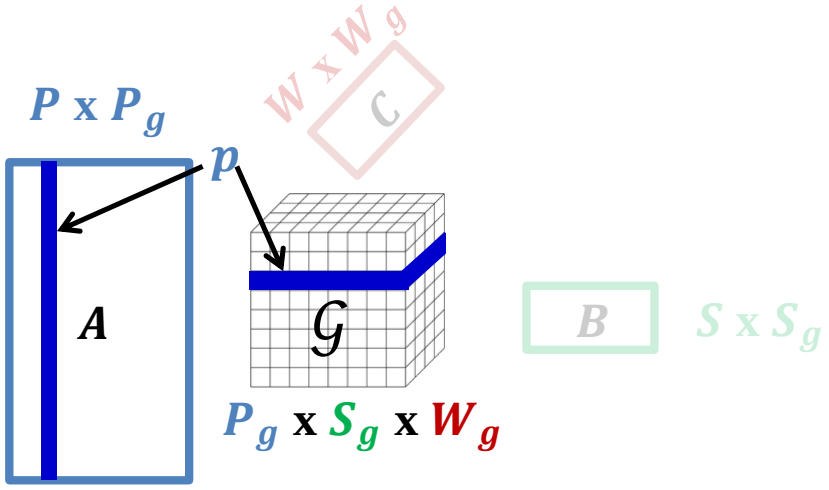
- No closed form solution, alternating minimization
- Block Coordinate Descent (Xu et al. 2013)

Creating Feature Matrix from SANTF Result

- Focus on patient group matrix and core tensor

$$w_p = \|\mathcal{G}_{p::}\|_2 = \sqrt{\sum_{q=1}^{S_g}\sum_{r=1}^{W_g}\mathcal{G}_{p\,q\,r}^2},\ 1 \le p \le P_g$$

- Multiply column p of A with w_p to get A^w

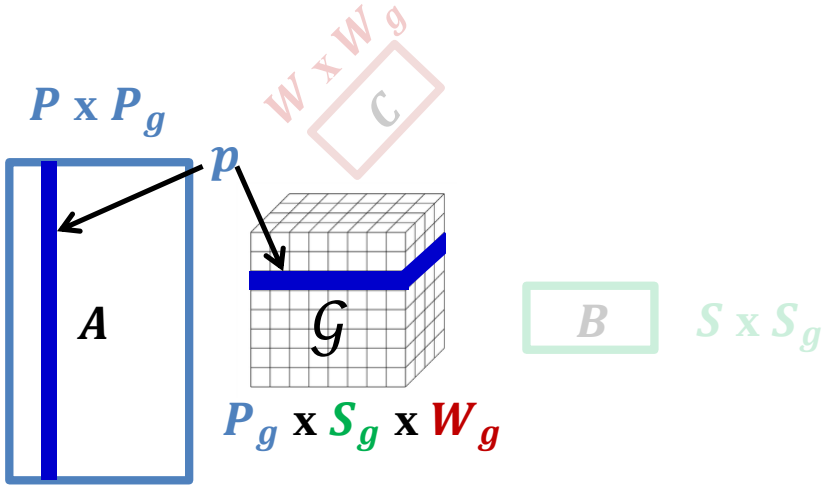


Creating Feature Matrix from SANTF Result

- Focus on patient group matrix and core tensor

$$w_p = \|\mathcal{G}_{p::}\|_2 = \sqrt{\sum_{q=1}^{S_g}\sum_{r=1}^{W_g}\mathcal{G}_{p\,q\,r}^2},\ 1 \le p \le P_g$$

- Multiply column p of A with w_p to get A^w
- Intuition: A is column-wise normalized, reweighting with w_p takes into account the magnitude of patient groups interacting with subgraph and word groups



Lymphoma Subtypes – Clustering Experiment

- Comparison Models

NMF: Non-negative matrix factorization

[Methods:
NMF
k-means (Euclidean)
k-means (Cosine)] × [Features:
Word
Subgraph
Subgraph and word]

- SANTF as target
  - Translate A^w into clustering interpretation

Lymphoma	Number of Cases	Number of Training Cases	Number of Test Cases
DLBCL	589	305 (64.8%)	284 (66.7%)
Follicular	184	101 (21.4%)	83 (19.5%)
Hodgkin	124	65 (13.8%)	59 (13.8%)

Lymphoma Subtypes – Clustering Results

Methods	Avg. Precision	Avg. Recall	Avg. F-measure	Accuracy
(1) NMF pt × wd	0.492	0.495	0.428	0.626
(2) NMF pt × sg	0.621	0.765	0.601	0.605
(3) NMF pt × [sg wd]	0.637	0.787	0.615	0.614
(4) k-means (Euclidean) pt × wd	0.483	0.420	0.398	0.664
(5) k-means (Euclidean) pt × sg	0.700	0.602	0.584	0.708
(6) k-means (Euclidean) pt × [sg wd]	0.690	0.593	0.573	0.726
(7) k-means (Cosine) pt × wd	0.620	0.694	0.618	0.617
(8) k-means (Cosine) pt × sg	0.647	0.762	0.624	0.615
(9) k-means (Cosine) pt × [sg wd]	0.648	0.759	0.626	0.617
(10) SANTF pt × sg × wd	**0.720**$^{1\text{-}9}$	**0.849**$^{1\text{-}9}$	**0.743**$^{1\text{-}9}$	**0.751**$^{1\text{-}9}$

Lymphoma Subtypes – Clustering Results

Methods	Avg. Precision	Avg. Recall	Avg. F-measure	Accuracy
(1) NMF pt × wd	0.492	0.495	0.428	0.626
(2) NMF pt × sg	0.621	0.765	0.601	0.605
(3) NMF pt × [sg wd]	0.637	0.787	0.615	0.614
(4) k-means (Euclidean) pt × wd	0.483	0.420	0.398	0.664
(5) k-means (Euclidean) pt × sg	0.700	0.602	0.584	0.708
(6) k-means (Euclidean) pt × [sg wd]	0.69			
(7) k-means (Cosine) pt × wd	0.62			
(8) k-means (Cosine) pt × sg	0.647	0.762	0.624	0.615
(9) k-means (Cosine) pt × [sg wd]	0.648	0.759	0.626	0.617
(10) SANTF pt × sg × wd	**0.720**[1-9]	**0.849**[1-9]	**0.743**[1-9]	**0.751**[1-9]

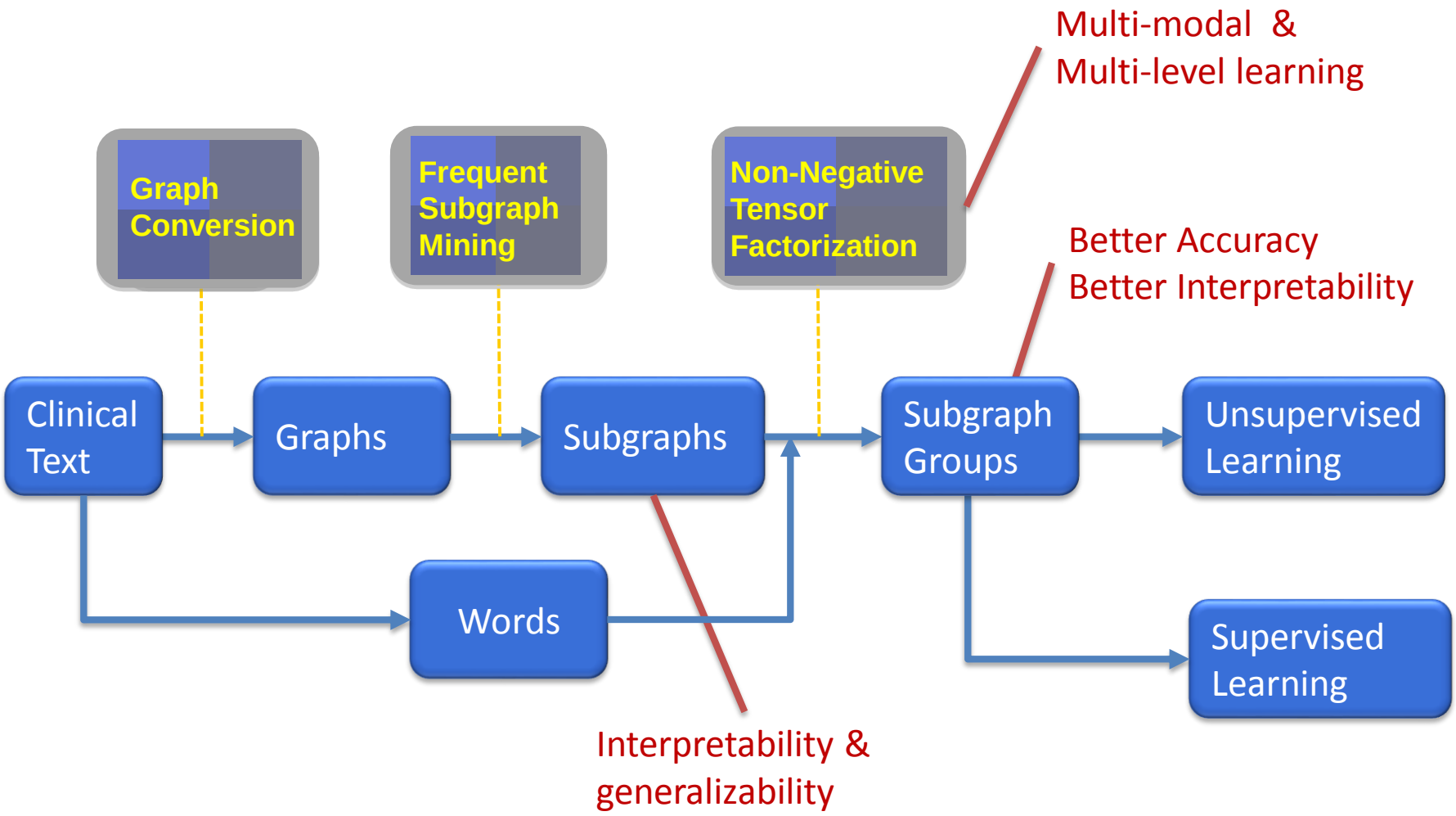
~12% absolute increase in Average F-measure from best baseline

A Closer Look into DLBCL First Group

	DLBCL 1st Subgraph Group
0.6640	**… atypical cells …**
0.0929	**… large lymphoid cells …**
0.0057	… show … positive cells …
0.0040	… large lymphoid cell with vesicular nuclei …
0.0025	… show the cells are … B-cells co-expressing …
0.0019	**… large cells predominate …**
0.0010	**… cells are CD30+, MUM1+ …**
0.0005	**… large cells stain for CD79a …**
0.0005	… admixed small lymphocytes …
0.0004	**… large cells stain positively for CD20 …**
0.0002	… large atypical cell with vesicular nuclei …

	… large cells predominate …	… large cells stain for CD79a …	… large cells stain positively for CD20 …	… large lymphoid cells …	… cells are CD30+, MUM1+ …	… atypical cells …
… large cells predominate …	1	0.5664	0.4741	0.5566	0.5415	0.5953
… large cells stain for CD79a …		1	0.6468	0.3281	0.2501	0.3521
… large cells stain positively for CD20 …			1	0.3145	0.3238	0.3314
… large lymphoid cells …				1	0.2518	0.8972
… cells are CD30+, MUM1+ …					1	0.3873
… atypical cells …						1

Contribution of SANTF



References

Liao Katherine P, Cai Tianxi, Gainer Vivian, Goryachev Sergey, Zeng-treitler Qing, Raychaudhuri Soumya et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis care & research, 62(8):1120–1127, 2010.

Conway Mike, Berg Richard, Carrel David, Denny Joshua, Kho Abel, Kullo Iftikhar et al. Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. In AMIA Annu Symp Proc, pages 274–283, 2011.

Denny Joshua C, Ritchie Marylyn D, Basford Melissa A, Pulley Jill M, Bastarache Lisa, Brown-Gentry Kristin et al. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics, 26(9):1205–1210, 2010.

Collier Nigel, Oellrich Anika and Groza Tudor. Concept selection for phenotypes and diseases using learn to rank. Journal of biomedical semantics, 6(1):1, 2015.

Zeng Qing T, Goryachev Sergey, Weiss Scott, Sordo Margarita, Murphy Shawn N and Lazarus Ross. Extracting principal diagnosis, comorbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC medical informatics and decision making, 6(1):1, 2006.

Bejan Cosmin Adrian, Xia Fei, Vanderwende Lucy, Wurfel Mark M and Yetisgen-Yildiz Meliha. Pneumonia identification using statistical feature selection. Journal of the American Medical Informatics Association, 19(5):817–823, 2012.

Herskovic Jorge R, Subramanian Devika, Cohen Trevor, Bozzo-Silva Pamela A, Bearden Charles F and Bernstam Elmer V. Graph-based signal integration for high-throughput phenotyping. BMC bioinformatics, 13(Suppl 13):S2, 2012.

References

Luo Yuan, Sohani Aliyah R, Hochberg Ephraim P and Szolovits Peter. Automatic lymphoma classification with sentence subgraph mining from pathology reports. Journal of the American Medical Informatics Association, 21(5):824–832, 2014.

Lehman Li-Wei H, Saeed Mohammed, Long William J, Lee Joon and Mark Roger G. Risk stratification of icu patients using topic models inferred from unstructured progress notes. In AMIA. Citeseer, 2012.

Luo Yuan, Xin Yu, Hochberg Ephraim, Joshi Rohit, Uzuner Ozlem and Szolovits Peter. Subgraph augmented non-negative tensor factorization (santf) for modeling clinical text. Journal of the American Medical Informatics Association, 22(5):1009–1019.

Peissig Peggy L, Rasmussen Luke V, Berg Richard L, Linneman James G, McCarty Catherine A, Waudby Carol et al. Importance of multimodal approaches to effectively identify cataract cases from electronic health records. Journal of the American Medical Informatics Association, 19(2):225–234, 2012.

Zhao Di and Weng Chunhua. Combining pubmed knowledge and ehr data to develop a weighted bayesian network for pancreatic cancer prediction. Journal of biomedical informatics, 44(5):859–868, 2011.

Kho Abel N, HayesMGeoffrey, Rasmussen-Torvik Laura, Pacheco Jennifer A, ThompsonWilliam K, Armstrong Loren L et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. Journal of the American Medical Informatics Association, 19(2):212–218, 2012.

Xu Hua, Fu Zhenming, Shah Anushi, Chen Yukun, Peterson Neeraja B, Chen Qingxia et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. In AMIA Annu Symp Proc, volume 2011, pages 1564–1572, 2011.

UC San Diego
SCHOOL OF MEDICINE
Department of BioMedical Informatics

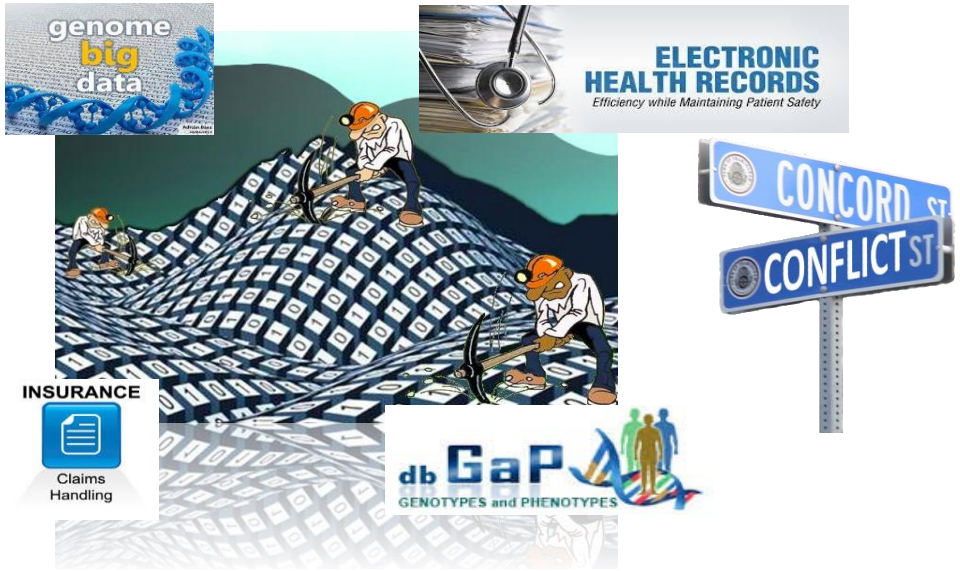
Xiaoqian Jiang

Computing phenotypes while respecting privacy

This research was supported by grant R21LM012060

Privacy Challenges in Biomedical informatics

Biomedical Data analyses



Data Privacy

- Biomedical data (of different cohorts and different modality) are hosted in different silos across many institutions
- Modern data analysis requires the ability to integrate and analyze such data to speedup research and promote discoveries

Biomedical data are complex and sensitive

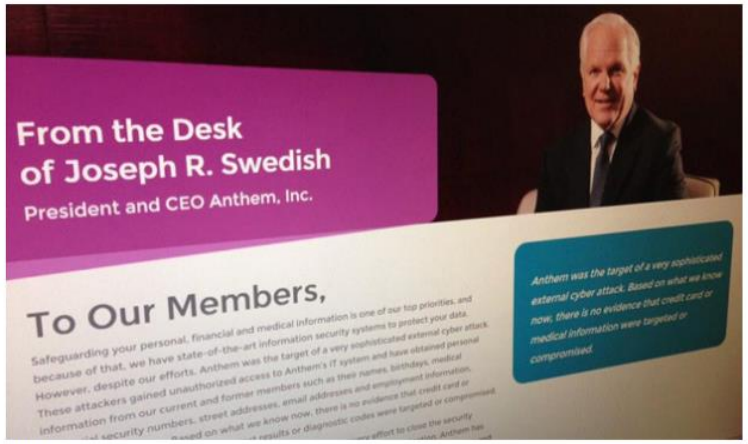
- Electronic health records are massive and sparse
- Genomic data are of high dimensionality
- Incremental collection of mobile sensor data is getting very large
- Free form text presents additional challenges to extract

112 Million Records breached In 2015 with billions of lost

FBI Investigating Anthem Security Breach Affecting 80M Customers

February 5, 2015 4:59 AM

View Comments



(credit: CBS)

Related Tags: Anthem, FBI, Health Insurance, Members, Security Breach

LOS ANGELES (AP) — Health insurer Anthem said hackers infiltrated its computer network and gained access to a host of personal information for customers and employees, including CEO Joseph Swedish.

The nation's second-largest health insurer said it was contacting customers affected by the "very sophisticated" cyberattack and was working to figure out how many people were affected.

The company said information the hackers gained access to included names, birthdates, email address, employment details, Social Security numbers, incomes and street addresses of people who are currently covered or have had coverage in the past.

Sponsored Links Advertise Here

LifeLock® Official Site
The Most Comprehensive ID Protection: LifeLock Ultimate Plus™
LifeLock.com

Psychology Programs
Apply to Argosy University℠ - Get Info on Psychology Programs.
visit.argosy.edu

http://losangeles.cbslocal.com/2015/02/05/fbi-investigating-anthem-security-breach-affecting-80m-customers/

THREAT INTELLIGENCE

Healthcare Suffers Estimated $6.2 Billion In Data Breaches

5/12/2016
12:01 AM

Kelly Jackson Higgins
News

Connect Directly

3 COMMENTS
COMMENT NOW

Login
100% 0%

Like 127
Tweet
Share
904
G+1 20

Nearly 90 percent of healthcare organizations were slammed by a breach in the past two years.

The 911 call has come in loud and clear for the healthcare industry: nearly 90% of all healthcare organizations suffered at least one data breach in the past two years with an average cost of $2.2 million per hack.

Despite heightened awareness and concern among the healthcare industry over its ability to thwart cybercrime, insider mistakes, and ransomware attacks, healthcare budgets for security have either dropped or remained the same in the past year, according to the newly released Sixth Annual Benchmark Study on Privacy & Security of Healthcare Data by the Ponemon Institute. Some 10% of budgets have declined, and more than half have remained static, and most believe they don't have the budget to properly protect data.

The Ponemon report, commissioned by ID Experts, estimates that data breaches cost the healthcare industry some $6.2 billion, as some 79% of healthcare organizations say they were hit with two or more data breaches in the past two years, and 45%, more than five breaches. Most of those exposed fewer than 500 data records, and thus don't get reported to the US Department of Health and Human Services nor are revealed to the media. Ponemon surveyed 91 healthcare organizations, mainly healthcare providers, and 84 healthcare business partner organizations, including pharmaceutical companies, IT and service providers, and medical device makers, and broke down the findings accordingly.

http://www.darkreading.com/threat-intelligence/healthcare-suffers-estimated-$62-billion-in-data-breaches/d/d-id/1325482

Top 10 Healthcare Data Breaches 2015

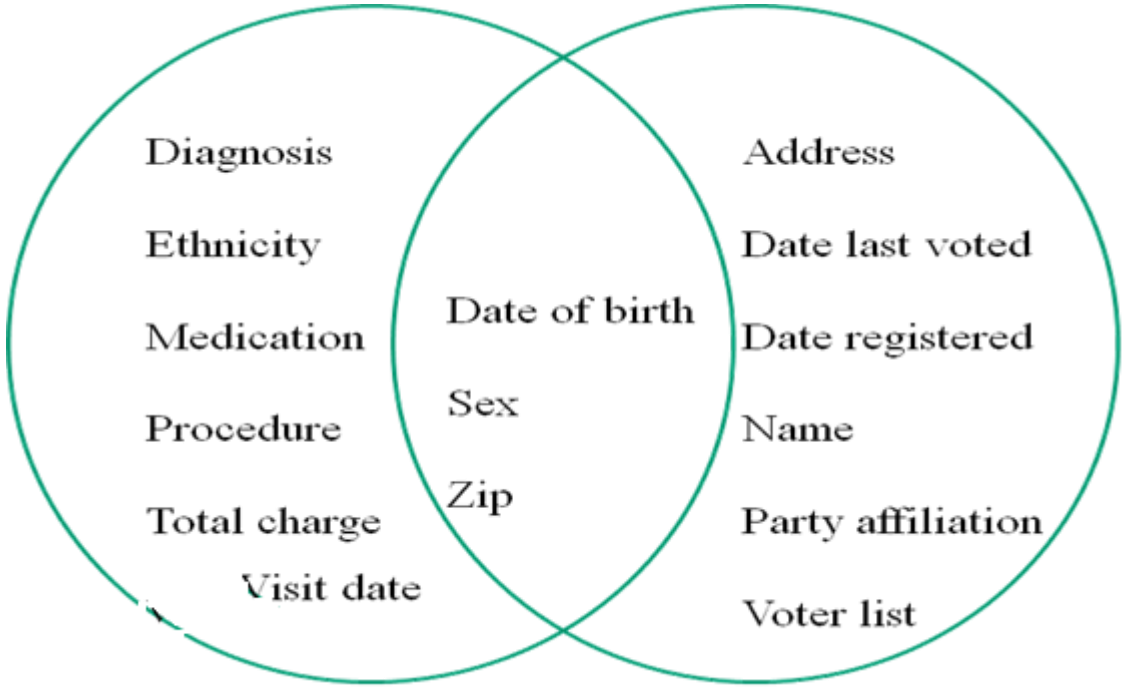
Organization	Records Breached	Type of Breach
Anthem	78,800,000	Hacking / IT Incident
Premera Blue Cross	11,000,000	Hacking / IT Incident
Excellus	10,000,000	Hacking / IT Incident
UCLA Health	4,500,000	Hacking / IT Incident
mie Medical Informatics Engineering	3,900,000	Hacking / IT Incident
CareFirst	1,100,000	Hacking / IT Incident
DMAS	697,586	Hacking / IT Incident
Georgia Department of Community Health	557,779	Hacking / IT Incident
Beacon Health System	306,789	Hacking / IT Incident
DJO Global	160,000	Laptop Theft
2015 Total	**111,022,154**	**(almost 35% U.S. population)**

http://www.forbes.com/sites/danmunro/2015/12/31/data-breaches-in-healthcare-total-over-112-million-records-in-2015/#561339197fd5

Identity vs. attribute disclosure

- **Identity disclosure:** allows identification of which record has which sensitive attribute (most well-known type of protection)
- **Attribute disclosure:** allows recipient to see that that a sub-population of records all have the same sensitive attributes
- **Inference disclosure:** partial release of data make it possible to determine the value of some characteristics of an individual more accurately than otherwise would have been possible

Identity disclosure



63-87% of USA estimated to be unique

[1] L. Sweeney. Uniqueness of Simple Demographics in the U.S Population. 2000.
[2] P. Golle. Revisiting the Uniqueness of US Population. ACM WPES. 2006.

A Common Practice

- A common practice is for organizations to release and receive person data with all explicit identifiers

Name
Bob
Gary
Doug
Alice
Cathy
Emily
Fiona

Patient table			
Job	**Sex**	**Age**	**Disease (sensitive)**
Engineer	Male	35	Fever
Engineer	Male	38	Fever
Lawyer	Male	38	Hepatitis
Musician	Female	30	Flu
Musician	Female	30	Hepatitis
Dancer	Female	30	Hepatitis
Dancer	Female	30	Hepatitis

Thanks Mr. Noman Mohammed for this slide.

Example: Healthcare Data

Name
Bob
Gary
Doug
Alice
Cathy
Emily
Fiona

Identifiers

Patient table			
Job	**Sex**	**Age**	**Disease**
Engineer	Male	35	Fever
Engineer	Male	38	Fever
Lawyer	Male	38	Hepatitis
Musician	Female	30	Flu
Musician	Female	30	Hepatitis
Dancer	Female	30	Hepatitis
Dancer	Female	30	Hepatitis

Quasi-identifiers (QID)

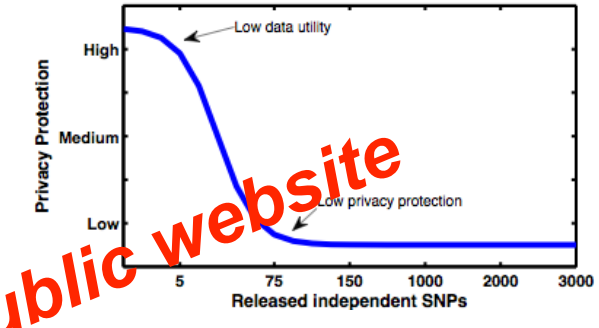
Sensitive Attribute

Voters' registry			
Name	**Job**	**Sex**	**Age**
Alice	Musician	Female	30
Bob	Engineer	Male	35
Cathy	Musician	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30

Thanks Mr. Noman Mohammed for this slide.

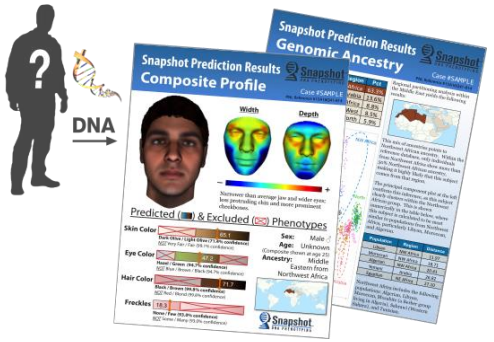
Inference disclosure

- ***Lin et. al. 2004 science:*** as few as 75 statistically independent SNPs (Single-nucleotide polymorphism) will be sufficient to identify a single person



- ***Gymrek et al. 2013 Science:*** surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome and querying recreational genetic genealogy databases

- ***Claes et. al. 2014 PLOS Genetics:*** DNA information can be used to reconstruct 3D human face

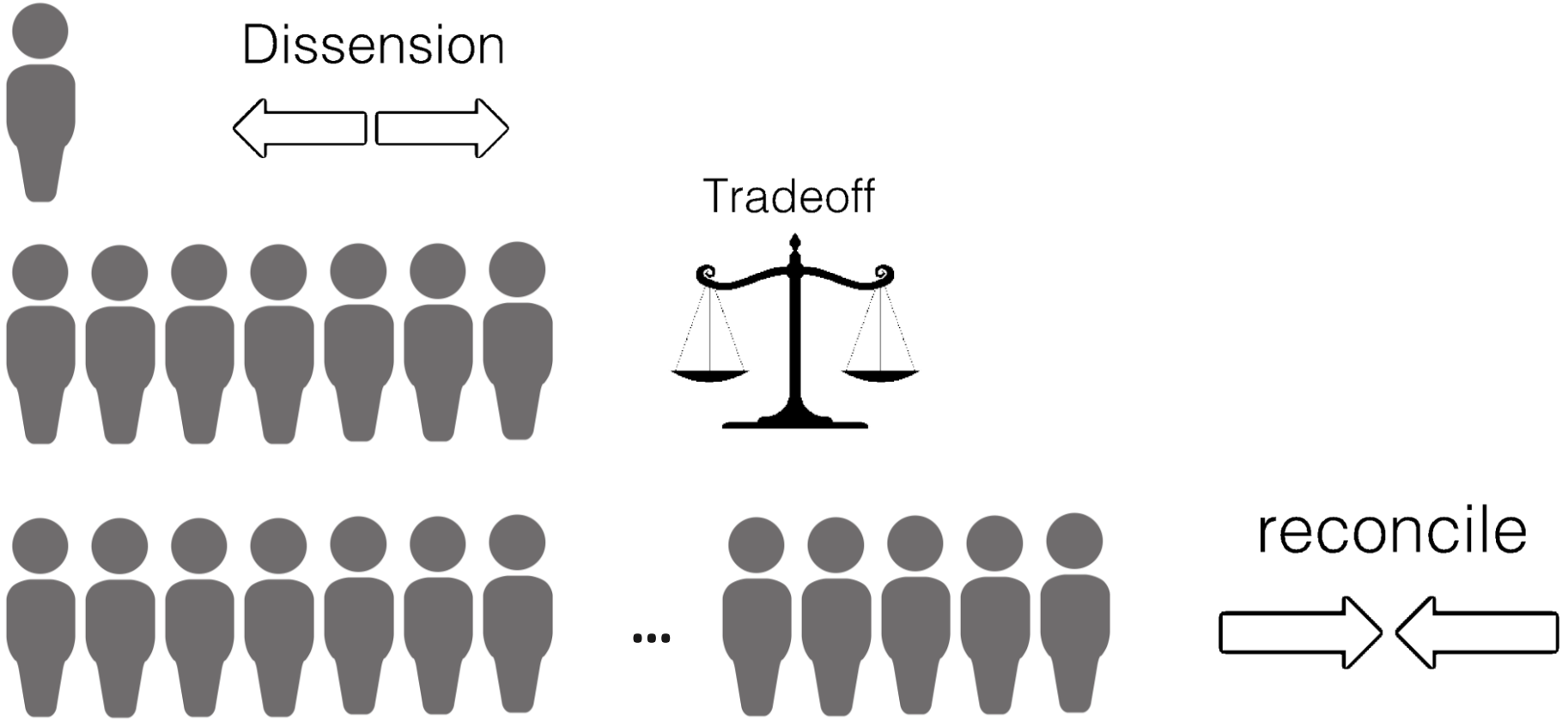


http://snapshot.parabon-nanolabs.com/img/snapshot-report_stack-72DPI.png

- ***Homer et. al. 2008 PLoS genetics:*** aggregated genome data (i.e., allele frequencies) can also be used for re-identifying an individual in a case group with a certain disease

NIH took off all p-values and statistics from their public website

Privacy and utility



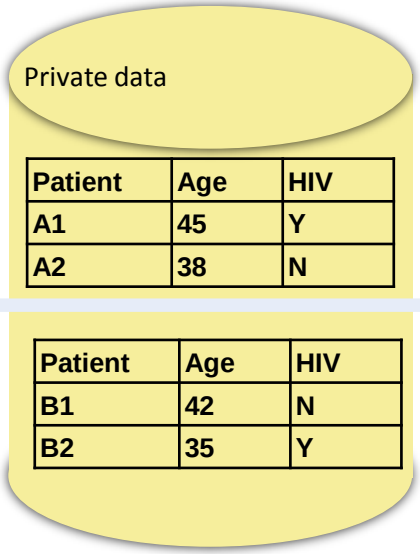
Tradeoff and compromise between privacy and utility are possible when the data analysis target is on the population rather than an individual

Distributed Data Analysis

Horizontally distributed data

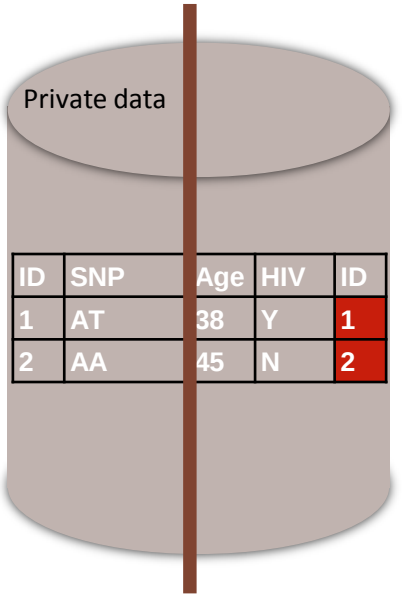
Institution A

Institution B



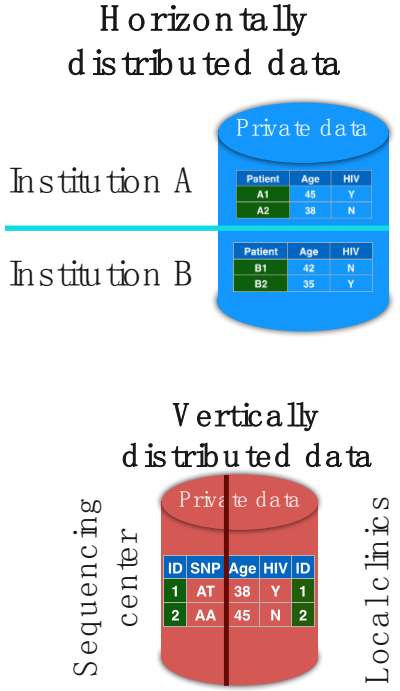
Vertically distributed data

Sequencing center



Local clinic

Common Distributed Data Analysis



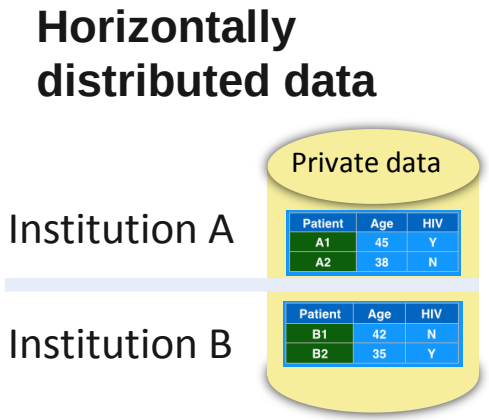
	Logistic Regression	Cox Model (Survival analysis)
Frequentist	GLORE[1,2,3] (2013, 2015)	WebDISCO[5] (2014, 2015)
Bayesian	EXPLORER[4] (2013)	To be investigated
Frequentist	VERTIGO[6] (2015)	DIADEM[WIP] (2016)
Bayesian	To be investigated	To be investigated

Other models

- Distributed computational phenotyping based on tensor factorization

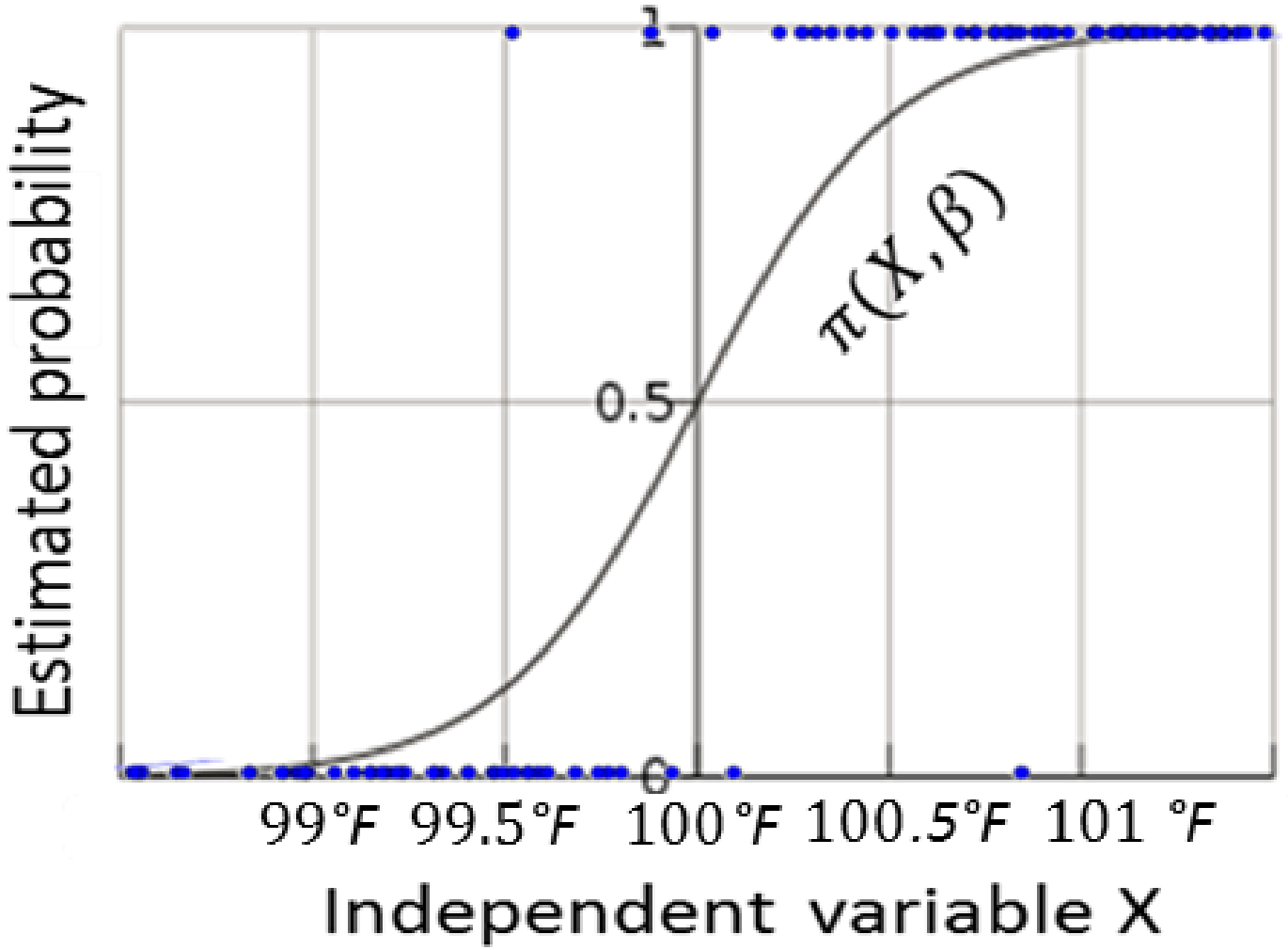
1. Jiang W, Li P, Wang S, *et al.* WebGLORE: a web service for Grid LOgistic REgression. *Bioinformatics* 2013;**29**:3238–40.
2. Wu Y, Jiang X, Wang S, *et al.* Grid multi-category response logistic models. *BMC Med Inform Decis Mak* 2015;**15**:10.
3. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. VERTIcal Grid lOgistic regression (VERTIGO). J Am Med Inform Assoc 2016 May;23(3):570–579. PMID: 26554428
4. Wang, S., Jiang, X., *et al.* . Expectation propagation logistic regression (explorer): distributed privacy-preserving online model learning. *Journal of biomedical informatics*, 46(3), 480-496.
5. Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, Ohno-Machado L. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015 Nov;22(6):1212–1219. PMID: 26159465
6. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. VERTIcal Grid lOgistic regression (VERTIGO). J Am Med Inform Assoc 2016 May;23(3):570–579. PMID: 26554428

GLORE: Grid LOgistic Regression (extends to generalized linear models)



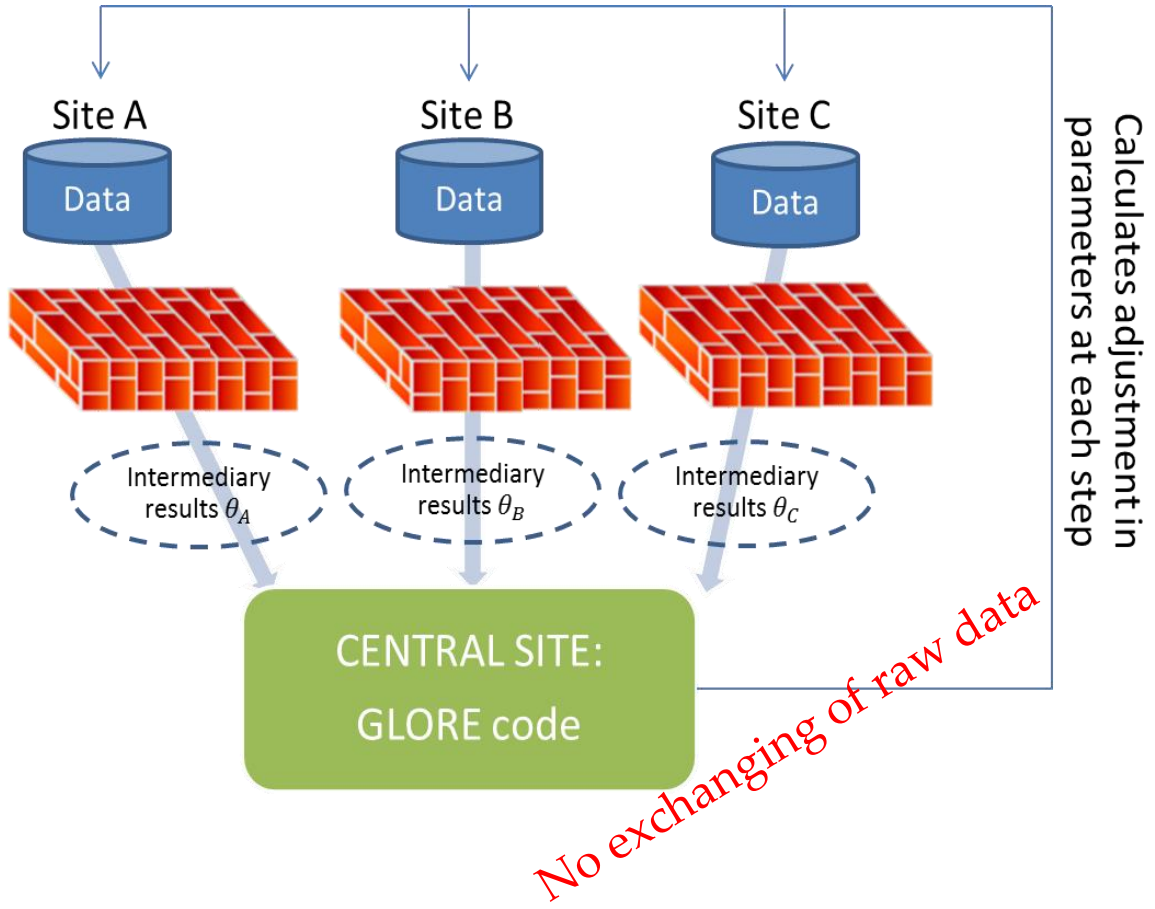
1. Jiang W, Li P, Wang S, *et al.* WebGLORE: a web service for Grid LOgistic REgression. *Bioinformatics* 2013;**29**:3238–40.
2. Wu Y, Jiang X, Wang S, *et al.* Grid multi-category response logistic models. *BMC Med Inform Decis Mak* 2015;**15**:10.
3. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. VERTIcal Grid lOgistic regression (VERTIGO). J Am Med Inform Assoc 2016 May;23(3):570–579. PMID: 26554428

Logistic Regression



Foundation of GLORE

- Support *m-1* features are consistent over *k* sites
- In each iteration, intermediary result of a *mxm* matrix and a *m*-dimensional vector are transmitted to *k-1* sites



Wu Y, Jiang X, Kim J, et al. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc 2012;2012:758–64.

Maximum Likelihood Estimation

- Estimated probability based on observations of a binary response Y and covariates X

$$P(Y = 1|X) = \pi(X, \beta) = \frac{1}{1 + e^{-X\beta}}$$

Binary response

Covariates

Logit function

Model parameter

- Likelihood function based on observed data (centralized)

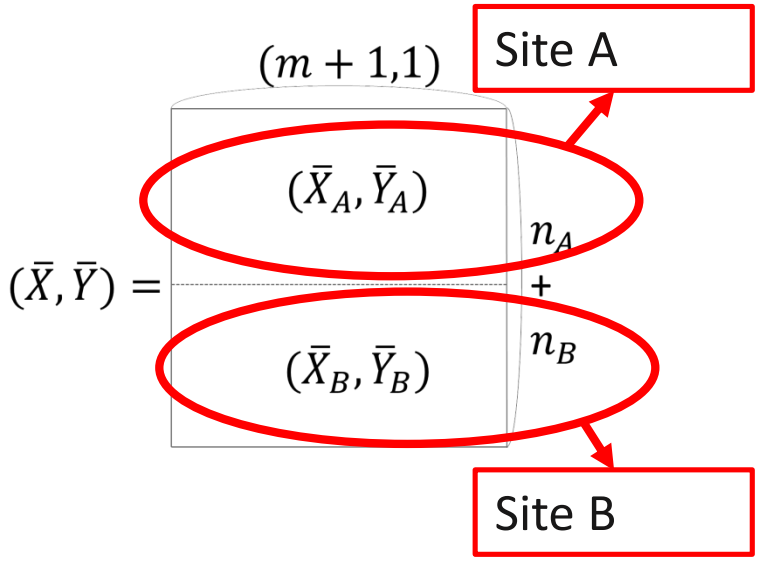
Number of records

$$l(\beta) = \sum_{1}^{n_A + n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

Maximum Likelihood Estimation

- Likelihood function based on observed data (distributed)

$$P(Y = 1|X) = \pi(X, \beta) = \frac{1}{1 + e^{-X\beta}}$$



Number of records held by site A

Number of records held by site B

$$l(\beta) = \sum_{1}^{n_A + n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

Maximum Likelihood Estimation

- Newton-Raphson algorithm for calculation

$$P(Y = 1|X) = \pi(X, \beta) = \frac{1}{1 + e^{-X\beta}}$$

$$(\bar{X}, \bar{Y}) = \begin{matrix} (m+1,1) \\ \begin{matrix} (\bar{X}_A, \bar{Y}_A) \\ \hline (\bar{X}_B, \bar{Y}_B) \end{matrix} \end{matrix} \begin{matrix} \\ n_A \\ + \\ n_B \end{matrix}$$

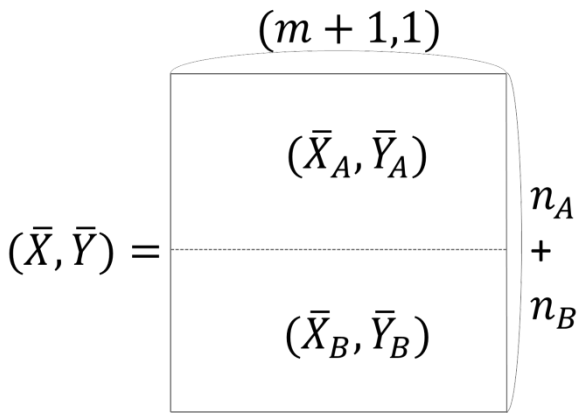
$l(\beta)$ is a concave function

$$\Rightarrow \quad l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$\beta^{(k+1)} = \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)^T}} \right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}}$$

$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$



$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$\beta^{(k+1)} = \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)^T}} \right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}}$$

$$= \beta^{(k)} + \left[\bar{X}^T W\left(\bar{X}, \beta^{(k)}\right) \bar{X} \right]^{-1} \bar{X}^T \left[\bar{Y} - \Pi\left(\bar{X}, \beta^{(k)}\right) \right]$$

Global variance-covariance matrix

Global prediction outcomes

$$(\bar{X}, \bar{Y}) = \begin{matrix} (m+1,1) \\ \left.\begin{matrix} (\bar{X}_A, \bar{Y}_A) \\ + \\ (\bar{X}_B, \bar{Y}_B) \end{matrix}\right. \end{matrix} \begin{matrix} \\ n_A \\ \\ n_B \end{matrix}$$

$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$\begin{aligned}
\beta^{(k+1)} &= \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)^T}} \right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}} \\
&= \beta^{(k)} + \left[\bar{X}^T W\left(\bar{X}, \beta^{(k)}\right) \bar{X} \right]^{-1} \bar{X}^T \left[\bar{Y} - \Pi\left(\bar{X}, \beta^{(k)}\right) \right] \\
&= \beta^{(k)} + \left[\bar{X}_A^T W_A\left(\bar{X}_A, \beta^{(k)}\right) \bar{X}_A + \bar{X}_B^T W_B\left(\bar{X}_B, \beta^{(k)}\right) \bar{X}_B \right]^{-1} \\
&\quad \cdot \{ \bar{X}_A^T [\bar{Y}_A - \Pi_A(\bar{X}_A, \beta)] + \bar{X}_B^T [\bar{Y}_B - \Pi_B(\bar{X}_B, \beta)] \}.
\end{aligned}$$

Local variance-covariance matrix

Local prediction outcomes

$$(\bar{X}, \bar{Y}) = \begin{matrix} (m+1,1) \\ \left.\begin{matrix} (\bar{X}_A, \bar{Y}_A) \\ + \\ (\bar{X}_B, \bar{Y}_B) \end{matrix}\right\} \begin{matrix} n_A \\ \\ n_B \end{matrix} \end{matrix}$$

$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$\begin{aligned}
\beta^{(k+1)} &= \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)^T}} \right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}} \\
&= \beta^{(k)} + \left[\bar{X}^T W\left(\bar{X}, \beta^{(k)}\right) \bar{X} \right]^{-1} \bar{X}^T \left[\bar{Y} - \Pi\left(\bar{X}, \beta^{(k)}\right) \right] \\
&= \beta^{(k)} + \left[\bar{X}_A^T W_A\left(\bar{X}_A, \beta^{(k)}\right) \bar{X}_A + \bar{X}_B^T W_B\left(\bar{X}_B, \beta^{(k)}\right) \bar{X}_B \right]^{-1} \\
&\quad \cdot \left\{ \bar{X}_A^T \left[\bar{Y}_A - \Pi_A(\bar{X}_A, \beta) \right] + \bar{X}_B^T \left[\bar{Y}_B - \Pi_B(\bar{X}_B, \beta) \right] \right\}.
\end{aligned}$$

$$(\bar{X}, \bar{Y}) = \begin{matrix} (m+1,1) \\ \left.\begin{matrix} (\bar{X}_A, \bar{Y}_A) \\ \hline (\bar{X}_B, \bar{Y}_B) \end{matrix}\right\} \begin{matrix} n_A \\ + \\ n_B \end{matrix} \end{matrix}$$

Local variance-covariance matrix

$$W_A(\bar{X}_A, \beta) = \begin{bmatrix} \pi(x_1, \beta)(1 - \pi(x_1, \beta)) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \pi(x_{n_A}, \beta)(1 - \pi(x_{n_A}, \beta)) \end{bmatrix},$$

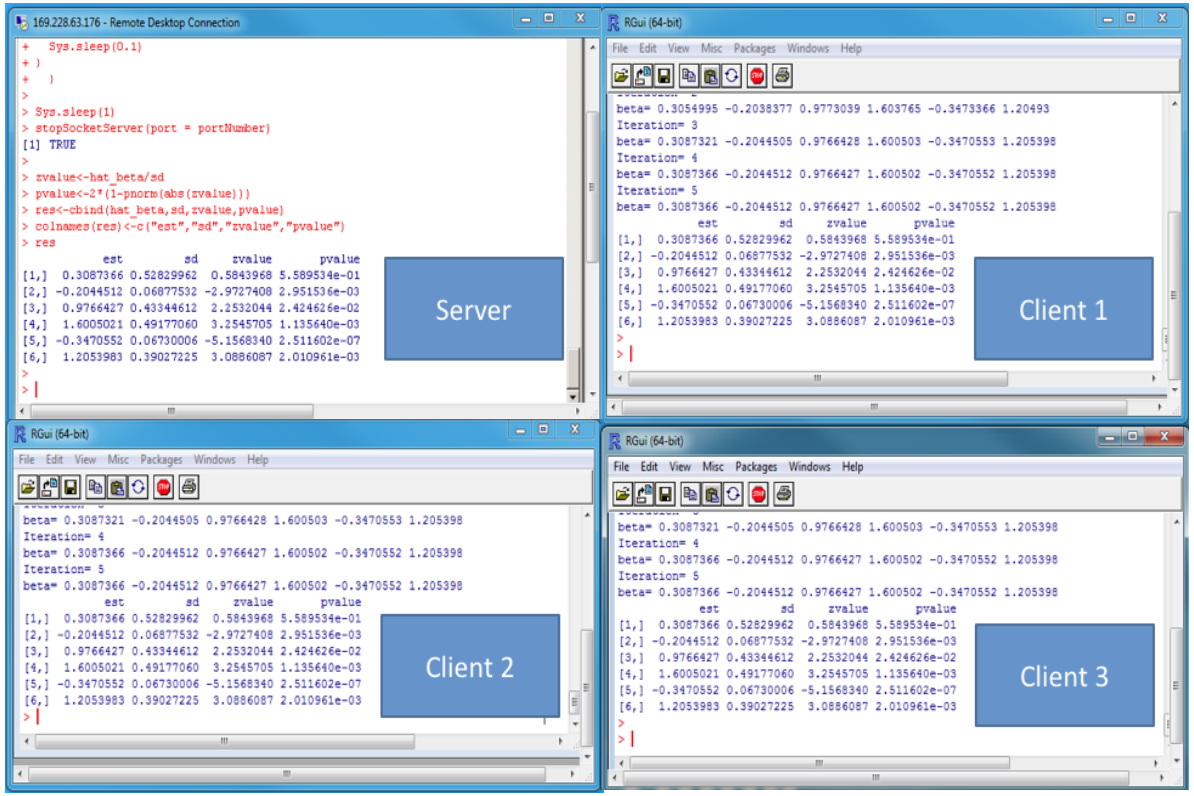
$$W_B(\bar{X}_B, \beta) = \begin{bmatrix} \pi(x_{n_A+1}, \beta)(1 - \pi(x_{n_A+1}, \beta) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \pi(x_{n_A+n_B}, \beta)(1 - \pi(x_{n_A+n_B}, \beta)) \end{bmatrix},$$

$$\Pi_A(\bar{X}_A, \beta) = \begin{bmatrix} \pi(x_1, \beta) \\ \vdots \\ \pi(x_{n_A}, \beta) \end{bmatrix}, \text{ and } \Pi_B(\bar{X}_B, \beta) = \begin{bmatrix} \pi(x_{n_A+1}, \beta) \\ \vdots \\ \pi(x_{n_A+n_B}, \beta) \end{bmatrix}.$$

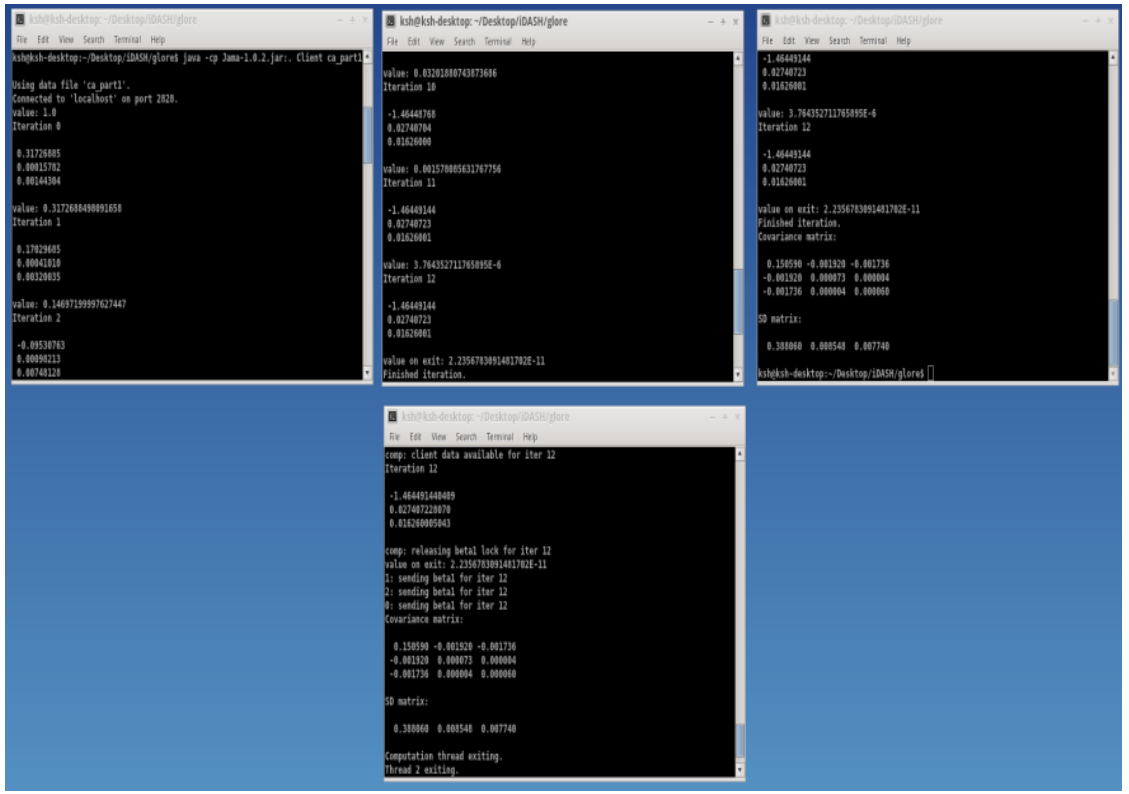
Local prediction outcomes

Backbone implementation

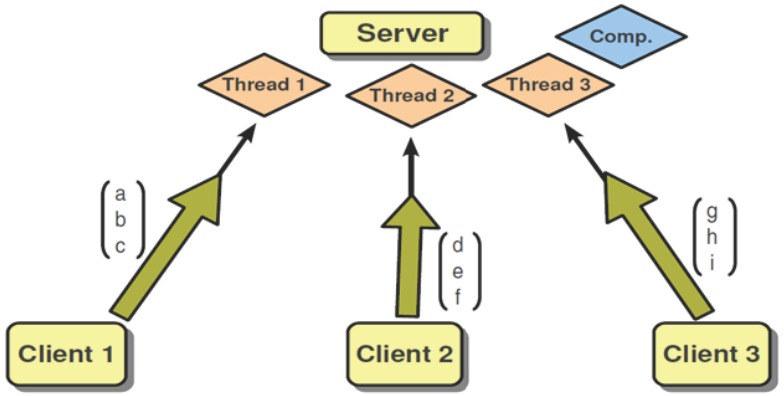
- R backbone



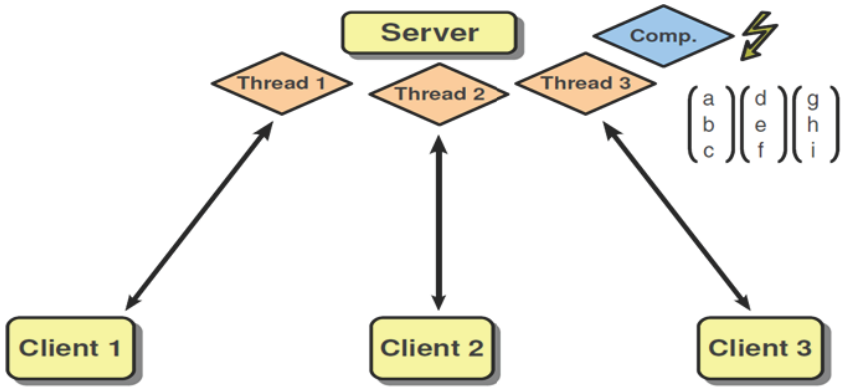
- JAVA backbone



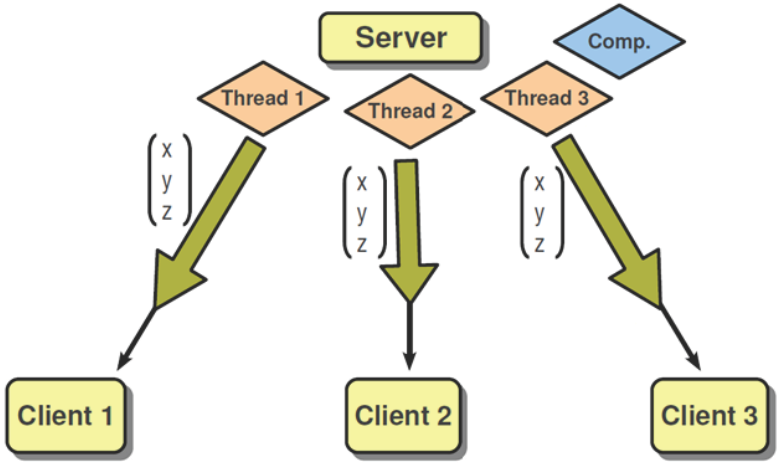
Applet-Servlet architecture



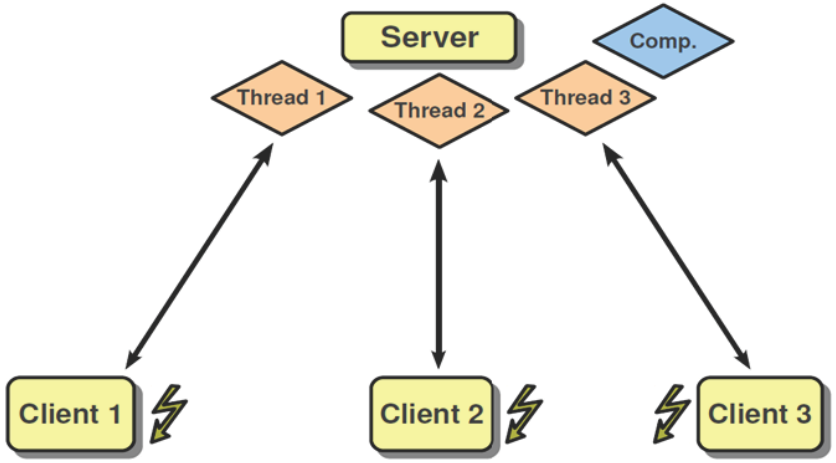
(a) Clients send partial results to server



(b) Server integrates the results

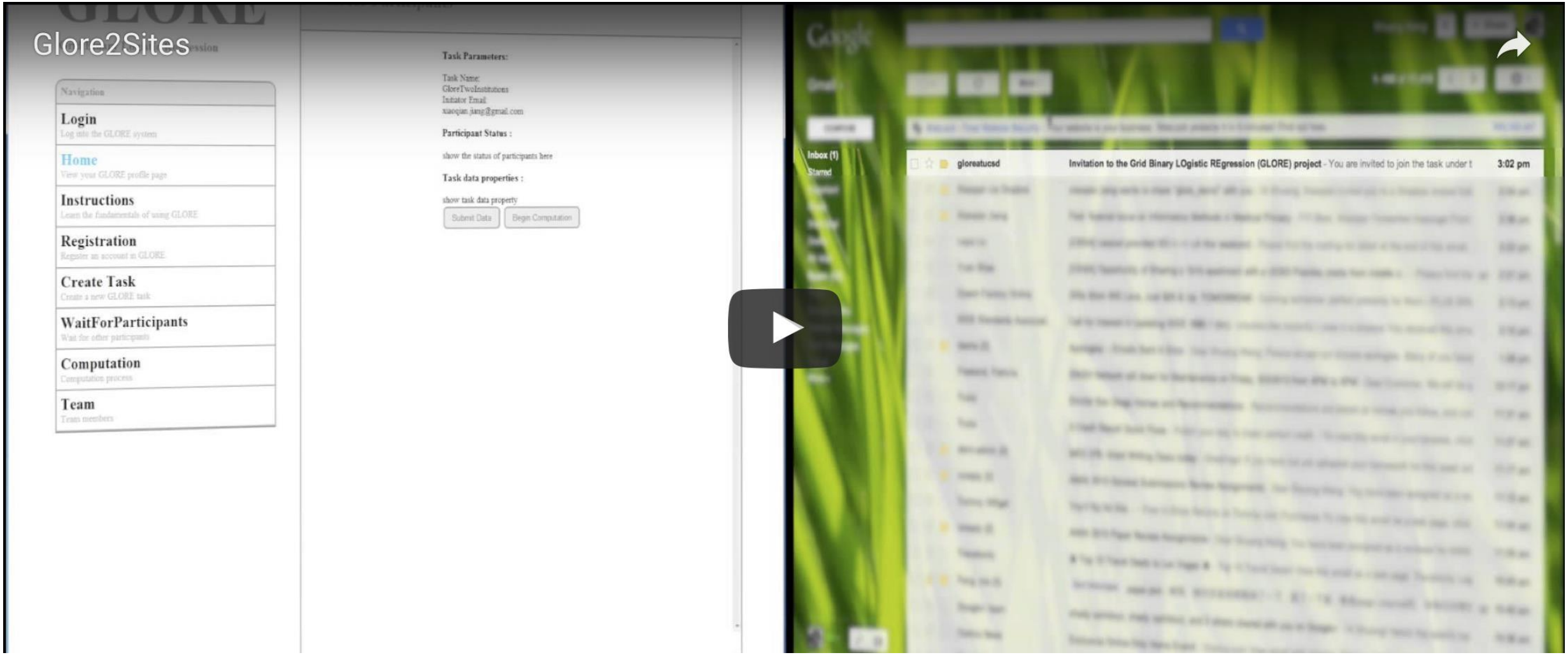


(c) Sever sends global statistics to clients



(d) Clients compute partial results locally

A quick demo


Glore2Sites
Navigation
Login
Log into the GLORE system
Home
View your GLORE profile page
Instructions
Learn the fundamentals of using GLORE
Registration
Register an account in GLORE
Create Task
Create a new GLORE task
WaitForParticipants
Wait for other participants
Computation
Computation process
Team
Team members
Task Parameters:
Task Name:
GloreTwoInstitutions
Initiator Email:
xiaoqian.jiang@gmail.com
Participant Status :
show the status of participants here
Task data properties :
show task data property
Submit Data
Begin Computation
Inbox (1)
Starred
gloreatucsd
Invitation to the Grid Binary LOgistic REgression (GLORE) project - You are invited to join the task under t
3:02 pm

Experiments

Breast cancer biomarkers (CA-19, CA-125)

	Estimate	Std. Error	Z-value	Pr(>|z|)
Intercept	-1.4645	0.3881	-3.7739	1.61E-04
CA19	0.0274	0.0085	3.2063	1.34E-03
CA125	0.0163	0.0077	2.1008	3.57E-02

H-L test p-value = 0.891

AUC =0.891

Edinburgh myocardial infraction data

	Estimate	Std. Error	Z-value	Pr(>|z|)
Intercept	-4.3485	0.2968	-14.6508	0.00E+00
Pain in left arm	0.1816	0.2680	0.6777	4.98E-01
Pain in right arm	0.1764	0.3061	0.5763	5.64E-01
Nausea	0.1323	0.3862	0.3426	7.32E-01
Hypoperfusion	2.2511	0.6590	3.4160	6.36E-04
ST elevation	5.5556	0.4404	12.6150	0.00E+00
New Q waves	4.1453	0.6747	6.1435	8.07E-10
ST depression	3.4173	0.2815	12.1392	0.00E+00
T wave inversion	1.2030	0.2635	4.5649	5.00E-06
Sweating	0.2721	0.2510	1.0837	2.79E-01

H-L test p-value = 0.439

AUC =0.699

Experiments

- Cincinnati data (ImproveCareNow! CDRN)

A quality improvement and research collaborative focused on improving the care and outcomes of children with Inflammatory Bowel Disease

Site 1 – 245 observations on 5 patients.
Site 2 – 563 observations on 24 patients.

Features

F1 – patient id
F2 – weeks to response
F3 – patient on biologics
F4 – days since diagnosis
F5 – gender
F6 – Race
F7 – Age in years at start of treatment
F8 – Extent of disease
F9 – patient on thiopurine
F10 – patient on methotrexate
F11 – patient on salicylate
F12 – patient on steroids
F13 – days since diagnosis (recorded variable)
F14 – gender (recorded variable)
F15 – race (recorded variable)
F16 – race (factor variable)
F17 – patient on steroid (factor variable)
F18 – patient on salicylate (factor variable)
F19 – patient on thiopurine (factor variable)
F20 – patient on methotrexate (factor variable)
F21 – patient diagnosis
F22 – patient diagnosis (factor variable)

Target

Target = responded to treatment (i.e., improvement in condition)

Experiments

Predictor	Beta	SE	Z-statistics	df	p	Odds ratio
Intercept	4.8802	2581.989	0.0019	1	0.9985	N/A
F1	0.0034	0.0016	2.1977	1	0.028	1.0035
F2	0.1143	0.0373	3.0652	1	0.0022	1.1211
F3	1.8766	0.9398	1.9969	1	0.0458	6.5311
F4	0.0027	0.0012	2.206	1	0.0274	1.0027
F5	-1.7232	1290.995	-0.0013	1	0.9989	0.1785
F6	-0.7147	0.4921	-1.4523	1	0.1464	0.4893
F7	-0.5522	0.1909	-2.8926	1	0.0038	0.5757
F8	0.0673	0.1231	0.5469	1	0.5845	1.0696
F9	-0.8537	2236.068	-0.0004	1	0.9997	0.4259
F10	0	3162.278	0	1	1	1
F11	0.5396	2236.068	0.0002	1	0.9998	1.7154
F12	0.3057	2236.068	0.0001	1	0.9999	1.3576
F13	0.0245	1.0657	0.023	1	0.9816	1.0248
F14	0.7519	1290.995	0.0006	1	0.9995	2.1211
F15	0.5949	2236.068	0.0003	1	0.9998	1.8128
F16	0.5949	2236.068	0.0003	1	0.9998	1.8128
F17	0.3057	2236.068	0.0001	1	0.9999	1.3576
F18	0.5396	2236.068	0.0002	1	0.9998	1.7154
F19	-0.8537	2236.068	-0.0004	1	0.9997	0.4259
F20	0	3162.278	0	1	1	1
F21	-0.3472	2236.068	-0.0002	1	0.9999	0.7066
F22	-0.3472	2236.068	-0.0002	1	0.9999	0.7066

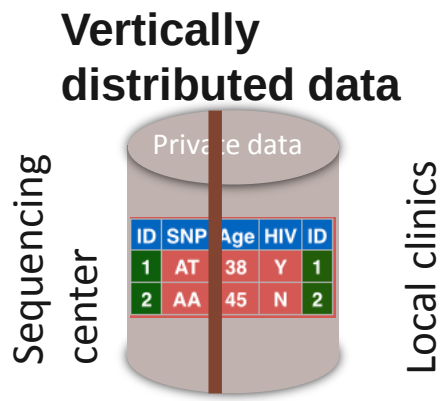
Calibration Error = 0.05

AUC = 0.744

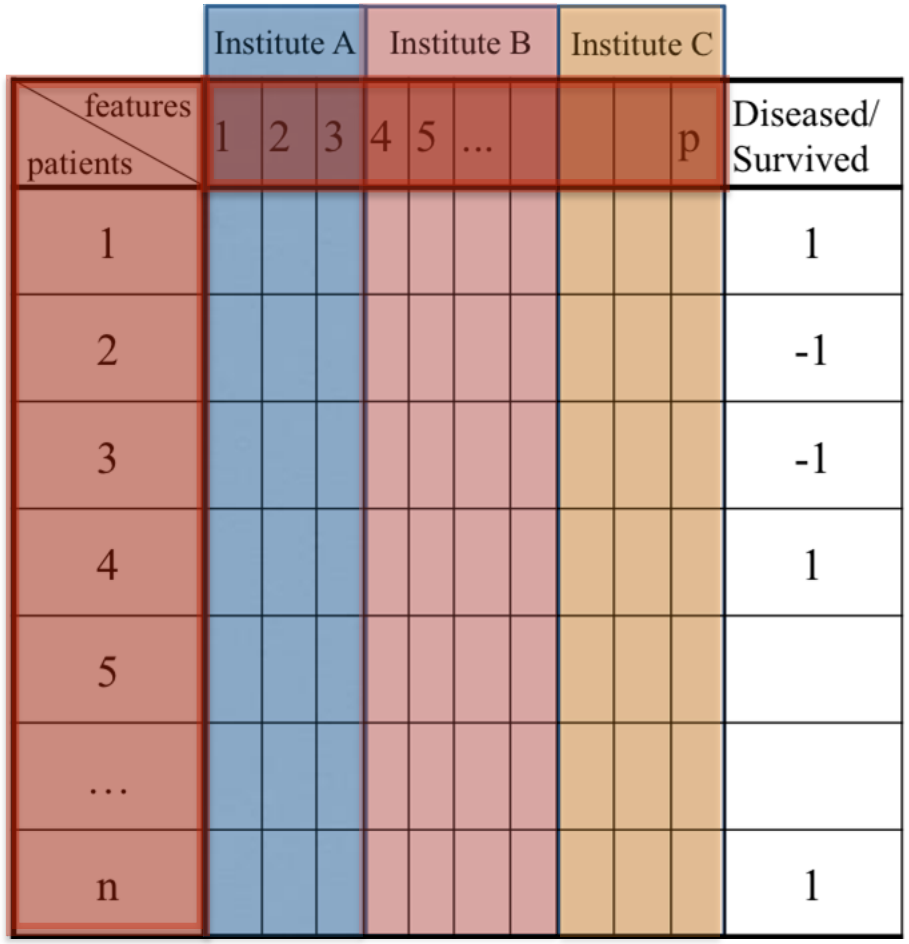
HL-C = 0.26

HL-H = 0.59

VERTIGO: VERTIcal Grid lOgistic regression

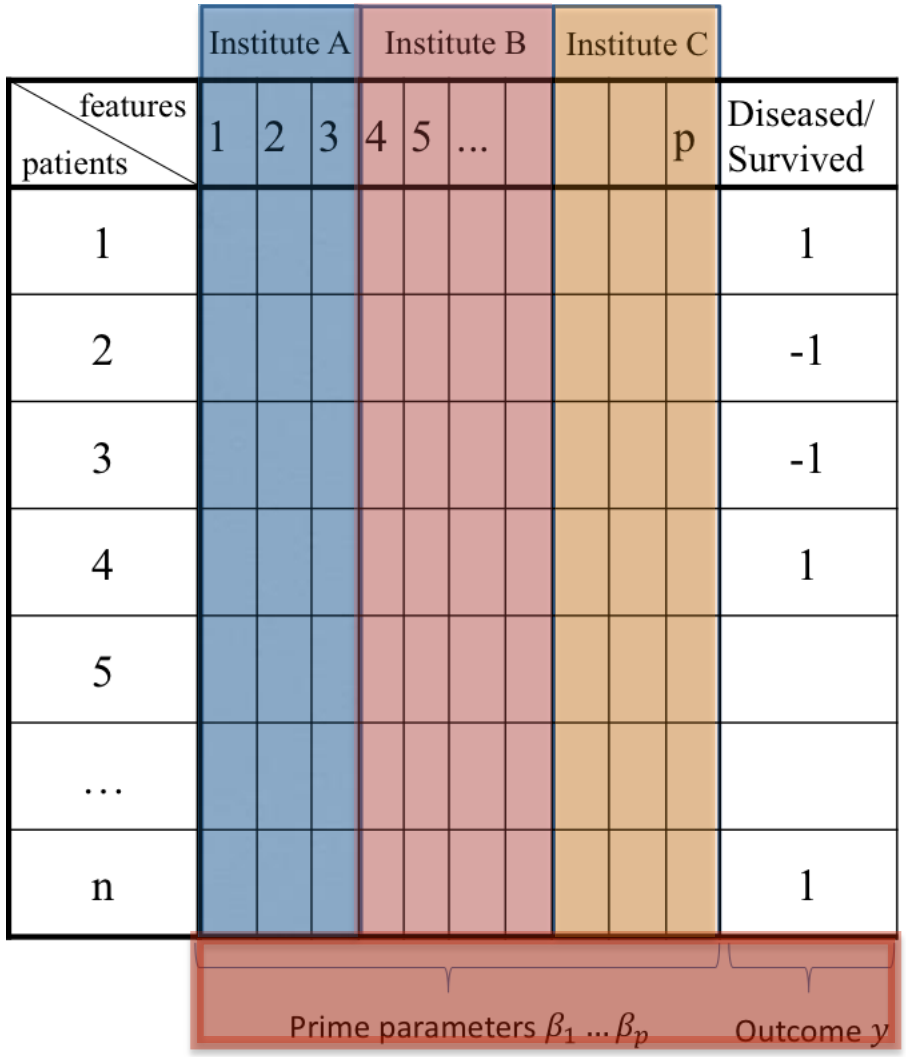


VERTIcal Grid lOgistic regression (VERTIGO)



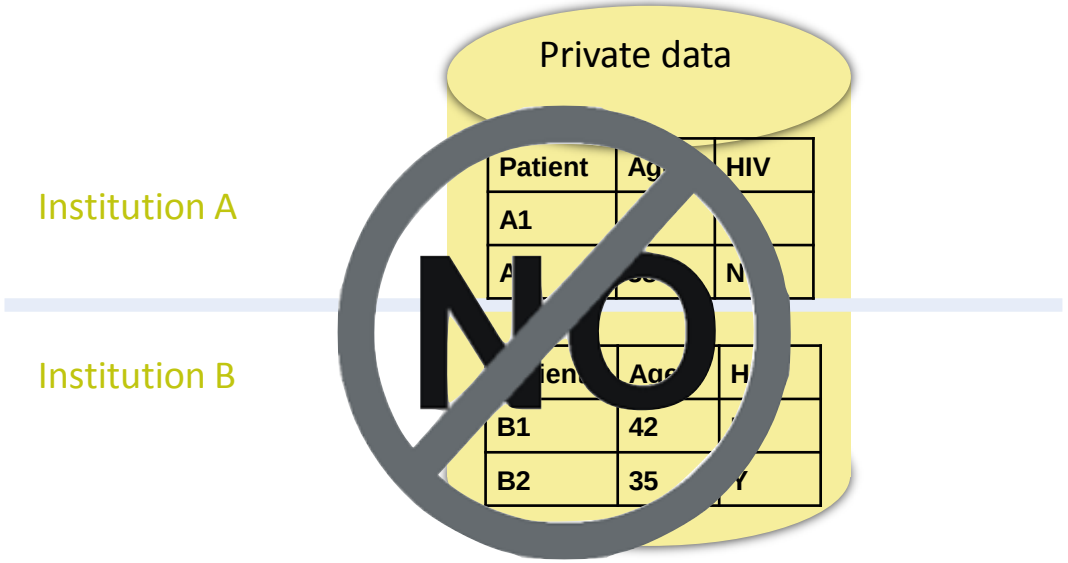
	Institute A			Institute B				Institute C			
features / patients	1	2	3	4	5	...				p	Diseased/ Survived
1											1
2											-1
3											-1
4											1
5											
…											
n											1

VERTIcal Grid lOgistic regression (VERTIGO)



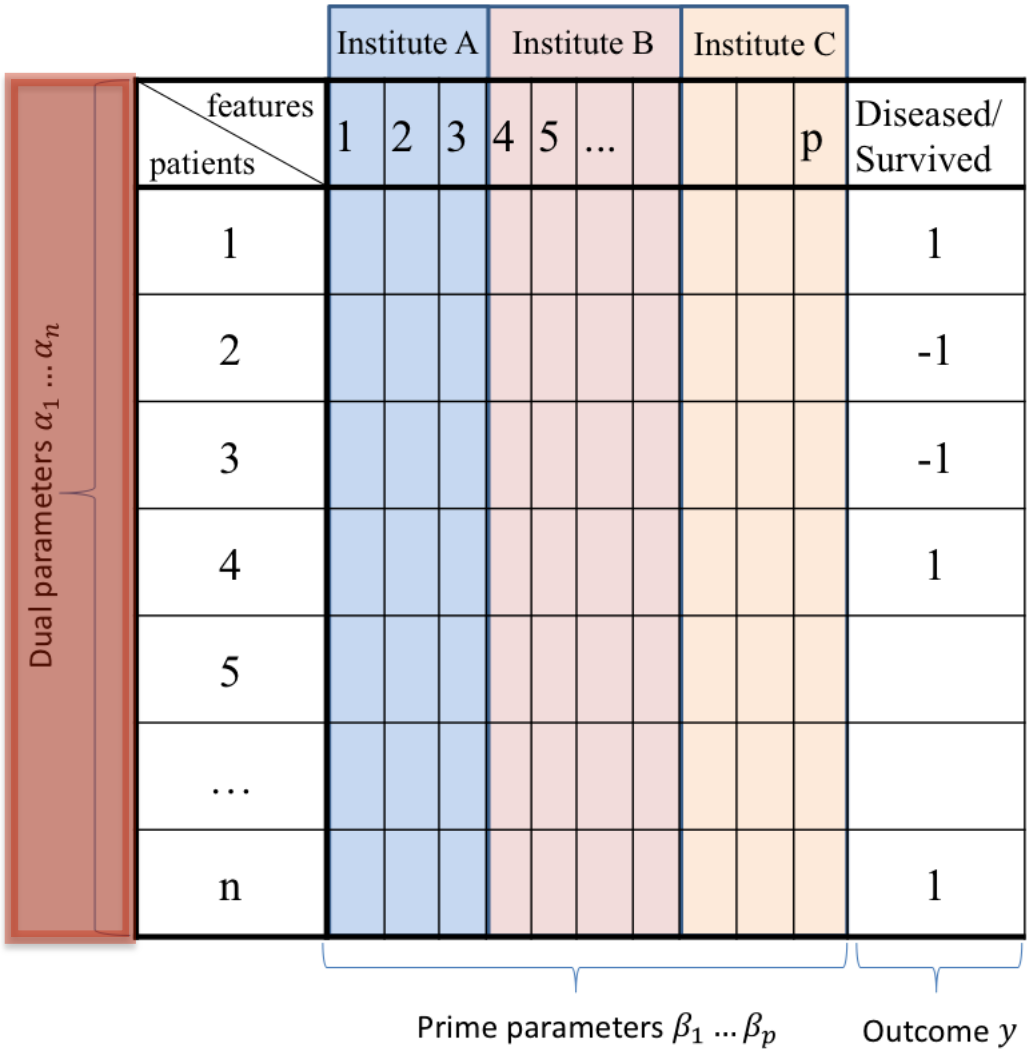
GLORE (horizontally distributed data)

Horizontally distributed data



1. Jiang W, Li P, Wang S, et al. WebGLORE: a web service for Grid LOgistic REgression. Bioinformatics 2013;29:3238–40.
2. Wu Y, Jiang X, Wang S, et al. Grid multi-category response logistic models. BMC Med Inform Decis Mak 2015;15:10.

VERTIcal Grid lOgistic regression (VERTIGO)



Primal form

GLORE (horizontally distributed data)

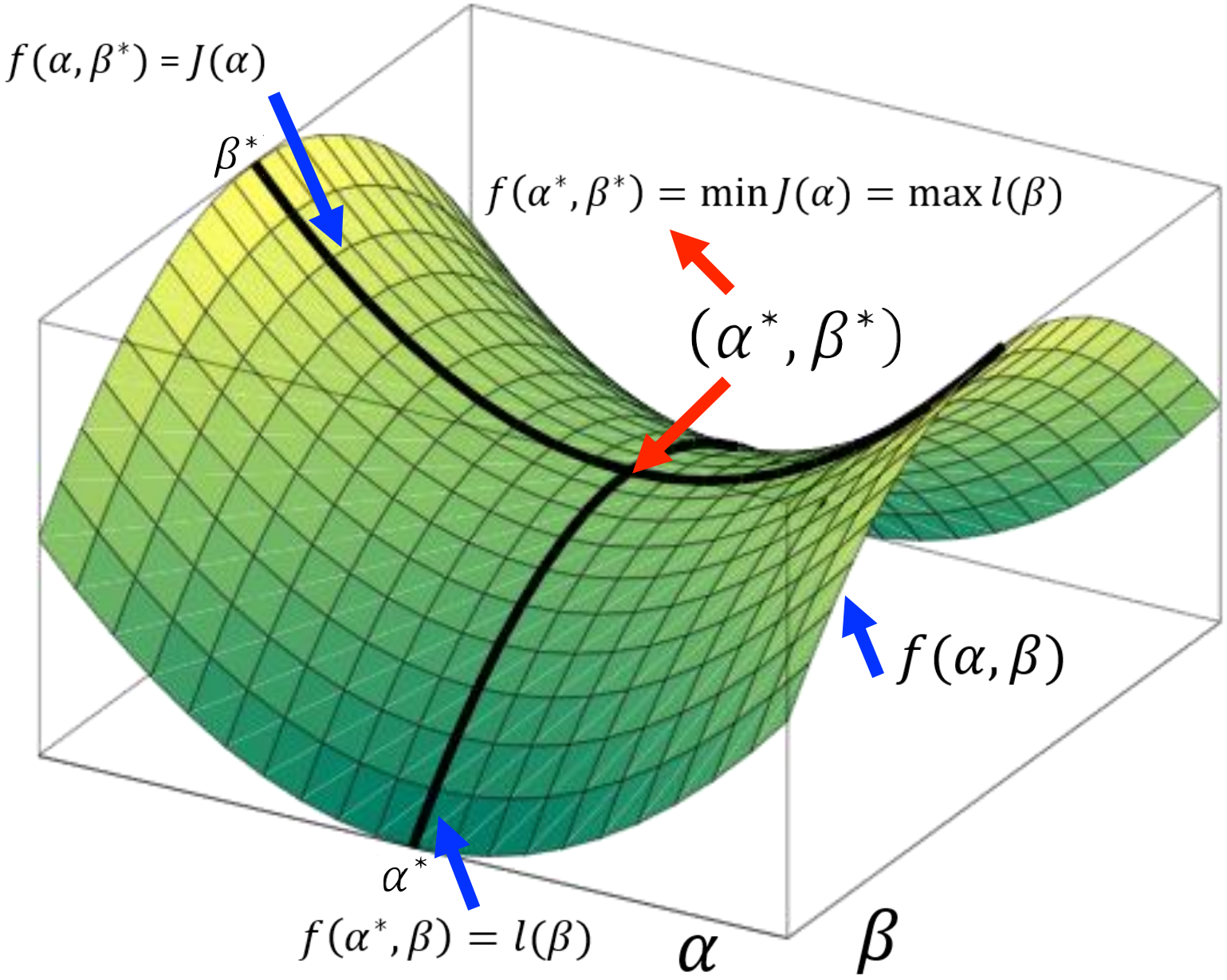
$$\text{Log-likelihood: } l(\beta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \beta^T z_i)) - \frac{\lambda}{2}\beta^T\beta$$

Prime parameter

$$\text{Objective: } \max_{\beta} l(\beta)$$

VERTIGO (Vertically distributed data)

VERTIcal Grid lOgistic regression (VERTIGO)



Primal form

Log-likelihood: $l(\beta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \beta^T z_i)) - \frac{\lambda}{2}\beta^T\beta$

Prime parameter

Objective: $\max_{\beta} l(\beta)$

Dual form

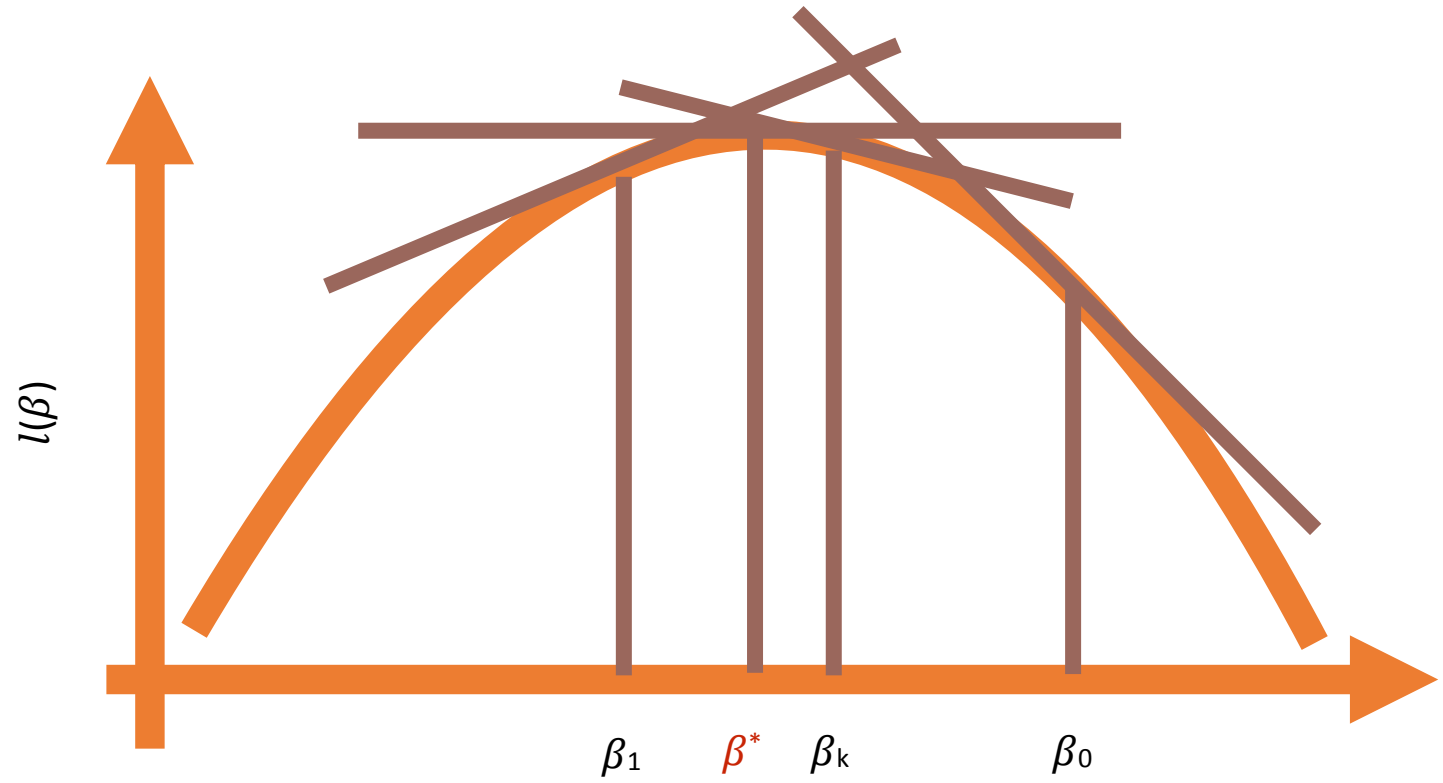
Log-likelihood: $J(\alpha) = \frac{1}{2\lambda}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j z_i^T z_j - \sum_{i=1}^{n} H(\alpha_i)$

$H(\alpha_i) = -\alpha_i \log \alpha_i - (1 - \alpha_i)\log(1 - \alpha_i)$

Dual parameter

Objective: $\min_{\alpha} J(\alpha)$

Newton-Raphson (NR) Algorithm



$$l(\boldsymbol{\beta}) = \sum_{i=1}^{D} \{\boldsymbol{\beta}^T \sum_{l \in \mathcal{D}_i} \mathbf{z}^l - d_i \log[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)]\}$$

$$\beta^{(k+1)} = \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)^T}}\right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}}$$

Distributed NR Algorithm for VERTIGO in Dual Form

	First derivative:	**Hessian matrix:**
Centralized	$J_i'(\alpha) = \lambda^{-1} y_i \sum_j \alpha_j y_j z_j^T z_i + log\frac{\alpha_i}{1-\alpha_i}$	$J_{i,j}''(\alpha) = \lambda^{-1}\mathrm{diag}(y) Z Z^T \mathrm{diag}(y) + \mathrm{diag}(\frac{\alpha_i}{1-\alpha_i})$
Distributed	$J_i'(\alpha) = \lambda^{-1} y_i \sum_j \alpha_j y_j \left(\sum_k \left(z_j^{\mathrm{site}_k}\right)^T \left(z_i^{\mathrm{site}_k}\right) \right) + \log\frac{\alpha_i}{1-\alpha_i}$	$J_{i,j}''(\alpha) = \lambda^{-1}\mathrm{diag}(y) \left(\sum_k \left(Z^{\mathrm{site}_k}\right)\left(Z^{\mathrm{site}_k}\right)^T \right) \mathrm{diag}(y) + \mathrm{diag}(\frac{\alpha_i}{1-\alpha_i})$

Local statistics from site k

Local statistics from site k

Distributed NR Algorithm for VERTIGO in Dual Form

First derivative:

Hessian matrix:

Distributed

$$J_i'(\alpha) = \lambda^{-1} y_i \sum_j \alpha_j y_j \left(\sum_k \left(z_j^{\text{site}_k}\right)^T \left(z_i^{\text{site}_k}\right) \right) + \log\frac{\alpha_i}{1-\alpha_i}$$

$$J_{i,j}''(\alpha) = \lambda^{-1}\,\text{diag}(y) \left(\sum_k (Z^{\text{site}_k})(Z^{\text{site}_k})^T \right) \text{diag}(y) + \text{diag}\left(\frac{\alpha_i}{1-\alpha_i}\right)$$

$$\left(z_j^{\text{site}_1}\right)^T \left(z_i^{\text{site}_1}\right) + \left(z_j^{\text{site}_2}\right)^T \left(z_i^{\text{site}_2}\right)$$

$$\left(z_j^{\text{site}_1}\right)^T \left(z_1^{\text{site}_1}\right)$$

Site 1			
Patients	Features		Response
	1	2	y
1: $(z_1^{\text{site}_1})^T$	3	87	1
2: $(z_2^{\text{site}_1})^T$	4	91	1
3: $(z_3^{\text{site}_1})^T$	4	100	-1
	Z^{site_1}		

j=1	3	87	X	3 87	=	7578
j=2	4	91	X	3 87	=	7929
j=3	4	100	X	3 87	=	8712

Site 2			
Patients	Features		Response
	3	4	y
1: $(z_1^{\text{site}_2})^T$	1	1	1
2: $(z_2^{\text{site}_2})^T$	0	1	1
3: $(z_3^{\text{site}_2})^T$	1	0	-1
	Z^{site_2}		

Distributed NR Algorithm for VERTIGO in Dual Form

First derivative:

Distributed

$$J_i'(\alpha) = \lambda^{-1} y_i \sum_j \alpha_j y_j \left(\sum_k \left(z_j^{\text{site}_k}\right)^T \left(z_i^{\text{site}_k}\right) \right) + \log \frac{\alpha_i}{1-\alpha_i}$$

$$\left(z_j^{\text{site}_1}\right)^T \left(z_i^{\text{site}_1}\right) + \left(z_j^{\text{site}_2}\right)^T \left(z_i^{\text{site}_2}\right)$$

$$\left(z_j^{\text{site}_2}\right)^T \left(z_1^{\text{site}_2}\right)$$

Hessian matrix:

$$J_{i,j}''(\alpha) = \lambda^{-1} \text{diag}(y) \left(\sum_k \left(Z^{\text{site}_k}\right)\left(Z^{\text{site}_k}\right)^T \right) \text{diag}(y) + \text{diag}\left(\frac{\alpha_i}{1-\alpha_i}\right)$$

Site 1			
Patients	Features		Response
	1	2	y
1: $(z_1^{\text{site}_1})^T$	3	87	1
2: $(z_2^{\text{site}_1})^T$	4	91	1
3: $(z_3^{\text{site}_1})^T$	4	100	-1
	Z^{site_1}		

j=1 [1 1] X [1; 1] = 2

j=2 [0 1] X [1; 1] = 1

j=3 [1 0] X [1; 1] = 1

Site 2			
Patients	Features		Response
	3	4	y
1: $(z_1^{\text{site}_2})^T$	1	1	1
2: $(z_2^{\text{site}_2})^T$	0	1	1
3: $(z_3^{\text{site}_2})^T$	1	0	-1
	Z^{site_2}		

Distributed NR Algorithm for VERTIGO in Dual Form

First derivative:

Distributed

$$J_i'(\alpha) = \lambda^{-1} y_i \sum_j \alpha_j y_j \left(\sum_k \left(z_j^{\text{site}_k}\right)^T \left(z_i^{\text{site}_k}\right) \right) + \log \frac{\alpha_i}{1-\alpha_i}$$

Hessian matrix:

$$J_{i,j}''(\alpha) = \lambda^{-1} \text{diag}(y) \left(\sum_k \left(Z^{\text{site}_k}\right)\left(Z^{\text{site}_k}\right)^T \right) \text{diag}(y) + \text{diag}\left(\frac{\alpha_i}{1-\alpha_i}\right)$$

$$\left(z_j^{\text{site}_1}\right)^T \left(z_i^{\text{site}_1}\right) + \left(z_j^{\text{site}_2}\right)^T \left(z_i^{\text{site}_2}\right)$$

$$\left(z_j^{\text{site}_1}\right)^T \left(z_1^{\text{site}_1}\right) + \left(z_j^{\text{site}_2}\right)^T \left(z_1^{\text{site}_2}\right)$$

Site 1			
Patients	Features		Response
	1	2	y
1: $(z_1^{\text{site}_1})^T$	3	87	1
2: $(z_2^{\text{site}_1})^T$	4	91	1
3: $(z_3^{\text{site}_1})^T$	4	100	-1
	Z^{site_1}		

j=1	7578	+	2
j=2	7929	+	1
j=3	8712	+	1

Site 2			
Patients	Features		Response
	3	4	y
1: $(z_1^{\text{site}_2})^T$	1	1	1
2: $(z_2^{\text{site}_2})^T$	0	1	1
3: $(z_3^{\text{site}_2})^T$	1	0	-1
	Z^{site_2}		

Only exchange aggregated local statistics

Distributed NR Algorithm for VERTIGO in Dual Form

First derivative:

Distributed

$$J_i'(\alpha) = \lambda^{-1} y_i \sum_j \alpha_j y_j \left(\sum_k \left(z_j^{\text{site}_k}\right)^T \left(z_i^{\text{site}_k}\right) \right) + \log\frac{\alpha_i}{1-\alpha_i}$$

$$\left(z_j^{\text{site}_1}\right)^T \left(z_i^{\text{site}_1}\right) + \left(z_j^{\text{site}_2}\right)^T \left(z_i^{\text{site}_2}\right)$$

Hessian matrix:

$$J_{i,j}''(\alpha) = \lambda^{-1}\,\text{diag}(y) \left(\sum_k \left(Z^{\text{site}_k}\right)\left(Z^{\text{site}_k}\right)^T \right) \text{diag}(y) + \text{diag}\left(\frac{\alpha_i}{1-\alpha_i}\right)$$

$$\left(Z^{\text{site}_1}\right)\left(Z^{\text{site}_1}\right)^T + \left(Z^{\text{site}_2}\right)\left(Z^{\text{site}_2}\right)^T$$

Site 1			
Patients	Features		Response
	1	2	y
1: $(z_1^{\text{site}_1})^T$	3	87	1
2: $(z_2^{\text{site}_1})^T$	4	91	1
3: $(z_3^{\text{site}_1})^T$	4	100	-1
	Z^{site_1}		

$$\left(Z^{\text{site}_1}\right)\left(Z^{\text{site}_1}\right)^T = \begin{bmatrix} 3 & 87 \\ 4 & 91 \\ 4 & 100 \end{bmatrix} X \begin{bmatrix} 3 & 4 & 4 \\ 87 & 91 & 100 \end{bmatrix}$$

$$= \begin{bmatrix} 7578 & 7929 & 8712 \\ 7929 & 8297 & 9116 \\ 8712 & 9116 & 10016 \end{bmatrix}$$

Site 2			
Patients	Features		Response
	3	4	y
1: $(z_1^{\text{site}_2})^T$	1	1	1
2: $(z_2^{\text{site}_2})^T$	0	1	1
3: $(z_3^{\text{site}_2})^T$	1	0	-1
	Z^{site_2}		

Distributed NR Algorithm for VERTIGO in Dual Form

First derivative:

Distributed

$$J_i'(\alpha) = \lambda^{-1} y_i \sum_j \alpha_j y_j \left(\sum_k \left(z_j^{\text{site}_k}\right)^T \left(z_i^{\text{site}_k}\right) \right) + \log \frac{\alpha_i}{1-\alpha_i}$$

$$\left(z_j^{\text{site}_1}\right)^T \left(z_i^{\text{site}_1}\right) + \left(z_j^{\text{site}_2}\right)^T \left(z_i^{\text{site}_2}\right)$$

Hessian matrix:

$$J_{i,j}''(\alpha) = \lambda^{-1} \text{diag}(y) \left(\sum_k \left(Z^{\text{site}_k}\right)\left(Z^{\text{site}_k}\right)^T \right) \text{diag}(y) + \text{diag}\left(\frac{\alpha_i}{1-\alpha_i}\right)$$

$$\left(Z^{\text{site}_1}\right)\left(Z^{\text{site}_1}\right)^T + \left(Z^{\text{site}_2}\right)\left(Z^{\text{site}_2}\right)^T$$

Site 1			
Patients	Features		Response
	1	2	y
1: $(z_1^{\text{site}_1})^T$	3	87	1
2: $(z_2^{\text{site}_1})^T$	4	91	1
3: $(z_3^{\text{site}_1})^T$	4	100	-1
	Z^{site_1}		

$$\left(Z^{\text{site}_2}\right)\left(Z^{\text{site}_2}\right)^T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \text{X} \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Site 2			
Patients	Features		Response
	3	4	y
1: $(z_1^{\text{site}_2})^T$	1	1	1
2: $(z_2^{\text{site}_2})^T$	0	1	1
3: $(z_3^{\text{site}_2})^T$	1	0	-1
	Z^{site_2}		

Distributed NR Algorithm for VERTIGO in Dual Form

First derivative:

Distributed

$$J_i'(\alpha) = \lambda^{-1} y_i \sum_j \alpha_j y_j \left(\sum_k \left(z_j^{\text{site}_k}\right)^T \left(z_i^{\text{site}_k}\right) \right) + \log \frac{\alpha_i}{1-\alpha_i}$$

$$\left(z_j^{\text{site}_1}\right)^T \left(z_i^{\text{site}_1}\right) + \left(z_j^{\text{site}_2}\right)^T \left(z_i^{\text{site}_2}\right)$$

Hessian matrix:

$$J_{i,j}''(\alpha) = \lambda^{-1} \text{diag}(y) \left(\sum_k \left(Z^{\text{site}_k}\right)\left(Z^{\text{site}_k}\right)^T \right) \text{diag}(y) + \text{diag}\left(\frac{\alpha_i}{1-\alpha_i}\right)$$

$$\left(Z^{\text{site}_1}\right)\left(Z^{\text{site}_1}\right)^T + \left(Z^{\text{site}_2}\right)\left(Z^{\text{site}_2}\right)^T$$

Site 1

Patients	Features		Response
	1	2	y
1: $(z_1^{\text{site}_1})^T$	3	87	1
2: $(z_2^{\text{site}_1})^T$	4	91	1
3: $(z_3^{\text{site}_1})^T$	4	100	-1
	Z^{site_1}		

7578	7929	8712
7929	8297	9116
8712	9116	10016

+

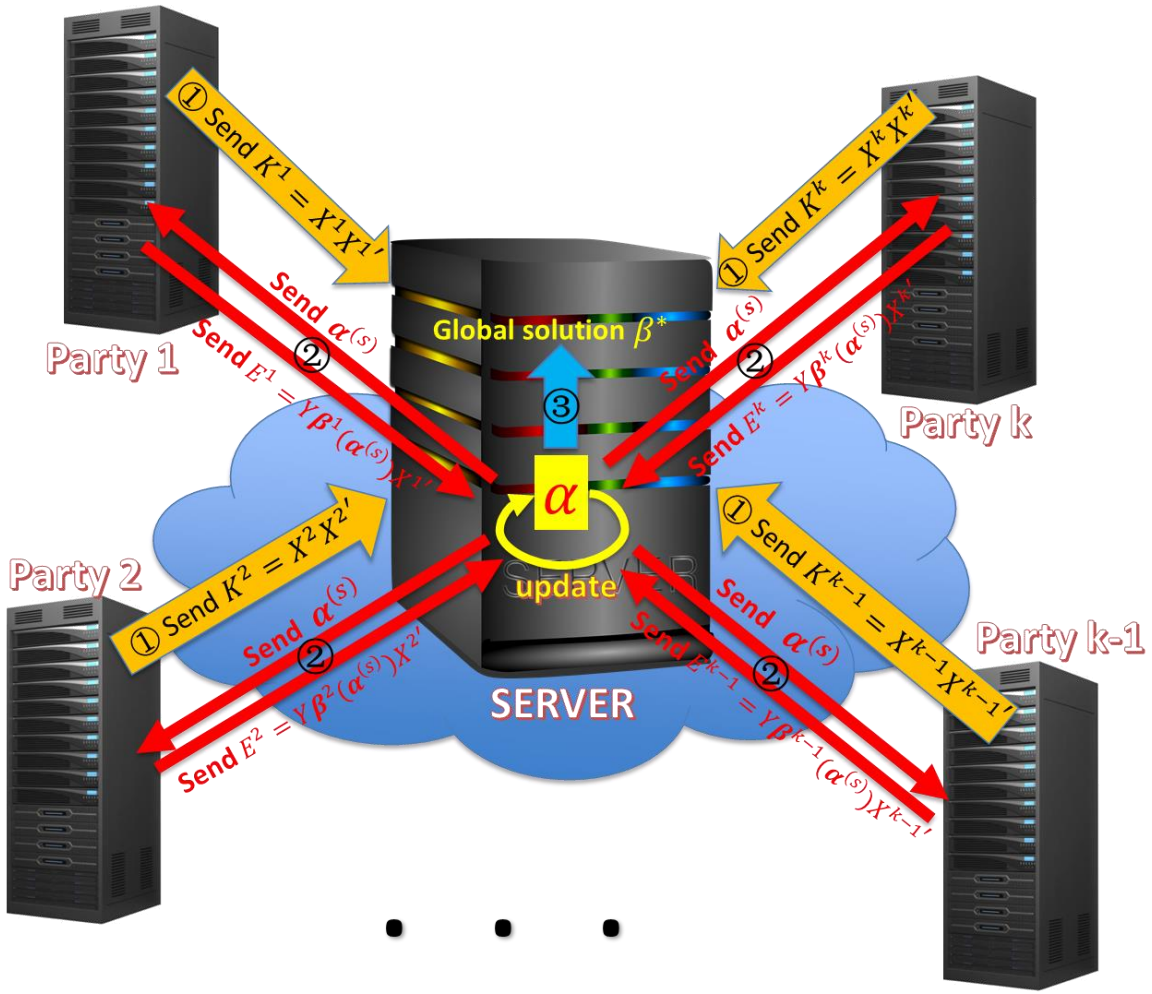
2	1	1
1	1	0
1	0	1

Site 2

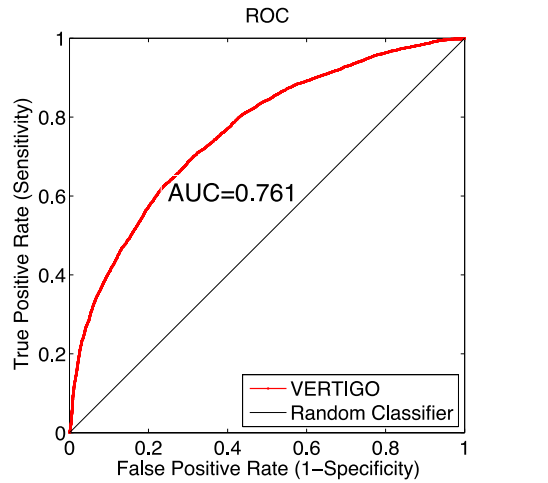
Patients	Features		Response
	3	4	y
1: $(z_1^{\text{site}_2})^T$	1	1	1
2: $(z_2^{\text{site}_2})^T$	0	1	1
3: $(z_3^{\text{site}_2})^T$	1	0	-1
	Z^{site_2}		

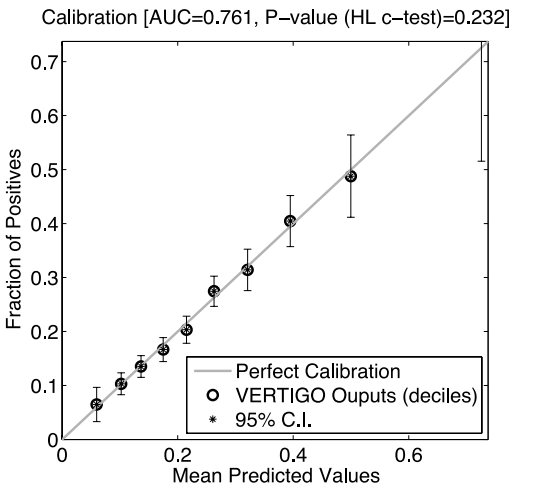
Only exchange aggregated local statistics

Experiments (estimation accuracy)



Only dot products of records are exchanged





Performance of VERTIGO on the MIMIC2

Matches exactly with centralized model

Experiments (estimation accuracy)

Coefficients*	Synthetic Data (20 variables)	Genome Data (41 variables)	Myocardial Infarction Data (36 variables)	MIMIC II Data (42 variables)
β_1	2.166e-13	9.163e-14	7.611e-10	-2.312e-12
β_2	6.964e-14	-2.683e-13	-1.042e-11	-3.914e-13
β_3	-5.208e-14	1.691e-14	-1.203e-11	8.951e-12
β_4	1.485e-13	-3.473e-15	-6.350e-11	-1.648e-13
β_5	1.734e-13	1.307e-14	9.765e-11	-2.846e-14
β_6	1.359e-13	9.983e-14	-4.881e-11	2.102e-13
β_7	3.142e-14	-2.453e-13	-2.541e-12	-5.365e-13
β_8	-1.104e-14	1.413e-13	3.205e-10	-1.666e-13
β_9	-3.232e-13	-8.706e-14	1.980e-10	5.939e-13
β_{10}	4.281e-14	-2.783e-14	1.872e-10	2.846e-12

*Averaged error between first 10 model coefficients of VERTIGO and Primal logistic regression over 50 trials

Experiments (Average Response Time)

# of total records	VERTIGO				Primal optimization
	Iterative Hessian on CPU	Fixed Hessian on CPU	Iterative Hessian on GPU	Fixed Hessian on GPU	
2,000	2.84	7.03	**1.69**	7.01	0.72
4,000	12.51	8.68	6.55	8.22	1.71
8,000	66.38	15.30	28.57	11.05	4.10
20,000	815.6	101.7	-	-	26.5

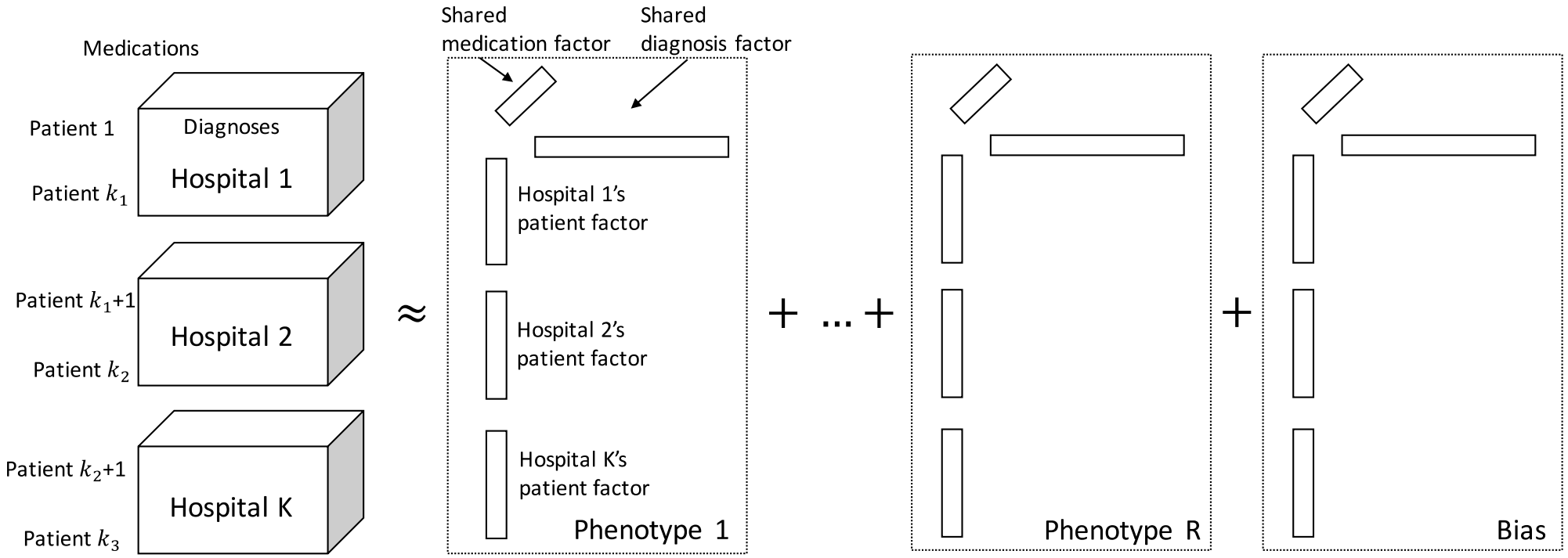
Average computing time (seconds) for training LR models (M=2). The red numbers are the best performers in each row

Experiments (Average Response Time)

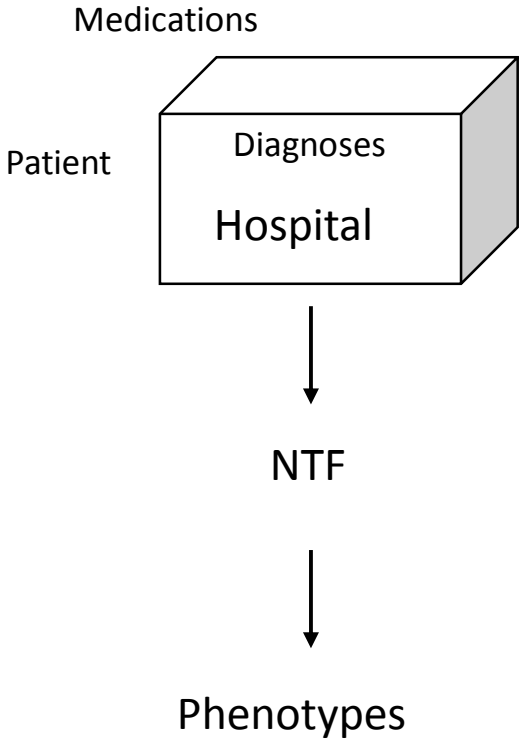
# of total records	VERTIGO				Primal optimization
	Iterative Hessian on CPU	Fixed Hessian on CPU	Iterative Hessian on GPU	Fixed Hessian on GPU	
2,000	2.84	7.03	**1.69**	7.01	0.72
4,000	12.51	8.68	6.55	8.22	1.71
8,000	66.38	15.30	28.57	11.05	4.10
20,000	815.6	101.7	-	-	26.5

Average computing time (seconds) for training LR models (M=2). The red numbers are the best performers in each row

Federated tensor factorization for computational phenotyping based on tensor factorization

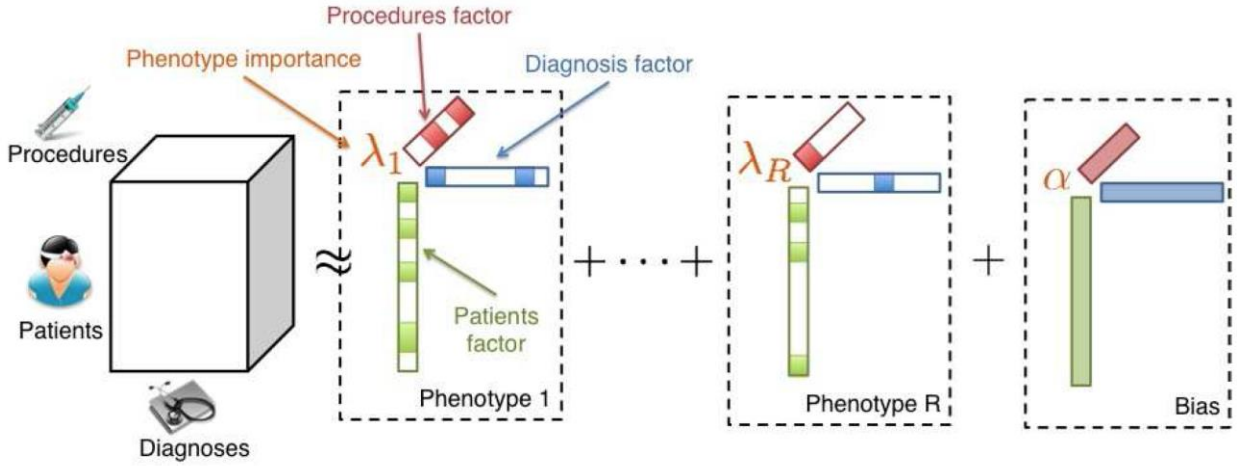


Backgrounds



EHR data in hospital contain 3-dimensional count data, so we can derive phenotypes using Nonnegative tensor factorization (NTF).

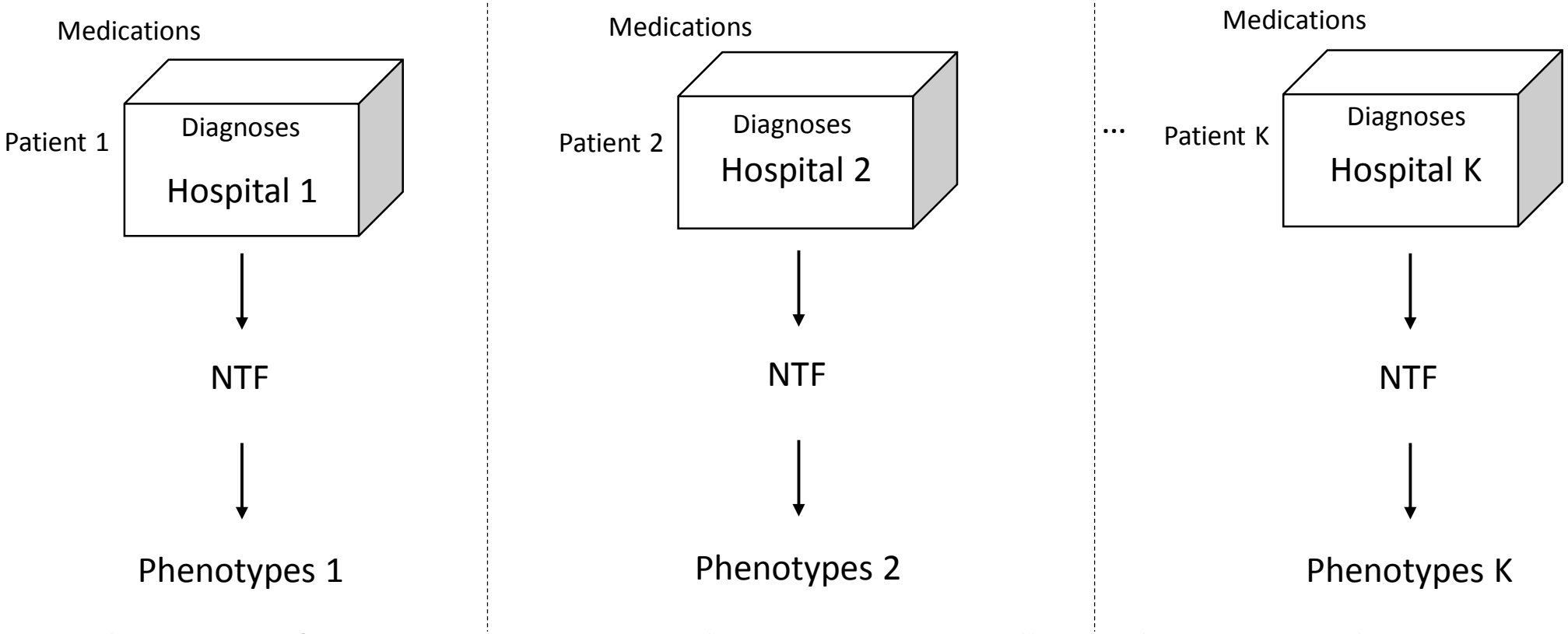
- Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization, JC Ho et al., KDD14,
- Rubik: Knowledge guided tensor factorization and completion for health data analytics, Y Wang et al., KDD15



Backgrounds

What if there are several hospitals?

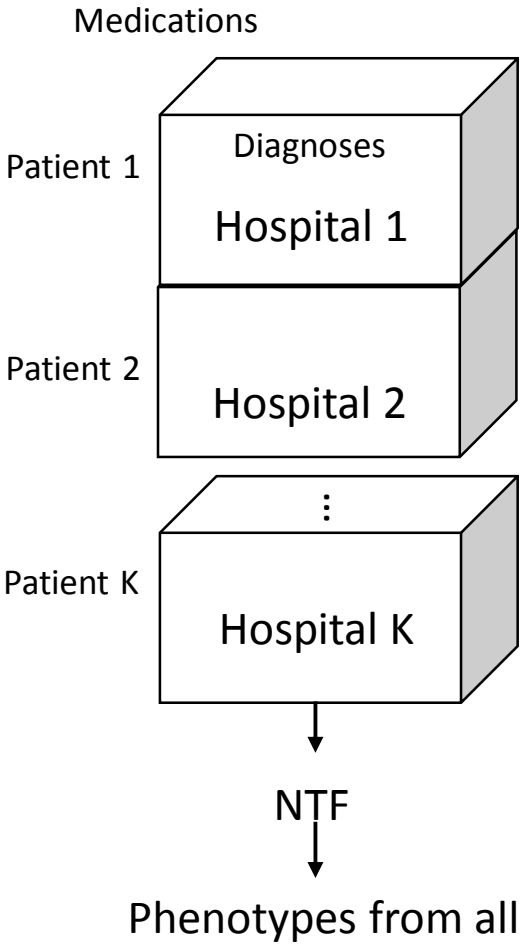
- Each hospital might run phenotyping and combine phenotypes



But, phenotypes from one institution are limited due to small sample size and inherent population bias.

Backgrounds

What if there are several hospitals?



Or, hospitals might want to combine data from all hospitals.

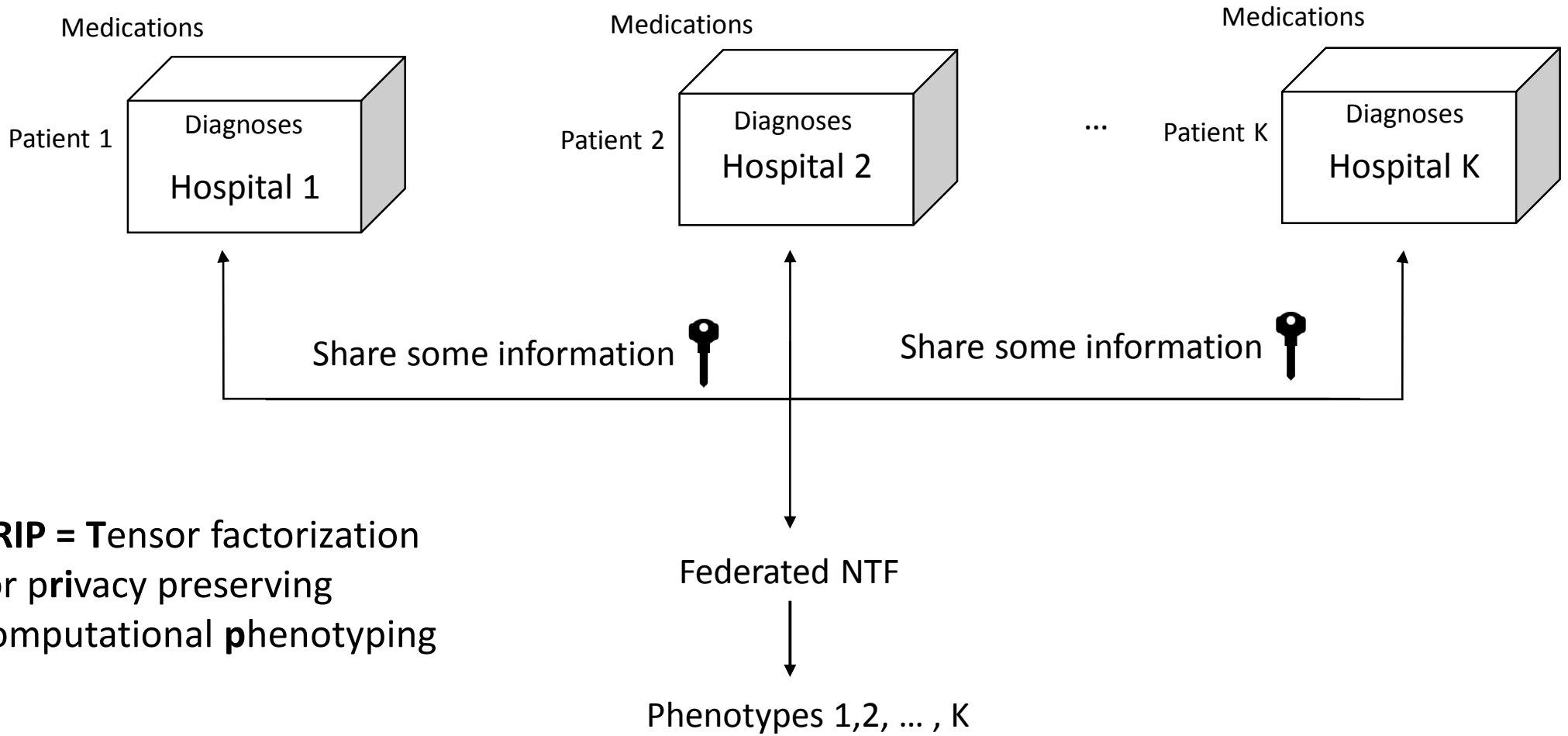
But, healthcare data sharing and exchange are impeded due to privacy concerns.
e.g., PCORnet, I2B2

pcornet® The National Patient-Centered Clinical Research Network

i2b2 A National Center for Biomedical Computing
Informatics for Integrating Biology & the Bedside

Objectives

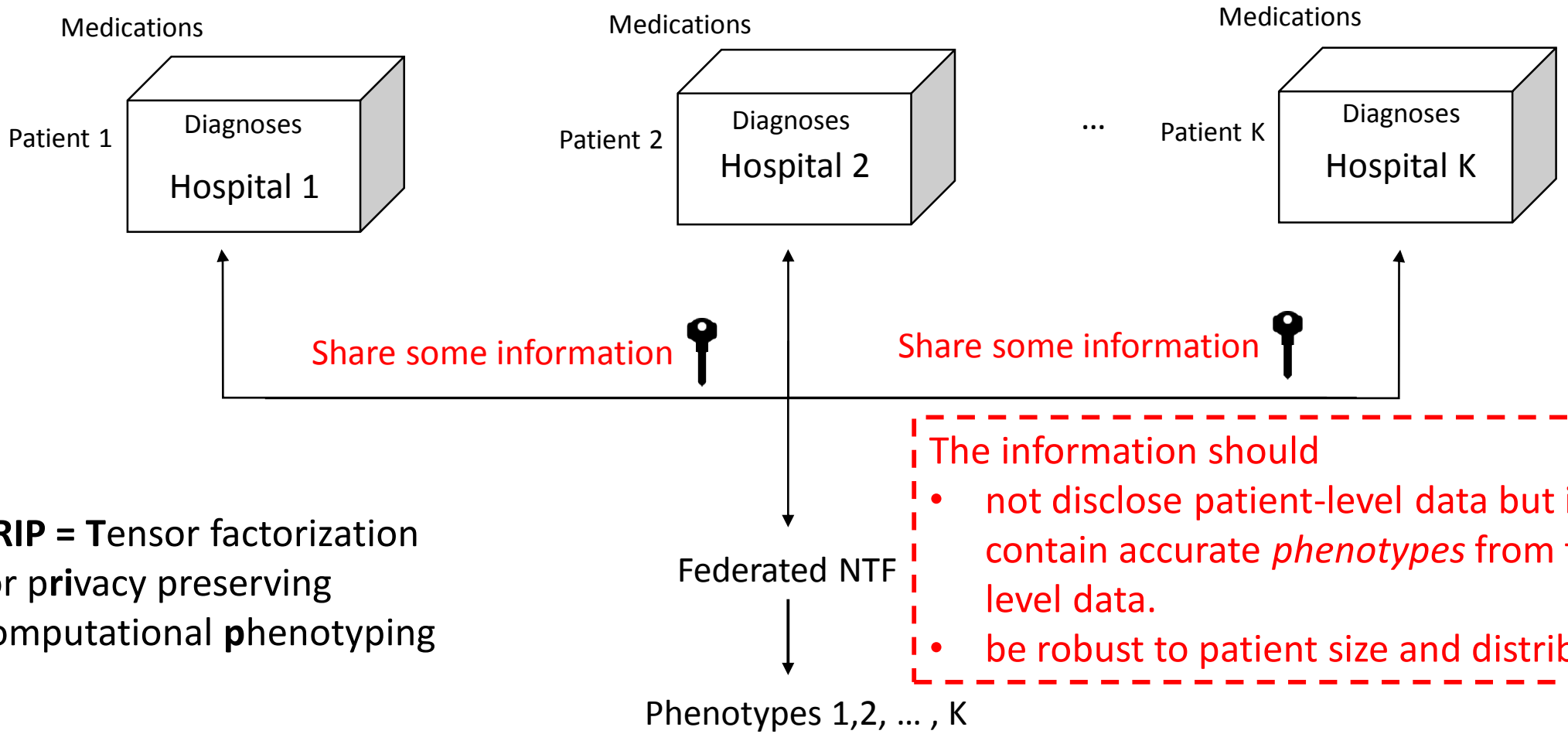
*We propose a **federated framework** (TRIP) that can derive phenotypes without sharing data*



TRIP = Tensor factorization
for p**ri**vacy preserving
computational **p**henotyping

Objectives

*We propose a **federated framework** (TRIP) that can derive phenotypes without sharing data*

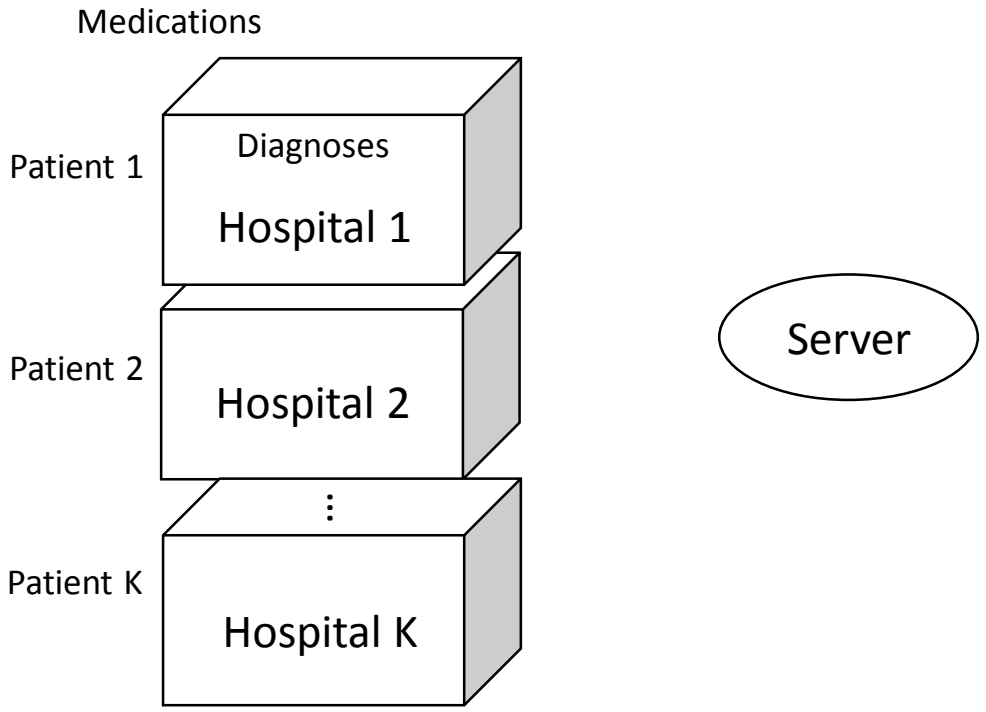


TRIP = Tensor factorization for p**ri**vacy preserving computational **p**henotyping

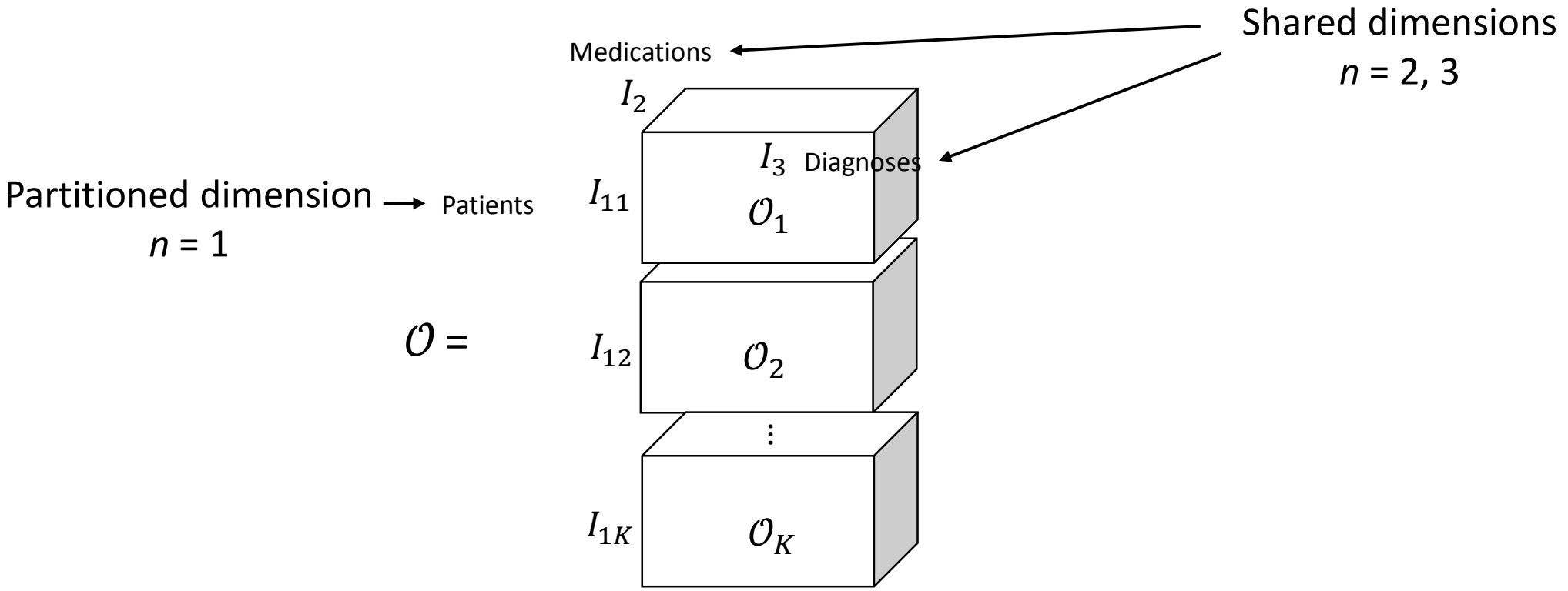
The information should
- not disclose patient-level data but instead, contain accurate *phenotypes* from the patient-level data.
- be robust to patient size and distribution

Assumptions

- Data are 3-order tensor of **patient, medication, and diagnosis** (*n* = 1, 2, 3, respectively).
- Data are **horizontally partitioned** along patient dimension. Hospitals (i.e., hospitals) have their own patient data on the same medical features.
- There are ***K* hospitals** and a central **server**.
- **Honesty-but-Curious adversary model**. The server and hospitals are curious on data of others but do not maliciously manipulate intermediate results.



Assumptions



CP tensor factorization

- Global data
  - $\mathcal{O}$ = Global observed tensor
  - $\mathcal{X}$ = Global estimated tensor

    $= \mathcal{T} + \mathcal{C}$
  - $\mathcal{T} = \sum_{r=1}^{R} \mathbf{A}^{(1)}(:, r) \circ \mathbf{A}^{(2)}(:, r) \circ \mathbf{A}^{(3)}(:, r)$

    where $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}$ are global factor matrices
  - $\mathcal{C} = \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ \mathbf{u}^{(3)}$

    where $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(3)}$ are global factor matrices
  - Obj. function $\min ||\mathcal{X} - (\mathcal{T} + \mathcal{C})||_F^2$

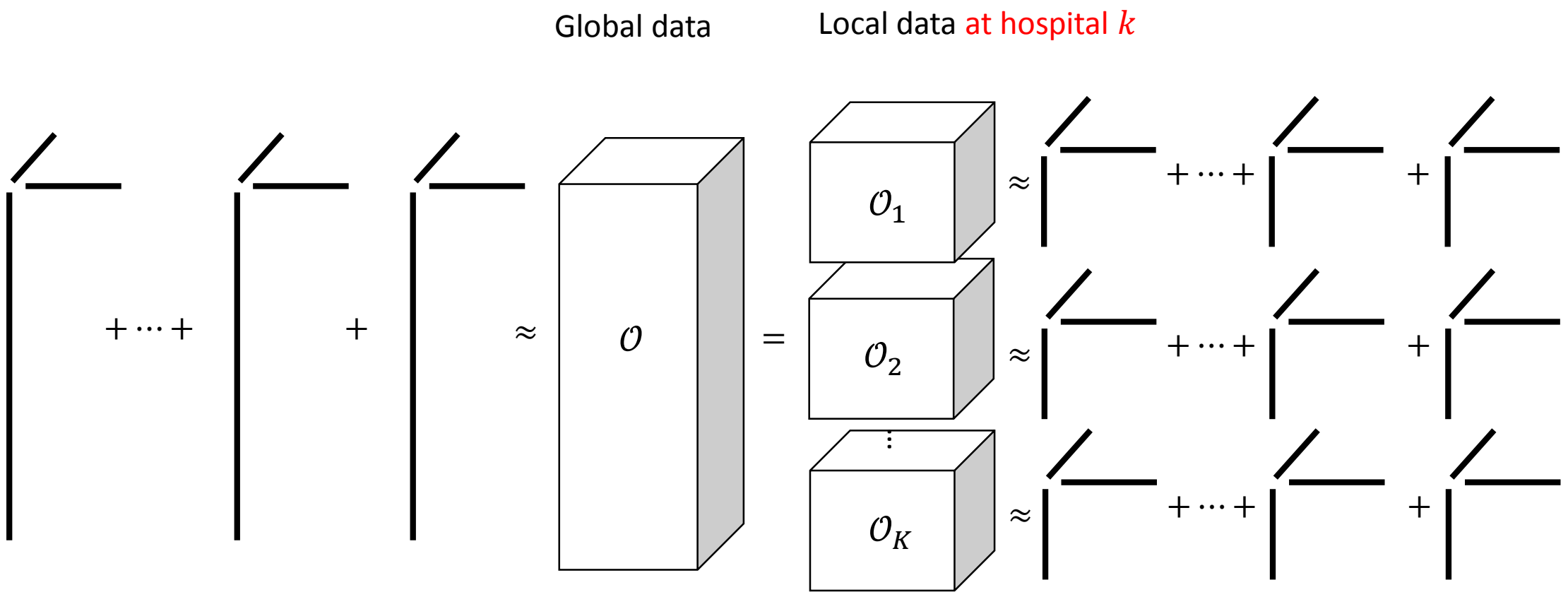
- Local data at hospital k
  - $\mathcal{O}_k$ = Local observed tensor
  - $\mathcal{X}_k$ = Local estimated tensor

    $= \mathcal{T}_k + \mathcal{C}_k$
  - $\mathcal{T}_k = \sum_{r=1}^{R} \mathbf{A}_k^{(1)}(:, r) \circ \mathbf{A}_k^{(2)}(:, r) \circ \mathbf{A}_k^{(3)}(:, r)$

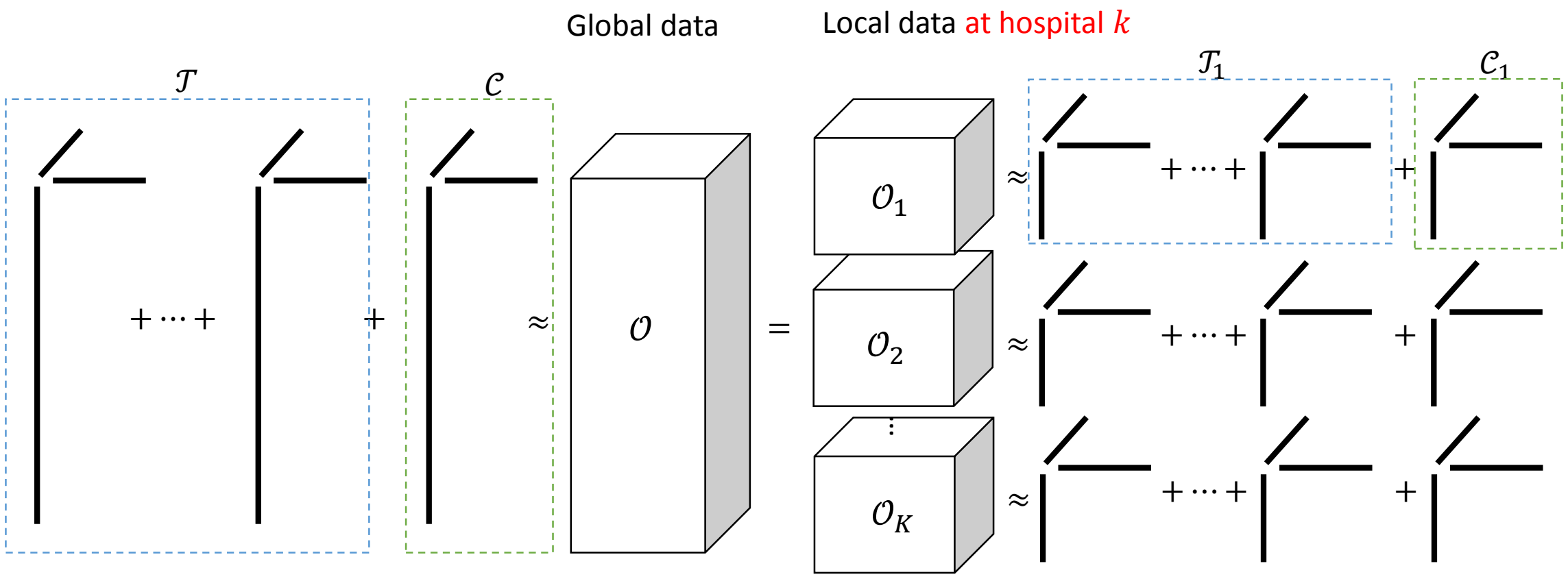
    where $\mathbf{A}_k^{(1)}, \mathbf{A}_k^{(2)}, \mathbf{A}_k^{(3)}$ are local factor matrices
  - $\mathcal{C}_k = \mathbf{u}_k^{(1)} \circ \mathbf{u}_k^{(2)} \circ \mathbf{u}_k^{(3)}$

    where $\mathbf{u}_k^{(1)}, \mathbf{u}_k^{(2)}, \mathbf{u}_k^{(3)}$ are local factor matrices
  - Obj. function $\min \sum_{k=1}^{K} ||\mathcal{X}_k - (\mathcal{T}_k + \mathcal{C}_k)||_F^2$

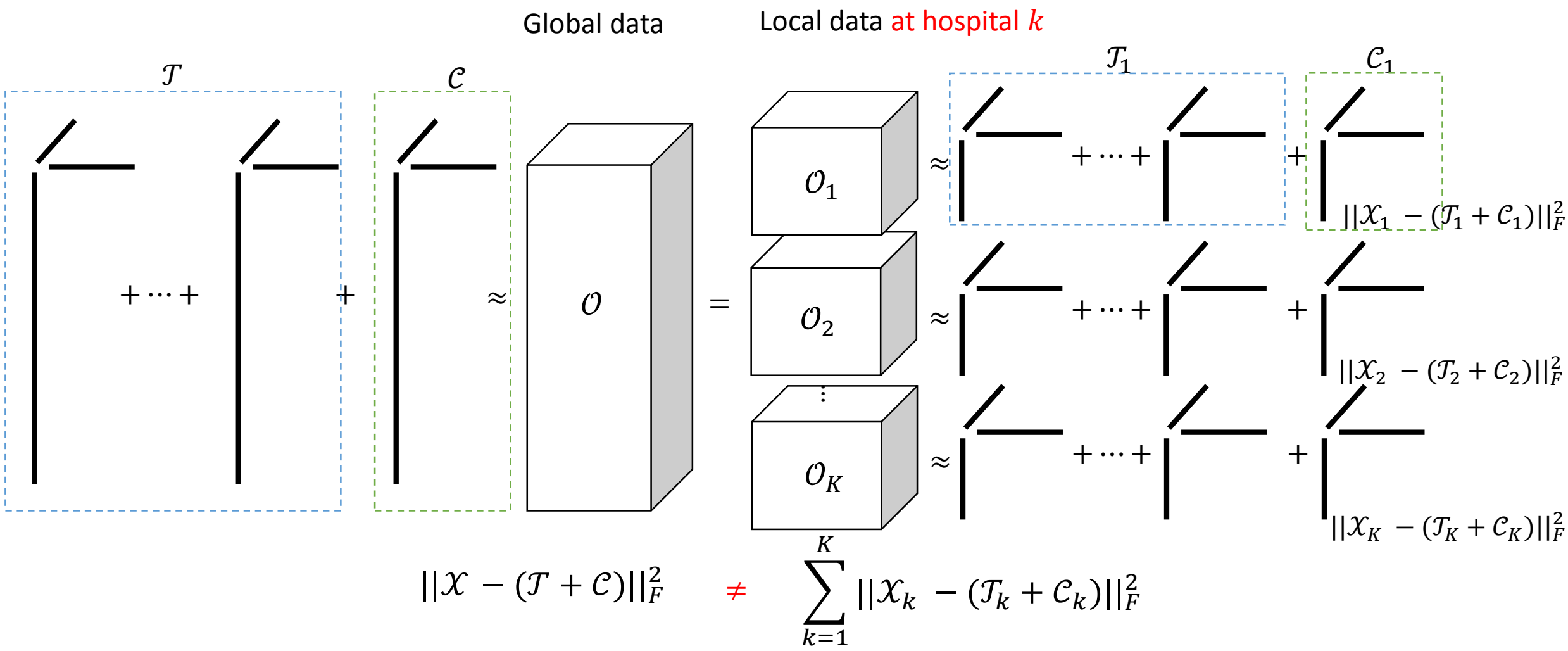
CP tensor factorization



CP tensor factorization



CP tensor factorization



$$||\mathcal{X} - (\mathcal{T} + \mathcal{C})||_F^2 \neq \sum_{k=1}^{K} ||\mathcal{X}_k - (\mathcal{T}_k + \mathcal{C}_k)||_F^2$$

Formulating objective function

- Assume
  - $\mathbf{A}^{(2)} = \mathbf{A}_1^{(2)} = \mathbf{A}_2^{(2)} = \cdots = \mathbf{A}_K^{(2)}$ (Shared medications dimension)
  - $\mathbf{A}^{(3)} = \mathbf{A}_1^{(3)} = \mathbf{A}_2^{(3)} = \cdots = \mathbf{A}_K^{(3)}$ (Shared diagnoses dimension)

for all hospitals K

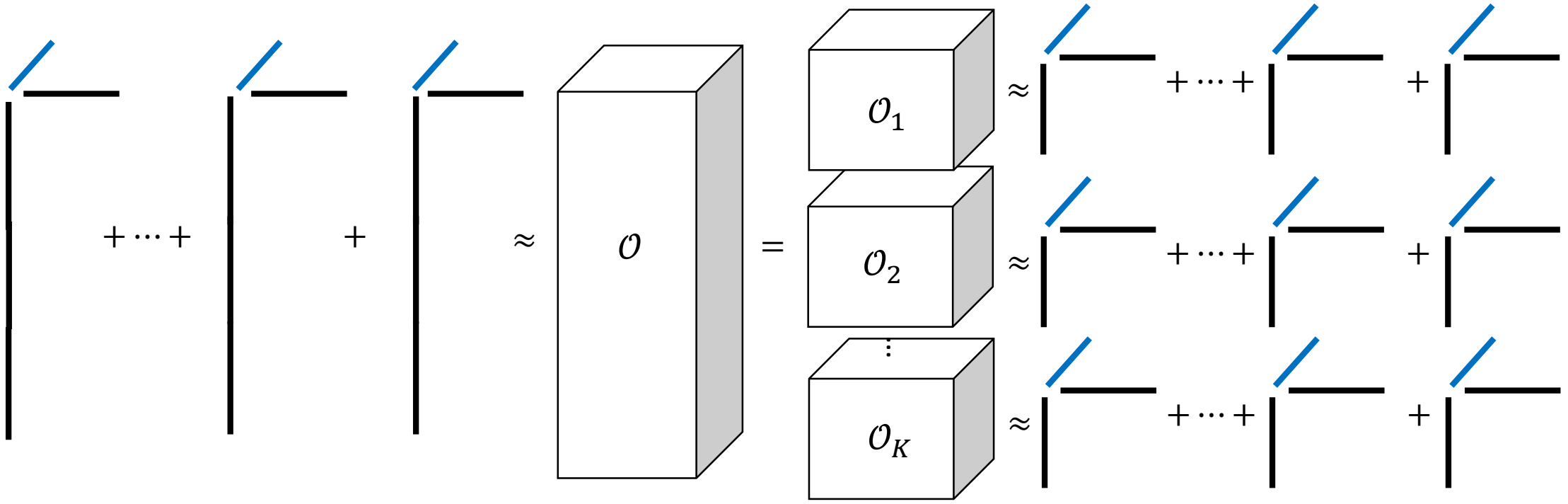
- Then, $\mathbf{A}^{(1)} = \begin{bmatrix} \mathbf{A}_1^{(1)}; \\ \vdots \\ \mathbf{A}_K^{(1)} \end{bmatrix}$ (Partitioned patient dimension)

(the same assumption applies to $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(3)}$ as well)

$$\therefore \min ||\mathcal{X} - (\mathcal{T} + \mathcal{C})||_F^2 = \sum_{k=1}^{K} ||\mathcal{X}_k - (\mathcal{T}_k + \mathcal{C}_k)||_F^2$$

Formulating objective function

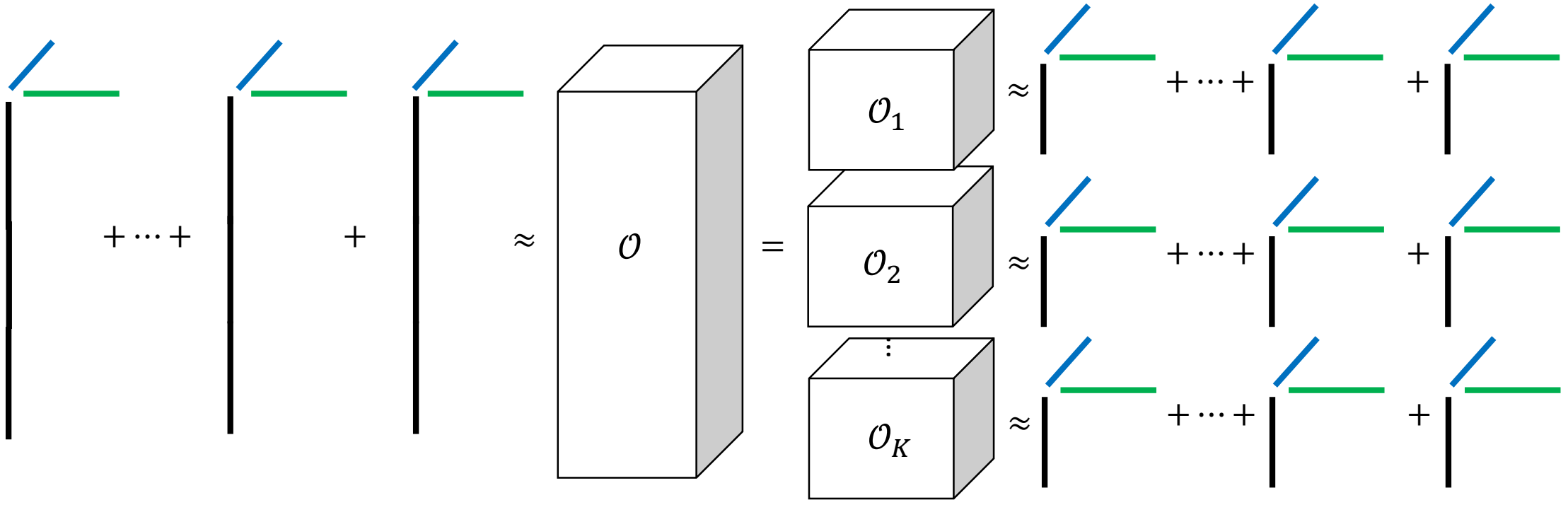
$\mathbf{A}^{(2)} = \mathbf{A}_1^{(2)} = \mathbf{A}_2^{(2)} = \cdots = \mathbf{A}_K^{(2)}$



Formulating objective function

$\mathbf{A}^{(2)} = \mathbf{A}_1^{(2)} = \mathbf{A}_2^{(2)} = \cdots = \mathbf{A}_K^{(2)}$

$\mathbf{A}^{(3)} = \mathbf{A}_1^{(3)} = \mathbf{A}_2^{(3)} = \cdots = \mathbf{A}_K^{(3)}$

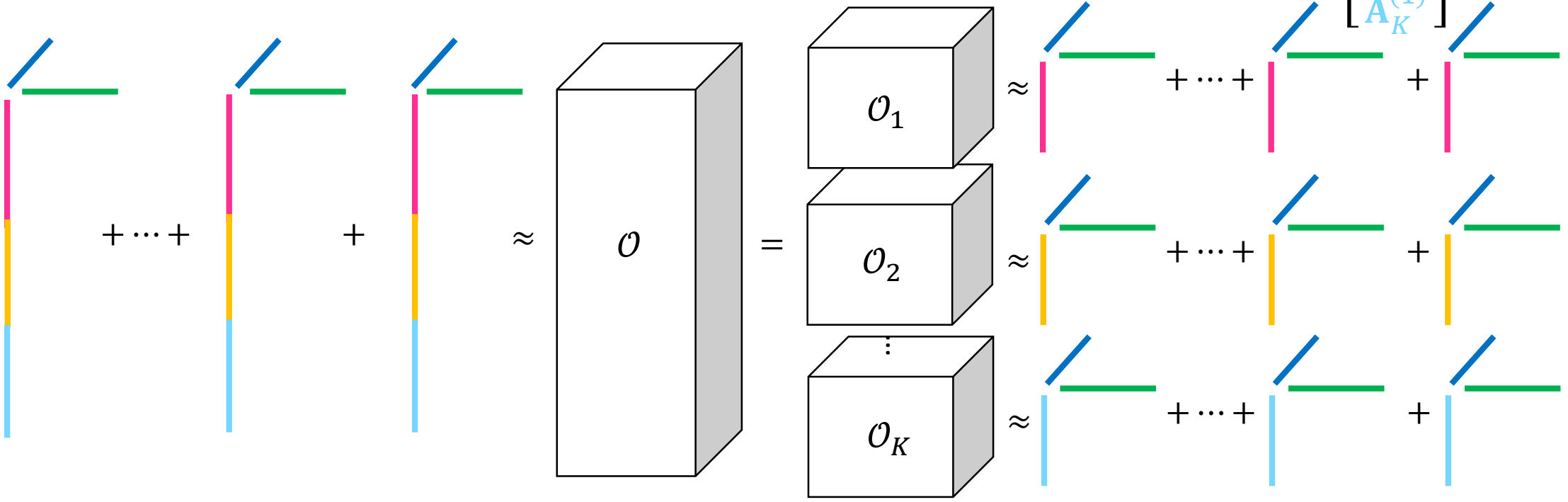


Formulating objective function

$\mathbf{A}^{(2)} = \mathbf{A}_1^{(2)} = \mathbf{A}_2^{(2)} = \cdots = \mathbf{A}_K^{(2)}$

$\mathbf{A}^{(3)} = \mathbf{A}_1^{(3)} = \mathbf{A}_2^{(3)} = \cdots = \mathbf{A}_K^{(3)}$

$\therefore \ \mathbf{A}^{(1)} = \begin{bmatrix} \mathbf{A}_1^{(1)}; \\ \mathbf{A}_2^{(1)}; \\ \mathbf{A}_K^{(1)} \end{bmatrix}$

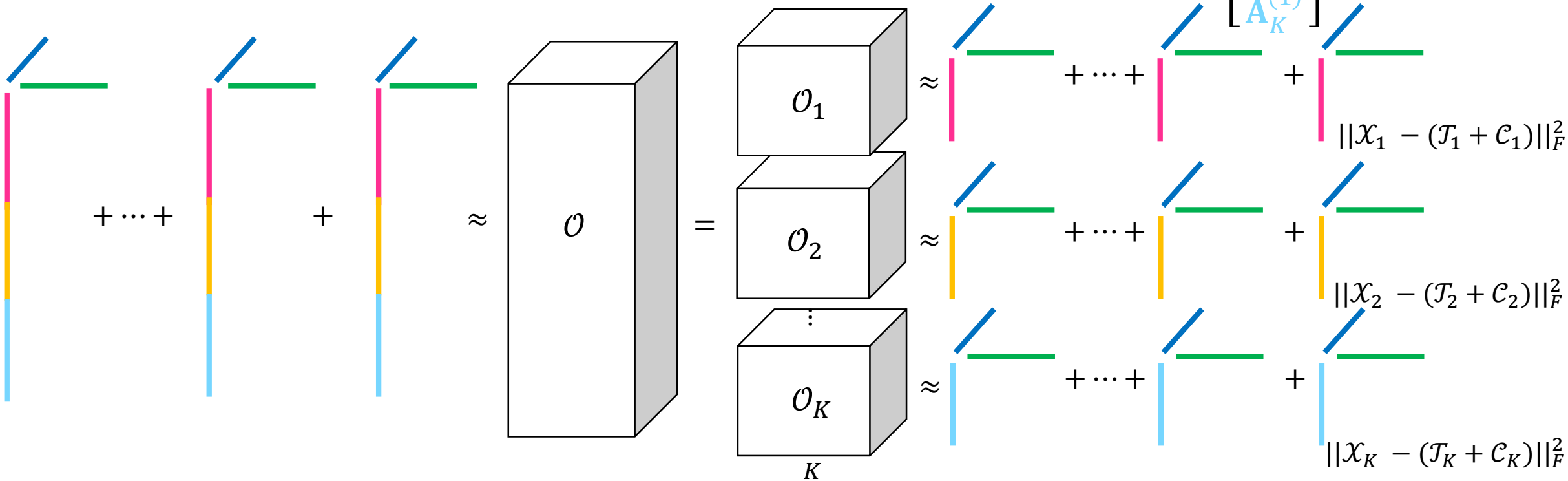


Formulating objective function

$\mathbf{A}^{(2)} = \mathbf{A}_1^{(2)} = \mathbf{A}_2^{(2)} = \cdots = \mathbf{A}_K^{(2)}$

$\mathbf{A}^{(3)} = \mathbf{A}_1^{(3)} = \mathbf{A}_2^{(3)} = \cdots = \mathbf{A}_K^{(3)}$

$\therefore \ \mathbf{A}^{(1)} = \begin{bmatrix} \mathbf{A}_1^{(1)}; \\ \mathbf{A}_2^{(1)}; \\ \mathbf{A}_K^{(1)} \end{bmatrix}$



$$||\mathcal{X} - (\mathcal{T} + \mathcal{C})||_F^2 = \sum_{k=1}^{K} ||\mathcal{X}_k - (\mathcal{T}_k + \mathcal{C}_k)||_F^2$$

Formulating objective function

$$\min_{\mathcal{X}_k, \mathcal{T}_k, \mathcal{C}_k} \Psi = \sum_{k=1}^{K} ||\mathcal{X}_k - (\mathcal{C}_k + \mathcal{T}_k)||_F^2$$

$$s.t. \quad \mathbf{A}^{(n)} = \mathbf{A}_k^{(n)} \forall k, \forall n \neq 1$$

$$\mathbf{u}^{(n)} = \mathbf{u}_k^{(n)} \forall k, \forall n \neq 1$$

$$\mathcal{P}_\Omega(\mathcal{X}_k) = \mathcal{P}_\Omega(\mathcal{O}_k)$$

$\leftarrow \mathcal{X}_k$ is initialized with observed values of $\mathcal{O}_k$

Formulating objective function

Orthogonality term

$$\min_{\mathcal{X}_k, \mathcal{T}_k, \mathcal{C}_k} \Psi = \sum_{k=1}^{K} ||\mathcal{X}_k - (\mathcal{C}_k + \mathcal{T}_k)||_F^2 + \sum_{n=2}^{N} \frac{\lambda_q}{2} ||\mathbf{I} - \mathbf{B}^{(n)T}\mathbf{A}^{(n)}||_F^2$$

$$s.t. \quad \mathbf{A}^{(n)} = \mathbf{A}_k^{(n)} \forall k, \forall n \neq 1$$

$$\mathbf{u}^{(n)} = \mathbf{u}_k^{(n)} \forall k, \forall n \neq 1$$

$$\mathcal{P}_\Omega(\mathcal{X}_k) = \mathcal{P}_\Omega(\mathcal{O}_k)$$

$$\mathbf{B}^{(k)} = \mathbf{A}^{(k)}$$

$$\mathbf{v}^{(n)} = \mathbf{u}^{(n)}$$

← $\mathcal{X}_k$ is initialized with observed values of $\mathcal{O}_k$

← to make the orthogonality term convex

Formulating objective function

- Augmented Lagrangian function is

$$\begin{aligned}\mathcal{L} =& \Psi + \sum_{k=1}^{K}\sum_{n=2}^{N}\left[(\mathbf{A}^{(n)} - \mathbf{A}_k^{(n)})^T\mathbf{H}_k^{(n)} + \frac{\omega}{2}||\mathbf{A}^{(n)} - \mathbf{A}_k^{(n)}||_F^2\right] \\ &+ \sum_{k=1}^{K}\sum_{n=2}^{N}\left[(\mathbf{u}^{(n)} - \mathbf{u}_k^{(n)})^T\mathbf{g}_k^{(n)} + \frac{\gamma}{2}||\mathbf{u}^{(n)} - \mathbf{u}_k^{(n)}||_F^2\right] \\ &+ \sum_{n=1}^{N}\left[(\mathbf{B}^{(n)} - \mathbf{A}^{(n)})^T\mathbf{Y}^{(n)} + \frac{\mu}{2}||\mathbf{B}^{(n)} - \mathbf{A}^{(n)}||_F^2\right] \\ &+ \sum_{n=1}^{N}\left[(\mathbf{v}^{(n)} - \mathbf{u}^{(n)})^T\mathbf{p}^{(n)} + \frac{\eta}{2}||\mathbf{v}^{(n)} - \mathbf{u}^{(n)}||_F^2\right]\end{aligned}$$

Formulating objective function

- Augmented Lagrangian function is

$$\mathcal{L} = \Psi + \sum_{k=1}^{K}\sum_{n=2}^{N}\left[(\mathbf{A}^{(n)} - \mathbf{A}_k^{(n)})^T \mathbf{H}_k^{(n)} + \frac{\omega}{2}||\mathbf{A}^{(n)} - \mathbf{A}_k^{(n)}||_F^2\right]$$
$$+ \sum_{k=1}^{K}\sum_{n=2}^{N}\left[(\mathbf{u}^{(n)} - \mathbf{u}_k^{(n)})^T \mathbf{g}_k^{(n)} + \frac{\gamma}{2}||\mathbf{u}^{(n)} - \mathbf{u}_k^{(n)}||_F^2\right]$$
$$+ \sum_{n=1}^{N}\left[(\mathbf{B}^{(n)} - \mathbf{A}^{(n)})^T \mathbf{Y}^{(n)} + \frac{\mu}{2}||\mathbf{B}^{(n)} - \mathbf{A}^{(n)}||_F^2\right]$$
$$+ \sum_{n=1}^{N}\left[(\mathbf{v}^{(n)} - \mathbf{u}^{(n)})^T \mathbf{p}^{(n)} + \frac{\eta}{2}||\mathbf{v}^{(n)} - \mathbf{u}^{(n)}||_F^2\right]$$

Penalty terms
to improve convergence
(for *method of multiplier)*

Lagrangian multipliers

Alternating Direction Methods of Multipliers

- A distributed convex optimization method that divides the objective problem into decentralized sub-problems by introducing auxiliary variables.
  - ADMM problem form (with f, g convex)

$$\begin{array}{ll} \text{minimize} & f(x) + g(z) \\ \text{subject to} & Ax + Bz = c \end{array}$$

Introduce auxiliary variables z to decompose the objective function

    - two sets of variables, with separable objective

  - $L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + (\rho/2)\|Ax + Bz - c\|_2^2$

Penalty term to Improve convergence

  - ADMM:

$$\begin{array}{lcll} x^{k+1} & := & \operatorname{argmin}_x L_\rho(x, z^k, y^k) & // \ x\text{-minimization} \\ z^{k+1} & := & \operatorname{argmin}_z L_\rho(x^{k+1}, z, y^k) & // \ z\text{-minimization} \\ y^{k+1} & := & y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) & // \ \text{dual update} \end{array}$$

x and z are updated in an alternating fashion

Consensus ADMM

$$\text{minimize} \quad f(x) = \sum_{i=1}^{N} f_i(x)$$

This problem can be rewritten with local variables x_i and a common global variable z:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{N} f_i(x_i) \\ &\text{subject to} && x_i - z = 0, \quad i = 1, \ldots, N \end{aligned}$$

Introduce local variables x_i to decompose the objective function

Augmented Lagrangian function is

$$L_\rho(x_1, \ldots, x_N, z, y) = \sum_{i=1}^{N} \left(f_i(x_i) + y_i^T (x_i - z) + (\rho/2) \|x_i - z\|_2^2 \right)$$

Penalty term to Improve convergence

The resulting ADMM algorithm is the following:

$$x_i^{k+1} := \underset{x_i}{\operatorname{argmin}} \left(f_i(x_i) + y_i^{kT} (x_i - z^k) + (\rho/2) \|x_i - z^k\|_2^2 \right)$$

$$z^{k+1} := \frac{1}{N} \sum_{i=1}^{N} \left(x_i^{k+1} + (1/\rho) y_i^k \right)$$

$$y_i^{k+1} := y_i^k + \rho (x_i^{k+1} - z^{k+1}).$$

x and z are updated in an alternating fashion

A simple consensus ADMM Example

We have a optimization problem:

$\min f(x) = 3x^2$

$\Leftrightarrow \min f(x) = x_1^2 + x_2^2 + x_3^2$

Subject to

$$x_1 - z = 0$$
$$x_2 - z = 0$$
$$x_3 - z = 0$$

Augmented Lagrangian function

$$L_\rho(x_1, x_2, x_3, z, y)$$
$$= x_1^2 + y_1^T(x_1 - z) + {}^{\rho}/_{2}\ ||x_1 - z||^2$$
$$+ x_2^2 + y_2^T(x_2 - z) + {}^{\rho}/_{2}\ ||x_2 - z||^2$$
$$+ x_3^2 + y_3^T(x_3 - z) + {}^{\rho}/_{2}\ ||x_3 - z||^2$$

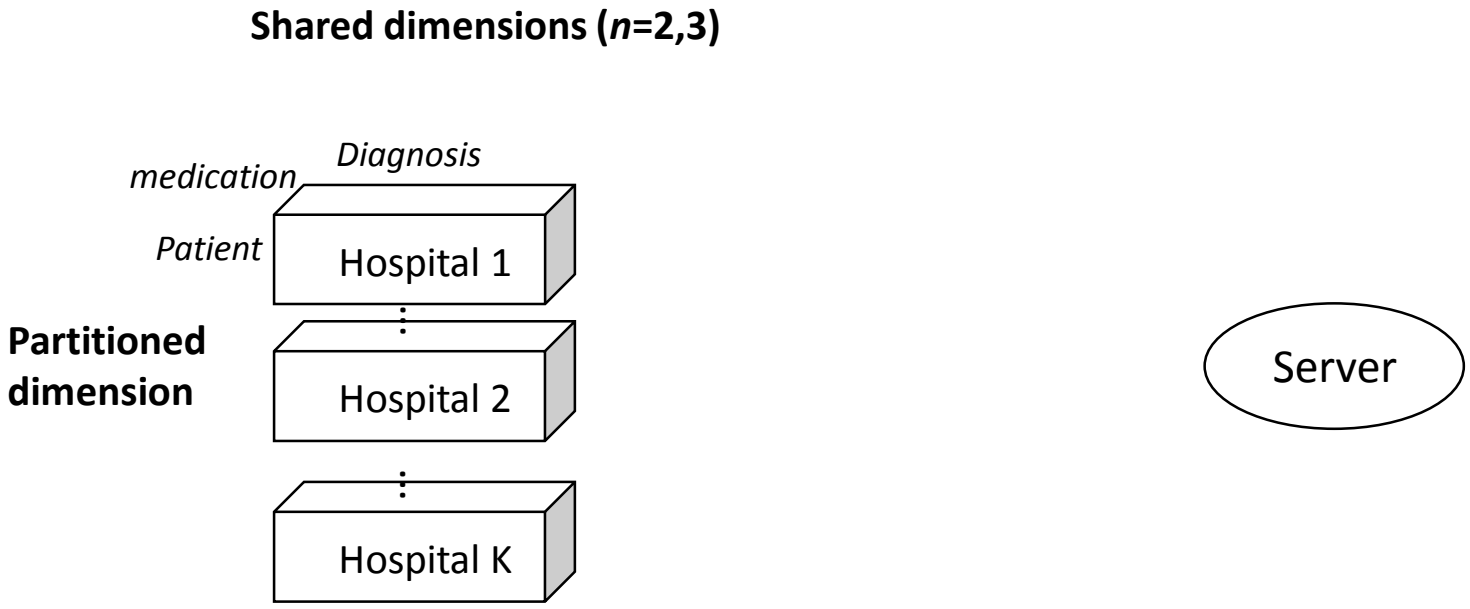
ADMM algorithm

$$x_1 = argmin_{x_1}(x_1^2 + y_1^T(x_1 - z) + {}^{\rho}/_{2}\ ||x_1 - z||^2)$$
$$x_2 = argmin_{x_1}(x_2^2 + y_2^T(x_2 - z) + {}^{\rho}/_{2}\ ||x_2 - z||^2)$$
$$x_3 = argmin_{x_3}(x_3^2 + y_3^T(x_3 - z) + {}^{\rho}/_{2}\ ||x_3 - z||^2)$$

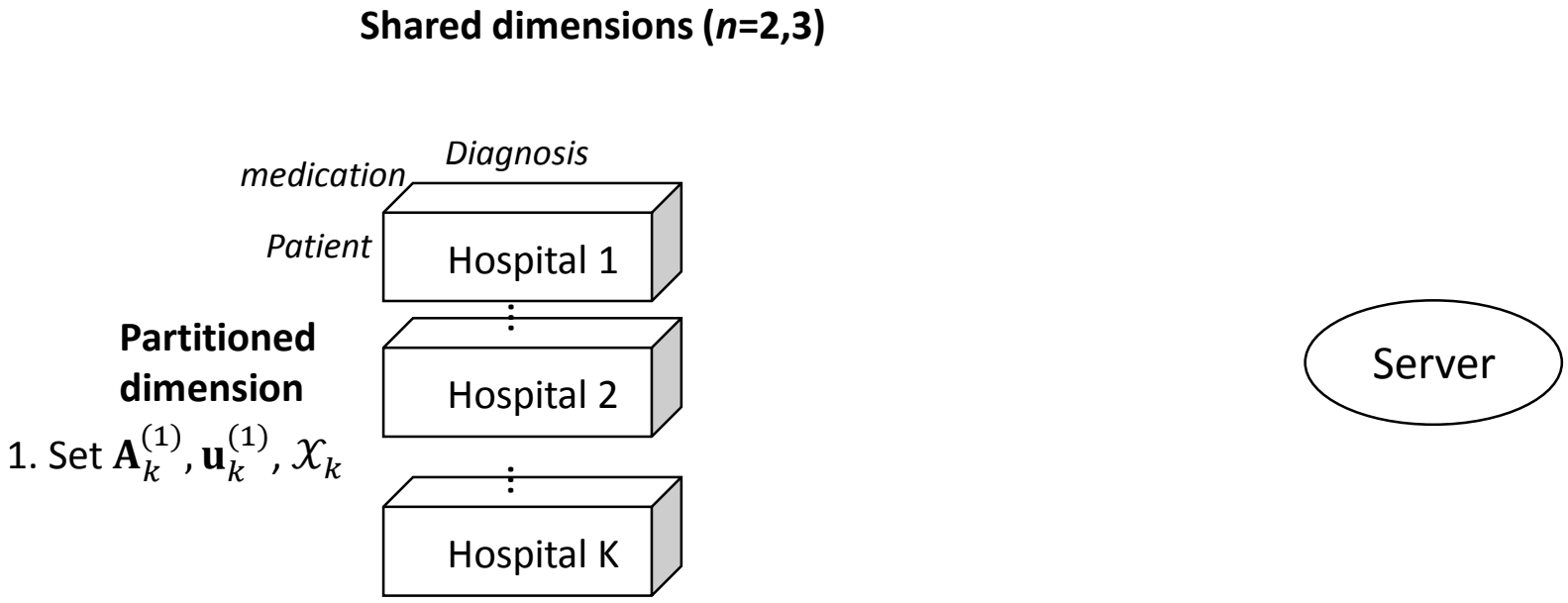
$$z = \frac{1}{3}(x_1 + \frac{1}{\rho}y_1 + x_2 + \frac{1}{\rho}y_2 + x_3 + \frac{1}{\rho}y_3)$$

$$y_1 = y_1 + \rho(x_1 - z)$$
$$y_2 = y_2 + \rho(x_2 - z)$$
$$y_3 = y_3 + \rho(x_3 - z)$$

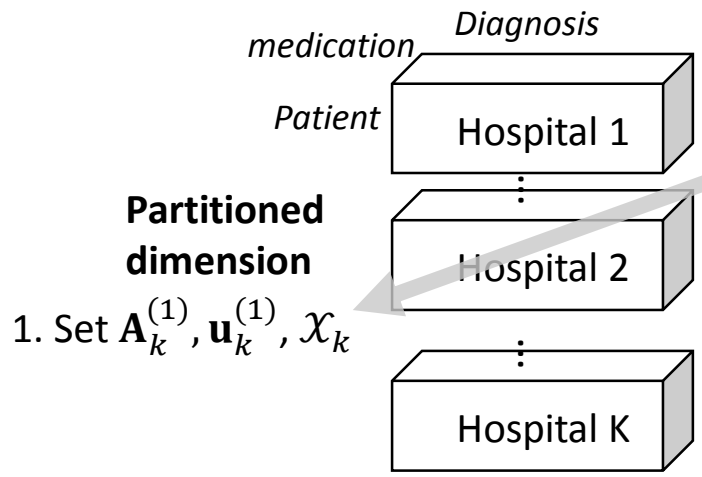
Solving the objective function



Solving the objective function



Solving the objective function



$$\mathbf{A}_k^{(1)} = \left\{2\mathbf{R}_{(1)k}\Pi^{(1)}\right\}\left\{2\Pi^{(1)T}\Pi^{(1)}\right\}^{-1}$$

$$\mathbf{u}_k^{(1)} = \left\{2\mathbf{E}_{(1)k}\Lambda^{(1)}\right\}\left\{2\Lambda^{(1)T}\Lambda^{(1)}\right\}^{-1}$$

$$\mathcal{X}_k = \mathcal{P}_{\Omega^c}(\mathcal{T}_k + \mathcal{C}_k) + \mathcal{P}_{\Omega}(\mathcal{O}_k)$$

where

$$\Pi^{(1)} = \mathbf{A}^{(3)} \odot \mathbf{A}^{(2)},$$

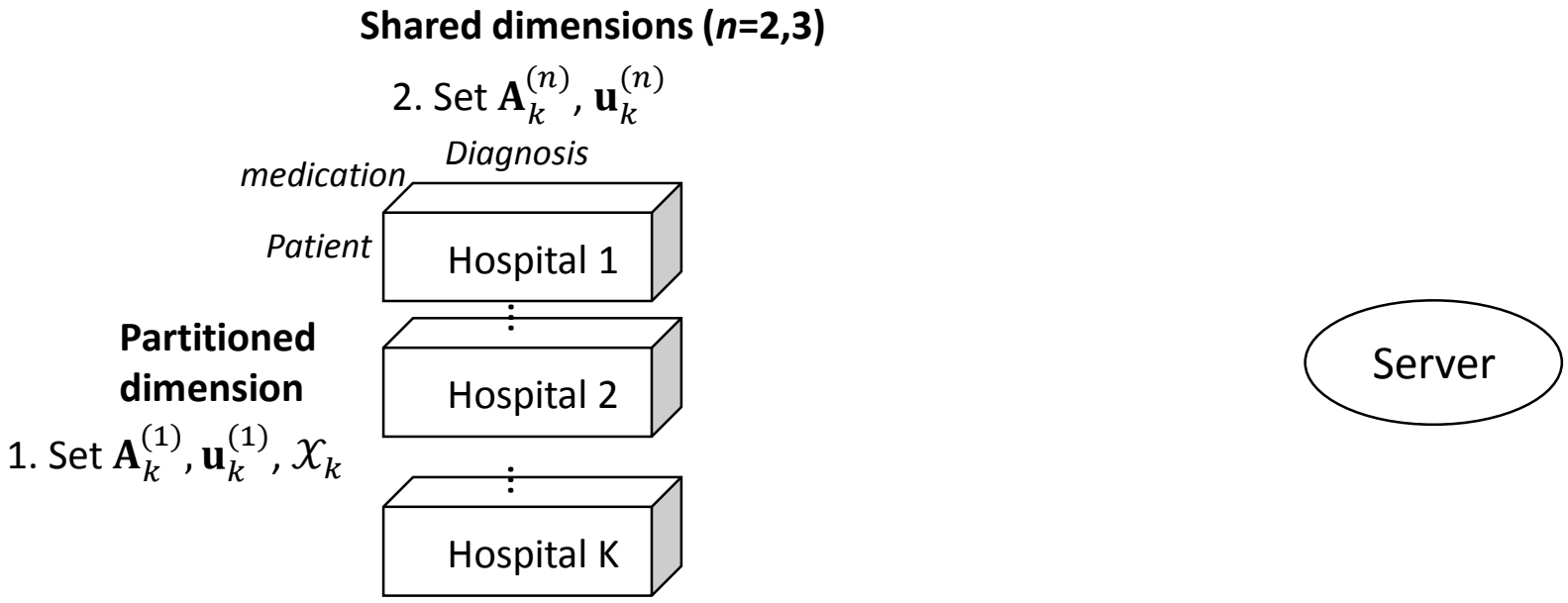
$$\Lambda^{(1)} = \mathbf{u}^{(3)} \odot \mathbf{u}^{(2)},$$

$$\mathcal{R}_k = \mathcal{X}_k - \mathcal{C}_k,$$

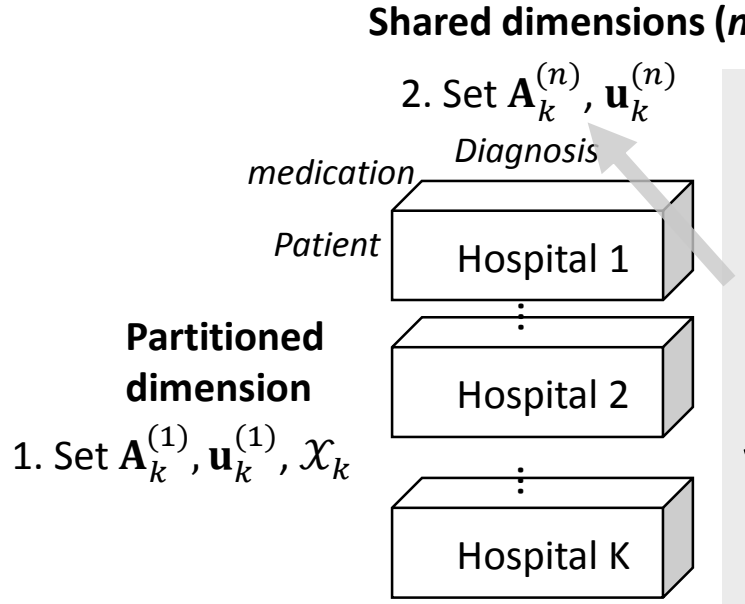
$$\mathcal{E}_k = \mathcal{X}_k - \mathcal{T}_k.$$

Server

Solving the objective function



Solving the objective function



$$\mathbf{A}_k^{(n)} = \left\{2\mathbf{R}_{(n)k}\Pi_k^{(n)} + \omega\mathbf{A}^{(n)} + \mathbf{H}_k^{(n)}\right\}\left\{2\Pi_k^{(n)T}\Pi_k^{(n)} + \omega\mathbf{I}\right\}^{-1}$$

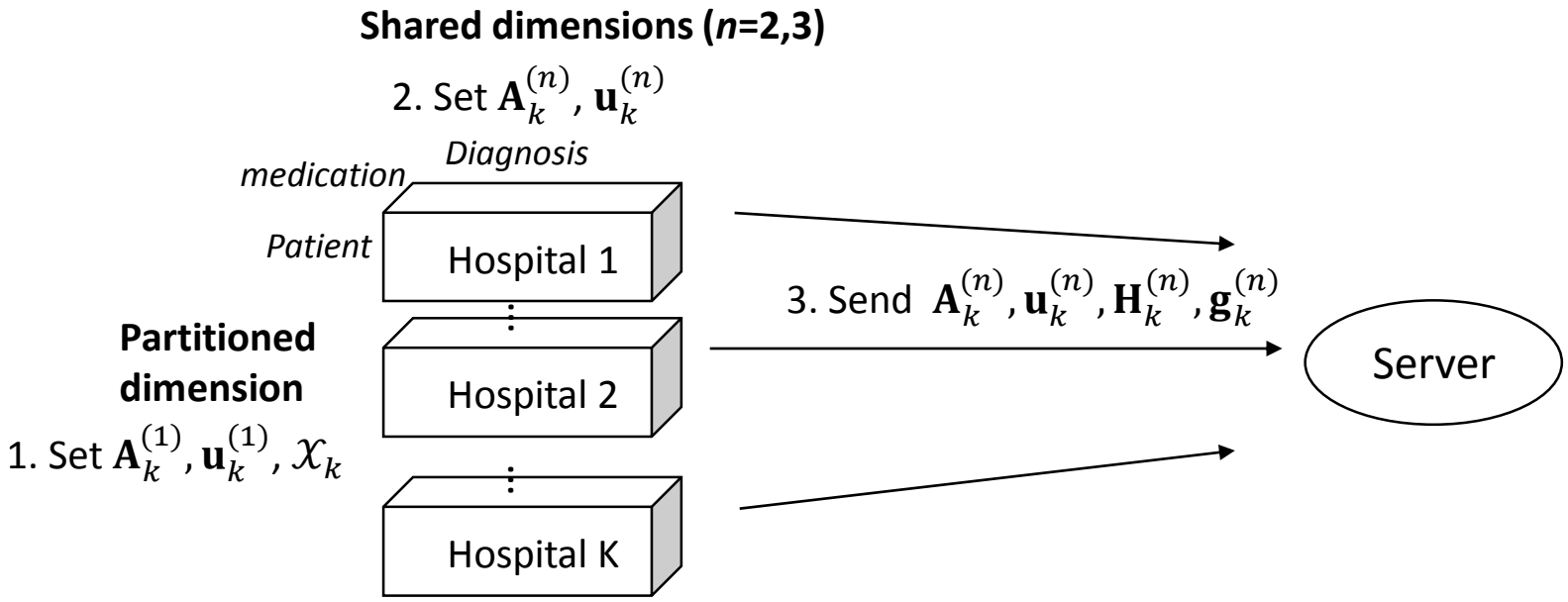
$$\mathbf{u}_k^{(n)} = \left\{2\mathbf{E}_{(n)k}\Lambda_k^{(n)} + \gamma\mathbf{u}^{(n)} + \mathbf{g}_k^{(n)}\right\}\left\{2\Lambda_k^{(n)T}\Lambda_k^{(n)} + \gamma\mathbf{I}\right\}^{-1}$$

where

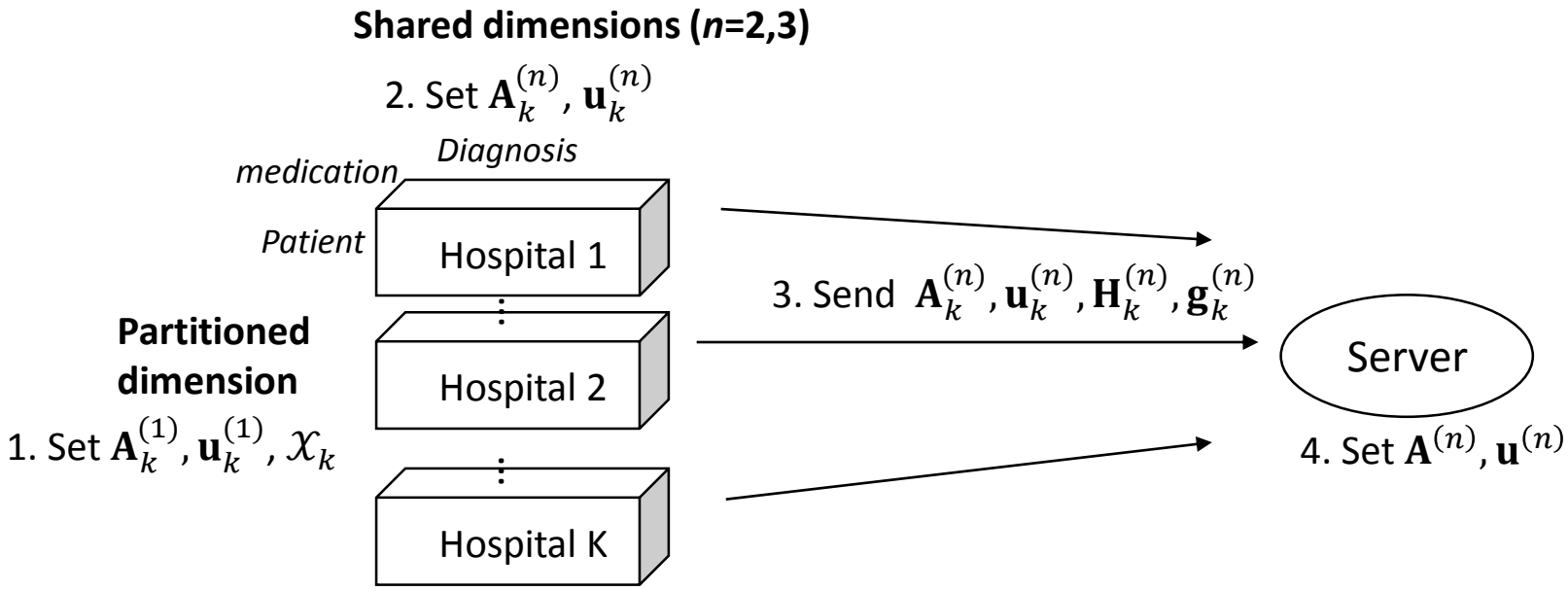
$$\Pi_k^{(2)} = \mathbf{A}_k^{(3)} \odot \mathbf{A}_k^{(1)}, \Pi_k^{(3)} = \mathbf{A}_k^{(2)} \odot \mathbf{A}_k^{(1)}$$

$$\Lambda_k^{(2)} = \mathbf{u}_k^{(3)} \odot \mathbf{u}_k^{(1)}, \Lambda_k^{(3)} = \mathbf{u}_k^{(2)} \odot \mathbf{u}_k^{(1)}$$

Solving the objective function



Solving the objective function



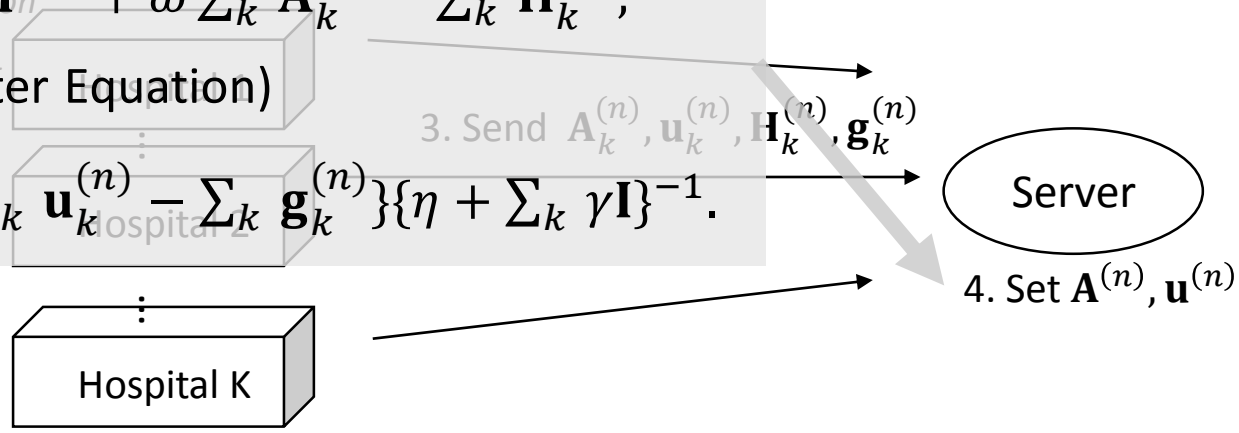
Solving the objective function

$$\mathbf{A}^{(n)}(\mu\mathbf{I} + K\omega\mathbf{I}) + (\lambda_q \mathbf{B}^{(n)}\mathbf{B}^{(n)T})\mathbf{A}^{(n)}$$

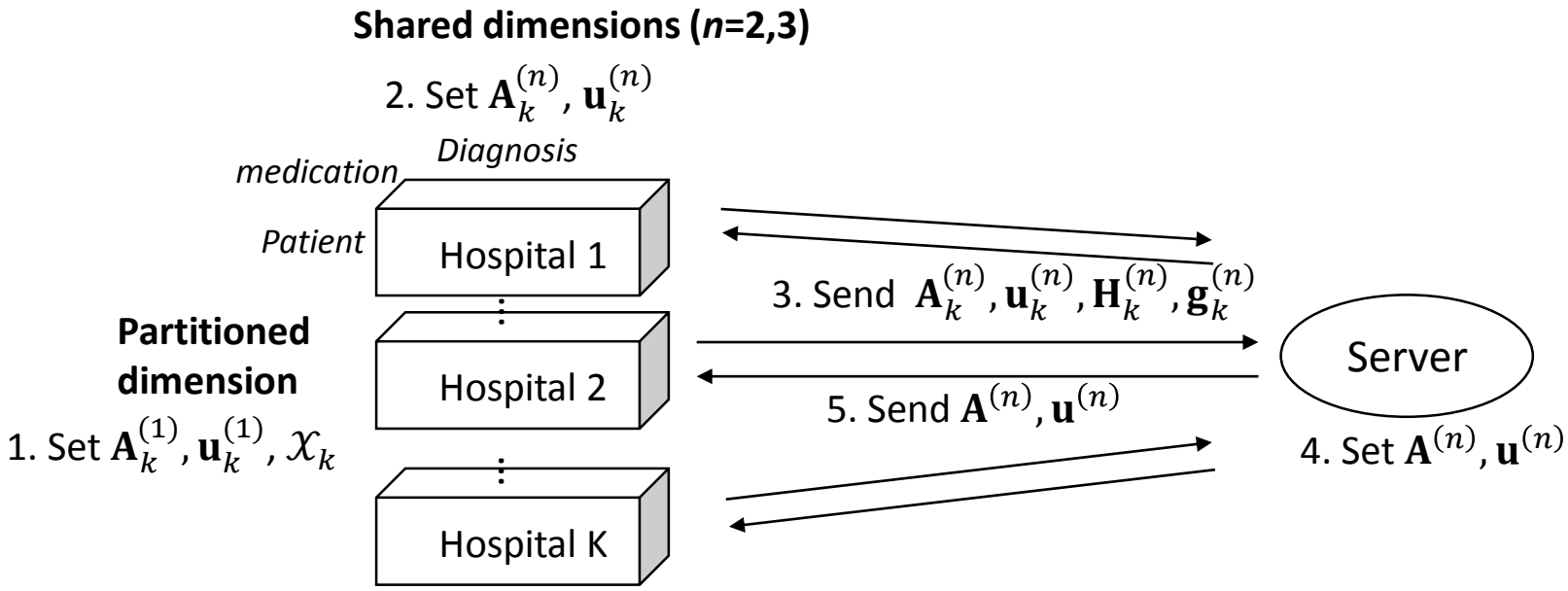
$$= \lambda_q \mathbf{B}^{(n)} + \mu\mathbf{B}^{(n)} + \mathbf{Y}^{(n)} + \omega \sum_k \mathbf{A}_k^{(n)} - \sum_k \mathbf{H}_k^{(n)},$$

(Sylvester Equation)

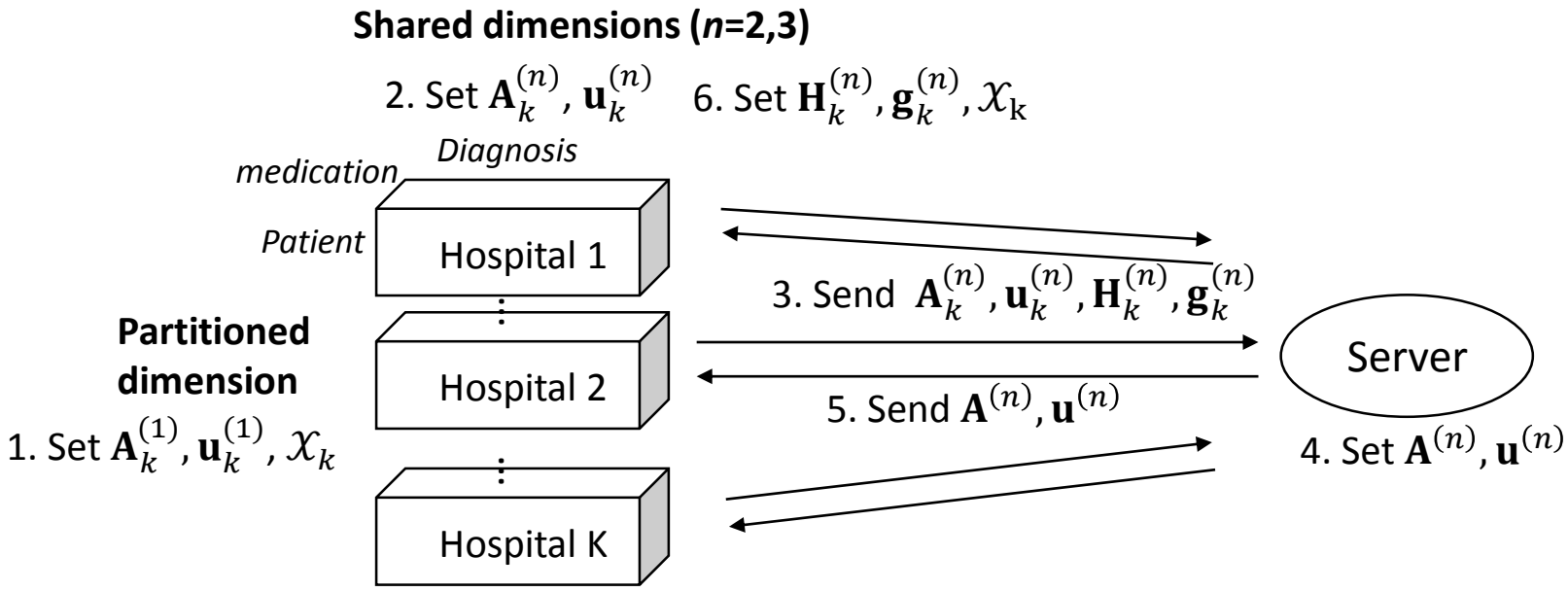
$$\mathbf{u}^{(n)} = \{\eta\mathbf{v}^{(n)} + \mathbf{p}^{(n)} + \gamma \sum_k \mathbf{u}_k^{(n)} - \sum_k \mathbf{g}_k^{(n)}\}\{\eta + \sum_k \gamma\mathbf{I}\}^{-1}.$$



Solving the objective function



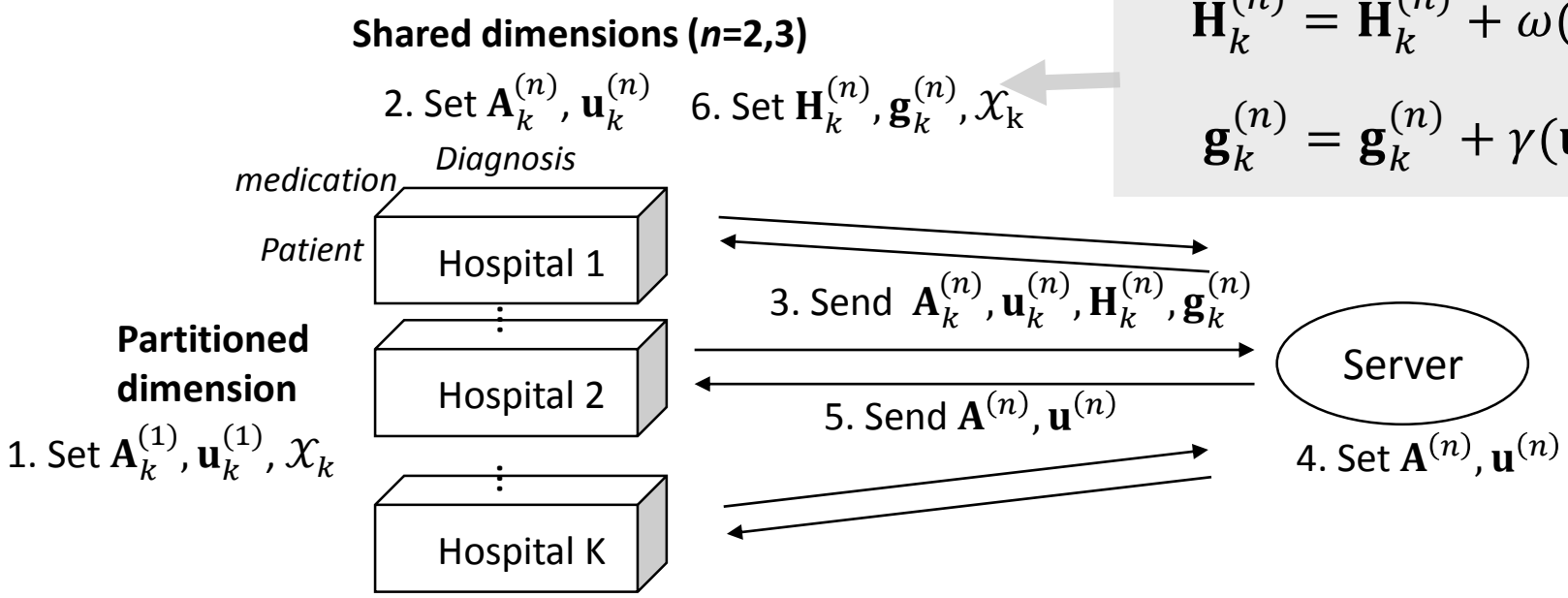
Solving the objective function



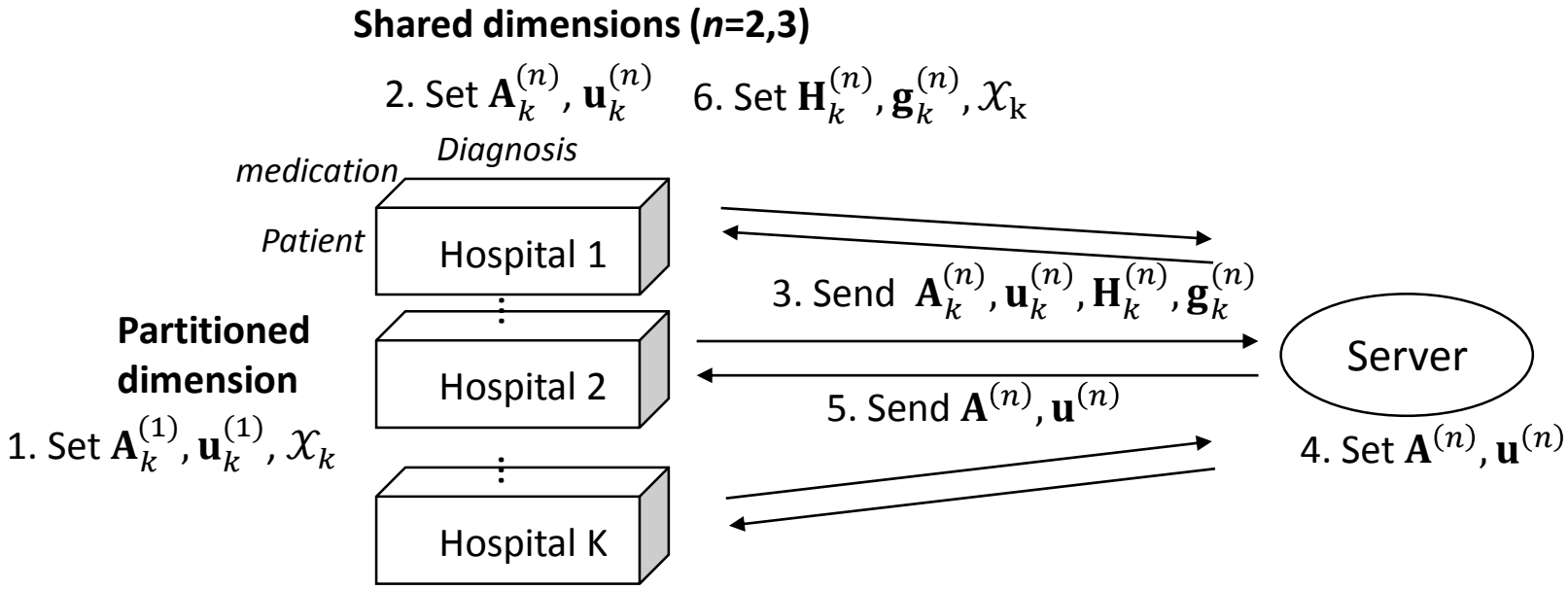
Solving the objective function

$$\mathbf{H}_k^{(n)} = \mathbf{H}_k^{(n)} + \omega(\mathbf{A}^{(n)} - \mathbf{A}_k^{(n)})$$

$$\mathbf{g}_k^{(n)} = \mathbf{g}_k^{(n)} + \gamma(\mathbf{u}^{(n)} - \mathbf{u}_k^{(n)}).$$



Solving the objective function



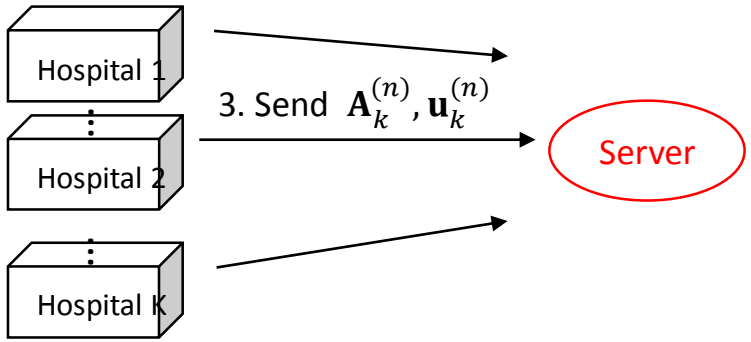
Repeat step 1 to 6 until convergence

Privacy Analysis

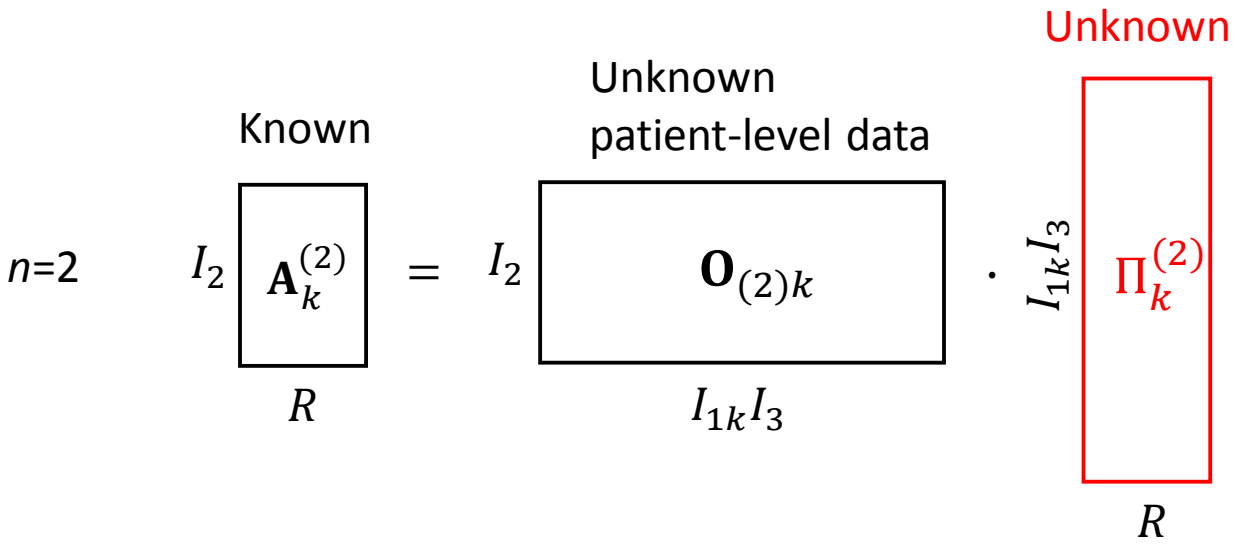
Server might try to do reverse-engineering with the local factor matrix received from hospitals.

$$\mathbf{A}_k^{(n)} = \left\{2\mathbf{R}_{(n)k}\Pi_k^{(\mathrm{n})} + \omega\mathbf{A}^{(n)} + \mathbf{H}_k^{(n)}\right\}\left\{2\Pi_k^{(\mathrm{n})T}\Pi_k^{(\mathrm{n})} + \omega\mathbf{I}\right\}^{-1}$$

$$\mathbf{A}_k^{(n)} = \mathbf{O}_{(n)k}\Pi_k^{(\mathrm{n})} \text{ (Simplified form after removing known values)}$$



But, patient-level data in $\mathbf{O}_{(n)k}$ is not revealed because



i) server cannot access to $\Pi_k^{(\mathrm{n})}$

($\because \Pi_{\mathrm{k}}^{(2)} = \mathbf{A}_{\mathrm{k}}^{(3)} \odot \mathbf{A}_{\mathrm{k}}^{(1)}$,

$\Pi_{\mathrm{k}}^{(3)} = \mathbf{A}_{\mathrm{k}}^{(2)} \odot \mathbf{A}_{\mathrm{k}}^{(1)}$,

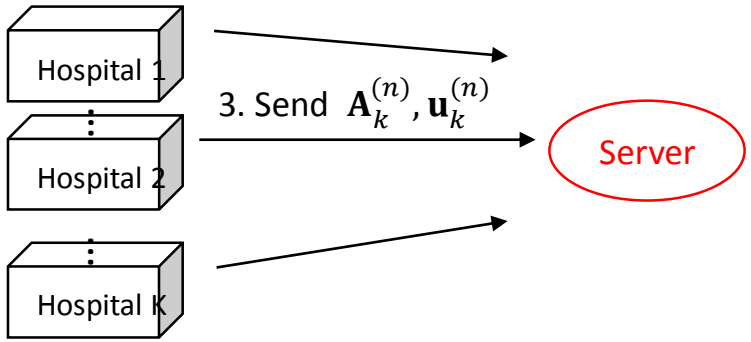
and $\mathbf{A}_{\mathrm{k}}^{(1)}$ is unknown)

Privacy Analysis

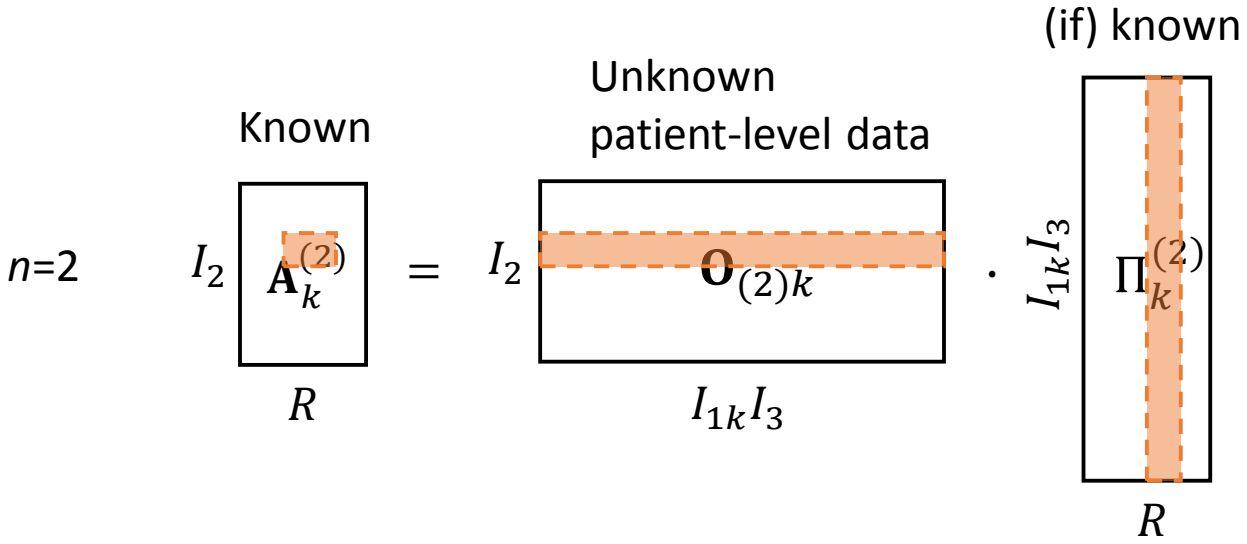
Server might try to do reverse-engineering with the local factor matrix received from hospitals.

$$\mathbf{A}_k^{(n)} = \left\{2\mathbf{R}_{(n)k}\Pi_k^{(n)} + \omega\mathbf{A}^{(n)} + \mathbf{H}_k^{(n)}\right\}\left\{2\Pi_k^{(n)T}\Pi_k^{(n)} + \omega\mathbf{I}\right\}^{-1}$$

$$\mathbf{A}_k^{(n)} = \mathbf{O}_{(n)k}\Pi_k^{(n)} \text{ (Simplified form after removing known values)}$$



But, patient-level data in $\mathbf{O}_{(n)k}$ is not revealed because



ii) Server cannot recover unknown

$I_{1k}I_2I_3$ values from I_2R equations

$(I_{1k}I_2I_3 \gg I_2R)$

Hospitals cannot do reverse-engineering by the same reasons.

Secure set alignment

- Medications and diagnoses dimensions might not be aligned

  Hospital 1, Y_1={COPD, diabetes, hypertension}

  Hospital 2, Y_2={asthenia, diabetes, hypertension, sickle cell}

- The index should refer to the same elements.

	$Y_1 \cap Y_2$		$Y_1 \cap Y_2^c$	$Y_1^c \cap Y_2$	
Y_1=	diabetes	hypertension	COPD	-	-
Y_2=	diabetes	hypertension	-	asthenia	Sickle cell

- We need a secure set alignment method by which hospitals do not reveal their elements and get an integrated view on the elements.

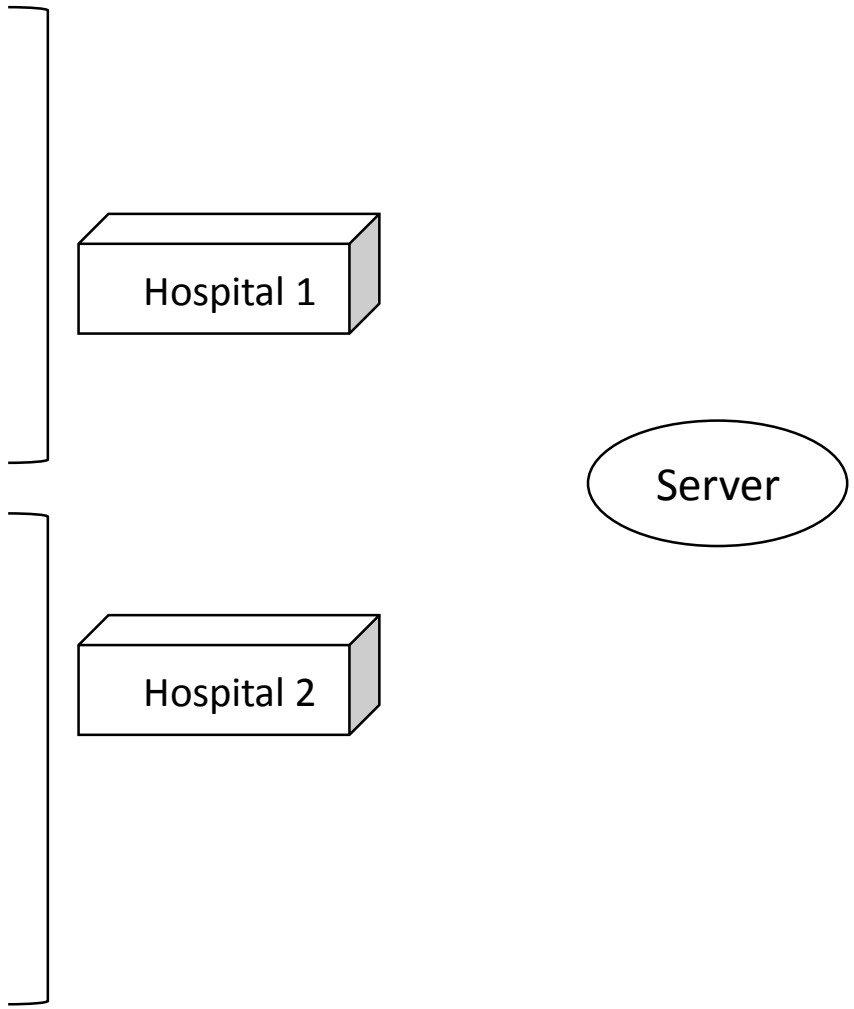
Secure set alignment

- **Lemma 1** *A polynomial function of y that represents set of elements $Y_k = \{y_k\}_{k=1}^{I}$ at hospital k is $f_k(y) = (y - y_{1k})(y - y_{2k}) \cdots (y - y_{Ik}) = \sum_{i=0}^{I} a_i y^i$. A y_k is an element of Y_k ($y_k \in Y_k$) if and only if $f_k(y_k) = 0$.*
- **Lemma 2** *A polynomial function that represents intersection of Y_k and $Y_{k'}$ ($Y_k \cap Y_{k'}$) is $f_k * r + f_{k'} * s$ where r, s are polynomial functions with $gcd(r, s) = 1$. Given $f_k * r + f_{k'} * s$, one cannot learn individual elements on Y_1 and Y_2 other than elements in $Y_1 \cap Y_2$.*

Secure set alignment

Y_1 ={COPD, diabetes, hypertension} ={2, 3, 4}

Y_2 ={asthenia, diabetes, hypertension, sickle cell} ={1, 3, 4, 5}



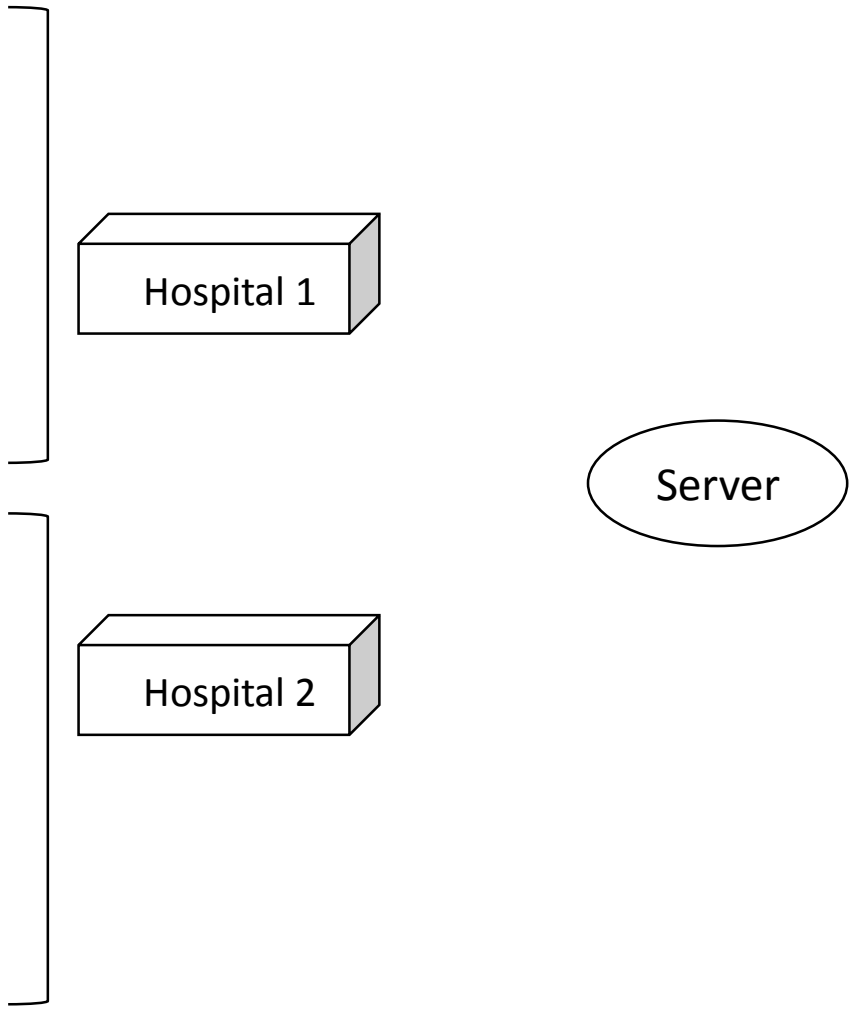
Secure set alignment

Y_1={COPD, diabetes, hypertension} ={2, 3, 4}

$\Leftrightarrow f_1(y) = (y-2)(y-3)(y-4)(y-\alpha_1) \Leftrightarrow f_1(y)\%P$

Y_2={asthenia, diabetes, hypertension, sickle cell} ={1, 3, 4, 5}

$\Leftrightarrow f_2(y) = (y-1)(y-3)(y-4)(y-5)(y-\alpha_2) \Leftrightarrow f_2(y)\%P$



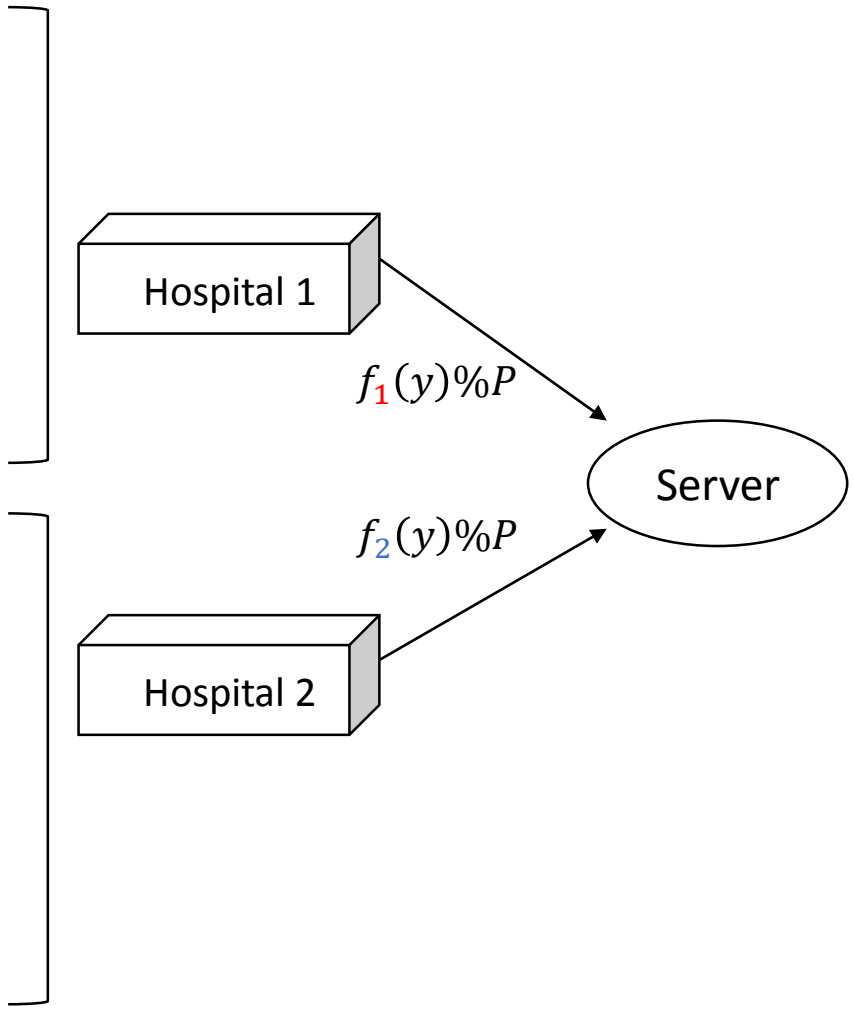
Secure set alignment

Y_1={COPD, diabetes, hypertension} ={2, 3, 4}

$\Leftrightarrow f_1(y) = (y-2)(y-3)(y-4)(y-\alpha_1) \Leftrightarrow f_1(y)\%P$

Y_2={asthenia, diabetes, hypertension, sickle cell} ={1, 3, 4, 5}

$\Leftrightarrow f_2(y) = (y-1)(y-3)(y-4)(y-5)(y-\alpha_2) \Leftrightarrow f_2(y)\%P$



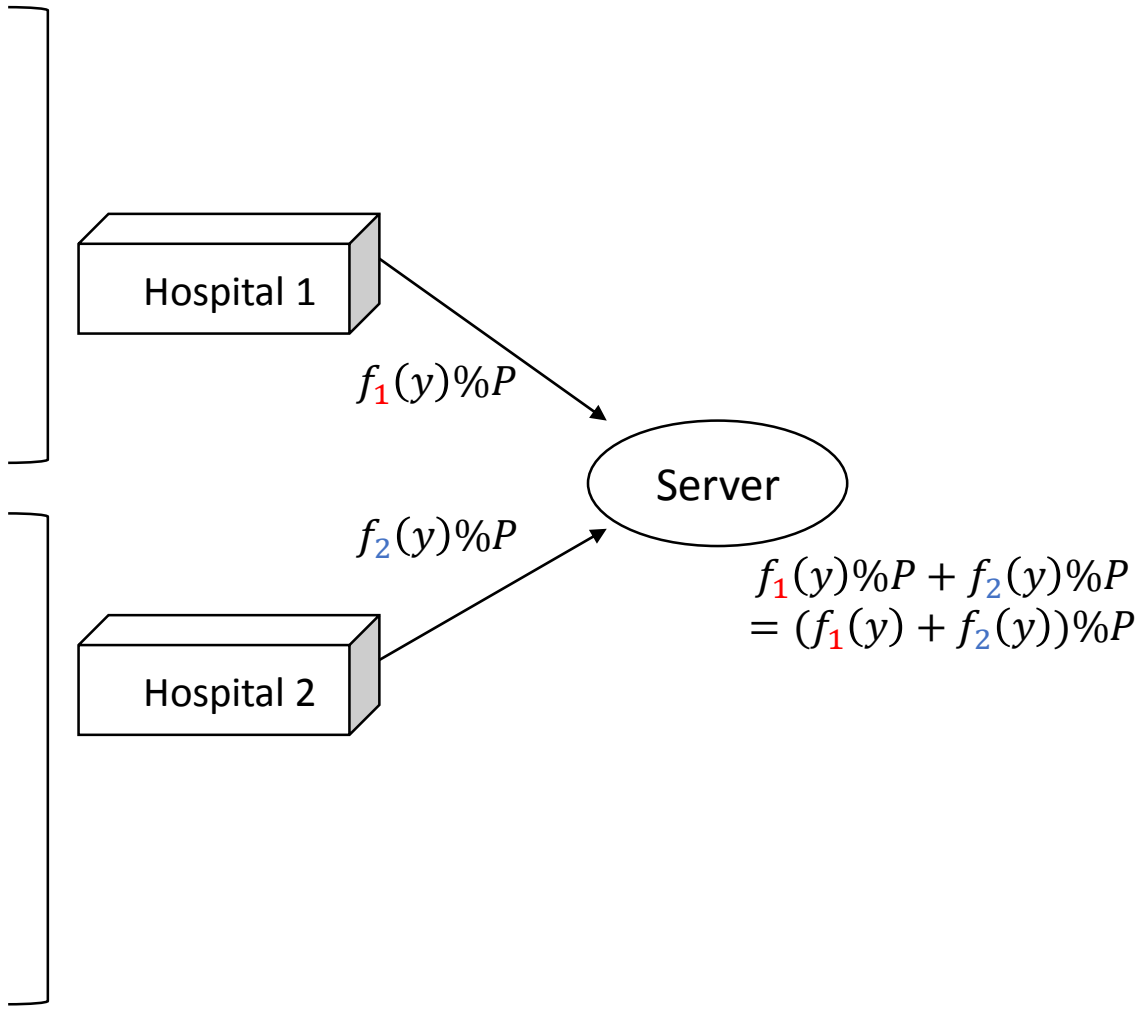
Secure set alignment

Y_1={COPD, diabetes, hypertension} ={2, 3, 4}

$\Leftrightarrow f_1(y) = (y-2)(y-3)(y-4)(y-\alpha_1) \Leftrightarrow f_1(y)\%P$

Y_2={asthenia, diabetes, hypertension, sickle cell} ={1, 3, 4, 5}

$\Leftrightarrow f_2(y) = (y-1)(y-3)(y-4)(y-5)(y-\alpha_2) \Leftrightarrow f_2(y)\%P$



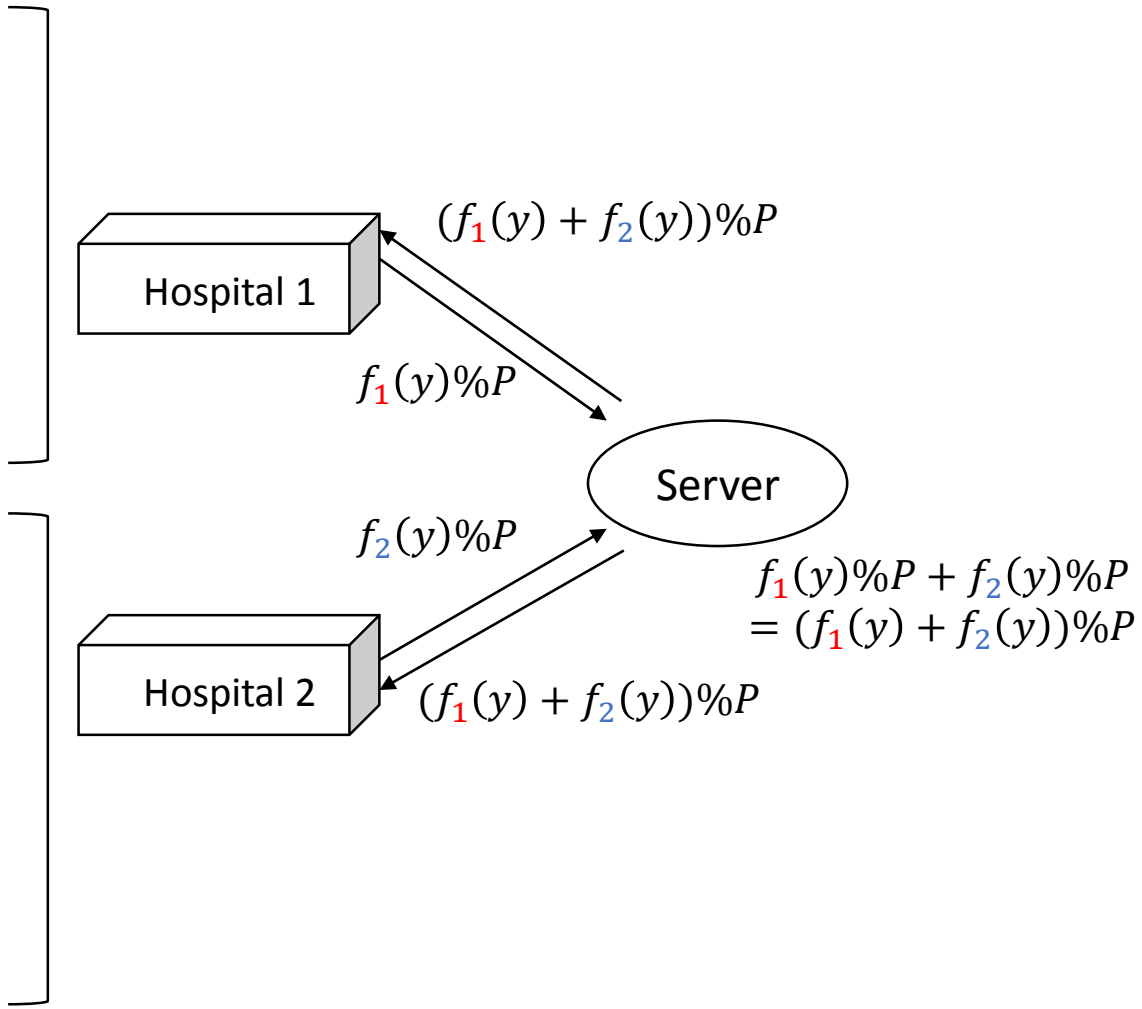
Secure set alignment

Y_1={COPD, diabetes, hypertension} ={2, 3, 4}

$\Leftrightarrow f_1(y) = (y-2)(y-3)(y-4)(y-\alpha_1) \Leftrightarrow f_1(y)\%P$

Y_2={asthenia, diabetes, hypertension, sickle cell} ={1, 3, 4, 5}

$\Leftrightarrow f_2(y) = (y-1)(y-3)(y-4)(y-5)(y-\alpha_2) \Leftrightarrow f_2(y)\%P$



Secure set alignment

Y_1={COPD, diabetes, hypertension} ={2, 3, 4}

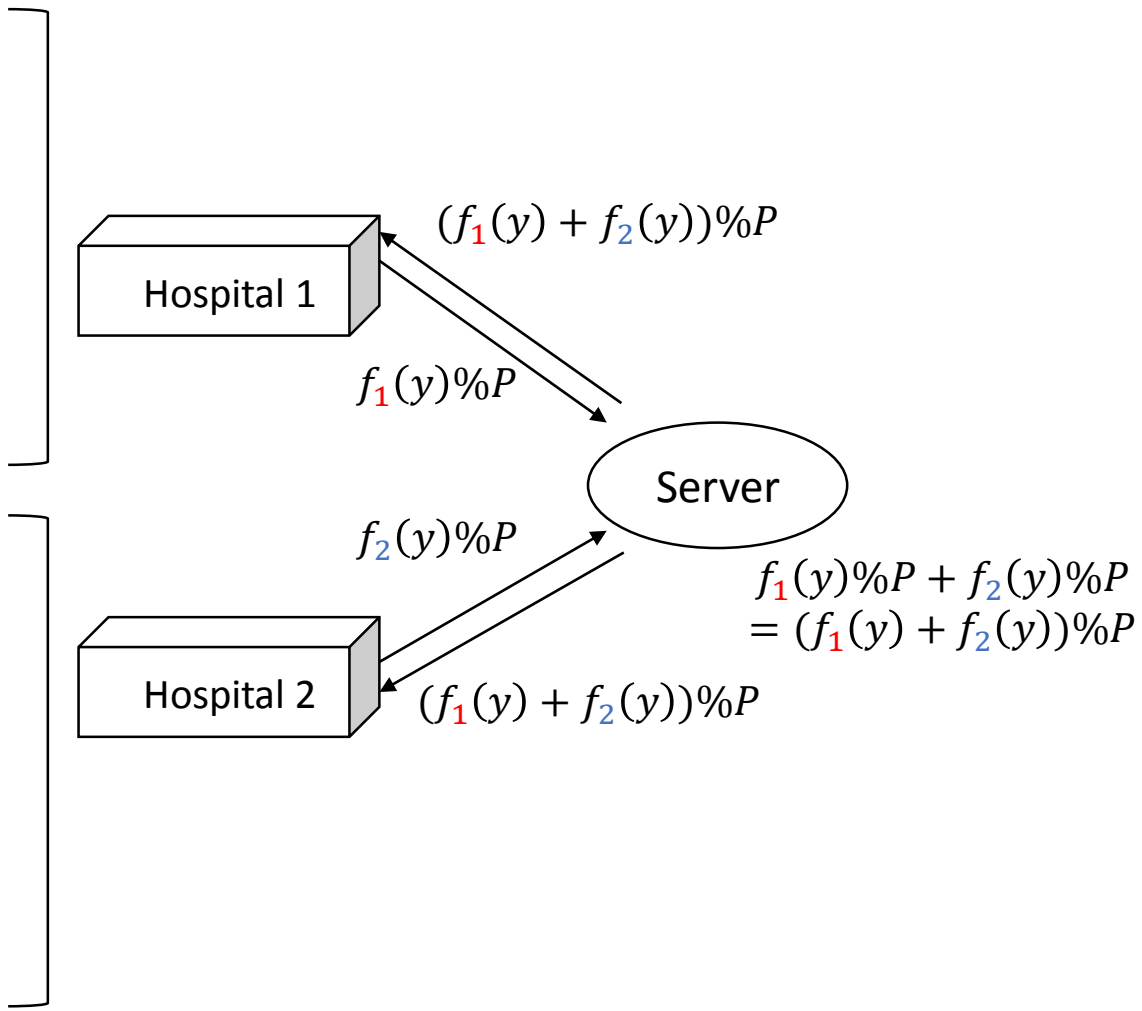
$\Leftrightarrow f_1(y) = (y-2)(y-3)(y-4)(y-\alpha_1) \Leftrightarrow f_1(y)\%P$

$(f_1(2) + f_2(2))\%P \neq 0,$
$(f_1(3) + f_2(3))\%P = 0,$
$(f_1(4) + f_2(4))\%P = 0,$

Y_2={asthenia, diabetes, hypertension, sickle cell} ={1, 3, 4, 5}

$\Leftrightarrow f_2(y) = (y-1)(y-3)(y-4)(y-5)(y-\alpha_2) \Leftrightarrow f_2(y)\%P$

$(f_1(1) + f_2(1))\%P \neq 0,$
$(f_1(3) + f_2(3))\%P = 0,$
$(f_1(4) + f_2(4))\%P = 0,$
$(f_1(5) + f_2(5))\%P \neq 0,$



Secure set alignment

Y_1={COPD, diabetes, hypertension} ={2, 3, 4}

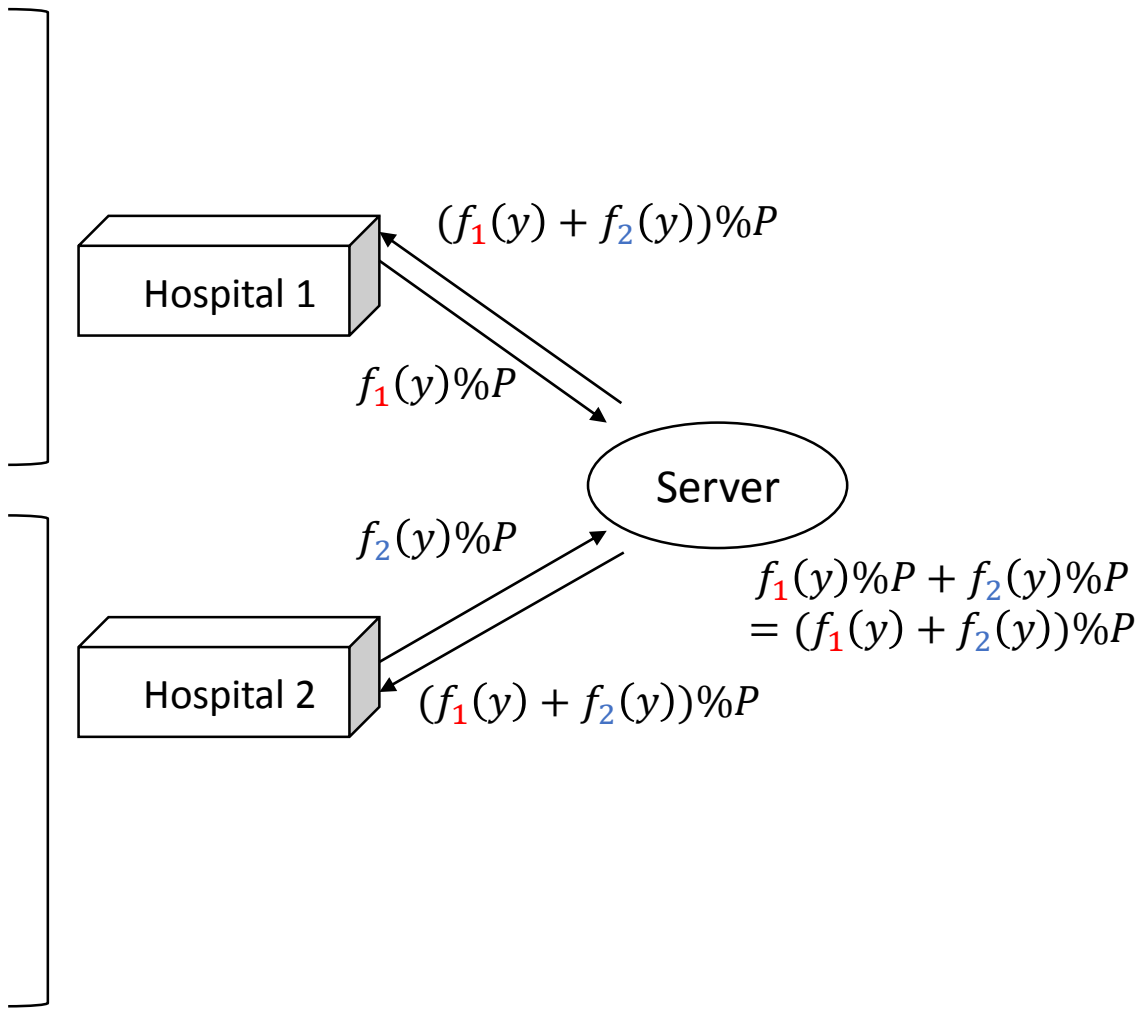
$\Leftrightarrow f_1(y) = (y-2)(y-3)(y-4)(y-\alpha_1) \Leftrightarrow f_1(y)\%P$

$(f_1(2) + f_2(2))\%P \neq 0, \quad Y_1 \cap Y_2^C = \{2\}, |Y_1 \cap Y_2^C| = 1,$
$(f_1(3) + f_2(3))\%P = 0, \quad Y_1 \cap Y_2 = \{3,4\}, \ |Y_1 \cap Y_2| = 2$
$(f_1(4) + f_2(4))\%P = 0,$

Y_2={asthenia, diabetes, hypertension, sickle cell} ={1, 3, 4, 5}

$\Leftrightarrow f_2(y) = (y-1)(y-3)(y-4)(y-5)(y-\alpha_2) \Leftrightarrow f_2(y)\%P$

$(f_1(1) + f_2(1))\%P \neq 0, \quad Y_1^C \cap Y_2 = \{1,5\}, |Y_1^C \cap Y_2| = 2,$
$(f_1(3) + f_2(3))\%P = 0, \quad Y_1 \cap Y_2 = \{3,4\}, \ |Y_1 \cap Y_2| = 2$
$(f_1(4) + f_2(4))\%P = 0,$
$(f_1(5) + f_2(5))\%P \neq 0,$



Secure set alignment

Y_1={COPD, diabetes, hypertension} ={2, 3, 4}

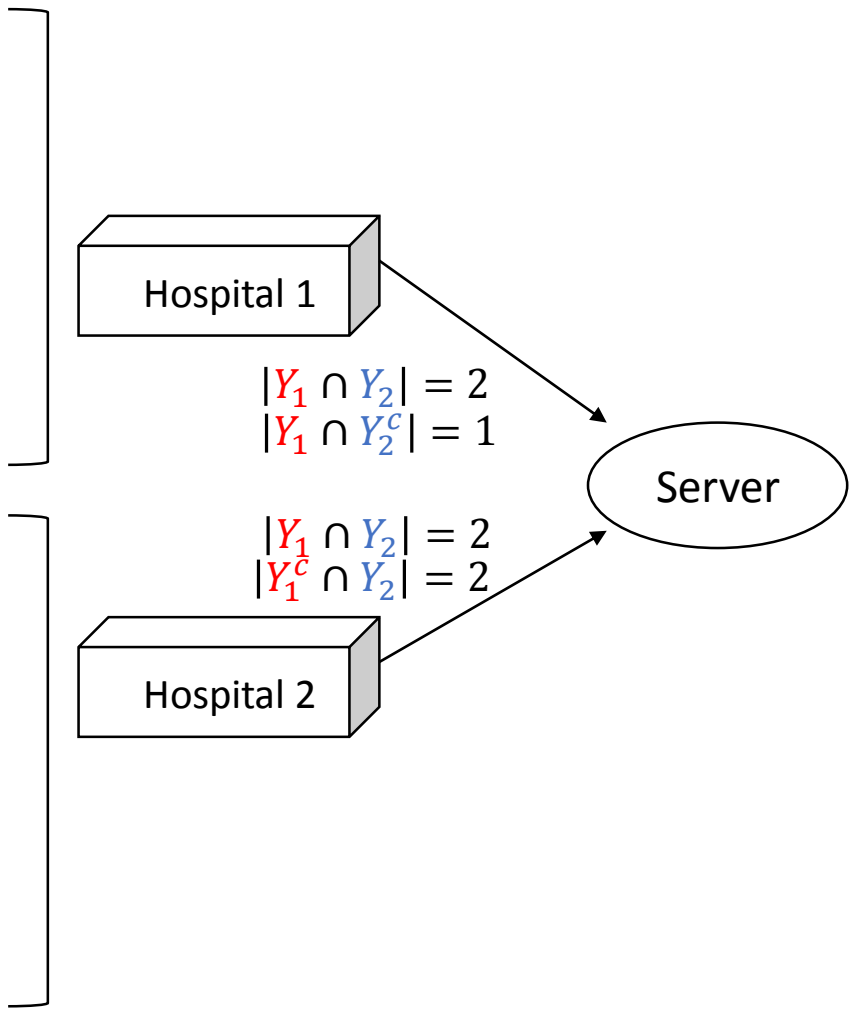
$\Leftrightarrow f_1(y) = (y-2)(y-3)(y-4)(y-\alpha_1) \Leftrightarrow f_1(y)\%P$

$(f_1(2) + f_2(2))\%P \neq 0, \quad Y_1 \cap Y_2^C = \{2\}, |Y_1 \cap Y_2^C| = 1,$
$(f_1(3) + f_2(3))\%P = 0, \quad Y_1 \cap Y_2 = \{3,4\}, \ |Y_1 \cap Y_2| = 2$
$(f_1(4) + f_2(4))\%P = 0,$

Y_2={asthenia, diabetes, hypertension, sickle cell} ={1, 3, 4, 5}

$\Leftrightarrow f_2(y) = (y-1)(y-3)(y-4)(y-5)(y-\alpha_2) \Leftrightarrow f_2(y)\%P$

$(f_1(1) + f_2(1))\%P \neq 0, \quad Y_1^C \cap Y_2 = \{1,5\}, |Y_1^C \cap Y_2| = 2,$
$(f_1(3) + f_2(3))\%P = 0, \quad Y_1 \cap Y_2 = \{3,4\}, \ |Y_1 \cap Y_2| = 2$
$(f_1(4) + f_2(4))\%P = 0,$
$(f_1(5) + f_2(5))\%P \neq 0,$



Secure set alignment

Y_1={COPD, diabetes, hypertension} ={2, 3, 4}

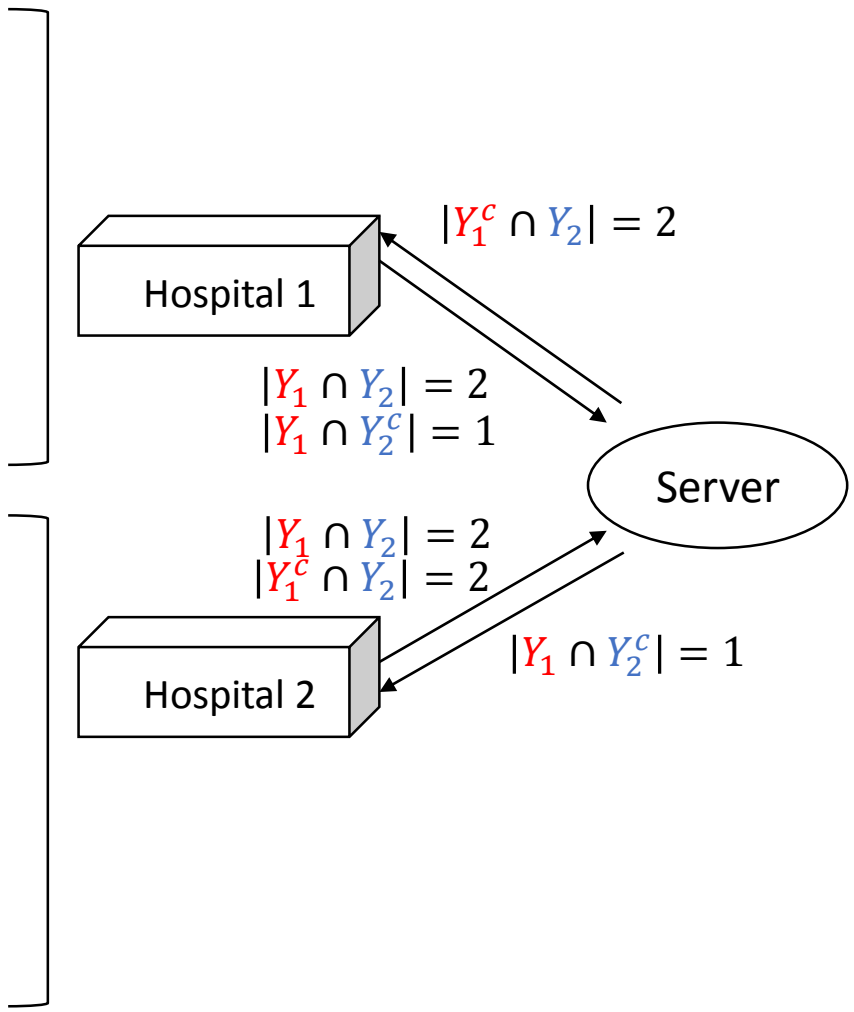
$\Leftrightarrow f_1(y) = (y-2)(y-3)(y-4)(y-\alpha_1) \Leftrightarrow f_1(y)\%P$

$(f_1(2) + f_2(2))\%P \neq 0, \quad Y_1 \cap Y_2^C = \{2\}, |Y_1 \cap Y_2^C| = 1,$
$(f_1(3) + f_2(3))\%P = 0, \quad Y_1 \cap Y_2 = \{3,4\}, \ |Y_1 \cap Y_2| = 2$
$(f_1(4) + f_2(4))\%P = 0,$

Y_2={asthenia, diabetes, hypertension, sickle cell} ={1, 3, 4, 5}

$\Leftrightarrow f_2(y) = (y-1)(y-3)(y-4)(y-5)(y-\alpha_2) \Leftrightarrow f_2(y)\%P$

$(f_1(1) + f_2(1))\%P \neq 0, \quad Y_1^C \cap Y_2 = \{1,5\}, |Y_1^C \cap Y_2| = 2,$
$(f_1(3) + f_2(3))\%P = 0, \quad Y_1 \cap Y_2 = \{3,4\}, \ |Y_1 \cap Y_2| = 2$
$(f_1(4) + f_2(4))\%P = 0,$
$(f_1(5) + f_2(5))\%P \neq 0,$



Secure set alignment

Y_1={COPD, diabetes, hypertension} ={2, 3, 4}

$\Leftrightarrow f_1(y) = (y-2)(y-3)(y-4)(y-\alpha_1) \Leftrightarrow f_1(y)\%P$

$(f_1(2) + f_2(2))\%P \neq 0, \quad Y_1 \cap Y_2^C = \{2\}, |Y_1 \cap Y_2^C| = 1,$
$(f_1(3) + f_2(3))\%P = 0, \quad Y_1 \cap Y_2 = \{3,4\}, \ |Y_1 \cap Y_2| = 2$
$(f_1(4) + f_2(4))\%P = 0,$

	$Y_1 \cap Y_2$		$Y_1 \cap Y_2^c$	$Y_1^c \cap Y_2$	
Y_1=	3	4	2	-	-

= {diabetes, hypertension, COPD, -, -}

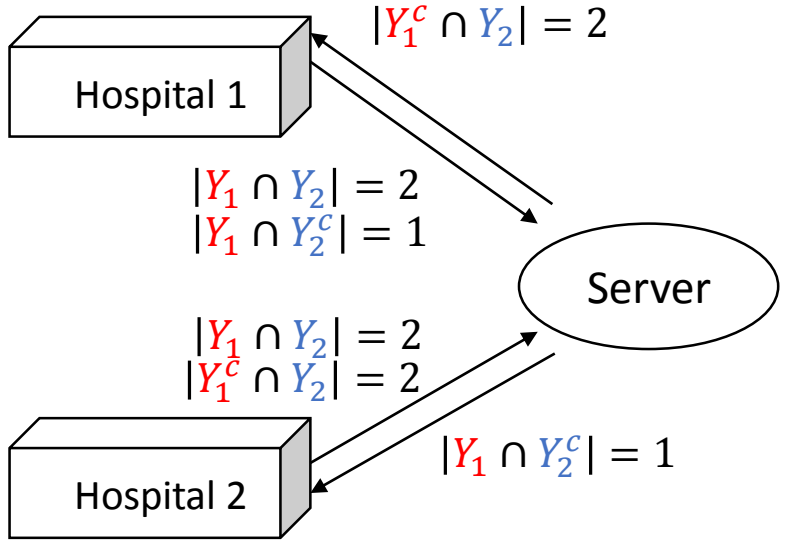
Y_2={asthenia, diabetes, hypertension, sickle cell} ={1, 3, 4, 5}

$\Leftrightarrow f_2(y) = (y-1)(y-3)(y-4)(y-5)(y-\alpha_2) \Leftrightarrow f_2(y)\%P$

$(f_1(1) + f_2(1))\%P \neq 0, \quad Y_1^C \cap Y_2 = \{1,5\}, |Y_1^C \cap Y_2| = 2,$
$(f_1(3) + f_2(3))\%P = 0, \quad Y_1 \cap Y_2 = \{3,4\}, \ |Y_1 \cap Y_2| = 2$
$(f_1(4) + f_2(4))\%P = 0,$
$(f_1(5) + f_2(5))\%P \neq 0,$

	$Y_1 \cap Y_2$		$Y_1 \cap Y_2^c$	$Y_1^c \cap Y_2$	
Y_2=	3	4	-	1	5

= {diabetes, hypertension, -, asthenia, asthenia}



Experimental Settings

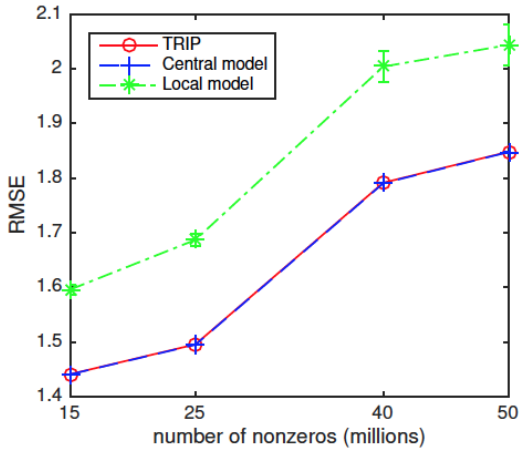
- Baselines
  - Central model: Traditional model in which all data is combined [KDD15*]
  - Local model: An intuitive model by which hospitals run central model and average the results
- Evaluation measures
  - RMSE and time w.r.t. #nonzeros, #hospitals, data skewness
  - Phenotype discovery
- Datasets
  - Co-occurrence of patients, medication, diagnosis from MIMIC-III with 38,035 patients * 3,229 medications * 304 lab results (RMSE and time)
  - UCSD Medical Center 8,022 patients * 748 medications * 299 diagnosis (Phenotype discovery)

* Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun. Rubik: Knowledge Guided Tensor Factorization and Completion for Health Data Analytics. KDD ’15

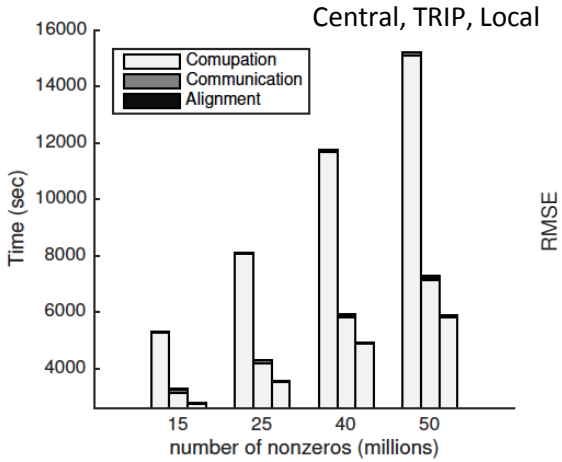
Experimental Results

i) RMSE and time with respect to the number of nonzeros

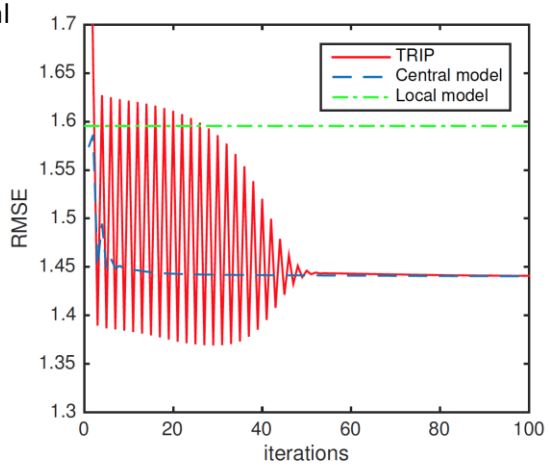
As a result, *TRIP* has low RMSE as much as central model and reduces computation time by distributing updating procedures to de-centralized hospitals.



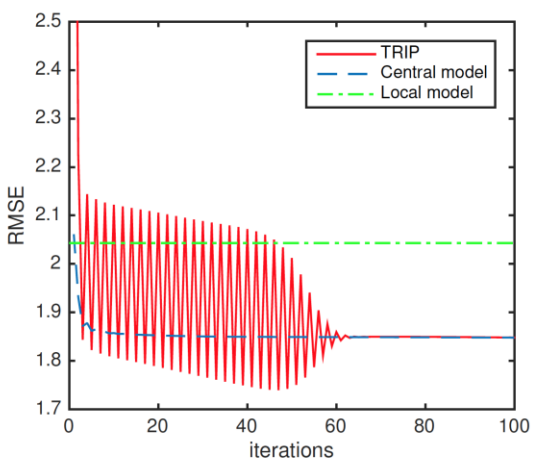
(a) RMSE vs # nonzero



(b) Time vs. # nonzero



(c) MIMIC-III 15M

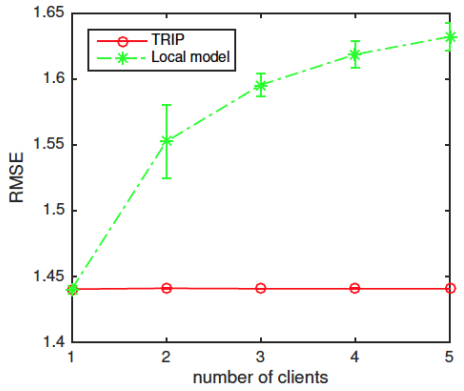


(d) MIMIC-III 50M

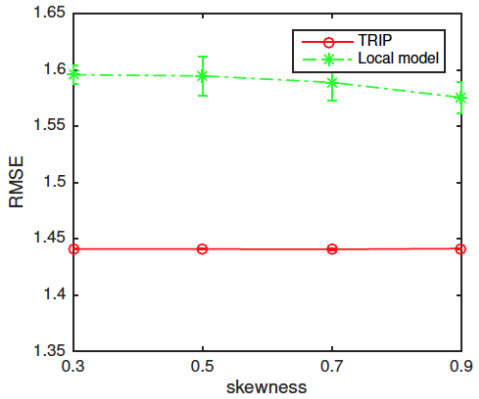
Experimental Results

ii) RMSE and time with respect to the number of hospitals and skewness

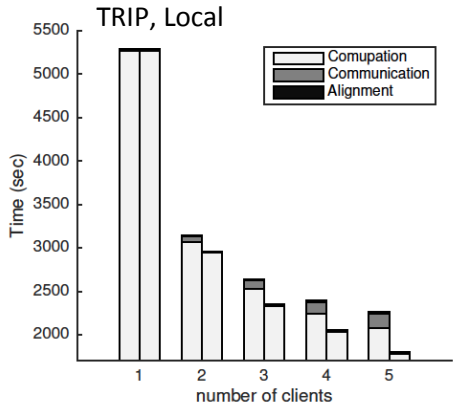
RMSE of *TRIP* is stable when the number of hospitals increases or patients are distributed unevenly, and is similar to RMSE of central model → TRIP is robust on the finely split or skewed data.



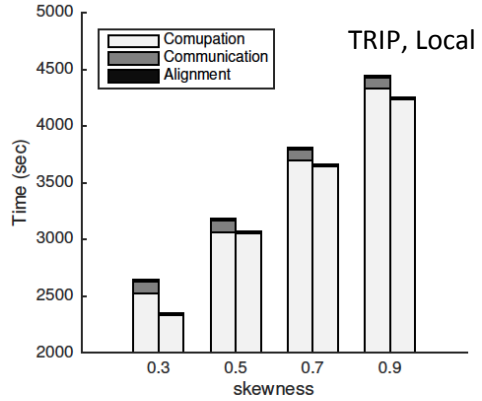
(a) RMSE vs. # clients



(b) RMSE vs. skewness



(c) Time vs. # clients



(d) Time vs. skewness

Experimental Results. Phenotype discovery

TRIP discovers phenotypes as the central model with combined patient data does

Table 2: Phenotypes from TRIP, central model, and each hospitals in UCSD. Phenotypes from UCSD1, UCSD2, and both are marked as blue, red, and purple, respectively.

Rank	TRIP	Central model	UCSD1	UCSD2
1	Coronary artery disease with diabetes & hypertension	Diabetic with hypertension	Diabetic with hypertension	Cystic fibrosis with pancreatic involvement
2	Diabetic with hypertension	Cystic fibrosis with pancreatic involvement	Coronary artery disease with diabetes & hypertension	Cystic fibrosis with pulmonary exacerbation
3	Chronic obstructive pulmonary disease (COPD) exacerbation	Coronary artery disease with diabetes & hypertension	COPD exacerbation	Neurogenic bladder with abdominal pain
4	Constipation	Hypertension	Decompensated cirrhosis	Non-specific gastrointestinal complaints
5	Cystic fibrosis with pancreatic involvement	COPD exacerbation	Non-specific gastrointestinal complaints	Diabetes
6	Decompensated cirrhosis	Constipation	Non-specific complaints	Constipation
7	Non-specific gastrointestinal complaints	Decompensated cirrhosis	COPD w/o exacerbation	Anxiety with gastrointestinal complaints
8	Cystic fibrosis with pulmonary exacerbation	Non-specific complaints	Acute on chronic pain	Cystic fibrosis with pneumonia
9	Sickle cell/chronic pain crisis	Sickle cell/chronic pain crisis	COPD with Pneumonia	Non-specific complaints
10	Non-specific complaints	Neurogenic bladder with abdominal pain	Anxiety with hypertension	Lymphoma

Conclusions

- We introduce federated tensor factorization for computational phenotyping without sharing patient-level data.
- We developed secure data harmonization and privacy preserving computation procedures based on ADMM, and analyzed that TRIP ensure the confidentiality of patient-level data.
- Well-designed distributed models can help derive useful phenotypes from EHR data to overcome policy barriers due to the privacy concerns.

UC San Diego

SCHOOL OF MEDICINE

Department of BioMedical Informatics

Questions?

Acknowledgements

- Yejin Kim
- Hwanjo Yu

This research was supported by grant R21LM012060