Distance Metric Learning in Data Mining (Part I)

Fei Wang and Jimeng Sun

IBM TJ Watson Research Center

Outline

Part I - Applications

- Motivation and Introduction
- Patient similarity application

Part II - Methods

- Unsupervised Metric Learning
- Supervised Metric Learning
- Semi-supervised Metric Learning
- Challenges and Opportunities for Metric Learning

Outline

Part I - Applications

- Motivation and Introduction
- Patient similarity application

Part II - Methods

- Unsupervised Metric Learning
- Supervised Metric Learning
- Semi-supervised Metric Learning
- Challenges and Opportunities for Metric Learning

What is a Distance Metric?

Suppose $\mathcal{X}$ is a sef ot data points, $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ are data vectors with the same dimensionality, then we call $\mathcal{D} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a **Distance Metric** if it satisfies the following four properties:

—Nonnegativity: $\mathcal{D}(\mathbf{x}, \mathbf{y}) \geqslant 0$

—Coincidence: $\mathcal{D}(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$

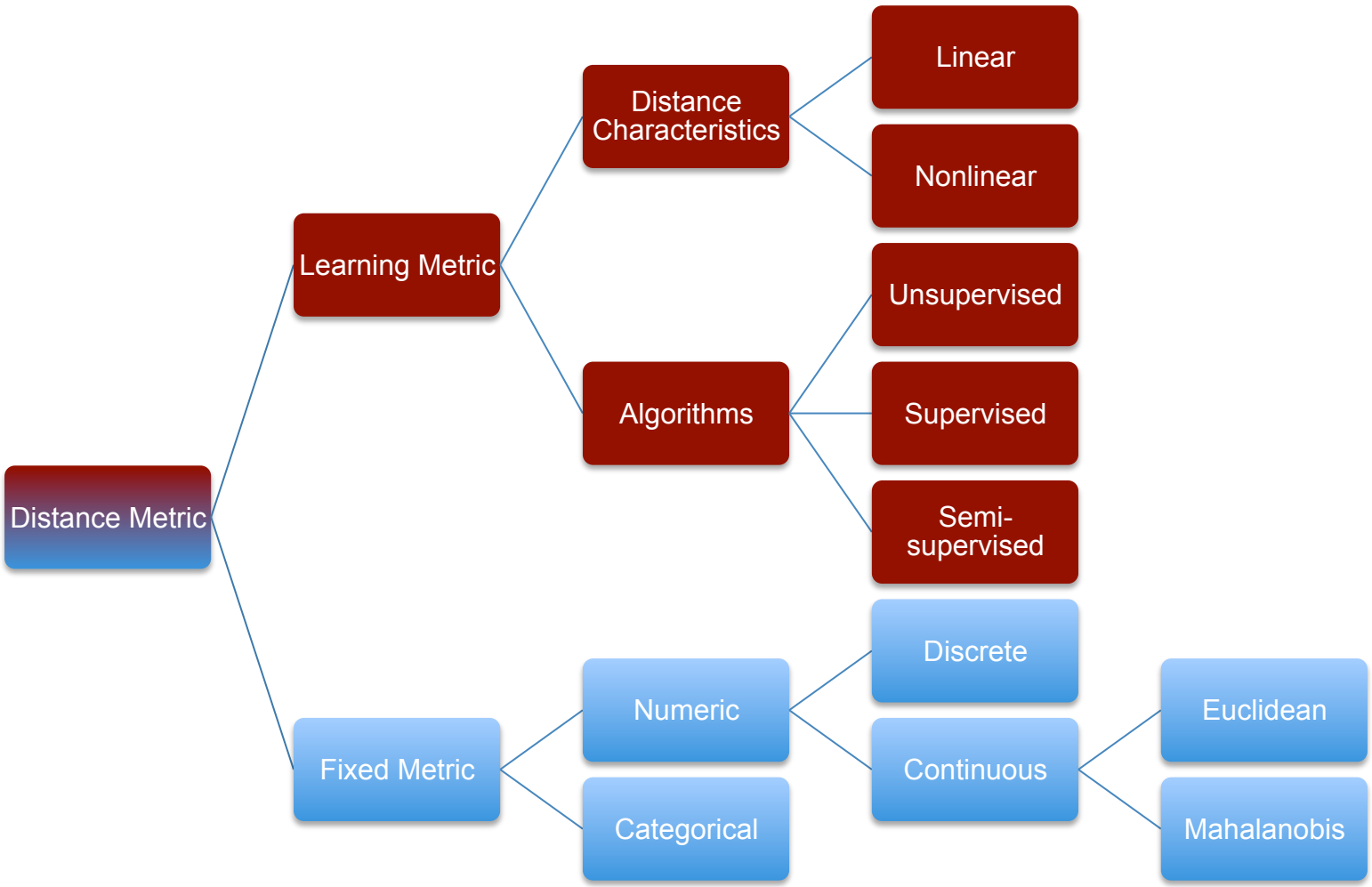
—Symmetry: $\mathcal{D}(\mathbf{x}, \mathbf{y}) = \mathcal{D}(\mathbf{y}, \mathbf{x})$

—Subadditivity: $\mathcal{D}(\mathbf{x}, \mathbf{y}) + \mathcal{D}(\mathbf{y}, \mathbf{z}) \geqslant \mathcal{D}(\mathbf{x}, \mathbf{z})$

If we relax the coincidence condition to *if* $\mathbf{x} = \mathbf{y} \Rightarrow \mathcal{D}(\mathbf{x}, \mathbf{y}) = 0$, then $\mathcal{D}$ is called a **Pseudo Metric**.

From wikipedia

Distance Metric Taxonomy



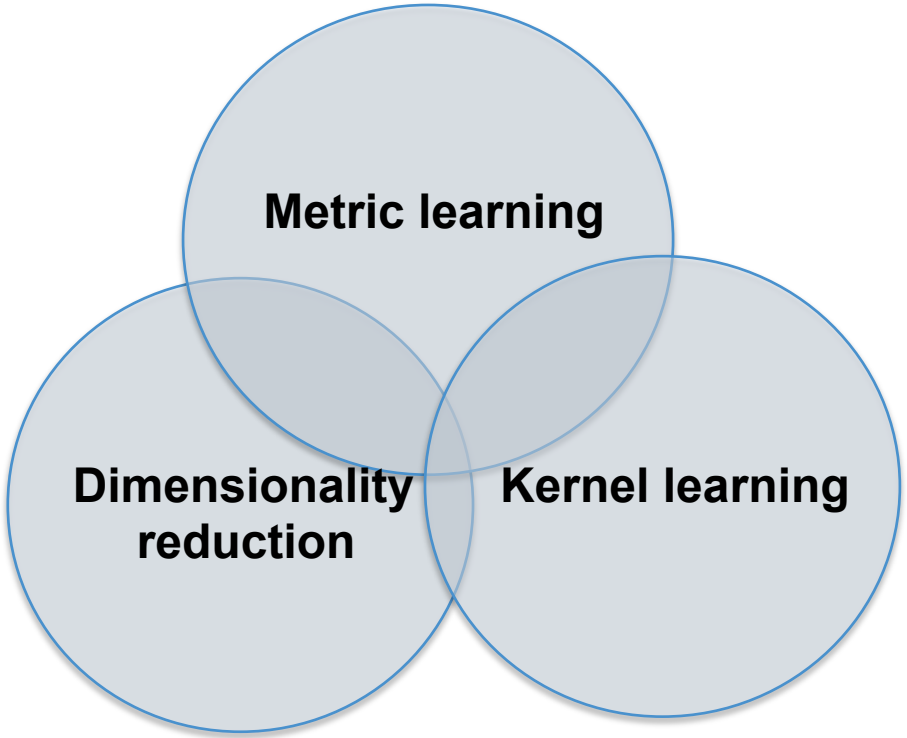
Categorization of Distance Metrics: Linear vs. Non-linear

- **Linear Distance**
  - First perform a **linear** mapping to project the data into some space, and then evaluate the pairwise data distance as their Euclidean distance in the projected space
  - Generalized Mahalanobis distance
    - $\mathcal{D}(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{x}-\mathrm{y})^\top \mathrm{M}(\mathrm{x}-\mathrm{y})}$
    - If M is symmetric, and positive definite, D is a distance metric;
    - If M is symmetric, and positive semi-definite, D is a pseudo-metric.
    - $\mathrm{M} = \mathbf{W}\mathbf{W}^\top$
    - $\mathcal{D}(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{x}-\mathrm{y})^\top \mathbf{W}\mathbf{W}^\top(\mathrm{x}-\mathrm{y})} = \sqrt{(\mathbf{W}^\top(\mathrm{x}-\mathrm{y}))^\top(\mathbf{W}^\top(\mathrm{x}-\mathrm{y}))}$
      $= \sqrt{(\tilde{\mathrm{x}}-\tilde{\mathrm{y}})^\top(\tilde{\mathrm{x}}-\tilde{\mathrm{y}})}$
- **Nonlinear Distance**
  - First perform a **nonlinear** mapping to project the data into some space, and then evaluate the pairwise data distance as their Euclidean distance in the projected space

Categorization of Distance Metrics: Global vs. Local

- **Global Methods**
  - The distance metric satisfies some global properties of the data set
    - PCA: find a direction on which the variation of the whole data set is the largest
    - LDA: find a direction where the data from the same classes are clustered while from different classes are separated
    - …
- **Local Methods**
  - The distance metric satisfies some local properties of the data set
    - LLE: find the low dimensional embeddings such that the local neighborhood structure is preserved
    - LSML: find the projection such that the local neighbors of different classes are separated
    - ….

Related Areas



Related area 1: Dimensionality reduction

- **Dimensionality Reduction** is the process of reducing the number of features through **Feature Selection** and **Feature Transformation**.

- **Feature Selection**
  - Find a subset of the original features
  - Correlation, mRMR, information gain…

- **Feature Transformation**
  - Transforms the original features in a lower dimension space
  - PCA, LDA, LLE, Laplacian Embedding…

- Each dimensionality reduction technique can map a distance metric, where we first perform dimensionality reduction, then evaluate the Euclidean distance in the embedded space

Related area 2: Kernel learning

What is Kernel?

Suppose we have a data set X with n data points, then the kernel matrix **K** defined on X is an nxn symmetric positive semi-definite matrix, such that its (i,j)-th element represents the similarity between the i-th and j-th data points

Kernel Leaning vs. Distance Metric Learning

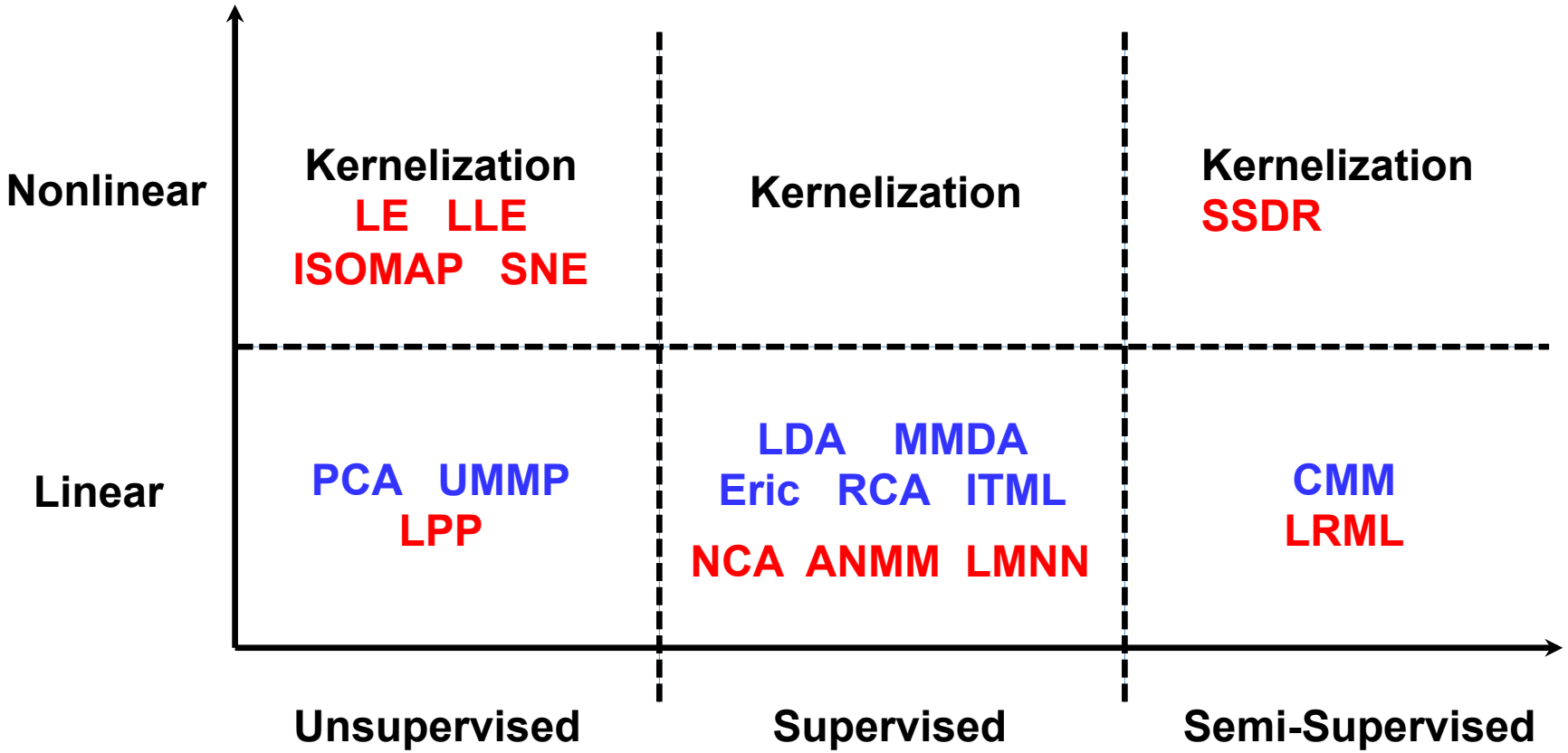
Kernel learning constructs a new kernel from the data, i.e., an inner-product function in some feature space, while distance metric learning constructs a distance function from the data

Kernel Leaning vs. Kernel Trick

- Kernel learning infers the n by n kernel matrix from the data,
- Kernel trick is to apply the predetermined kernel to map the data from the original space to a feature space, such that the nonlinear problem in the original space can become a linear problem in the feature space

B. Scholkopf, A. Smola. Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond. 2001.

Different Types of Metric Learning



- **Red** means local methods
- **Blue** means global methods

Applications of metric learning

Computer vision

- image classification and retrieval
- object recognition

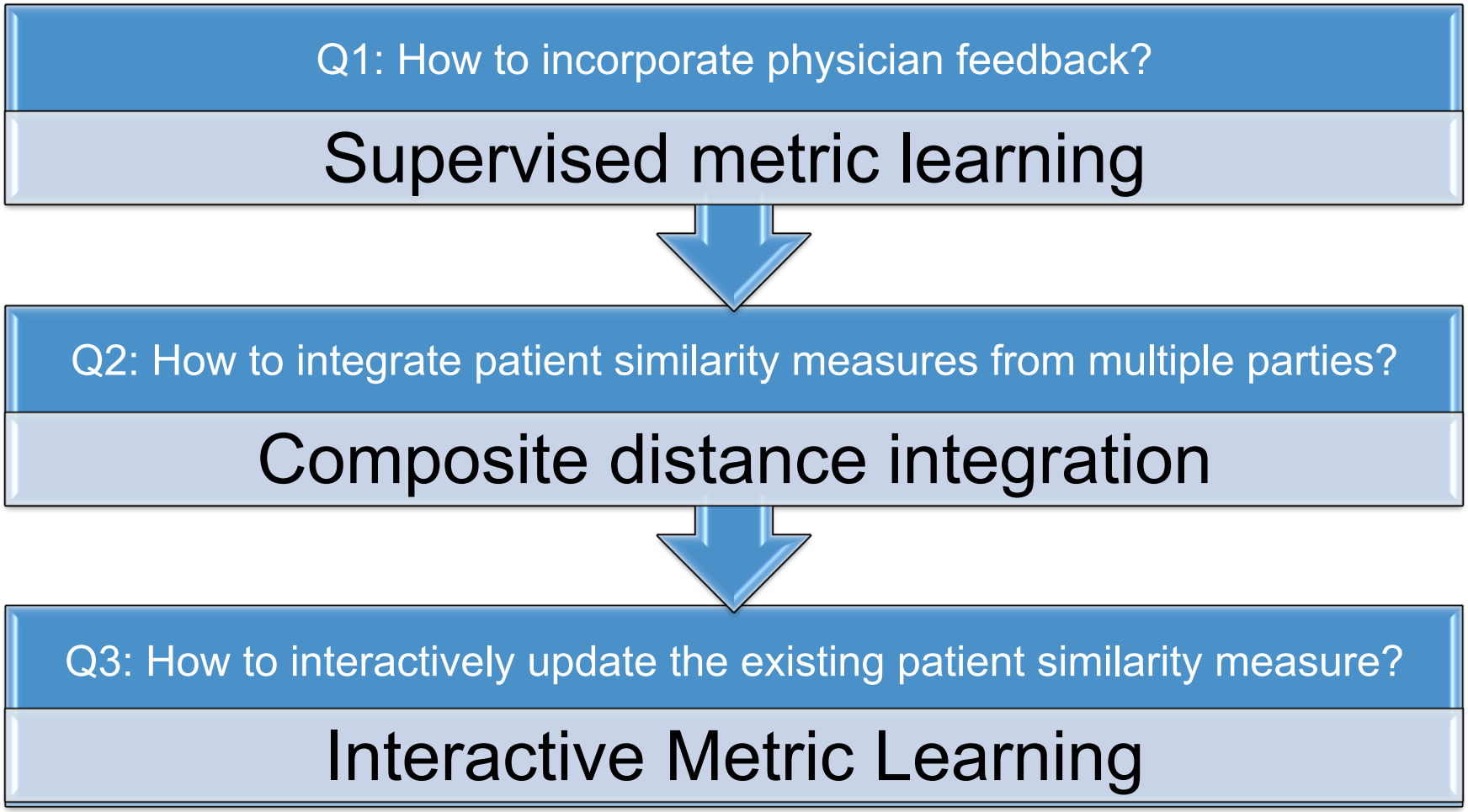
Text analysis

- document classification and retrieval

Medical informatics

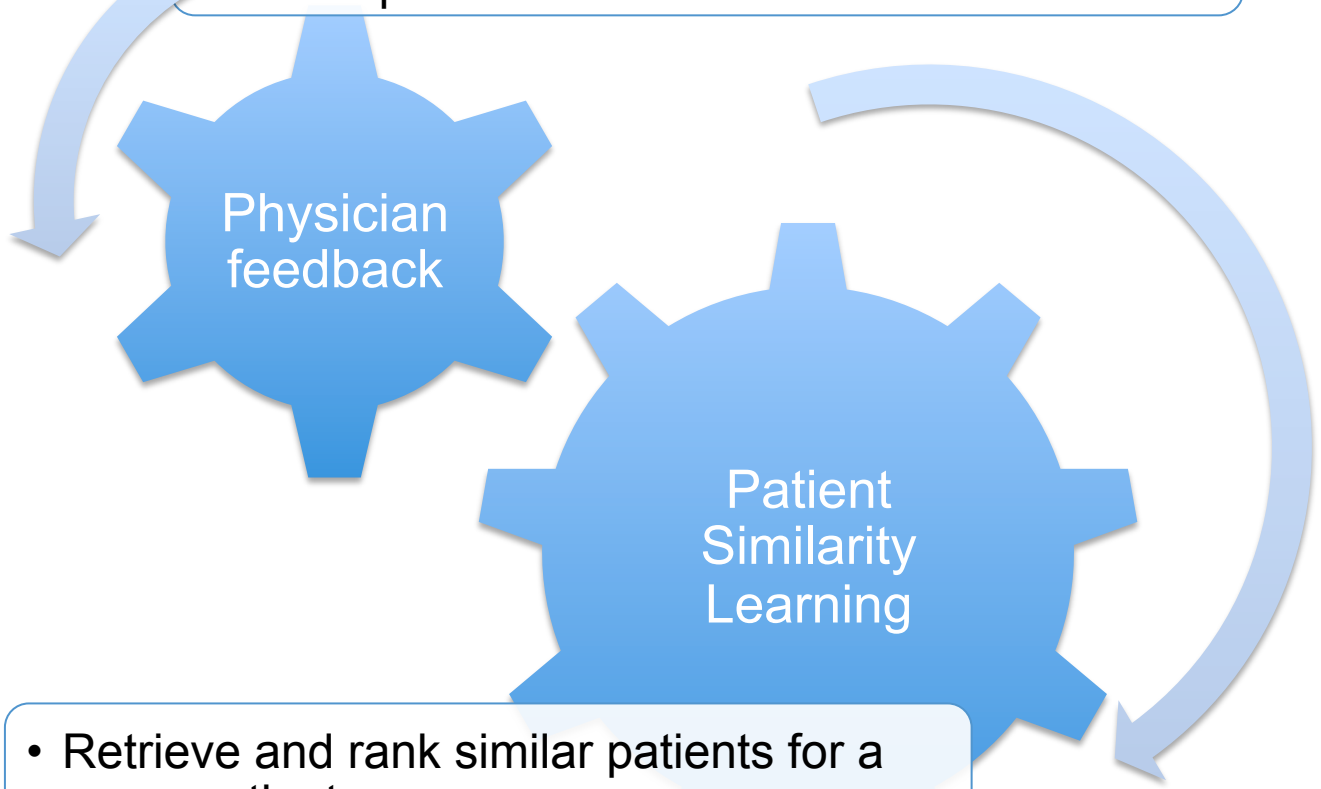
- patient similarity

Patient Similarity Applications



Q1: How to incorporate physician feedback?

- Leverage historical data about the similar patients to diagnose the query patient
- Provide positive/negative feedback about the similar patient results

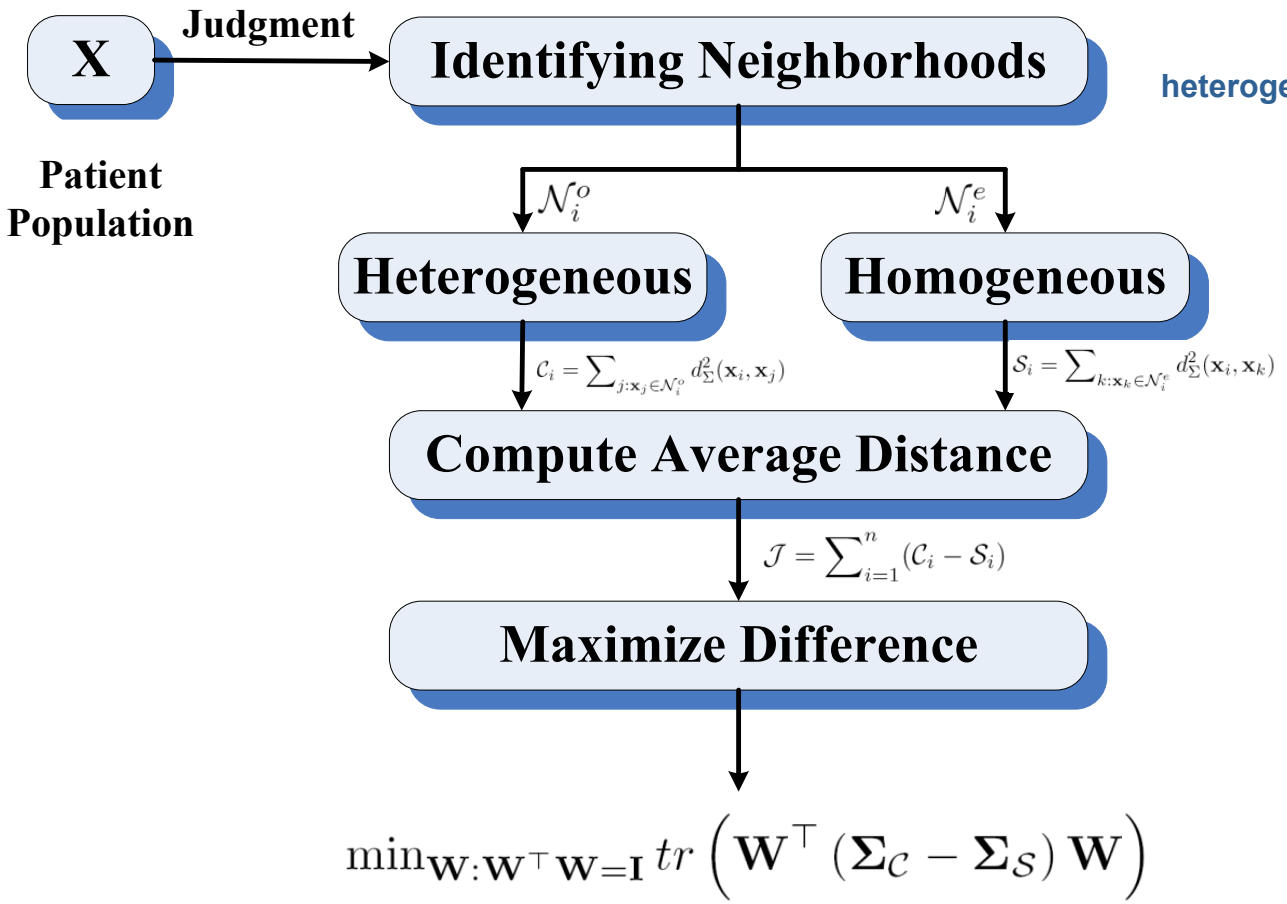


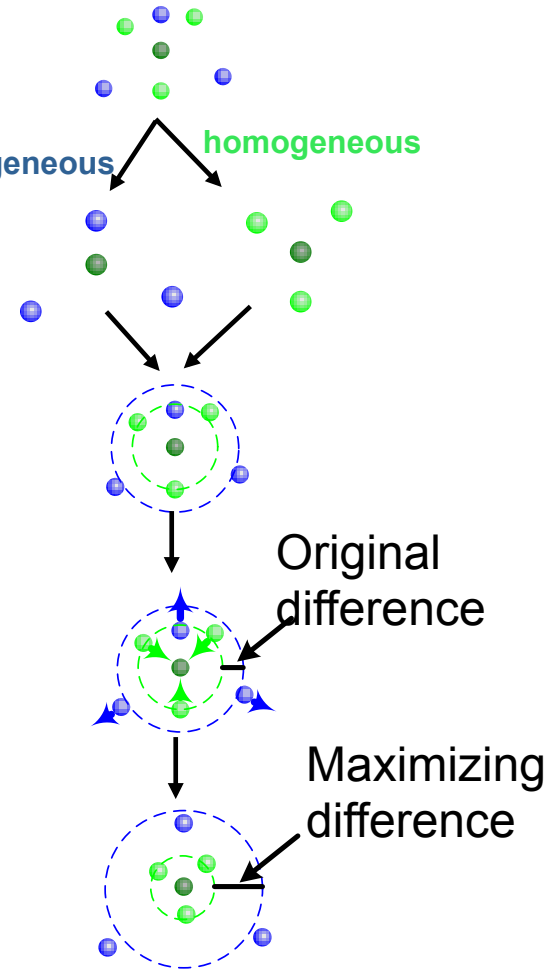
- Retrieve and rank similar patients for a query patient
- Explain why they are similar

Method: Locally Supervised Metric Learning (LSML)

- Goal: Learn a Mahalanobis distance $d_\Sigma(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\Sigma} (\mathbf{x}_i - \mathbf{x}_j)}$

$$\boldsymbol{\Sigma} = \mathbf{W}\mathbf{W}^\top$$





Fei Wang, Jimeng Sun, Tao Li, Nikos Anerousis: Two Heads Better Than One: Metric+Active Learning and its Applications for IT Service Classification. ICDM 2009: 1022-1027

Experiment Evaluation

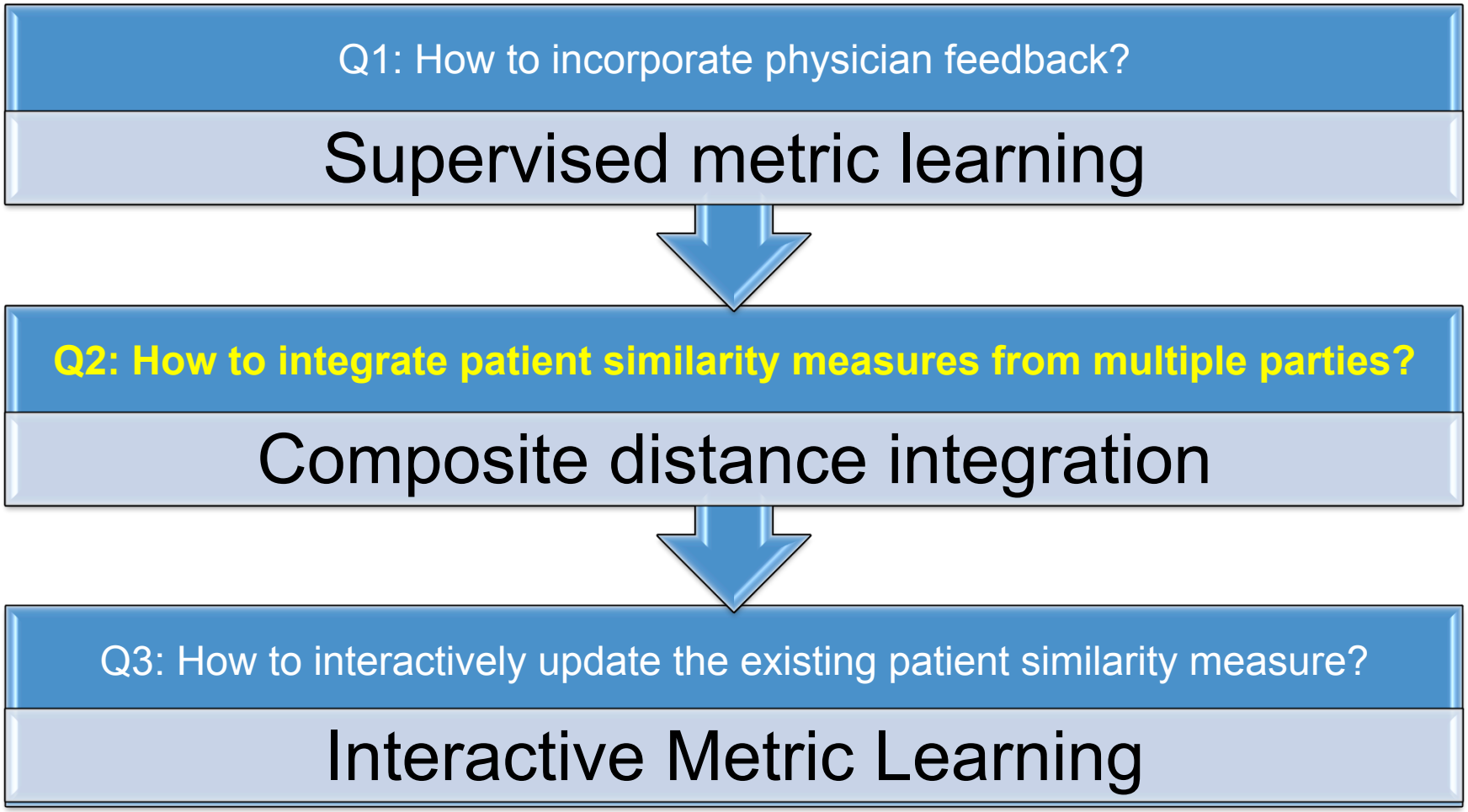
- Data:
  - 1500 patients downloaded from the MIMIC II database. Among the 1500 patients, 590 patients experienced at least one occurrence of Acute Hypotensive Episode (AHE), and 910 patients did not.
  - Physiological streams include mean ABP measure, systolic ABP, diastolic ABP, Sp02 and heart rate measurements. Every sensor is sampled at 1 minute intervals.
- Performance Metric: 1) classification accuracy, 2) Retrieval precision@10
- Baselines:
  - Challenge09: uses Euclidean distance of the variance of the mean ABP as suggested in [1];
  - PCA uses Euclidean distance over low-dimensional points after principal component analysis (PCA) (an unsupervised metri)
  - LDA uses Euclidean distance al points after linear discriminant analysis (LDA) (a global supervised metric learning algorithm);
- Results:

CLASSIFICATION AND RETRIEVAL ACCURACY

	Challenge09	PCA	LDA	**LSML**
Classification (accuracy)	0.4982	0.6325	0.7739	**0.8551**
Retrieval (precision@10)	0.5230	0.6902	0.7314	**0.7998**

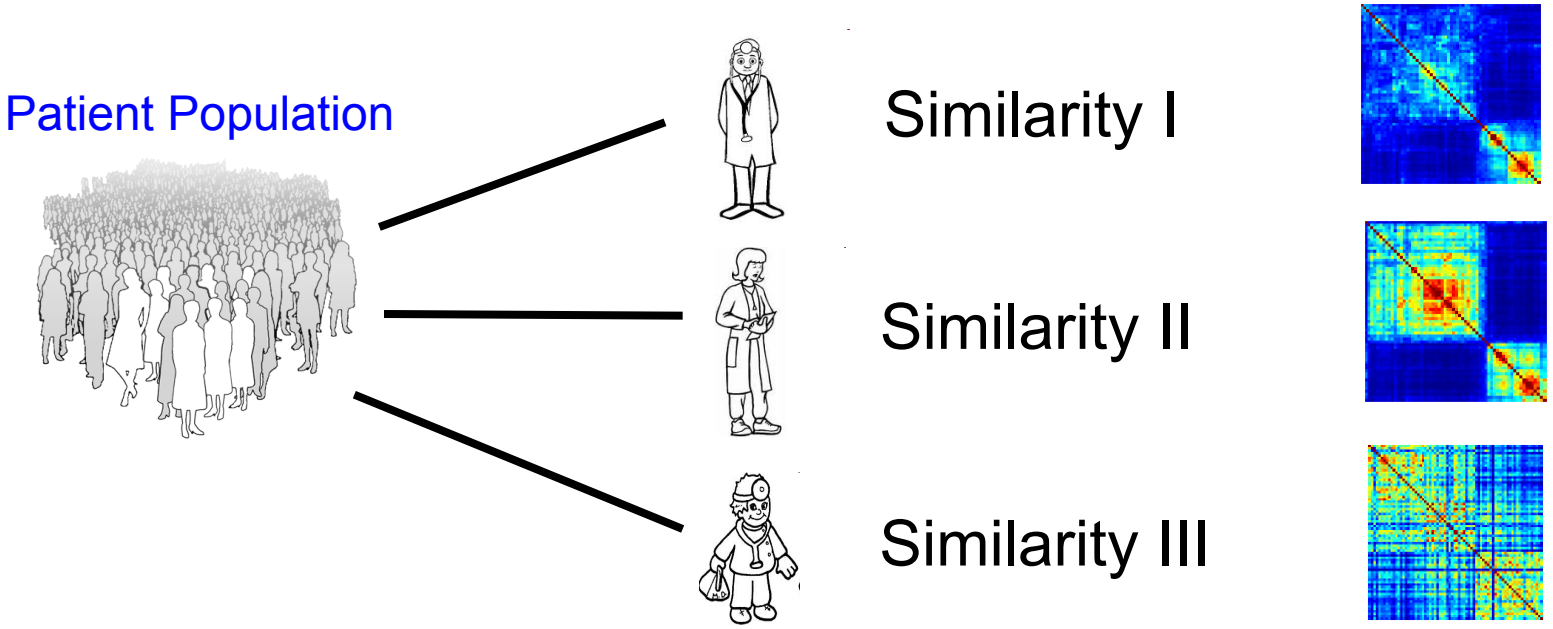
[1] X. Chen, D. Xu, G. Zhang, and R. Mukkamala. Forecasting acute hypotensive episodes in intensive care patients based on peripheral arterial blood pressure waveform. In Computers in Cardiology (CinC), 2009.

Patient Similarity Applications



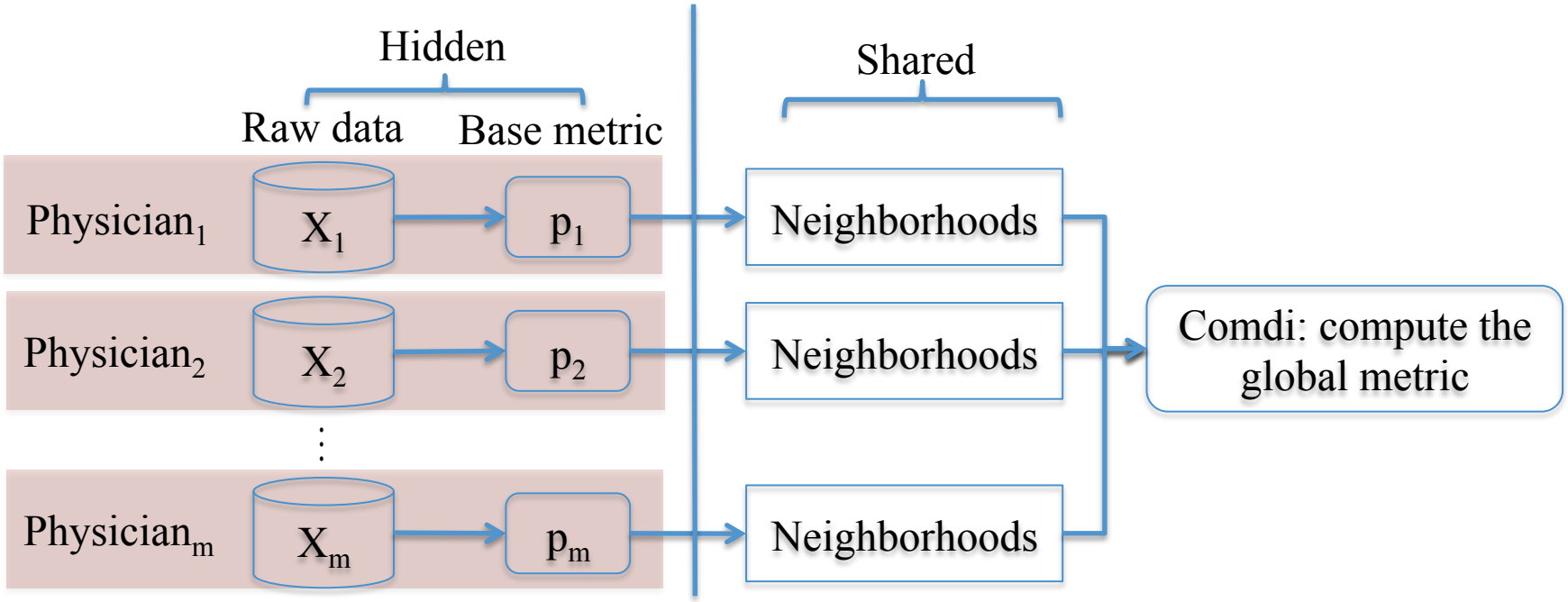
Q2: How to integrate patient similarity measures from multiple parties?

Different physicians have different similarity measures



- How to integrate these judgments from multiple experts to a consistent similarity measure?
- How to learn a meaningful distance metric considering the curse of dimensionality?

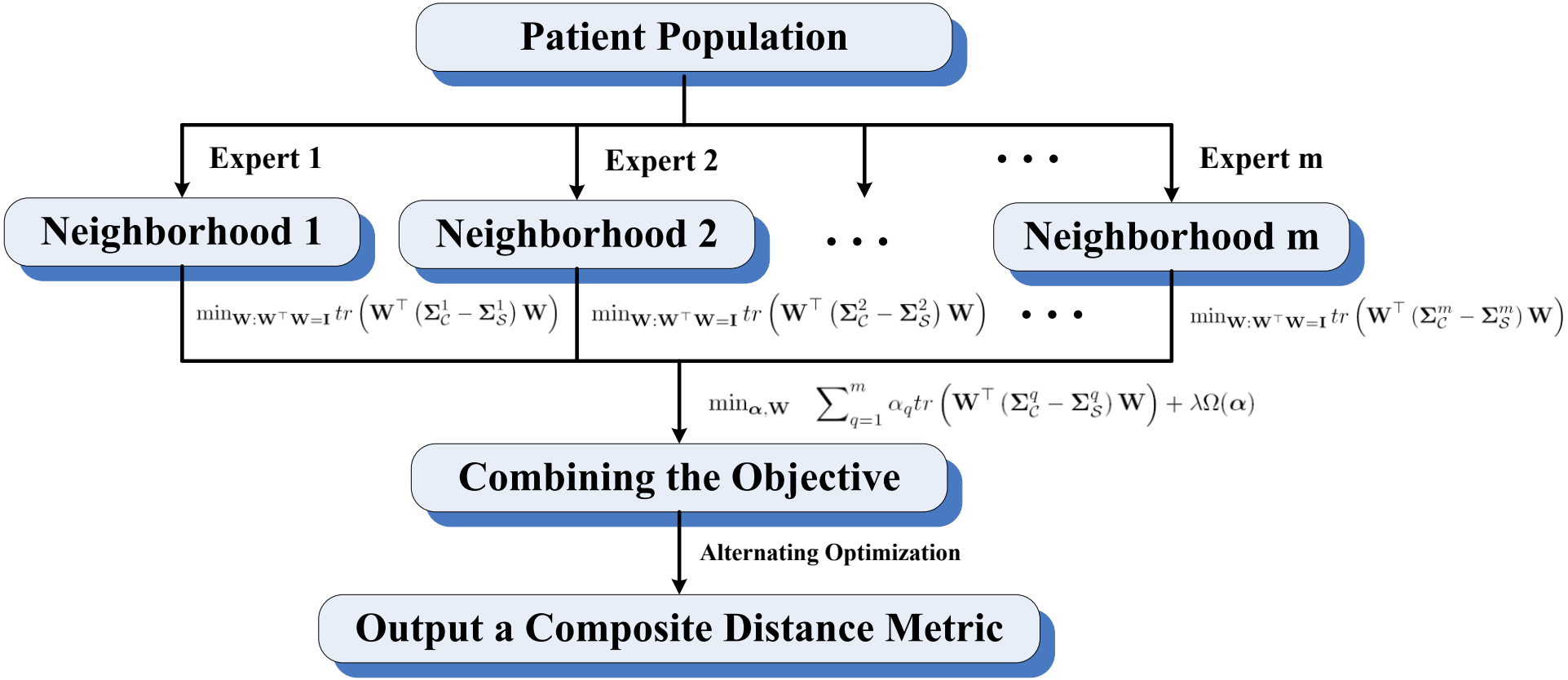
Comdi: Composite Distance Integration



- **Knowledge sharing:** integration of patient similarity from multiple physicians can compensate the experience of different physicians and achieve a better outcome.
- **Efficient collaboration:** Sharing the models across physicians can be an efficient way to collaborate across multiple hospitals and clinics.
- **Privacy preservation:** Due to the privacy rule regarding to PHI data, sharing individual patient data becomes prohibitively difficult. Therefore, model sharing provides an effective alternative.

Fei Wang, Jimeng Sun, Shahram Ebadollahi: Integrating Distance Metrics Learned from Multiple Experts and its Application in Inter-Patient Similarity Assessment. SDM 2011: 59-70 56

Composite Distance Integration: Formulation



- Individual distance metrics may be of arbitrary forms and scales

Composite Distance Integration: Alternating optimization method

$$\min_{\boldsymbol{\alpha},\mathbf{W}} \quad \sum_{q=1}^{m} \alpha_q tr\left(\mathbf{W}^\top \left(\boldsymbol{\Sigma}_{\mathcal{C}}^q - \boldsymbol{\Sigma}_{\mathcal{S}}^q\right)\mathbf{W}\right) + \lambda\Omega(\boldsymbol{\alpha})$$
$$s.t. \quad \boldsymbol{\alpha} \geqslant 0,\ \boldsymbol{\alpha}^\top \mathbf{e} = 1$$
$$\mathbf{W}^\top\mathbf{W} = \mathbf{I}$$

Optimization Problem

Fix $\mathbf{W}$ Solve $\boldsymbol{\alpha}$

$$\min_{\mathbf{W}} \quad \sum_{q=1}^{m} \alpha_q^{(t-1)} tr\left(\mathbf{W}^\top \left(\boldsymbol{\Sigma}_{\mathcal{C}}^q - \boldsymbol{\Sigma}_{\mathcal{S}}^q\right)\mathbf{W}\right) + \lambda\Omega(\boldsymbol{\alpha})$$
$$s.t. \quad \mathbf{W}^\top\mathbf{W} = \mathbf{I}$$

Eigenvalue Decomposition

Fix $\boldsymbol{\alpha}$ Solve $\mathbf{W}$

$$\min_{\boldsymbol{\alpha}} \quad \sum_{q=1}^{m} \alpha_q tr\left(\left(\mathbf{W}^{(t)}\right)^\top \left(\boldsymbol{\Sigma}_{\mathcal{C}}^q - \boldsymbol{\Sigma}_{\mathcal{S}}^q\right)\mathbf{W}^{(t)}\right) + \lambda\Omega(\boldsymbol{\alpha})$$
$$s.t. \quad \boldsymbol{\alpha} \geqslant 0,\ \boldsymbol{\alpha}^\top \mathbf{e} = 1$$

Euclidean Projection

Converge?

N

Y

Output the Metric

Experiment Evaluation

Data

- Data source: 135K patients over one year, consisting of diagnosis. procedure and etc.
- We construct 247 cohorts, one for each physician
- Labels (physician feedback): HCC019 - Diabetes with No or Unspecified Complications
- Features: All the remaining HCC diagnosis information.

Tasks

- Use 30 cohorts (select randomly) to train the models
- Performance metrics:

	Actual Value	
Predicted Value	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

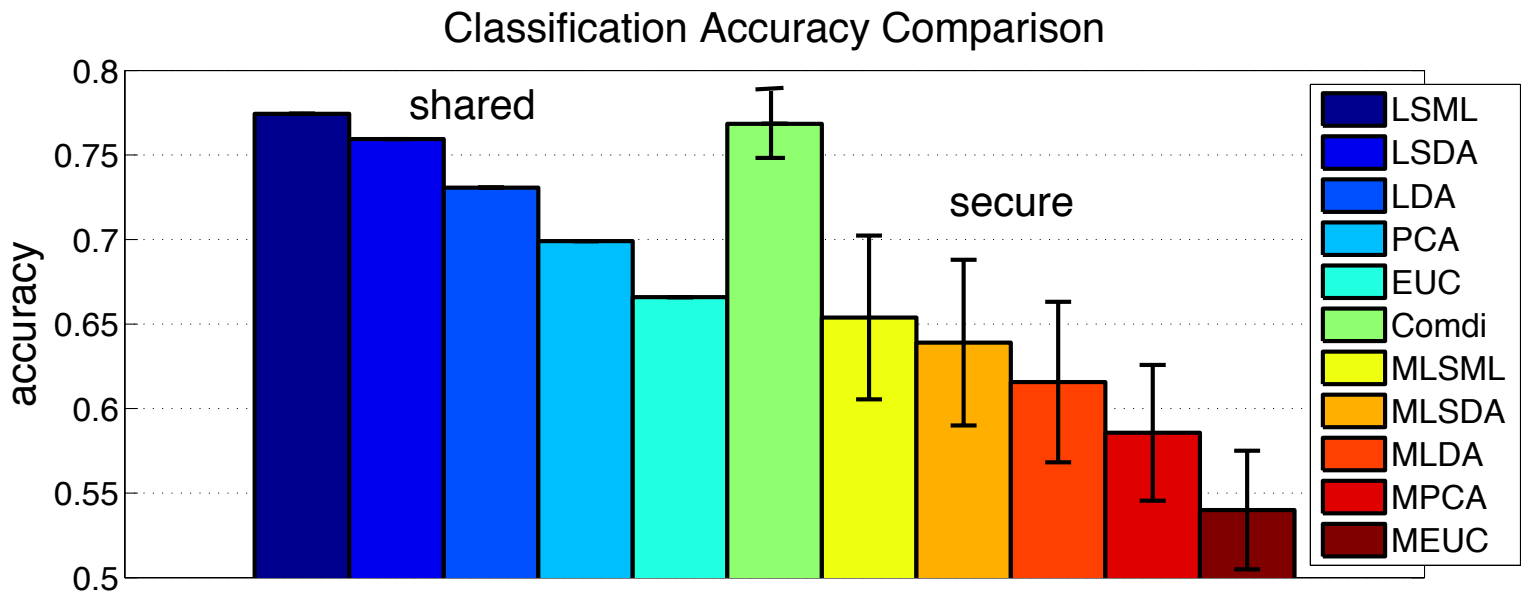
$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F1 Score} = \frac{2TP}{2TP+FN+FP}$$

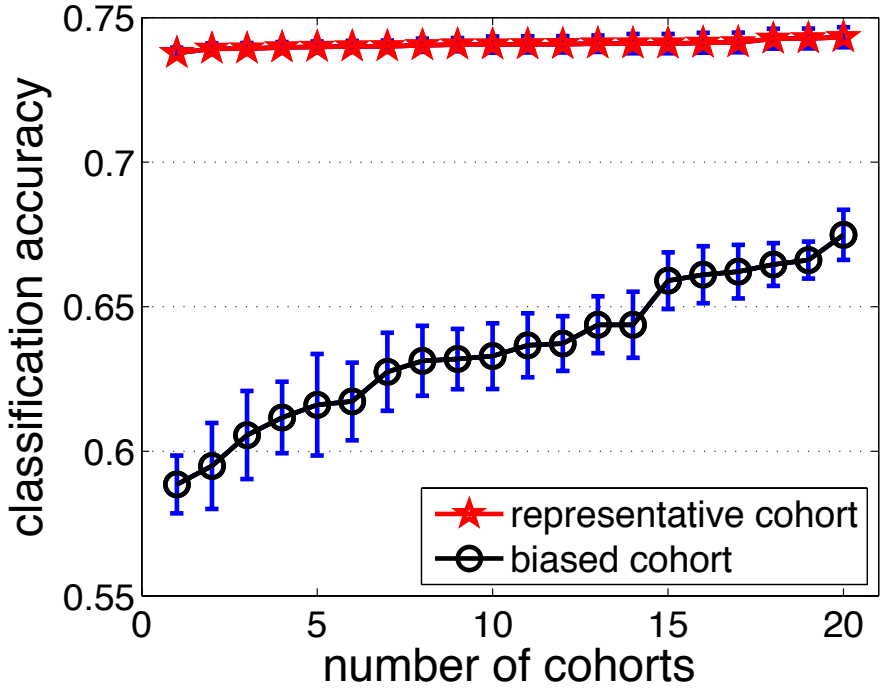
Results on Accuracy

- Share versions: learning on all 30 cohorts
- Secure versions: learning on 1 cohort
- Our method Comdi combines all 30 individual models
- Observations:
  - Shared version perform better than secure versions, which indicates sharing is better
  - Comdi is comparable to LSML, which is the best among sharing versions, which confirms its effectiveness in terms of combining multiple metrics

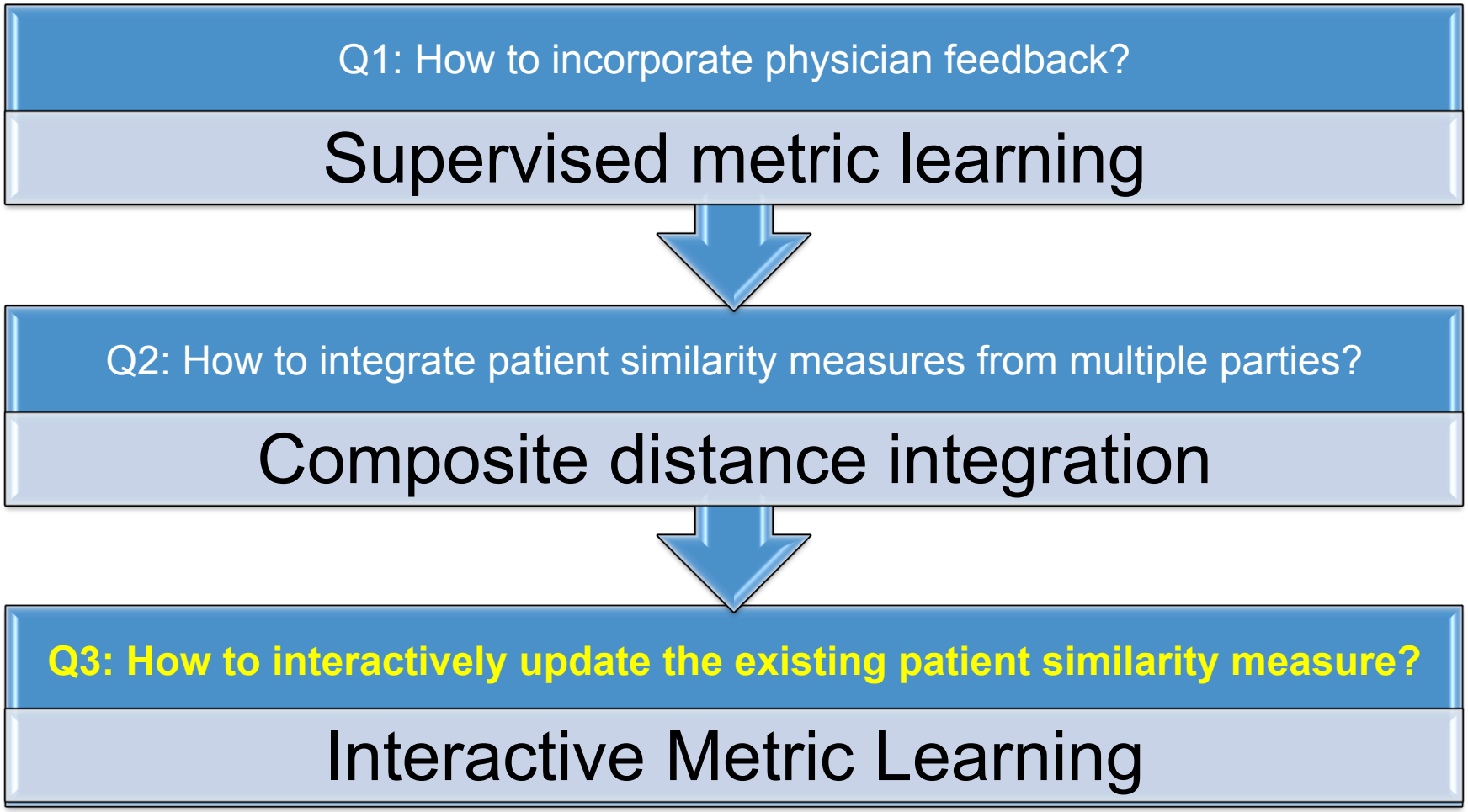


Positive effects on all cohorts

- **Representative cohort** is a good approximation of the entire patient distribution, which often leads to good base metric.
- **Biased cohort** is a bad approximation of the entire patient distribution, which often leads to bad base metric.
- The accuracy increases significantly for biased cohorts, and also still improves the accuracy for representative cohorts.



Patient Similarity Applications



iMet: Incremental Metric Learning Overview

Offline method

- Model building: build a patient similarity model from the historical data
- Model scoring: use the learned model to retrieve/score similar patients
- Disadvantage: physician feedback cannot be incorporated quickly.

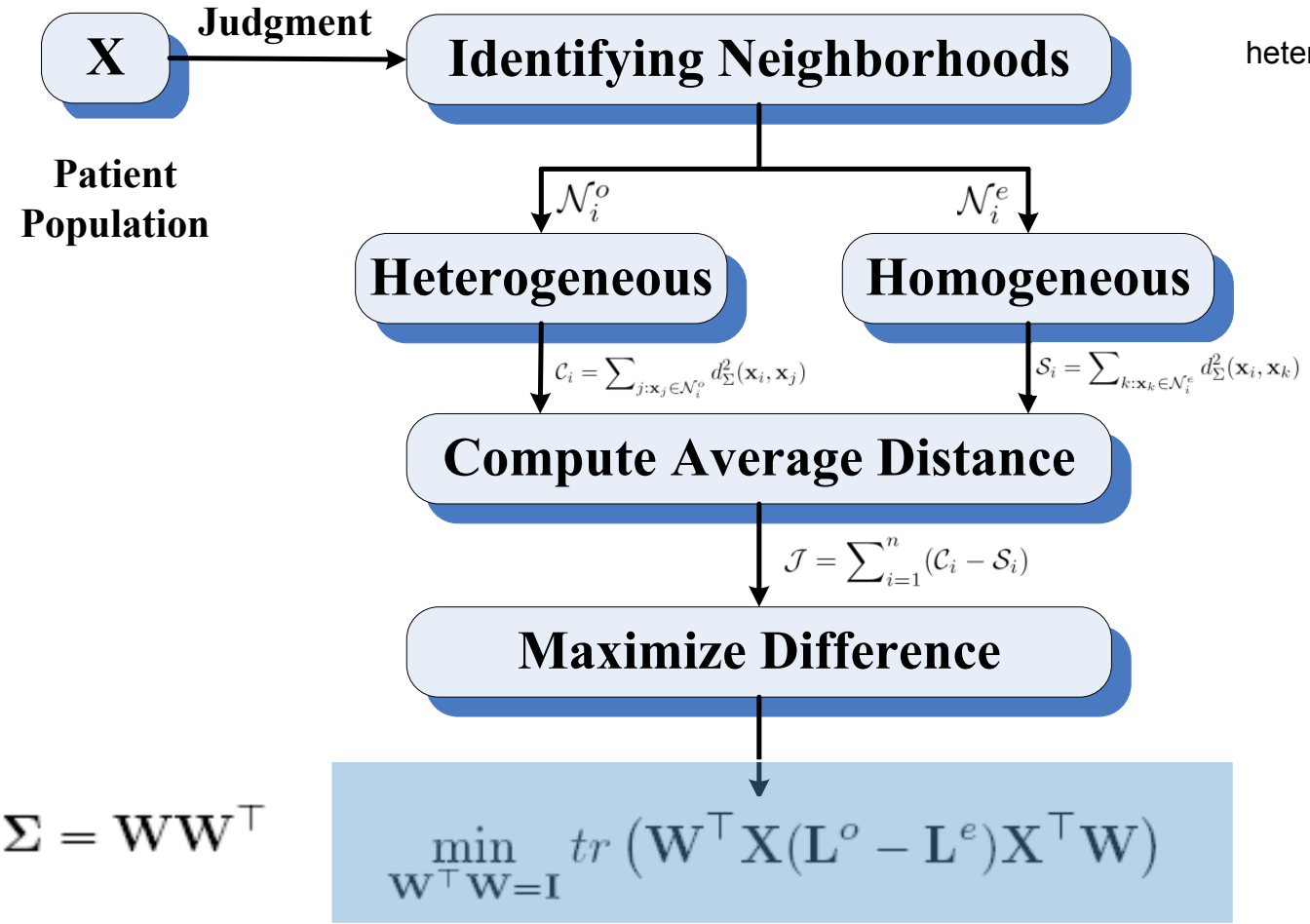
Interactive method

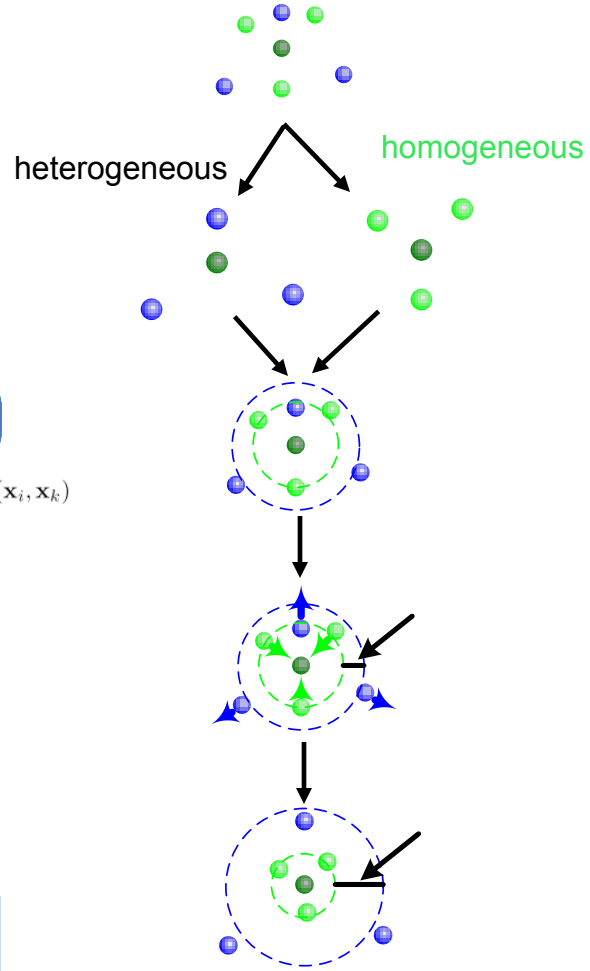
- Model update: receive updates in real-time to adjust the model parameters
- Update types: physician feedback modeled as label change

- How to adjust the learned patient similarity by incorporating physician's feedback in real time?
- How to adjust the learned patient similarity by incorporating patients' feature change efficiently?

Locally Supervised Metric Learning Revisit

Goal: $d_{\Sigma}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^{\top} \boldsymbol{\Sigma} (\mathbf{x}_i - \mathbf{x}_j)}$





iMet: Incremental Metric Learning Approach - Intuition

The optimal W can be achieved by doing eigenvalue decomposition on

$$\mathbf{X}(\mathbf{L}^o - \mathbf{L}^e)\mathbf{X}^\top$$

Any feedback can be viewed as a increment on the matrix

$$\mathbf{L} = \mathbf{L}^o - \mathbf{L}^e$$

How to efficiently update the eigensystem of $\mathbf{X}(\mathbf{L}^o - \mathbf{L}^e)\mathbf{X}^\top$ based on the increment on L?

Fei Wang, Jimeng Sun, Jianying Hu, Shahram Ebadollahi: iMet: Interactive Metric Learning in Healthcare Applications. SDM 2011: 944-955

iMet: Incremental Metric Learning Approach

Matrix → **Eigensystem** → **Perturbation**

$$\mathbf{X}\mathbf{L}\mathbf{X}^\top \quad \mathbf{X}(\mathbf{L}+\Delta\mathbf{L})\mathbf{X}^\top \;\Rightarrow\; (\lambda_i, \mathbf{w}_i) \quad (\widetilde{\lambda}_i, \widetilde{\mathbf{w}}_i) \;\Rightarrow\; \widetilde{\lambda}_i = \lambda_i + \Delta\lambda_i, \quad \widetilde{\mathbf{w}}_i = \mathbf{w}_i + \Delta\mathbf{w}_i$$

$$\mathbf{X}(\mathbf{L}+\Delta\mathbf{L})\mathbf{X}^\top(\mathbf{w}_i+\Delta\mathbf{w}_i) = (\lambda_i+\Delta\lambda_i)(\mathbf{w}_i+\Delta\mathbf{w}_i)$$

First-Order Perturbation

$$\mathbf{X}\mathbf{L}\mathbf{X}^\top\Delta\mathbf{w}_i + \mathbf{X}\Delta\mathbf{L}\mathbf{X}^\top\mathbf{w}_i = \lambda_i\Delta\mathbf{w}_i + \Delta\lambda_i\mathbf{w}_i$$

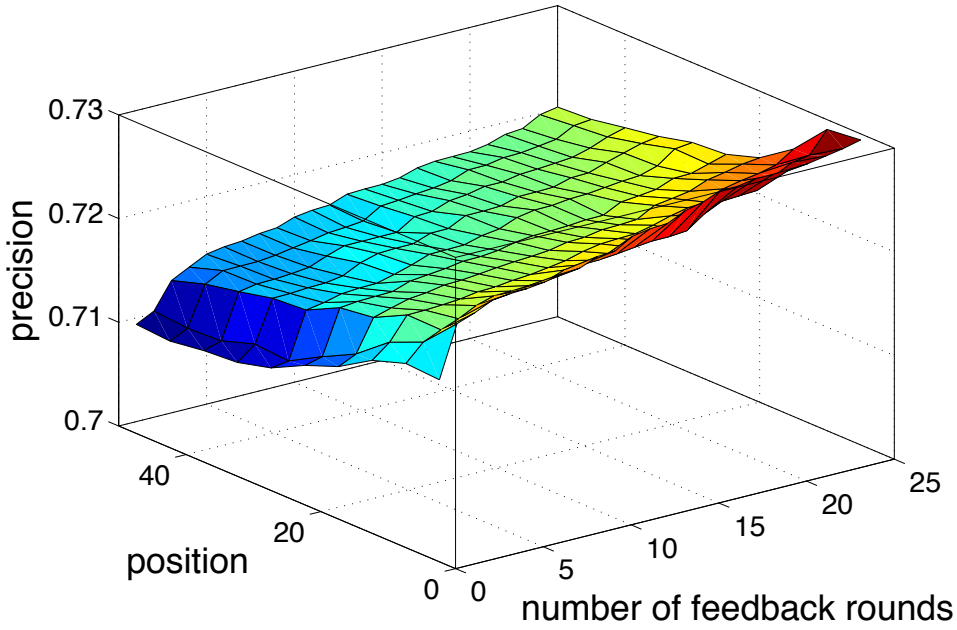
Solution

$$\Delta\lambda_i = \mathbf{w}_i^\top\mathbf{X}\Delta\mathbf{L}\mathbf{X}^\top\mathbf{w}_i$$

$$\Delta\mathbf{w}_i = -\frac{1}{2}\mathbf{w}_i + \sum_{j\neq i}\frac{\mathbf{w}_j^\top\mathbf{X}\Delta\mathbf{L}\mathbf{X}^\top\mathbf{w}_i}{\lambda_i-\lambda_j}\mathbf{w}_j$$

Experiment evaluation

- **Initial Metric:** The patient population was clustered into 10 clusters using Kmeans with the remaining 194 HCC features. An initial distance metric was then learned using LSML.
- **Feedback**: One of the key HCC is hold off as the simulated feedback. For each round of simulated feedback, an index patient was randomly selected and 20 similar patients were selected for feedback
- **Performance metric**: precision@position measure



Part I Summary

- Definition of distance metric learning
- Taxonomy of distance metric learning
- Related areas
  - Dimensionality reduction
  - Kernel learning
- Patient similarity application
  - Locally supervised metric learning
  - Multiple metric integration
  - Interactive metric learning