Multi-Task Learning: Theory, Algorithms, and Applications

Jiayu Zhou[1,2], Jianhui Chen[3], Jieping Ye[1,2]

[1] Computer Science and Engineering, Arizona State University, AZ

[2] Center for Evolutionary Medicine Informatics, Biodesign Institute, Arizona State University, AZ

[3] GE Global Research, NY

SDM 2012 Tutorial

Tutorial Goals

- Understand the basic concepts in multi-task learning
- Understand different approaches to model task relatedness
- Get familiar with different types of multi-task learning techniques
- Introduce the multi-task learning package: MALSAR

Tutorial Road Map

- Part I: Multi-task Learning (MTL) background and motivations
- Part II: MTL formulations
- Part III: Case study of real-world applications
  - Disease Progression
  - Dealing with Missing Value in Multiple Sources
  - Drosophila Image Analysis
- Part IV: MALSAR: Multi-task Learning via Structural Regularization Package
- Future directions

Multiple Tasks

- Examination Scores Prediction[1] (Argyriou et. al.'08)

School 1 - Alverno High School

Student id	Birth year	Previous score	…	School ranking	…
72981	1985	95	…	83%	…

student-dependent school-dependent

→

Exam score
?

…

School 138 - Jefferson Intermediate School

Student id	Birth year	Previous score	…	School ranking	…
31256	1986	87	…	72%	…

student-dependent school-dependent

→

Exam score
?

School 139 - Rosemead High School

Student id	Birth year	Previous score	…	School ranking	…
12381	1986	83	…	77%	…

student-dependent school-dependent

→

Exam score
?

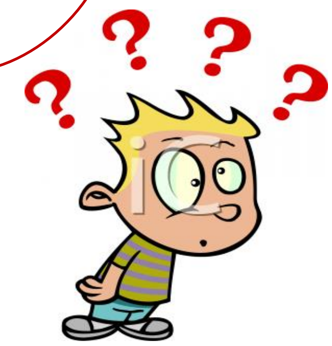
[1]The Inner London Education Authority (ILEA)

Learning Multiple Tasks

- Learning from the pool of all tasks

Student id	Birth year	Previous score	School ranking	…	Exam Score
72981	1985	95	83%	…	?
31256	1986	87	72%	…	?
12381	1987	83	77%	…	?
…	…	…	…	…	…
21901	1986	87	72%	…	?

Students with **same Features** but **different Exam Scores**



School A

School B

Learning Multiple Tasks

- Learning each task independently

School 1 - Alverno High School

Student id	Birth year	Previous score	School ranking	…
72981	1985	95	83%	…

→

Exam Score
?

⇨ task 1st

Excellent

…

School 138 - Jefferson Intermediate School

Student id	Birth year	Previous score	School ranking	…
31256	1986	87	72%	…

→

Exam Score
?

⇦ task 138th

Excellent

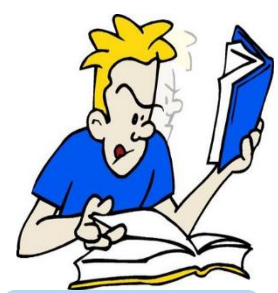
School 139 - Rosemead High School

Student id	Birth year	Previous score	School ranking	…
12381	1986	83	77%	…

→

Exam Score
?

⇨ task 139th



Excellent

Learning Multiple Tasks

- Leaning multiple tasks simultaneously

School 1 - Alverno High School

Student id	Birth year	Previous score	School ranking	…	→	Exam Score
72981	1985	95	83%	…		?

⇨ task 1st

…

School 138 - Jefferson Intermediate School

Student id	Birth year	Previous score	School ranking	…	→	Exam Score
31256	1986	87	72%	…		?

⇨ task 138th

School 139 - Rosemead High School

Student id	Birth year	Previous score	School ranking	…	→	Exam Score
12381	1986	83	77%	…		?

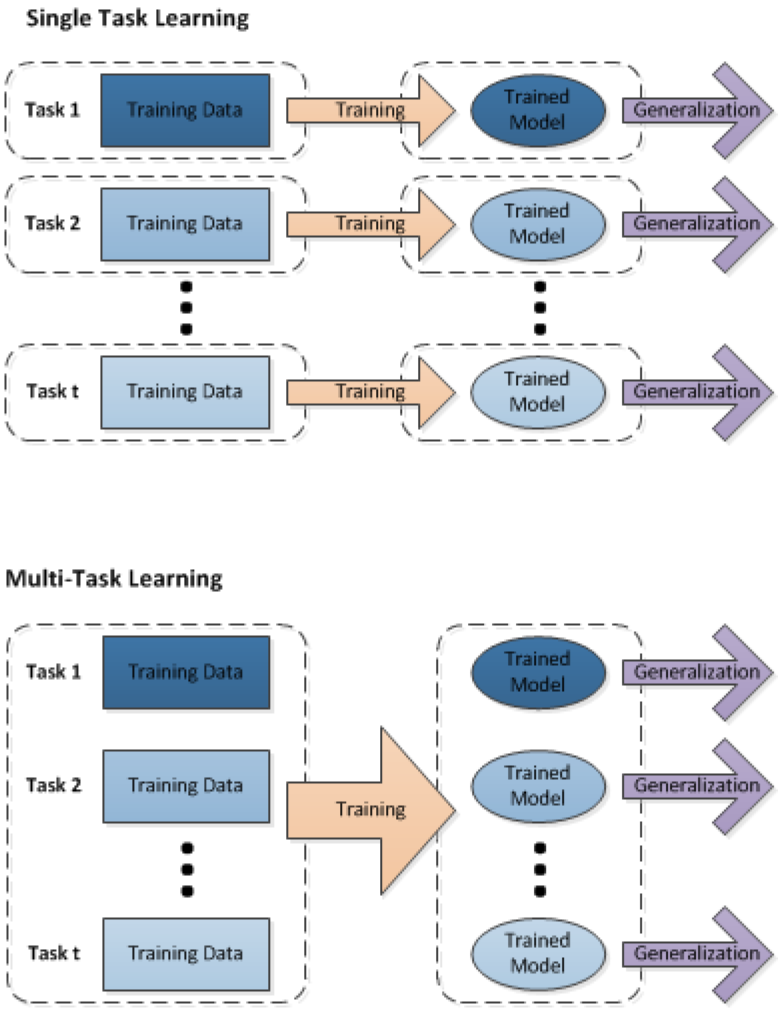
⇨ task 139th

Learn tasks simultaneously
Model the tasks relationship

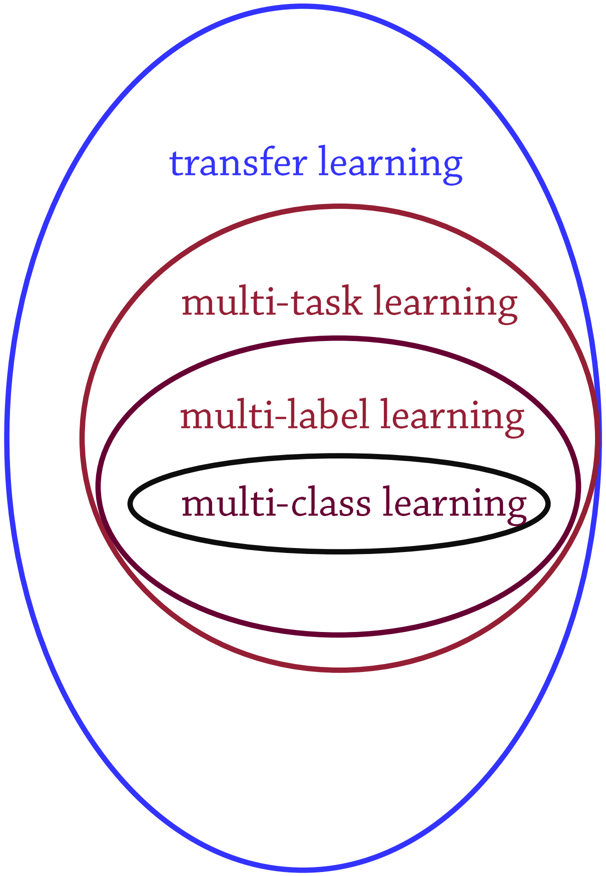
⇨ ……

Multi-Task Learning

- Multi-task Learning is different from single task learning in the training (induction) process.
- Inductions of multiple tasks are performed simultaneously to capture intrinsic relatedness.



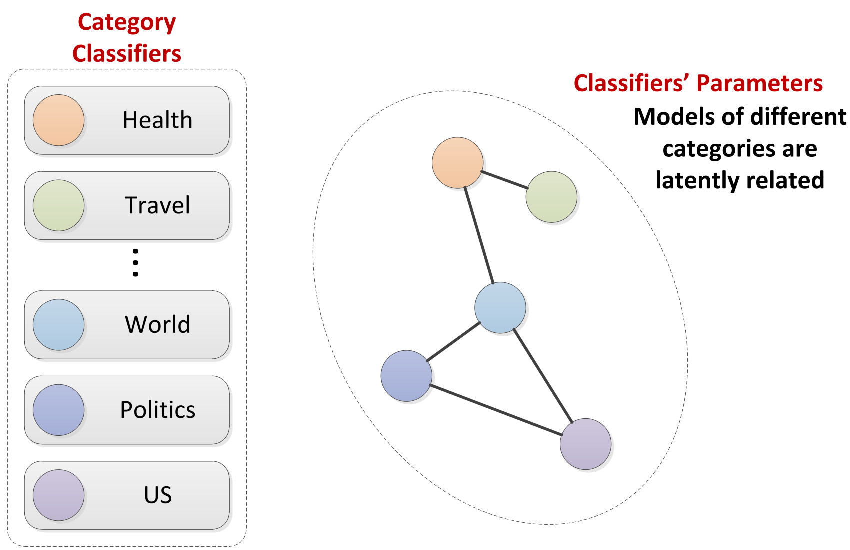
Learning Methods



- Transfer Learning
  - Define source & target domains
  - Learn on the source domain
  - Generalize on the target domain
- Multi-task Learning
  - Model the task relatedness
  - Learn all tasks simultaneously
  - Tasks may have different data/features
- Multi-label Learning
  - Model the label relatedness
  - Learn all labels simultaneously
  - Labels share the same data/features
- Multi-class Learning
  - Learn the classes independently
  - All classes are exclusive

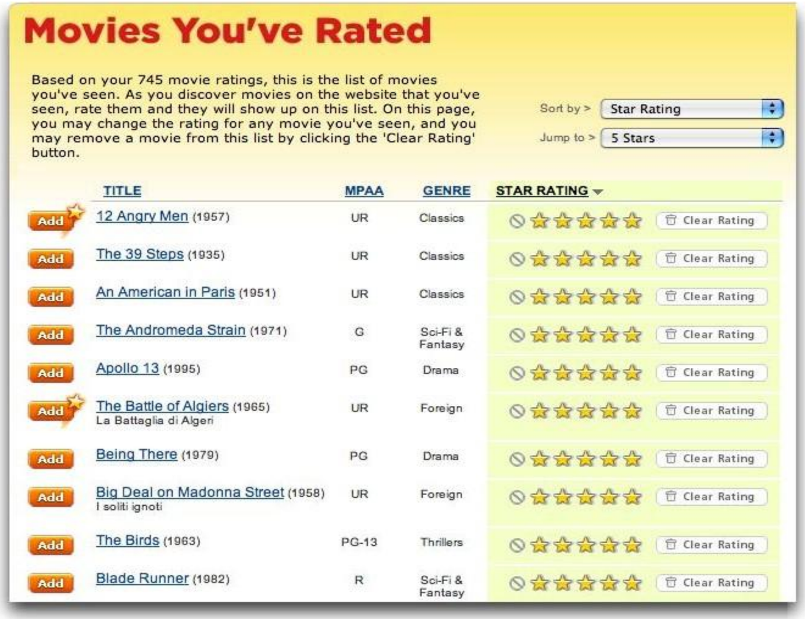
Web Pages Categorization

- Classify documents into categories
- The classification of each category is a task
- The tasks of predicting different categories may be latently related [Chen et.al. ICML 09]



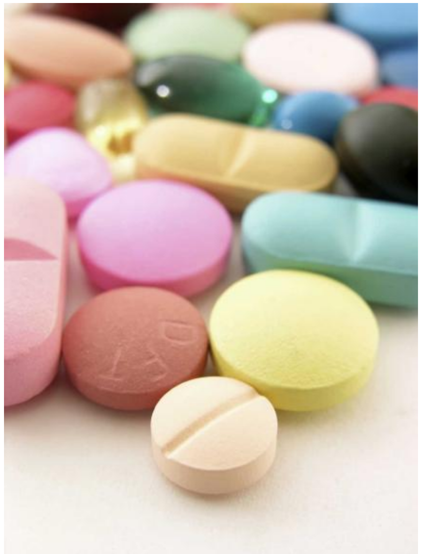
Collaborative Ordinal Regression

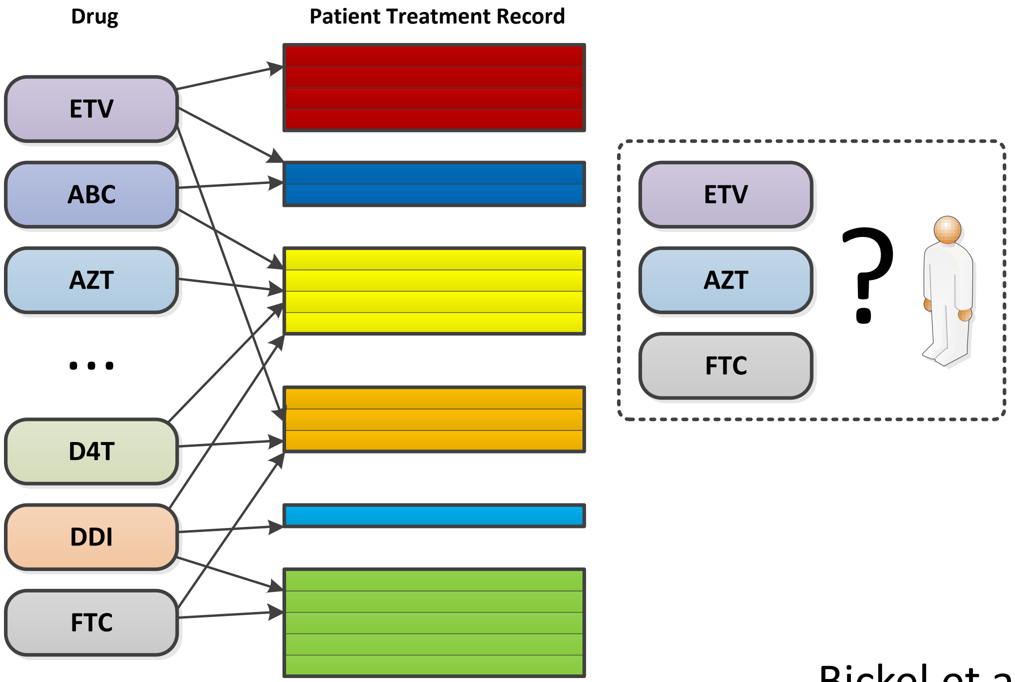
- The preference prediction of each user can be modeled using ordinal regression
- Some users have similar tastes and their predictions may also have similarities
- Simultaneously perform multiple prediction to use such similarity information [Yu et. al. NIPS 06]



MTL for HIV Therapy Screening

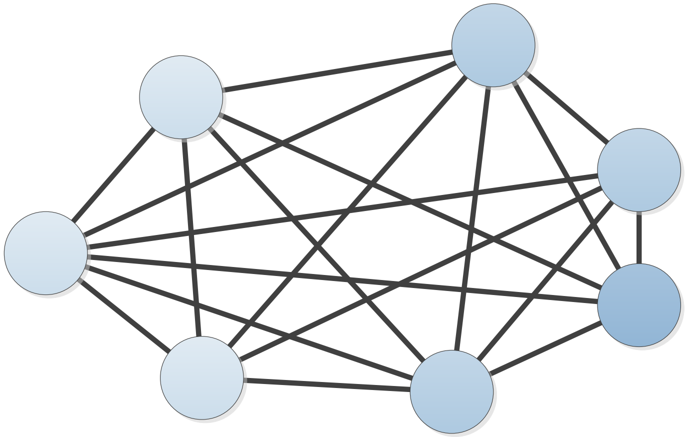
- Hundreds of possible combinations of drugs, some of which use similar biochemical mechanisms
- The sample available for each combination is limited.
- For a patient, the prediction of using one combination is a task
- Use the similarity information by simultaneously inference multiple tasks



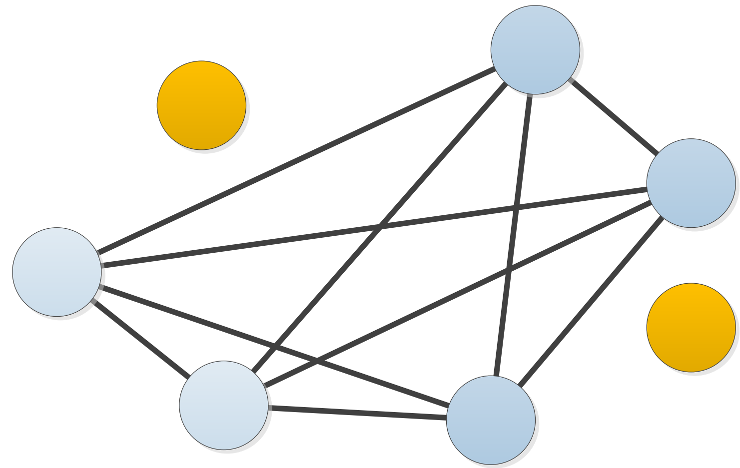


Bickel et al. ICML 08

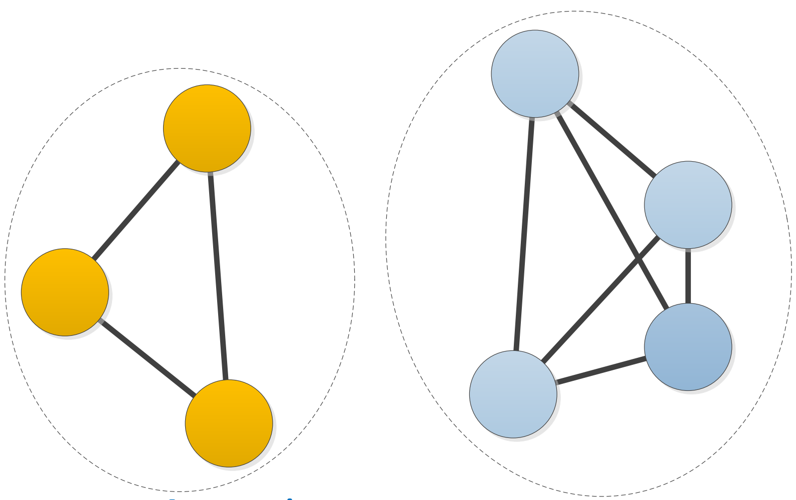
How to capture shared structures?



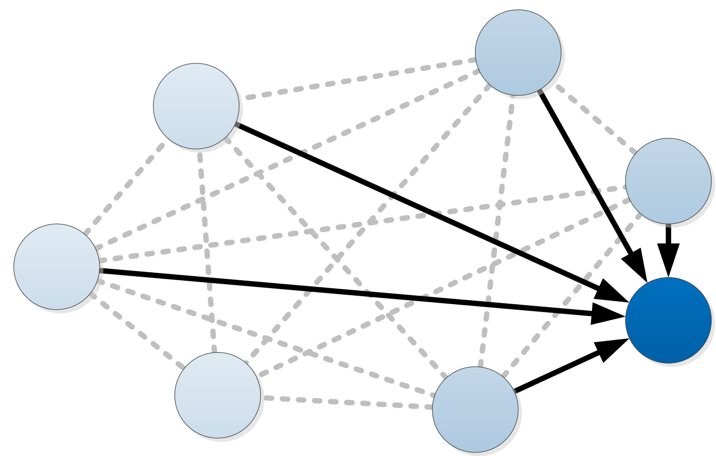
Assumption:
All tasks are related



Assumption:
There are outlier tasks

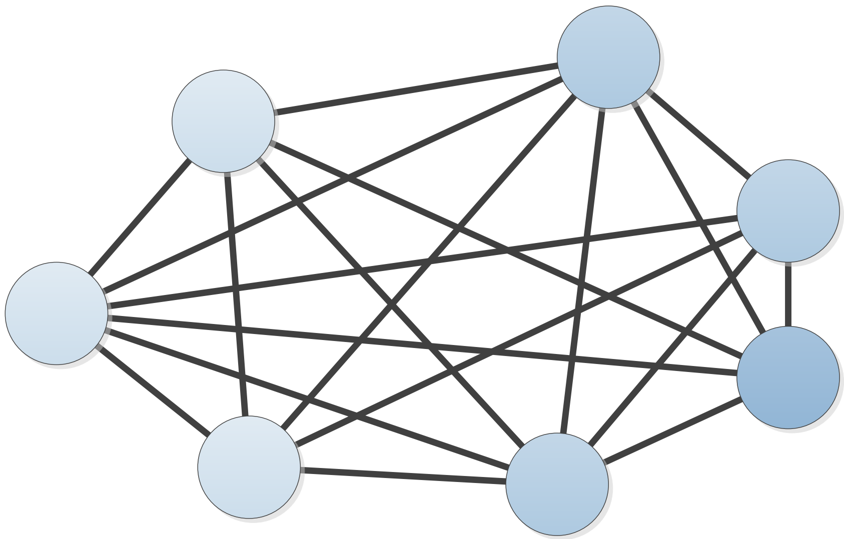


Assumption:
Tasks have group structures



Assumption:
The relationship is not symmetric

How to capture shared structures?

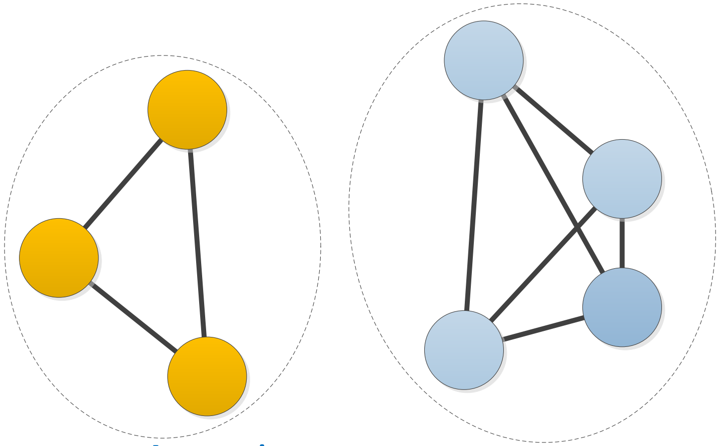


Assumption:
All tasks are related

Methods

- Mean-regularized MTL
- Joint feature learning
- Low rank regularized MTL
- alternating structural optimization (ASO)
- Shared Parameter Gaussian Process

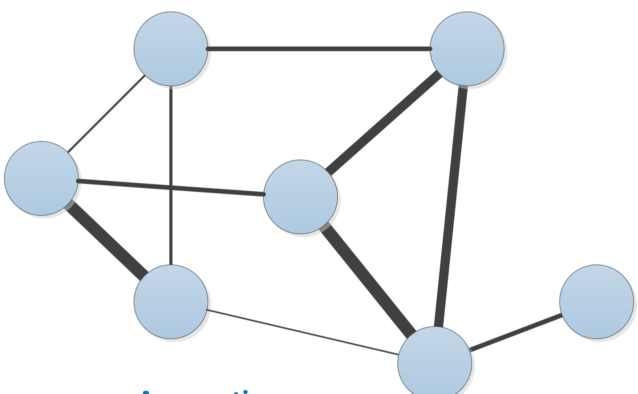
How to capture shared structures?



Assumption:
Tasks have group structures

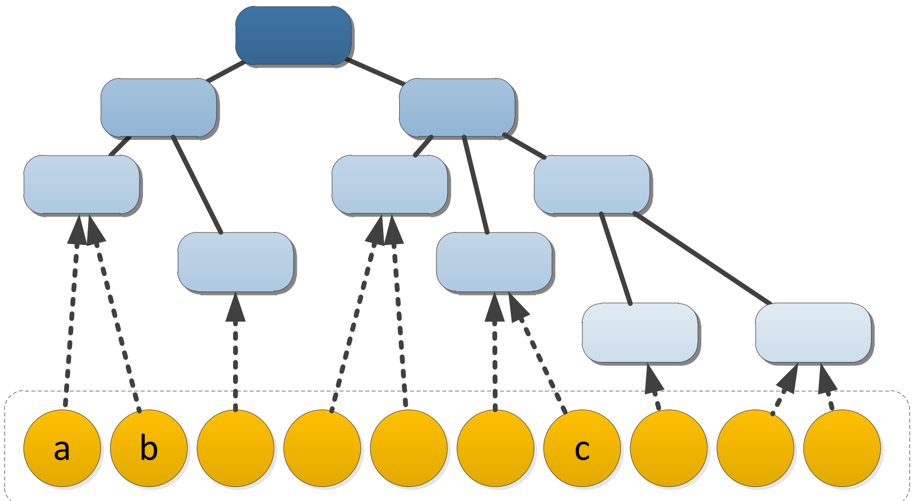
Methods

- Clustered MTL
- Tree MTL
- Network MTL



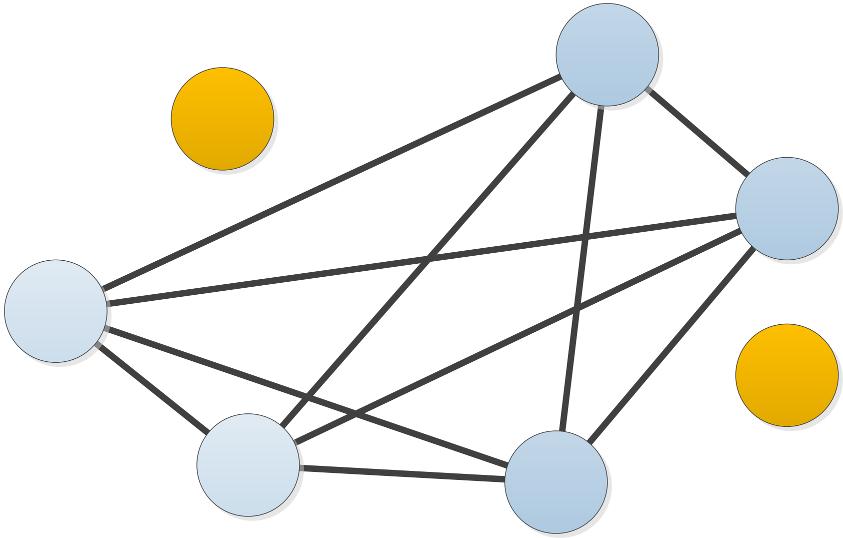
Assumption:
Tasks have graph/network structures

Assumption:
Tasks have tree structures



Models

How to capture shared structures?

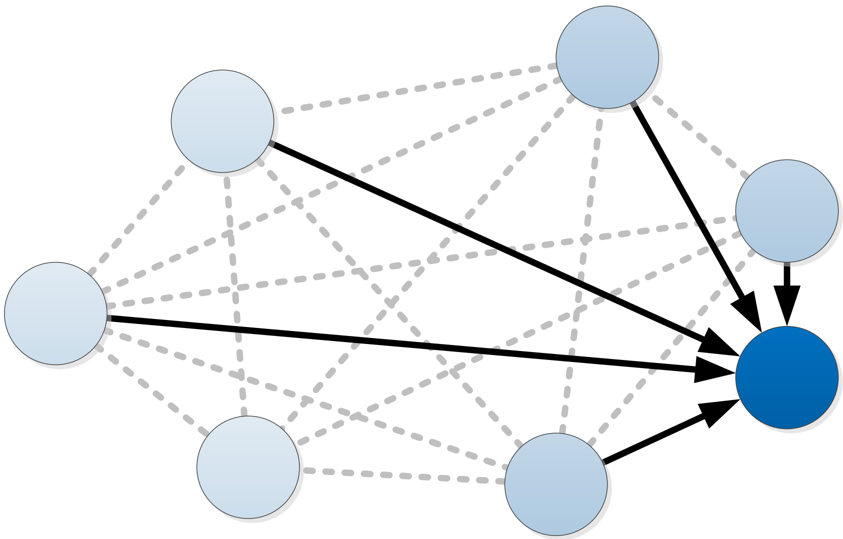


Assumption:
There are outlier tasks

Methods

- Robust MTL

How to capture shared structures?

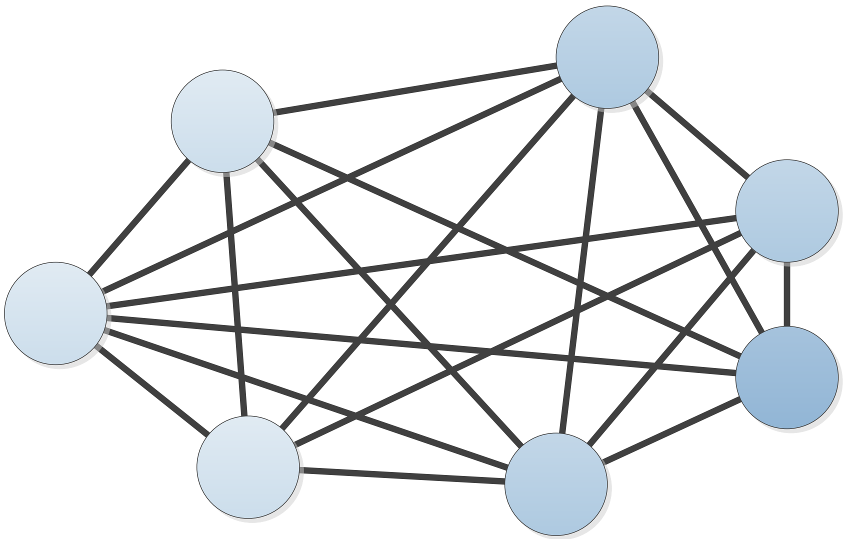


Assumption:
The relationship is not symmetric

Methods

- Asymmetric MTL

All tasks are related

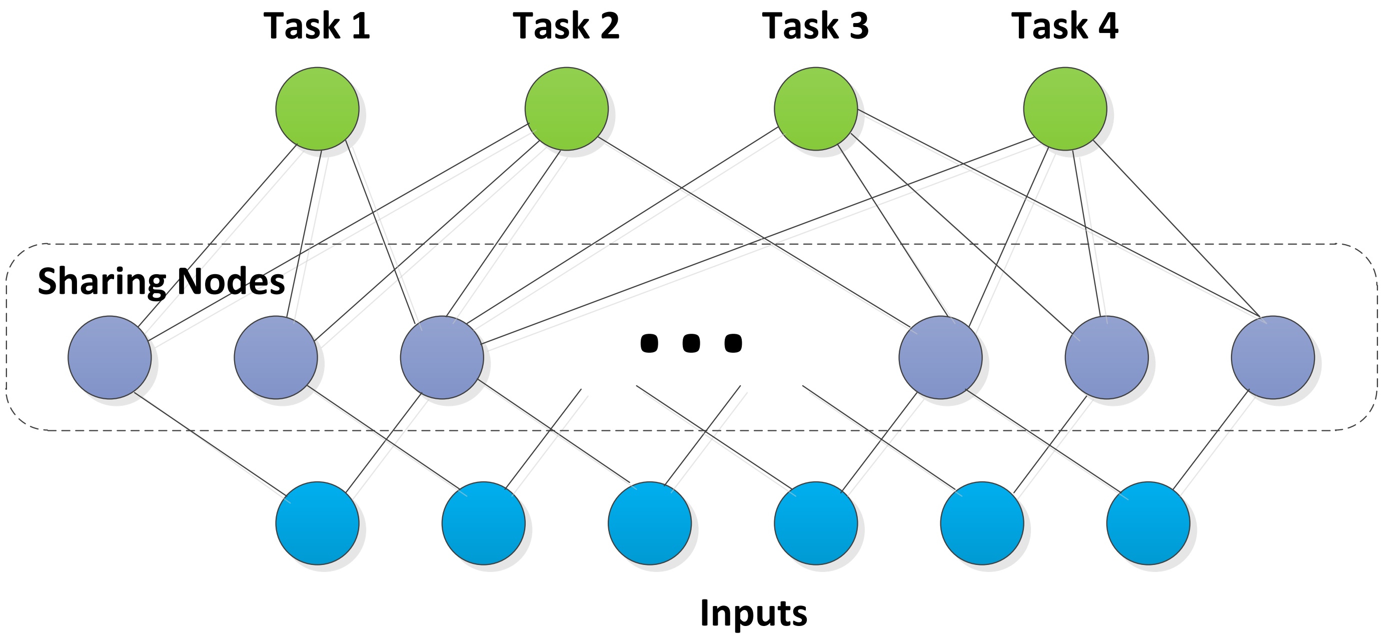


Assumption:
All tasks are related

- Shared Hidden Node in Neural Network
- Shared Parameter Gaussian Process
- Regularization-based MTL
  - Mean-regularized MTL
  - Joint feature learning
  - Low rank regularized MTL
  - alternating structural optimization

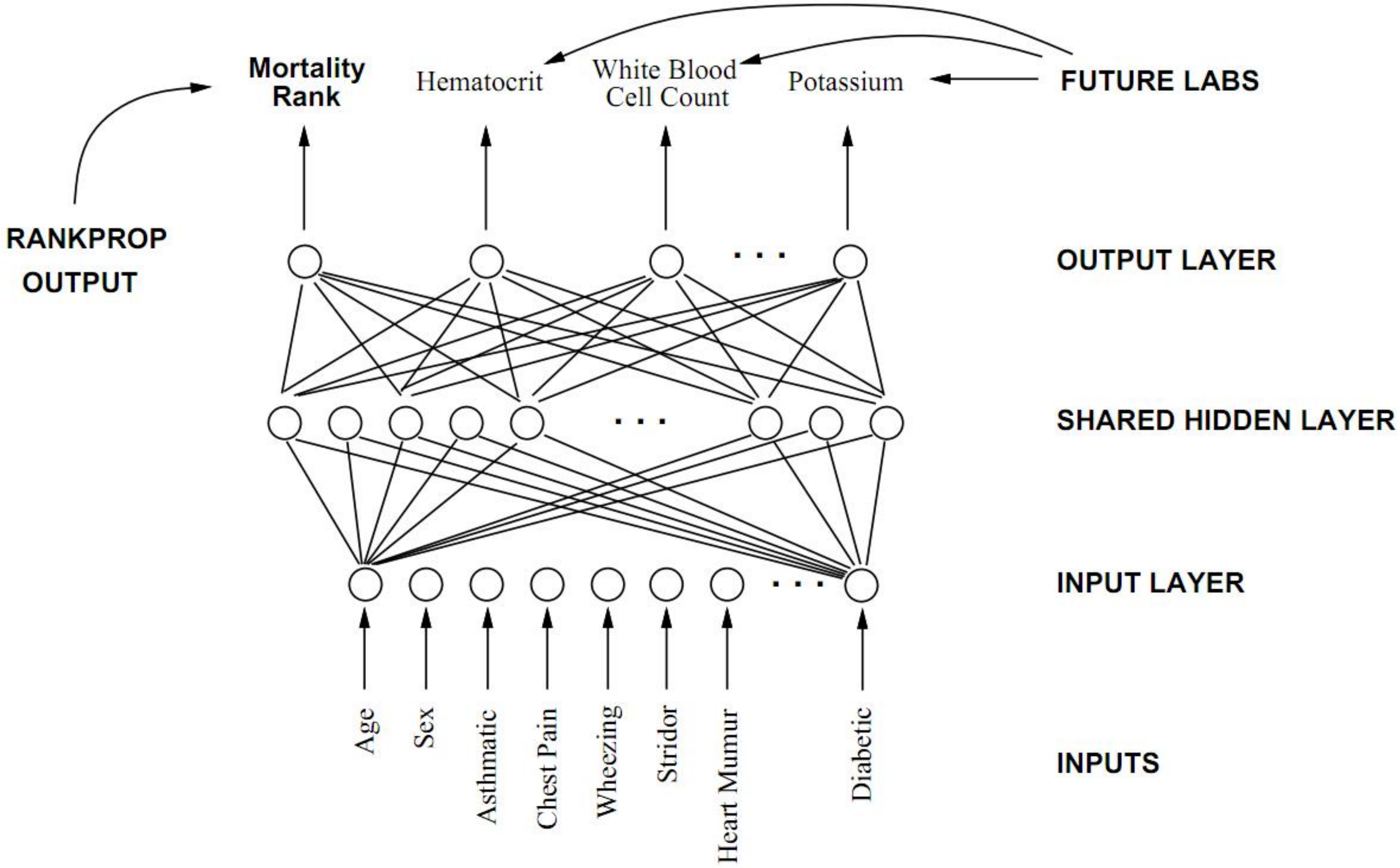
Sharing Hidden Nodes in Neural Network

- Neural network has been well studied for learning multiple related tasks for improved generalization performance.
- A set of hidden units are shared among multiple tasks for improved generalization (Caruana ML 97).



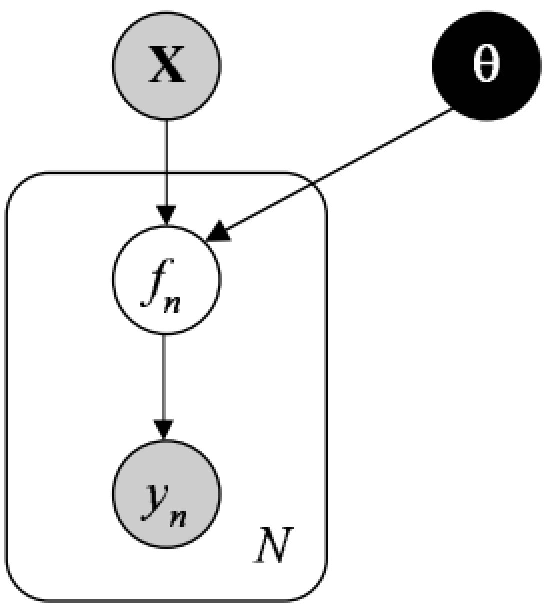
Mortality Rank

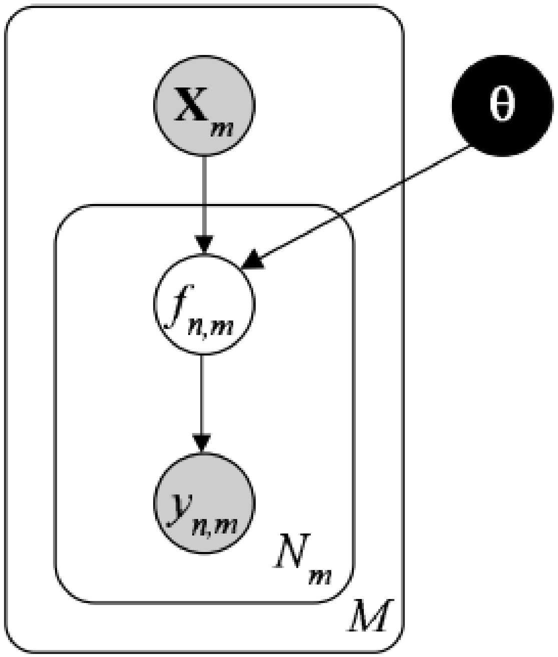
- Future lab results are used as extra outputs to bias learning for the main risk prediction task



Shared Parameter Gaussian Process

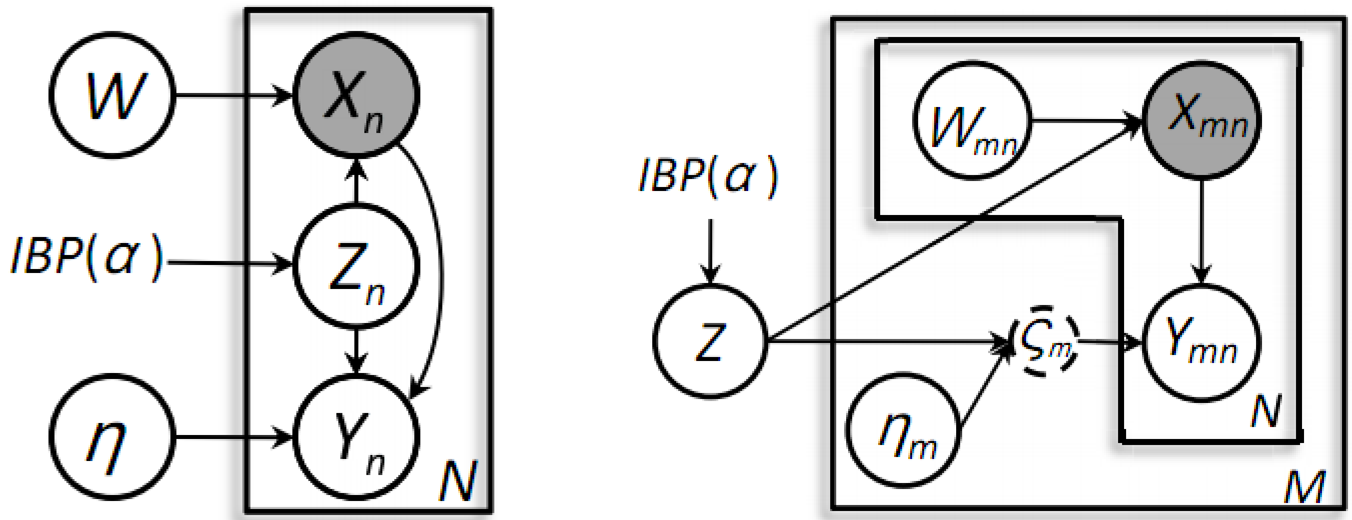
- In (Lawrence and Platt, ICML 04) an efficient method is proposed to learn the parameters (of a shared covariance function) for the Gaussian process.
- adopts the multi-task informative vector machine (IVM) to greedily select the most informative examples from the separate tasks and hence alleviate the computation cost.





Common Latent Representation in Nonparametric Bayesian Models

- Multi-Task Infinite Latent Support Vector Machines (Zhu, J. et al NIPS 11)



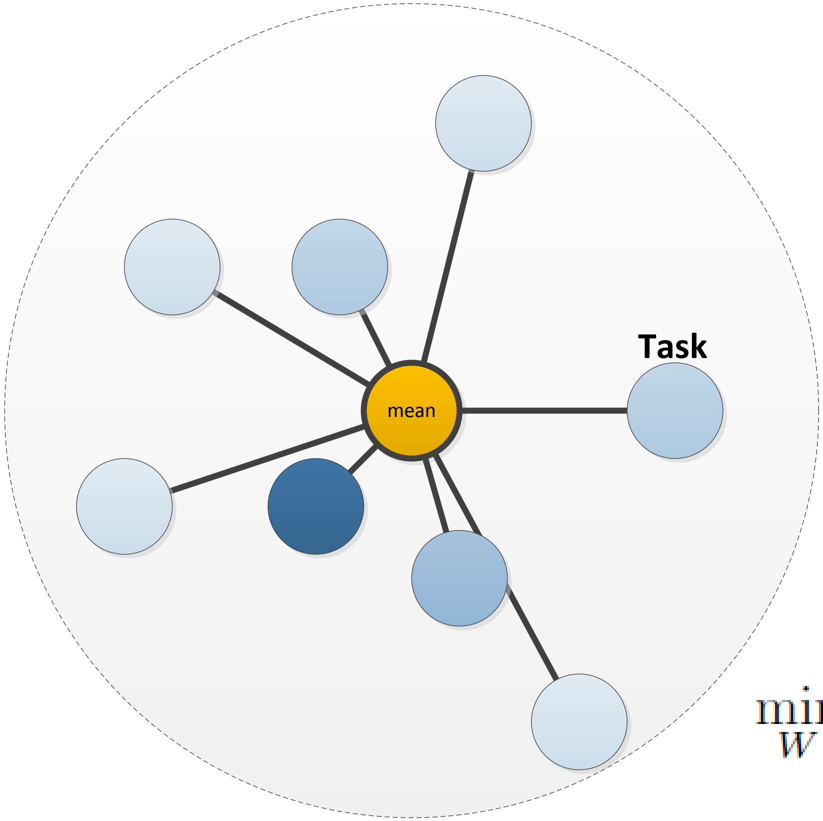
(a) (b)

Regularization-based Multi-task Learning

- All tasks are shared
  - regularized MTL, joint feature learning, low rank MTL, ASO
- Tasks form groups
  - clustered MTL, Network/Tree MTL
- Learning with outlier tasks: robust MTL
- Asymmetric MTL

Regularized Multi-Task Learning

- Assume all tasks are related in that the models of all tasks come from a particular distribution (Evgeniou & Pontil, KDD 04)

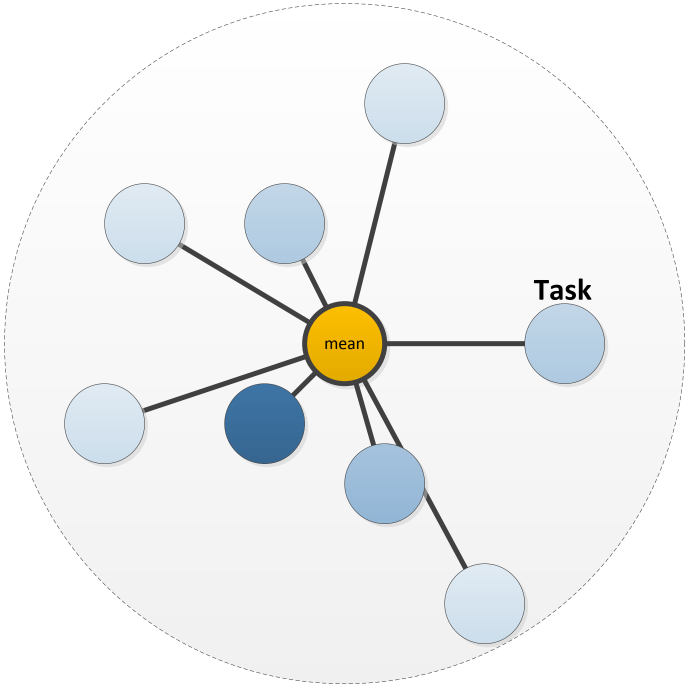


Regularization penalizes the deviation of each task from the mean

$$\min_{W} \mathrm{Loss}(W) + \lambda \sum_{t=1}^{T} \|W_t - \frac{1}{T}\sum_{s=1}^{T} W_s\|$$

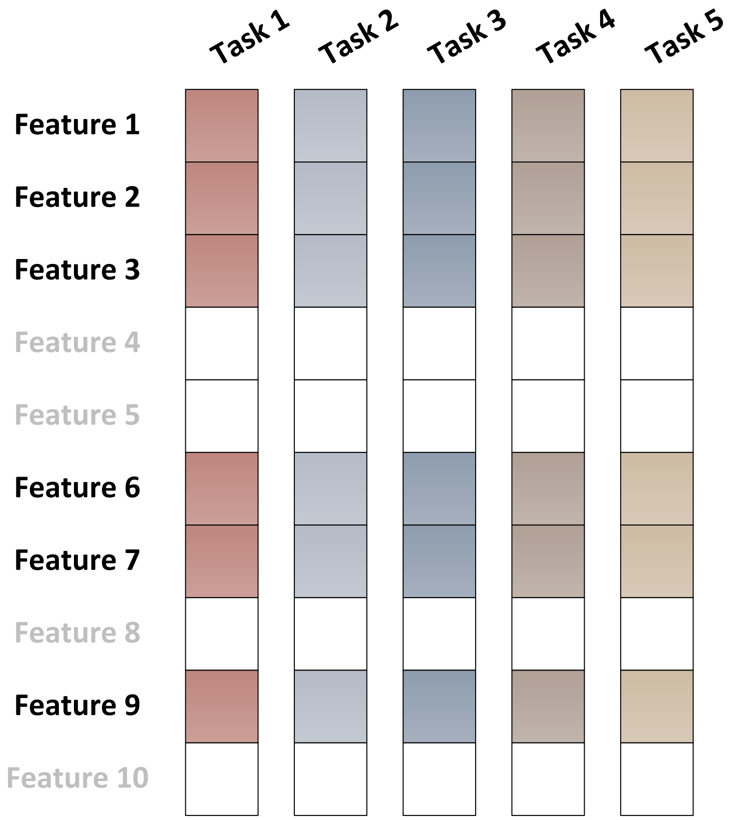
Regularized Multi-Task Learning

- Assumption: task parameter vectors of all tasks are close to each other.
  - Advantage: smooth objective, easy to optimize
  - Disadvantage: may not hold in real applications.



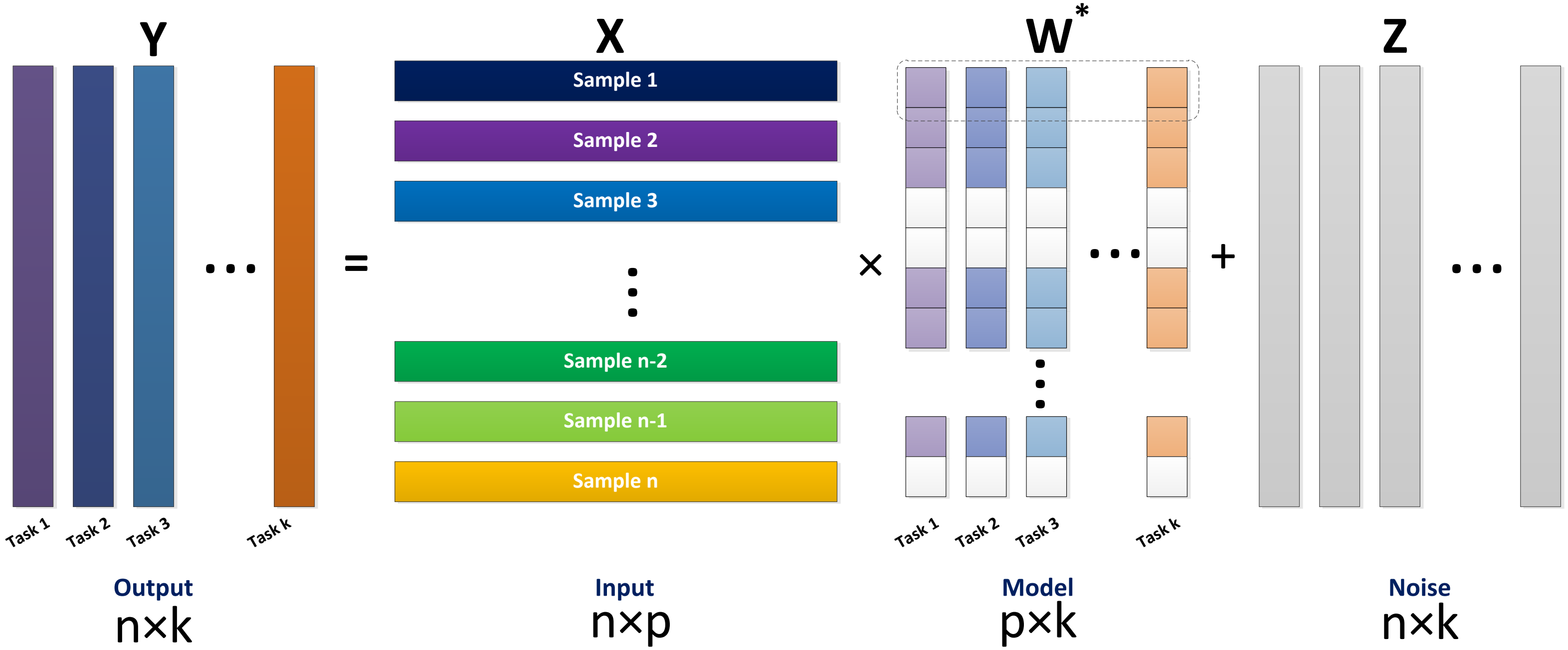
Multi-Task Learning with Joint Feature Learning

- One way to capture the task relatedness from multiple related tasks is to constrain all models to share a common set of features.
- For example, in school data, the scores from different schools may be determined by a similar set of features.



Multi-Task Learning with Joint Feature Learning

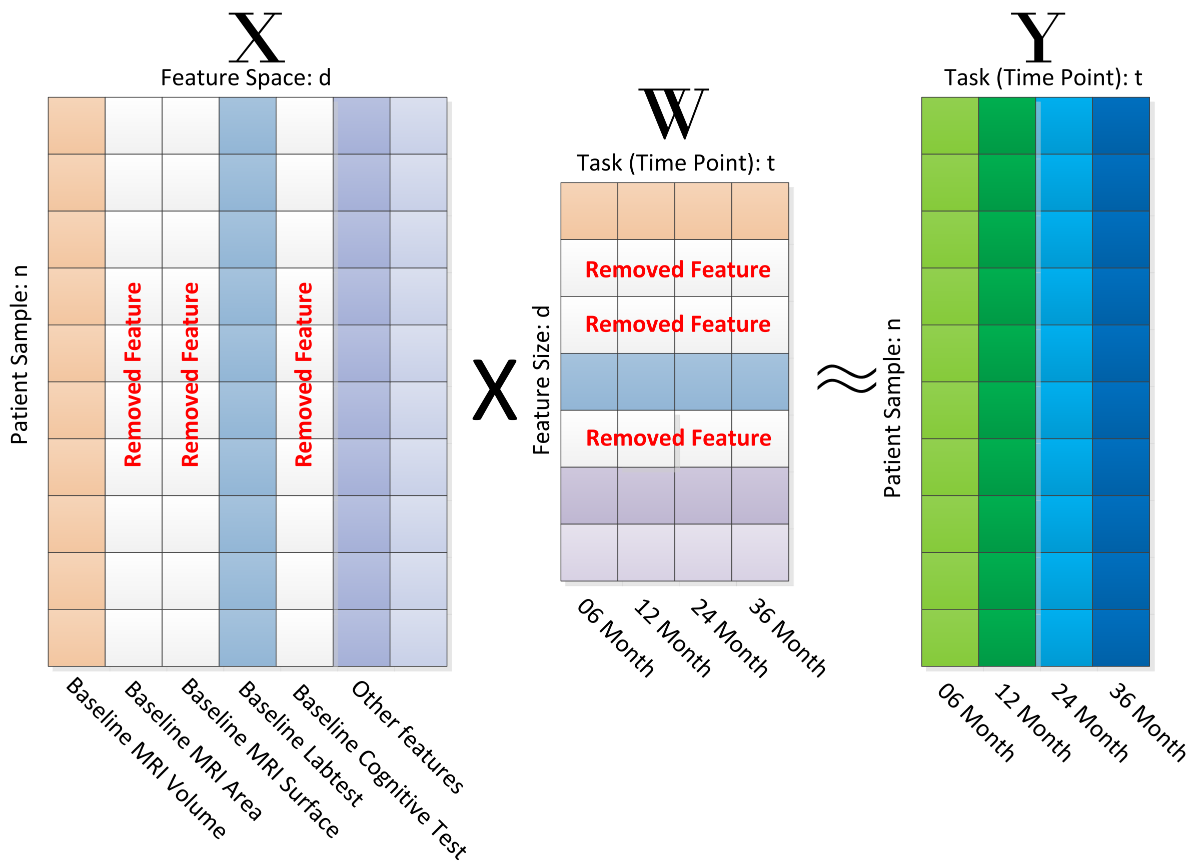
- Using group sparsity: ℓ_1/ℓ_2-norm regularization



$$\min_W \frac{1}{2}\|XW - Y\|_F^2 + \lambda \sum_{i=1}^{p} \|\mathbf{w}_i\|_2$$

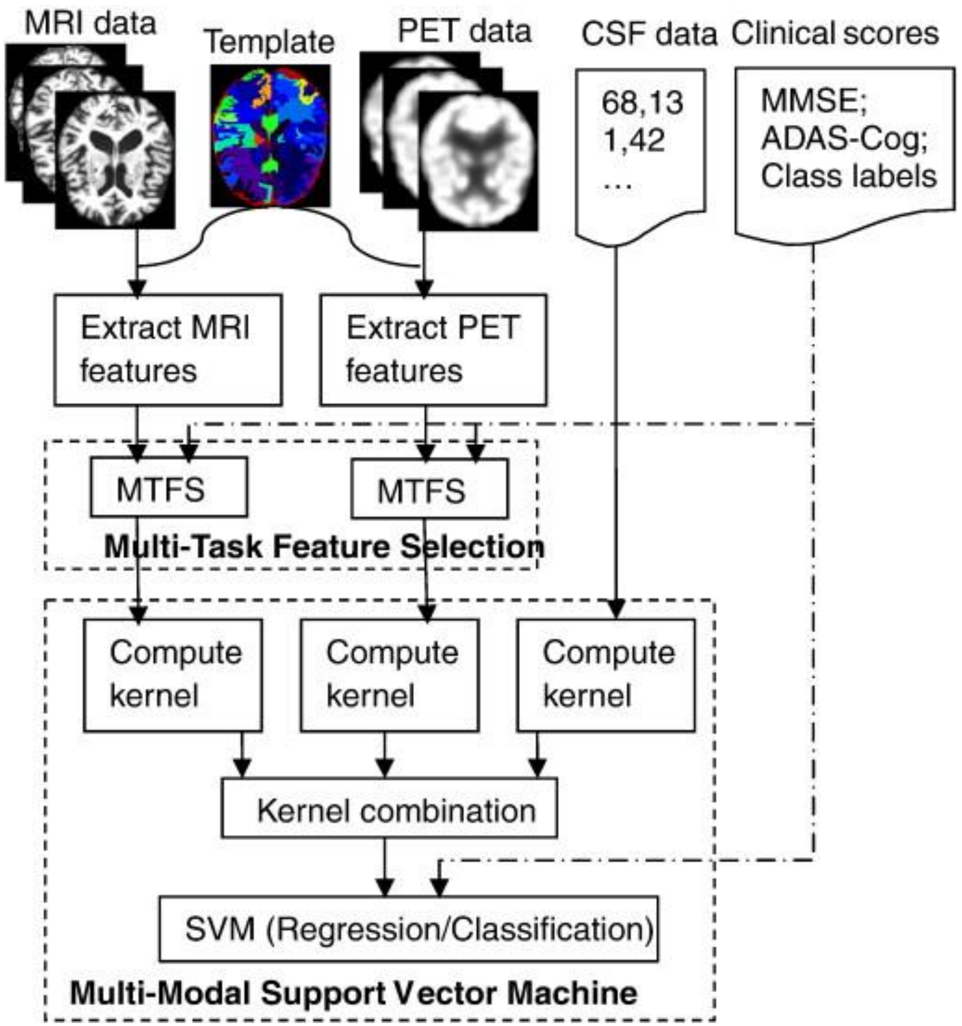
Joint Feature Selection in Disease Progression

- The progression of disease is assumed to involve the same set of features at different time points [Zhou et.al. KDD 11].



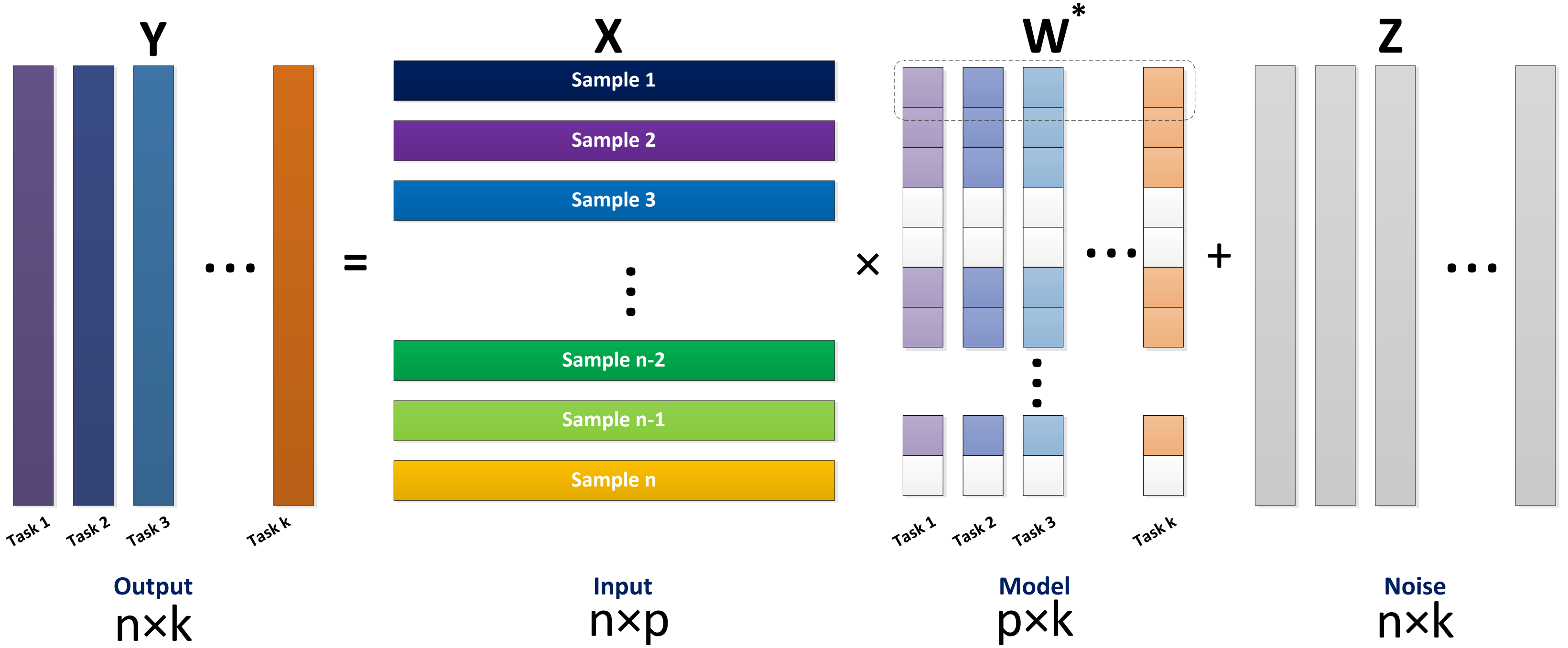
Joint Feature Selection in Disease Progression

- In predicting different cognitive scores, there may be shared features from different data sources.
- Multi-modal multi-task learning [Zhang, D. et.al. NeuroImage 12]



Multi-Task Learning with Joint Feature Learning – $L_1 L_q$

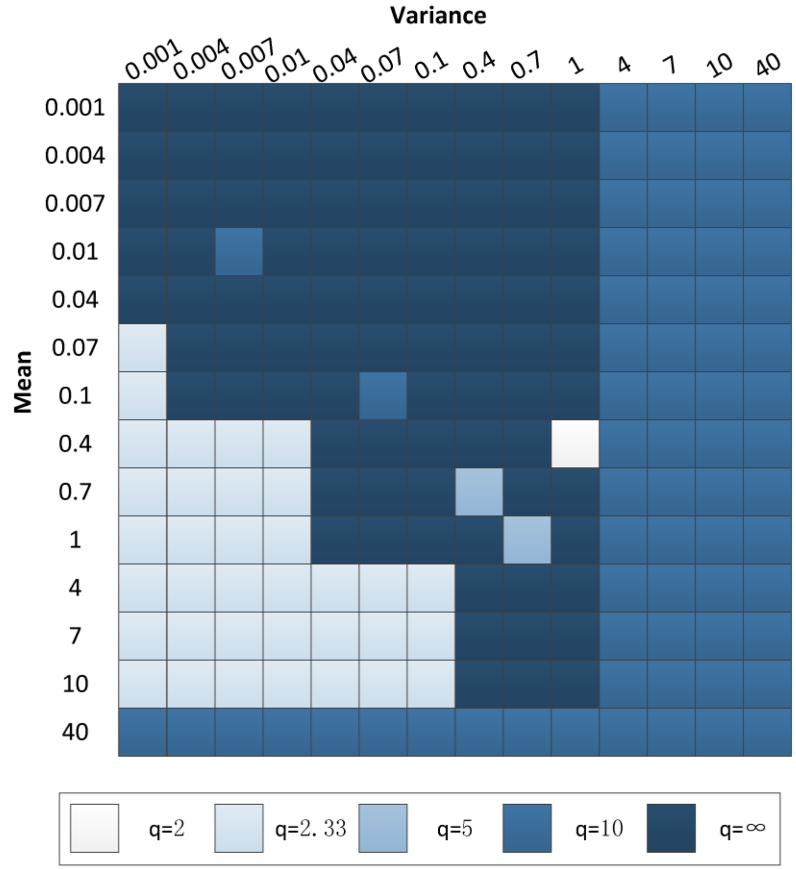
- More general ℓ_1/ℓ_q-norm regularization:



$$\min_{W} \frac{1}{2}\|XW - Y\|_F^2 + \lambda \sum_{i=1}^{p} \|\mathbf{w}_i\|_q$$

Multi-Task Learning with Joint Feature Learning – L_1L_q

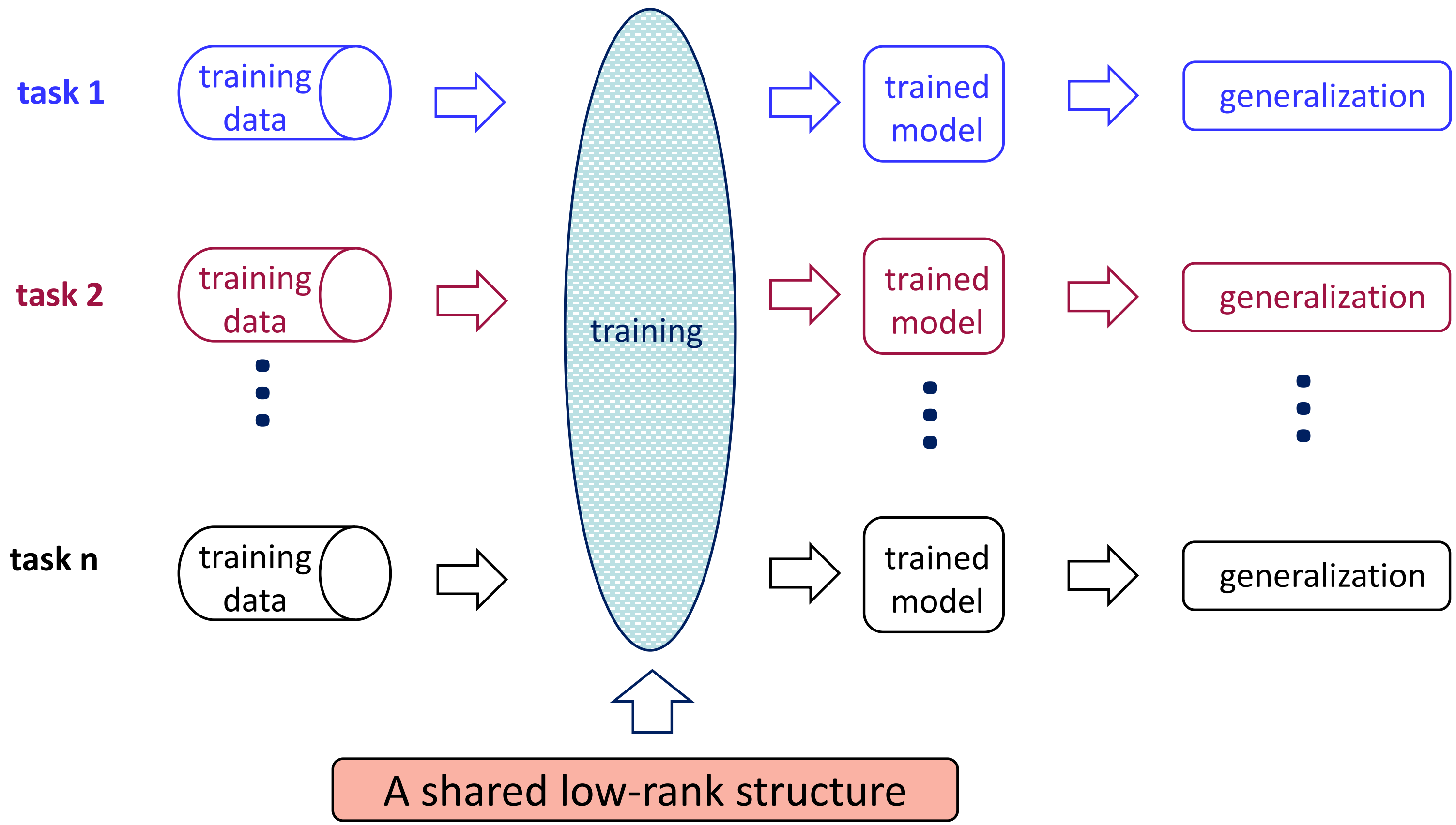
- The selection of *q* may depend on the distribution of the model:



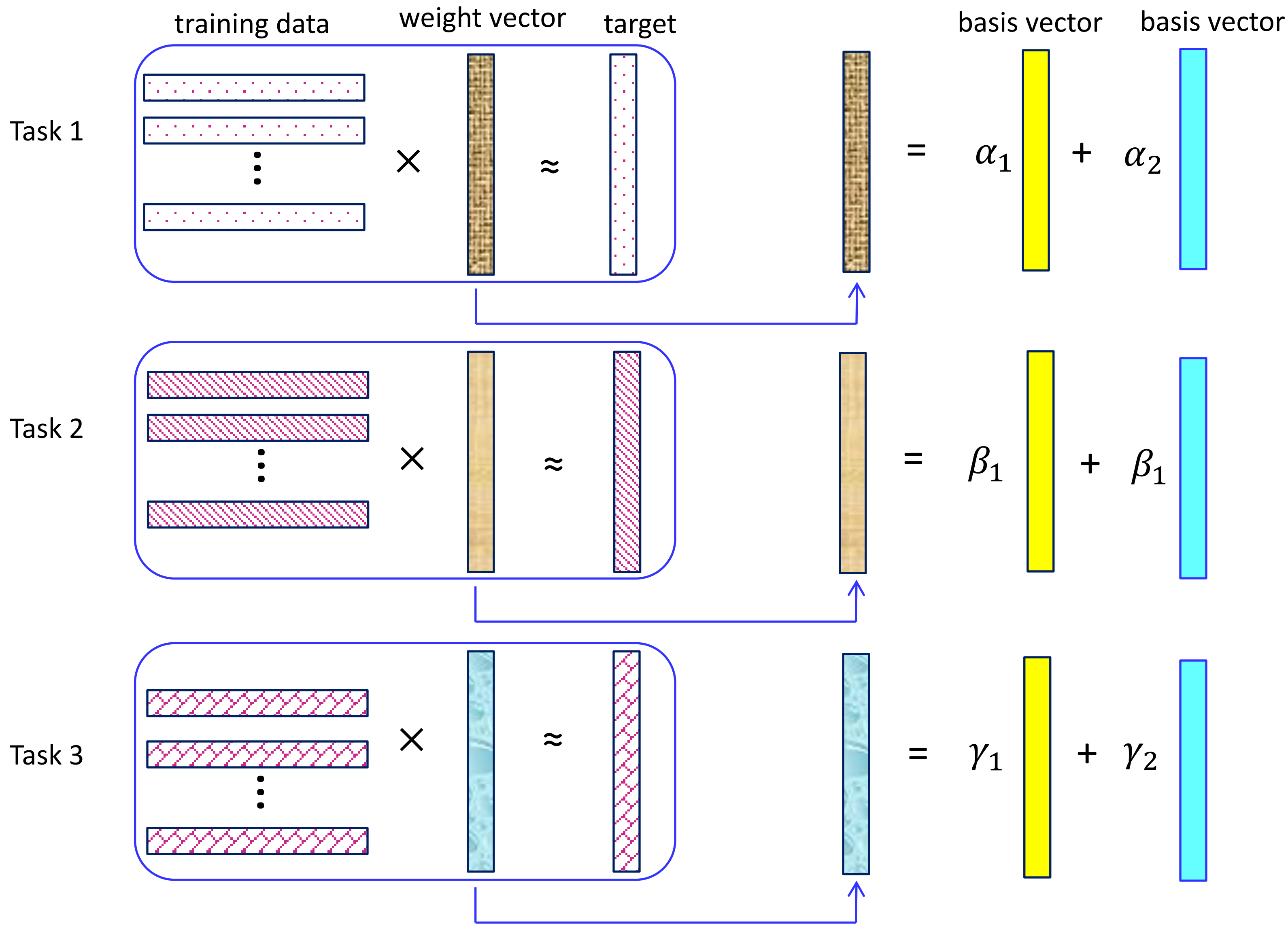
$W \sim N(\text{Mean, Variance})$

Trace-Norm Regularized MTL

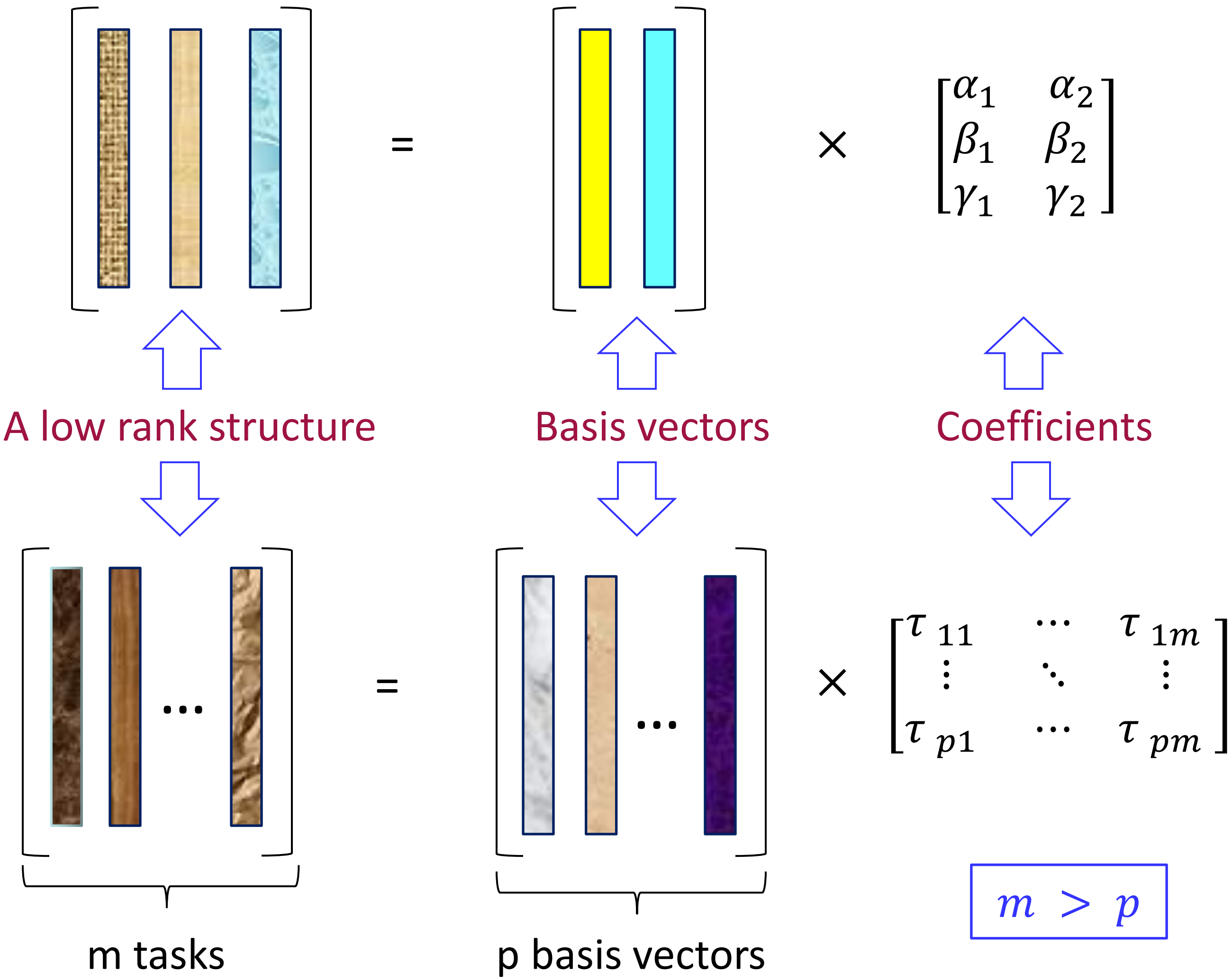
- Capture Task Relatedness via a Shared Low-Rank Structure



Low-Rank Structure for MTL



Low-Rank Structure for MTL

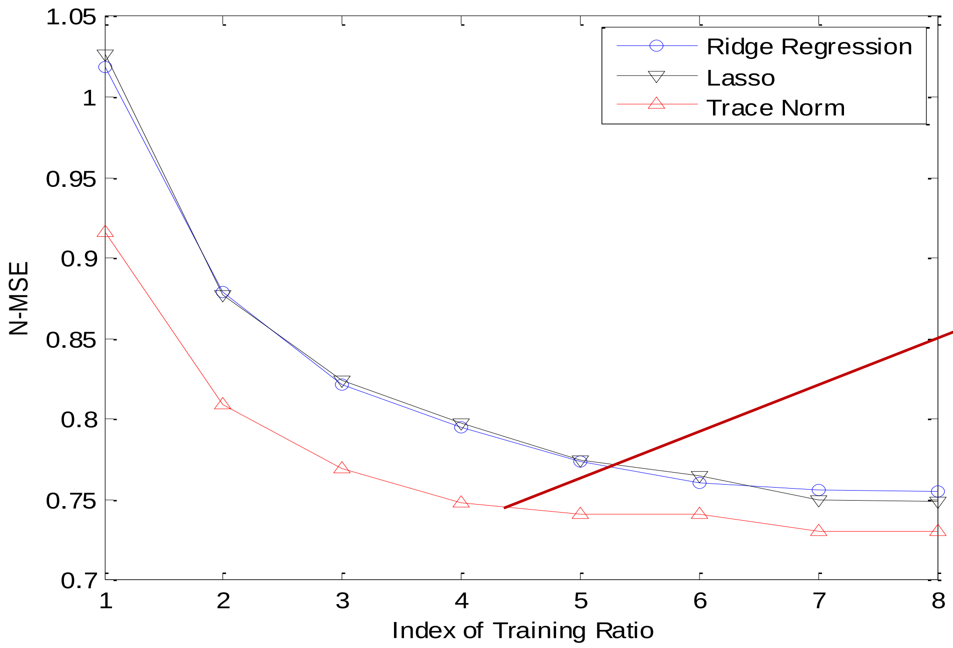


Low-Rank Structure for MTL

- Rank minimization formulation
  - $\min_{W} \text{Loss}(W) + \lambda \times \text{Rank}(W)$
  - Rank minimization is NP-Hard
- Convex relaxation: trace norm minimization
  - $\min_{W} \text{Loss}(W) + \lambda \times \|W\|_*$
  - Trace-norm minimization is the convex envelope of the rank minimization (Fazel et al., 2001).

Low-Rank Structure for MTL

- Evaluation on the *School* data[1]:
  - Predict exam scores for 15362 students from 139 schools
  - Describe each student by 27 attributes
  - Compare Ridge Regression, Lasso, and Trace Norm (for inducing a low-rank structure)



Performance measure:

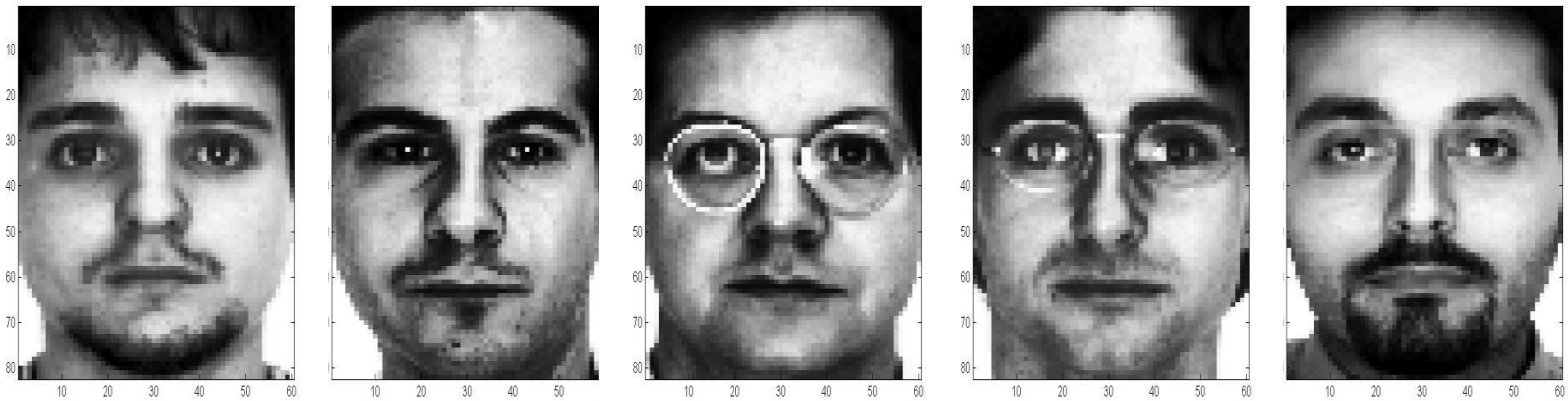
$$\text{N-MSE} = \frac{\text{mean squared error}}{\text{variance (target)}}$$

The Low-Rank Structure (induced via Trace Norm) leads to the smallest N-MSE.

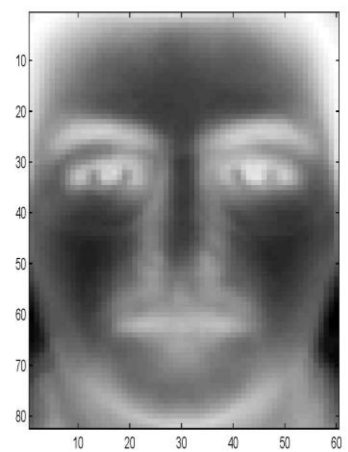
[1]http://ttic.uchicago.edu/~argyriou/code/

Low-Rank Structure for MTL

- Evaluation on the *Face* data[1]:
  - Trace Norm (low-rank structure)



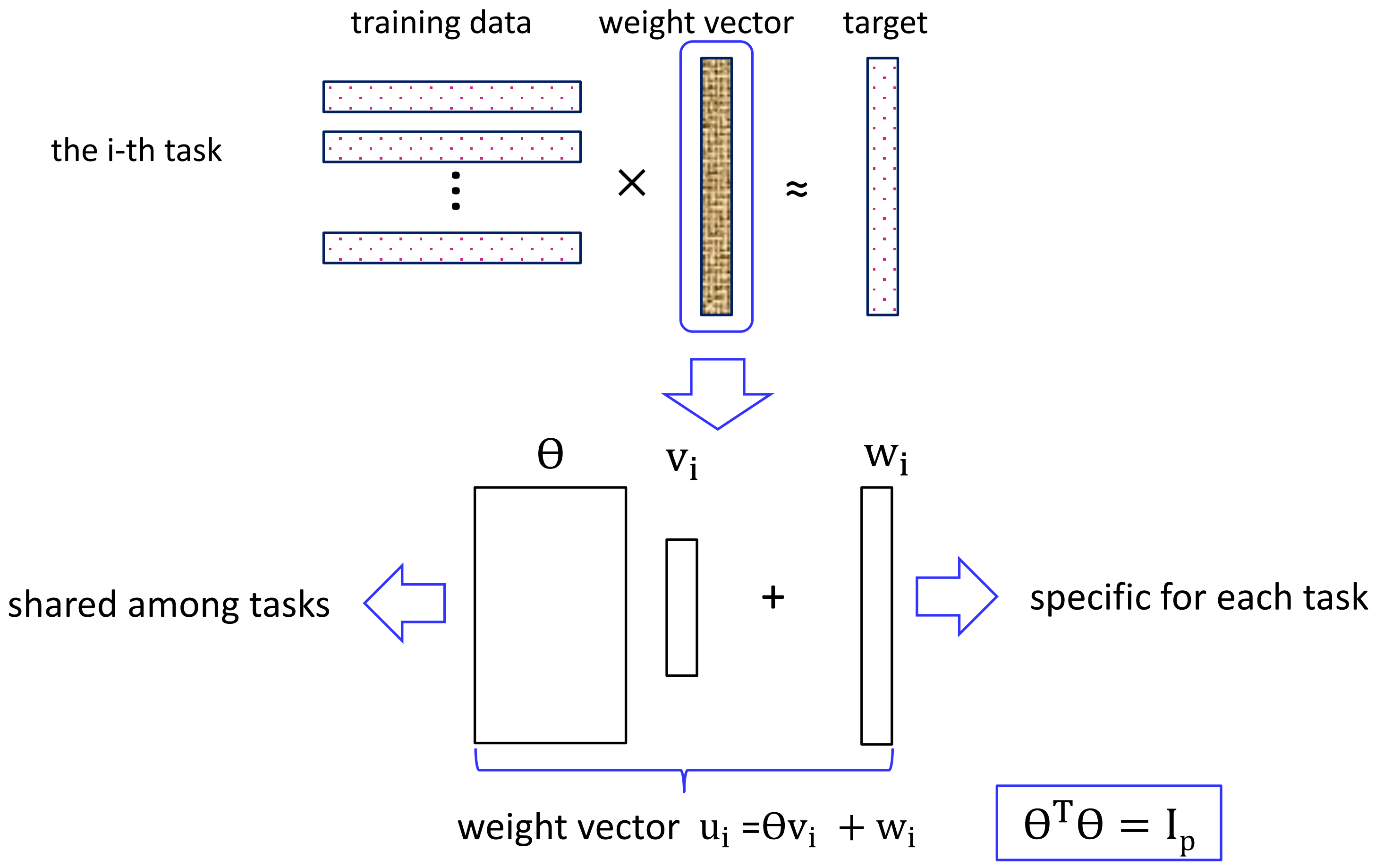
Plot of the 1st weight vector



Rough shape of the faces

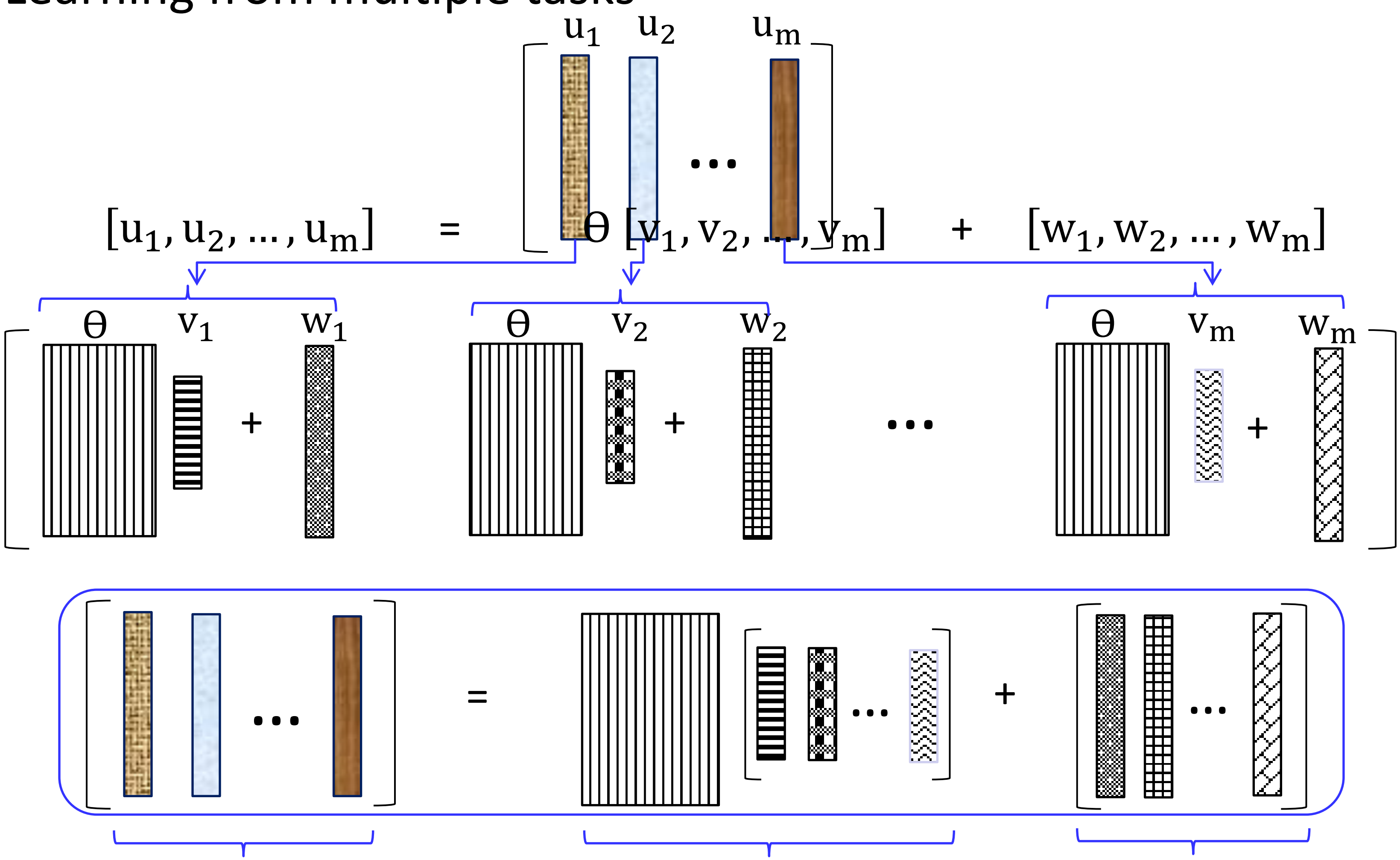
A shared Low-Rank Structure for MTL

- Learning from the i-th task (Ando et. al.'05, Chen et. al.'09)



A shared Low-Rank Structure for MTL

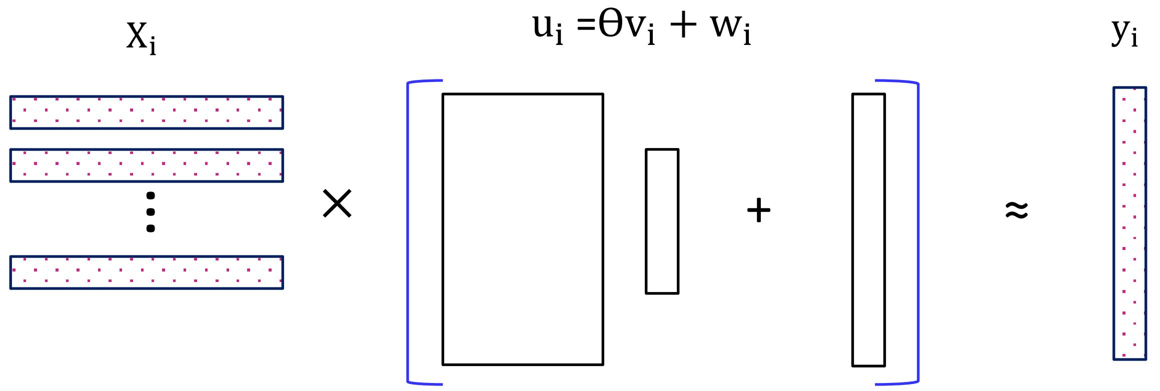
- Learning from multiple tasks



transformation matrix | a shared low rank structure | a task-specific structure

Empirical Loss

- Learning from the i-th task



- Empirical loss on the i-th task, for example,

$$\mathcal{L}_i(X_i(\Theta v_i + w_i), y_i) = \|X_i(\Theta v_i + w_i) - y_i\|^2$$

iASO Formulation

- iASO formulation

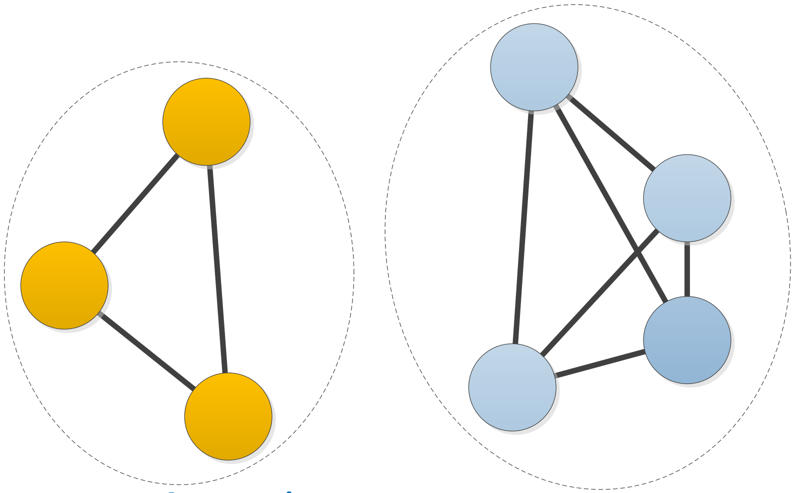
$$\underset{\Theta,\{\mathrm{v_i},\mathrm{w_i}\}}{\text{minimize}} \quad \sum_{i=1}^{m} \{\mathcal{L}_\mathrm{i}(\mathrm{X_i}(\Theta \mathrm{v_i} + \mathrm{w_i}), \mathrm{y_i}) + \alpha \|\Theta \mathrm{v_i} + \mathrm{w_i}\|^2 + \beta \, \|\mathrm{w_i}\|^2\}$$

$$\text{subject to} \quad \Theta^\mathrm{T}\Theta = \mathrm{I}$$

- control both model complexity and task relatedness
- subsume ASO (Ando et al.'05) and SVM as special cases
- naturally lead to a convex relaxation (Chen et al., 10)
- iASO and cASO are equivalent under certain conditions

Multi-Task Learning with Clustered Structure

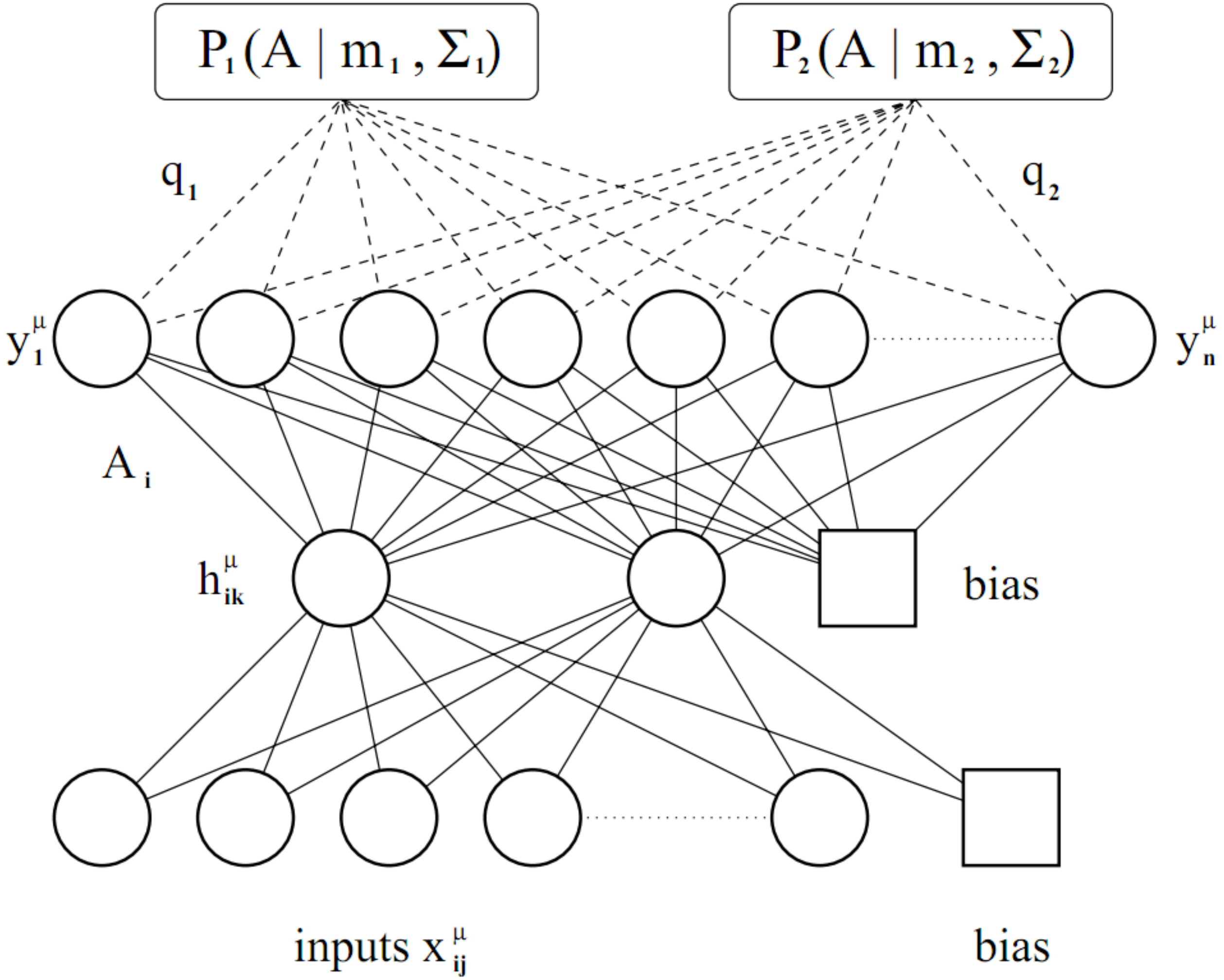
- Most MTL techniques assume all tasks are related
- Not true in many applications
- Clustered multi-task learning assumes
  - the tasks have group structures
  - the models of tasks from the same group are closer to each other than those from a different group



e.g. tasks in the yellow group are predictions of heart related diseases and in the blue group are brain related diseases.

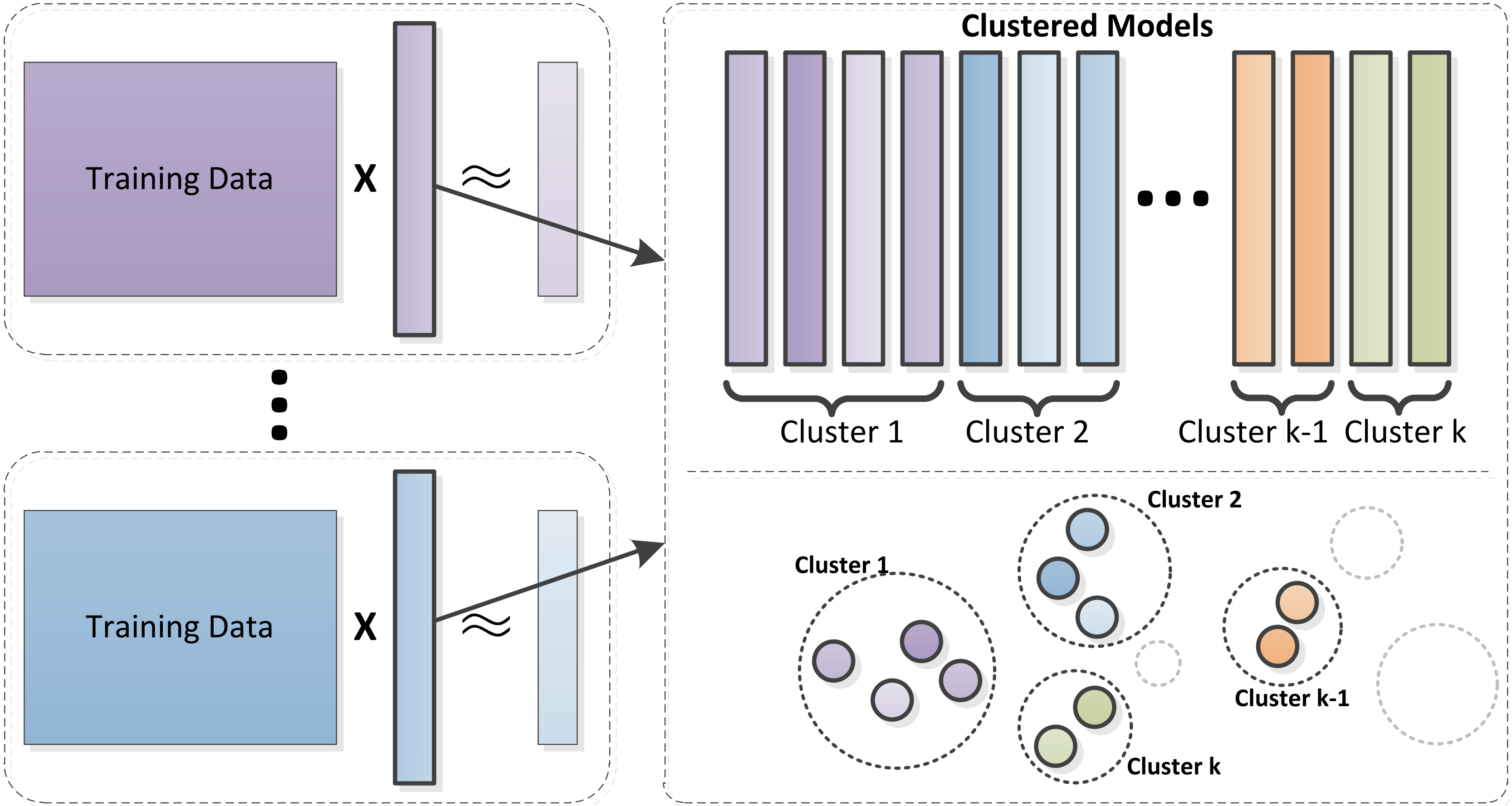
Task Clustering in Neural Network

- Bakker and Heskes JMLR 2003



Clustered Multi-Task Learning

- Use regularization to capture clustered structures.

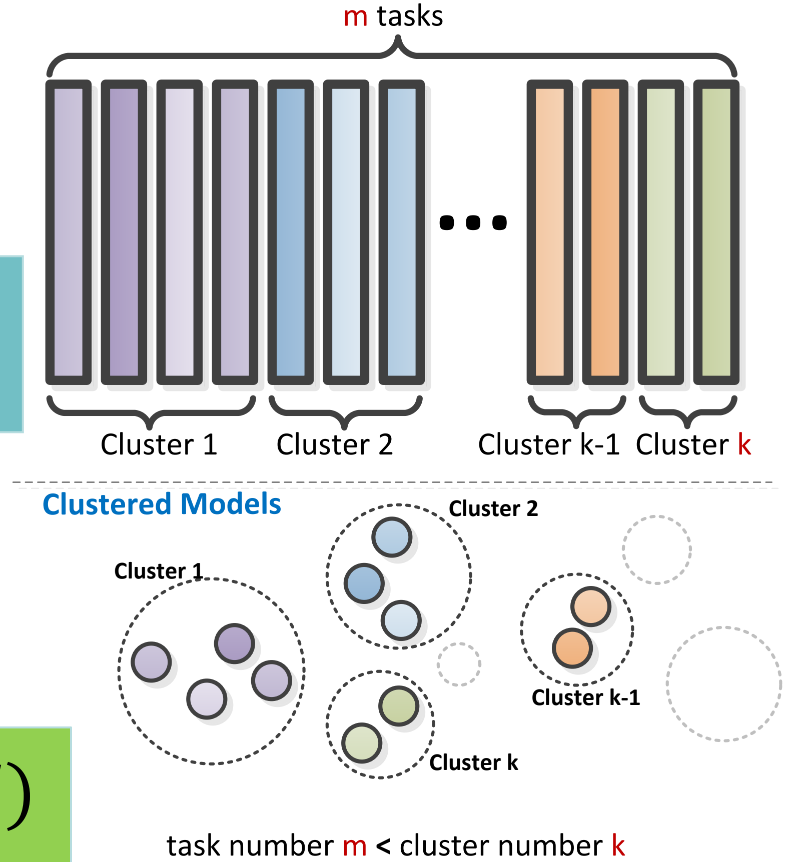


Clustered Multi-Task Learning

- Capture structures by minimizing sum-of-square error (SSE) in K-means clustering:

$$\min_{I} \sum_{j=1}^{k} \sum_{v \in I_j} \|w_v - \bar{w}_j\|_2^2$$

I_j index set of j^{th} cluster



Equivalent

$$\min_{F} \mathrm{tr}(W^T W) - \mathrm{tr}(F^T W^T W F)$$

F : $m \times k$ orthogonal cluster indicator matrix

$F_{i,j} = 1/\sqrt{n_j}$ if $i \in I_j$ and 0 otherwise

Clustered Multi-Task Learning

- Directly minimizing SSE is hard because of the non-linear constraint on F:

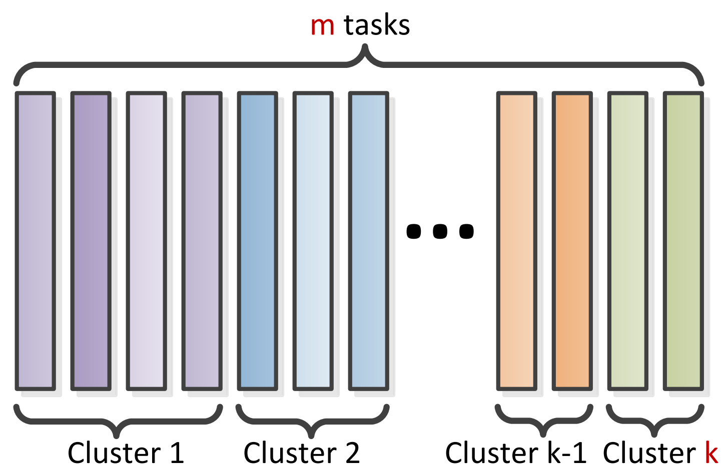
$$\min_{F} \operatorname{tr}(W^T W) - \operatorname{tr}(F^T W^T W F)$$

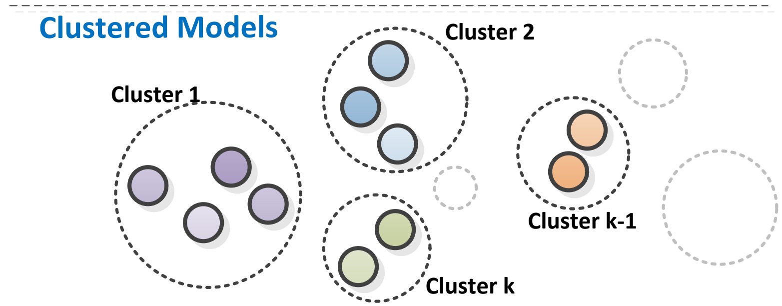
$F : m \times k$ orthogonal cluster indicator matrix
$F_{i,j} = 1/\sqrt{n_j}$ if $i \in I_j$ and 0 otherwise

Spectral Relaxation

$$\min_{F:F^T F = I_k} \operatorname{tr}(W^T W) - \operatorname{tr}(F^T W^T W F)$$

Zha et. al. 2001 NIPS





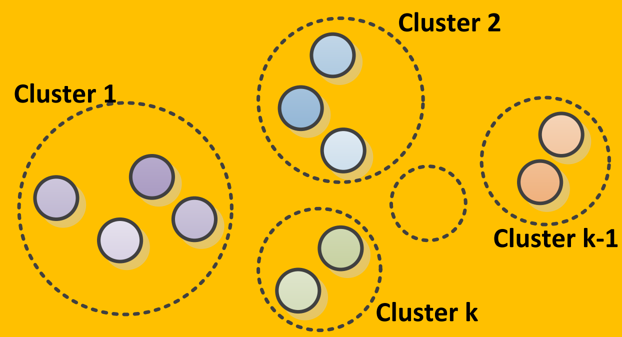
task number m < cluster number k

Clustered Multi-Task Learning

- Clustered multi-task learning (CMLT) formulation [Zhou et. al. NIPS 2011]

$$\min_{W,F:F^TF=I_k} \mathrm{Loss}(\mathrm{W}) + \alpha[\mathrm{tr}(W^TW) - \mathrm{tr}(F^TW^TWF)] + \beta\,\mathrm{tr}(W^TW)$$

capture cluster structures



Improves generalization performance

☹Non-Convex Optimization!

Convex Clustered Multi-Task Learning

$$\min_{W,F:F^TF=I_k} \mathrm{Loss(W)} + \alpha[\mathrm{tr}(W^TW) - \mathrm{tr}(F^TW^TWF)] + \beta\,\mathrm{tr}(W^TW)$$

Equivalent

$$\min_{W,F:F^TF=I_k} \mathrm{Loss(W)} + \alpha\eta(1+\eta)\mathrm{tr}(W(\eta I + FF^T)^{-1}W^T)$$

Convex Relaxation

Chen et al KDD 2009

Jacob et al NIPS 2009

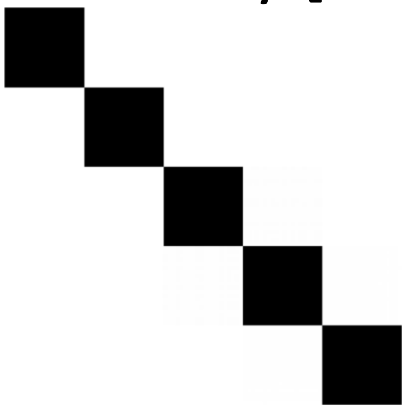
Zhou et al NIPS 2010

$$\min_{W,M} \mathrm{Loss(W)} + \alpha\eta(1+\eta)\mathrm{tr}(W(\eta I + M)^{-1}W^T)$$

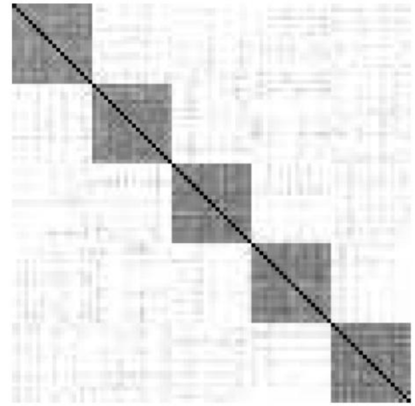
$$\text{subject to:}\quad \mathrm{tr}(M) = k, M \preccurlyeq I, M \in S_+^m$$

Convex Clustered Multi-Task Learning

- Synthetic Study [Zhou NIPS 2011]

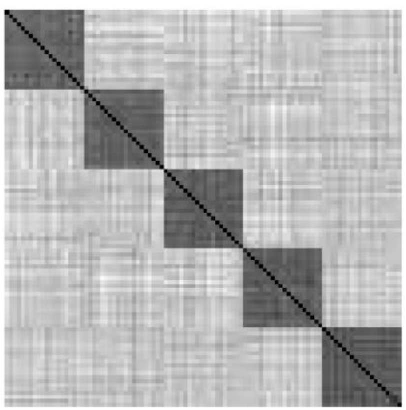


Ground Truth

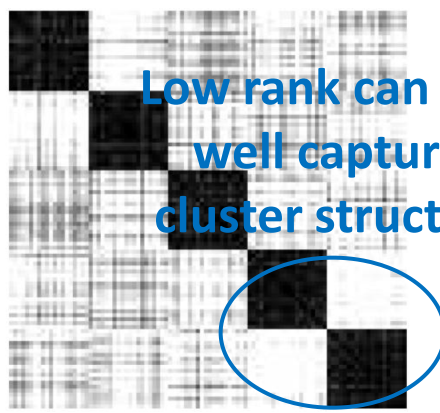


Single Task Learning

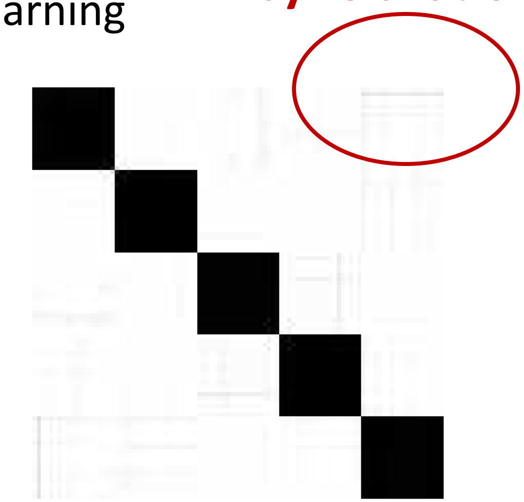
noise introduced by relaxations



Mean Regularized MTL



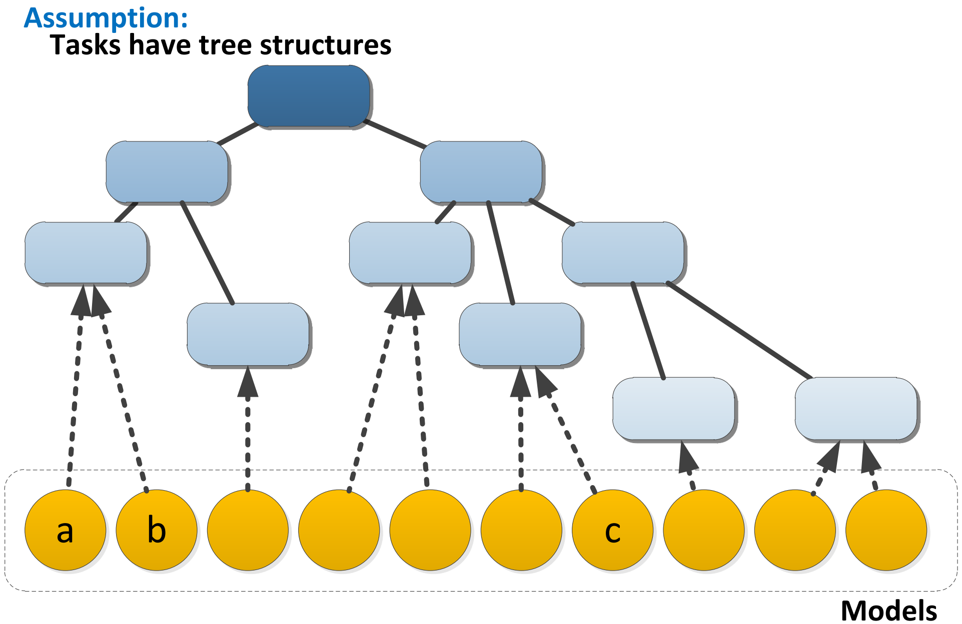
Trace Norm Regularized MTL



Convex Relaxed CMTL

Multi-Task Learning with Tree Structures

- In some scenarios, the tasks may be equipped with tree structures:
  - The tasks belong to the same node are similar to each other
  - The similarity between two nodes is structured and relates to the depth of the 'common' tree node

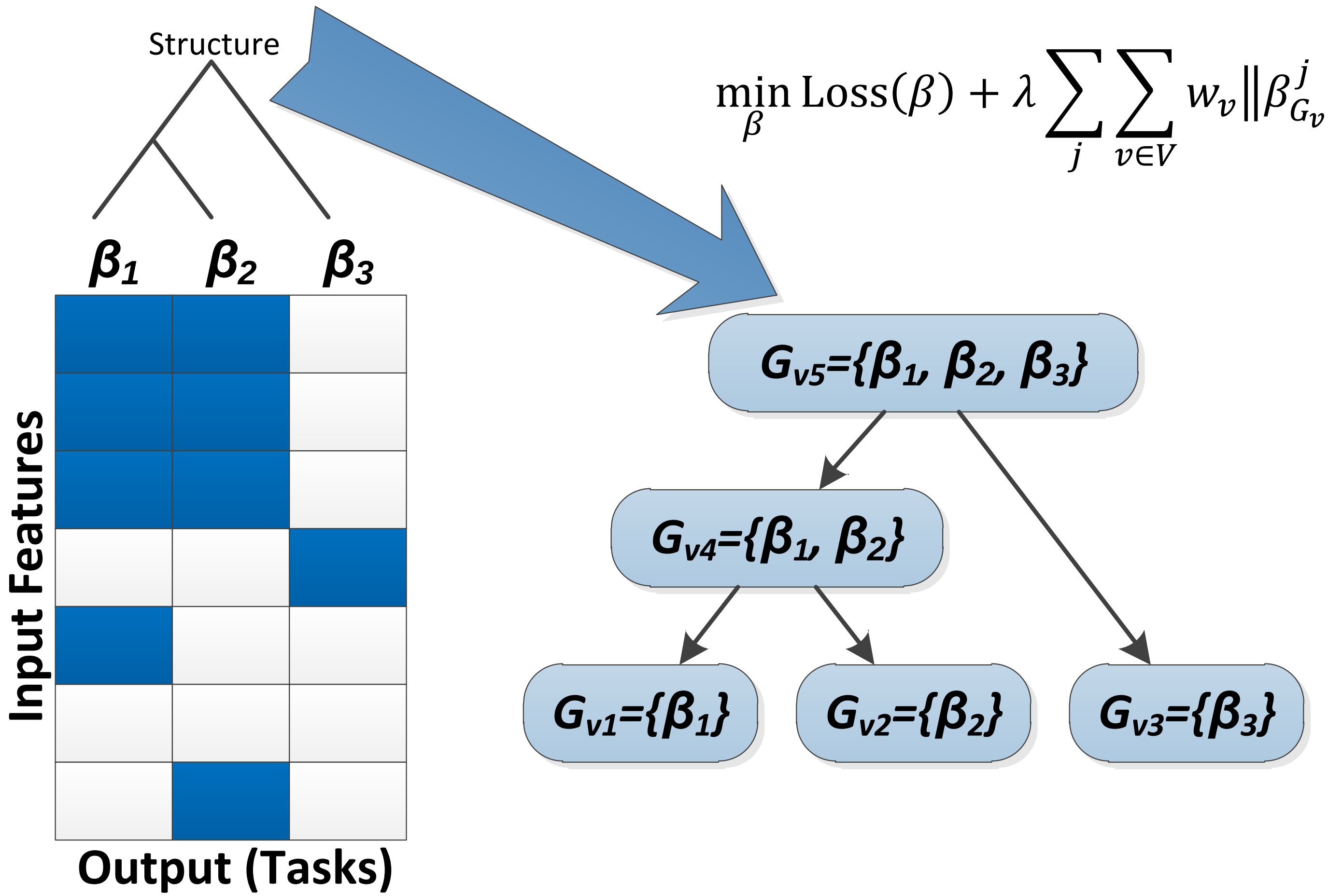


Task a is more similar with b, comparing to c

Multi-Task Learning with Tree Structures

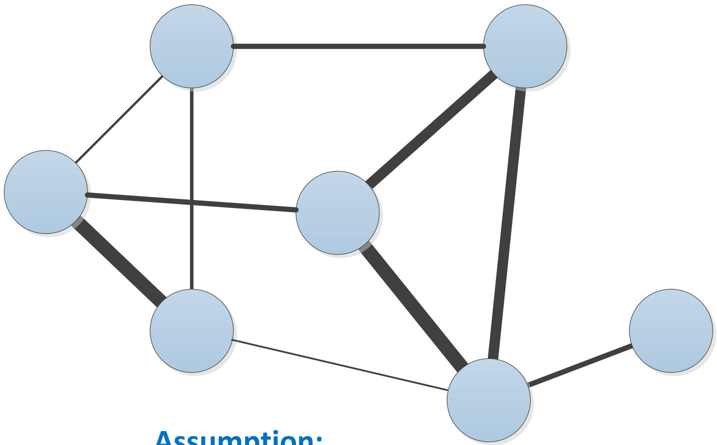
- Tree-Guided Group Lasso (Kim and Xing 2010 ICML)

$$\min_{\beta} \mathrm{Loss}(\beta) + \lambda \sum_{j} \sum_{v \in V} w_v \left\| \beta^{j}_{G_v} \right\|_2$$



Multi-Task Learning with Graph Structures

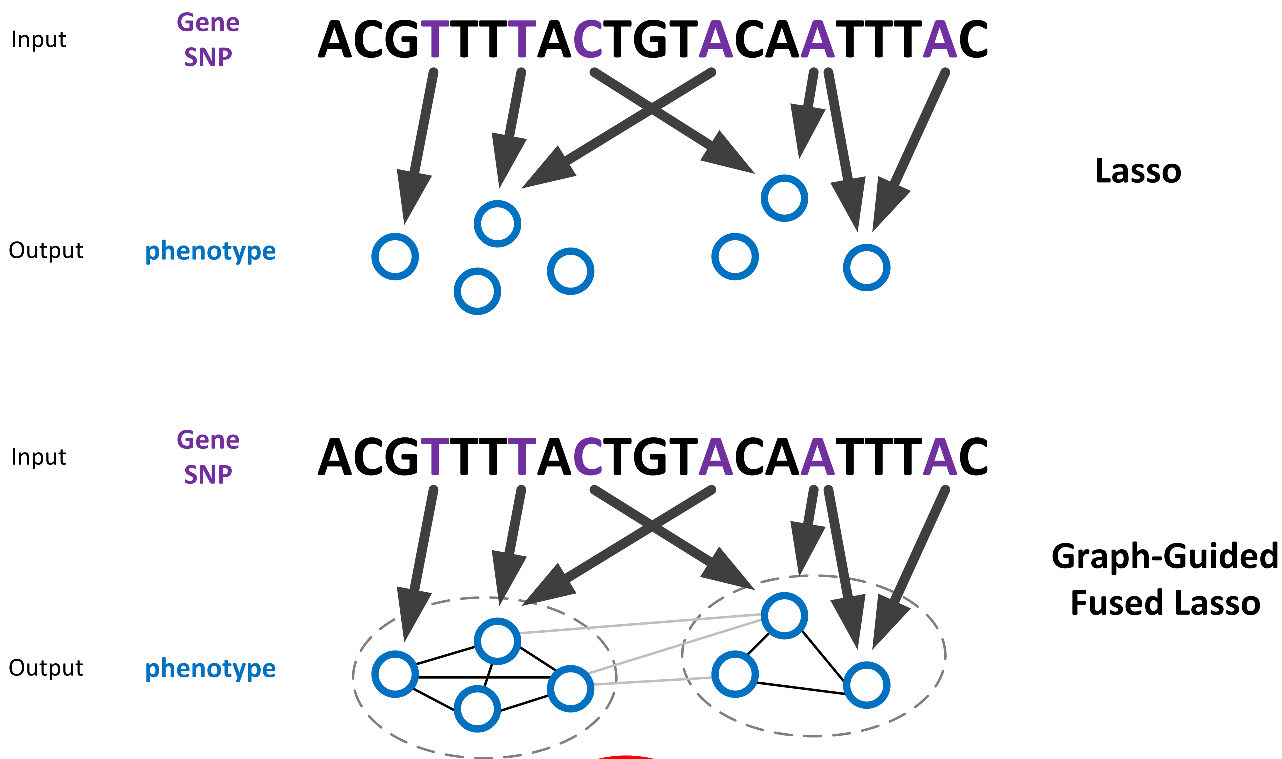
- In real applications, tasks involved in MTL may have graph structures
  - The two tasks are related if they are connected in a graph, i.e. the connected tasks are similar
  - The similarity of two related tasks can be represented by the weight of the connecting edge.



Assumption:
Tasks have graph/network structures

Multi-Task Learning with Graph Structures

- Graph-guided Fused Lasso (Chen et. al. UAI11)



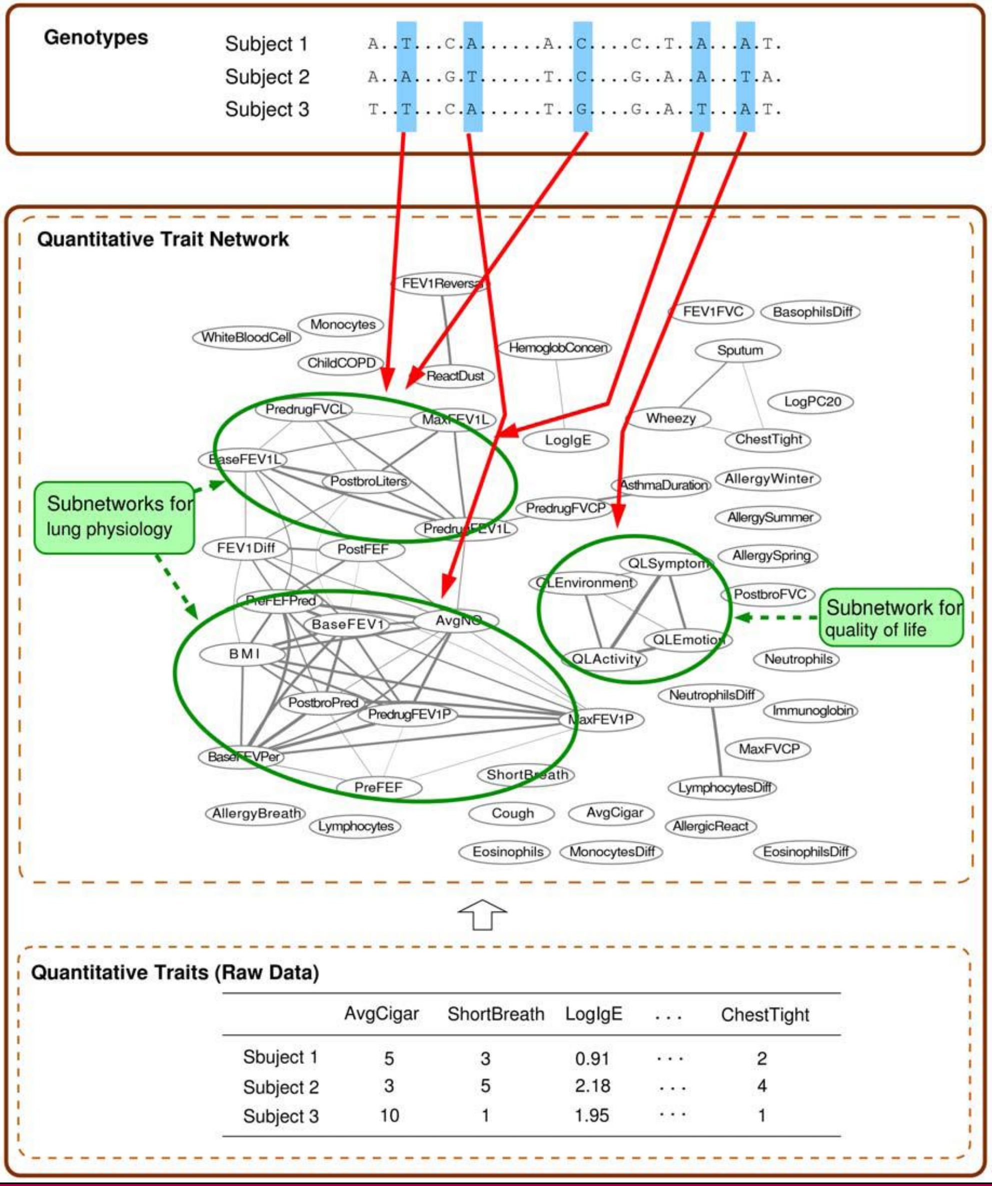
$$\min_{W} \text{Loss}(W) + \lambda \|W\|_1 + \Omega(W)$$

Graph-guided Fusion Penalty

$$\Omega(W) = \gamma \sum_{e=(m,l)\in E} \tau(r_{ml}) \sum_{j=1}^{J} \left| w_{jm} - sign(r_{ml}) w_{jl} \right|$$

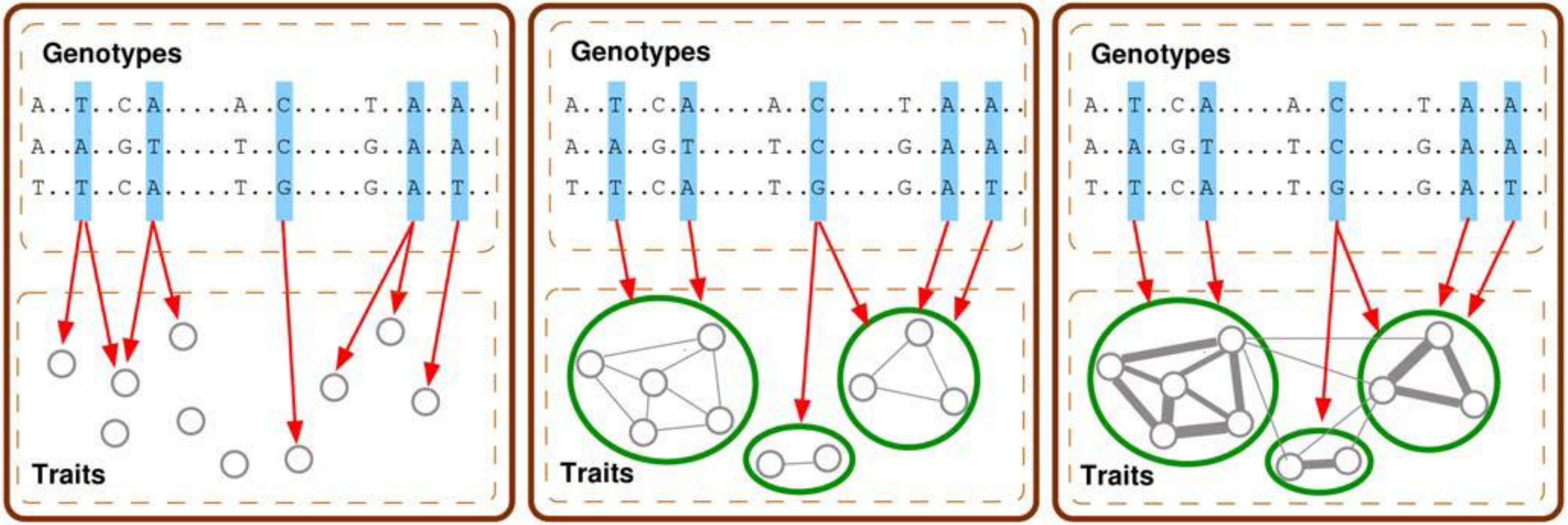
Quantitative Trait Network

- Linked Edge: the corresponding two traits are highly correlated.
- Thicknesses: strength of correlation.
- Identifying SNPs that are associated with a subnetwork of clinical traits (Kim and Xing 2009).



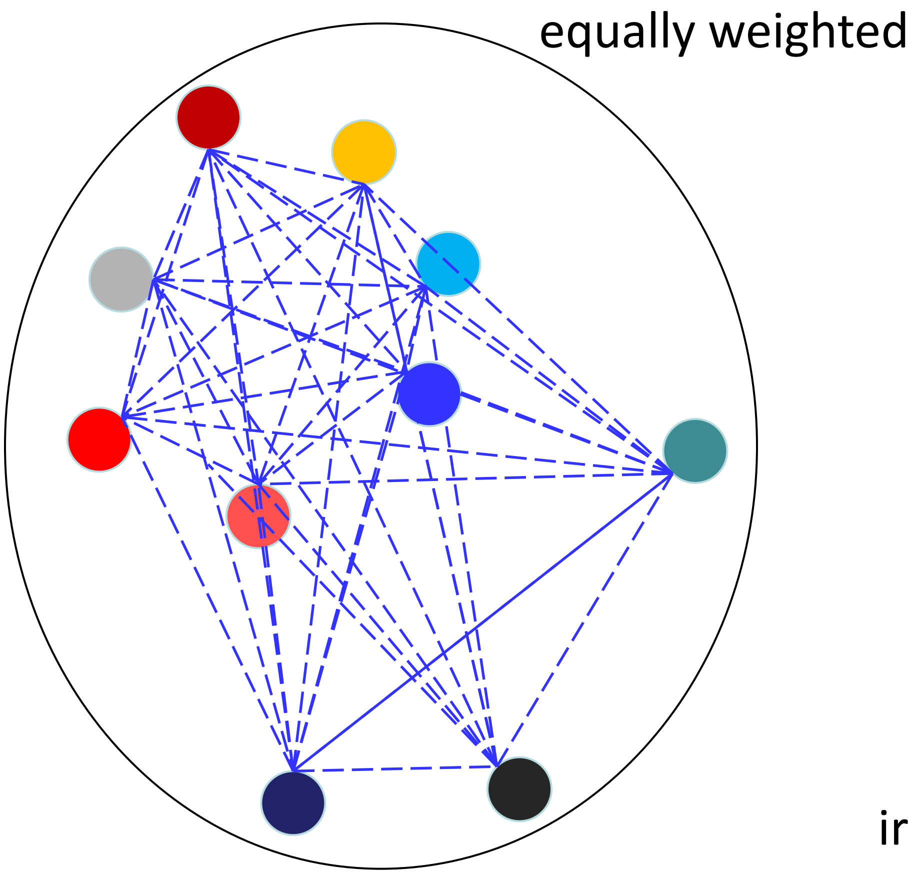
Graph-Weighted Fused Lasso

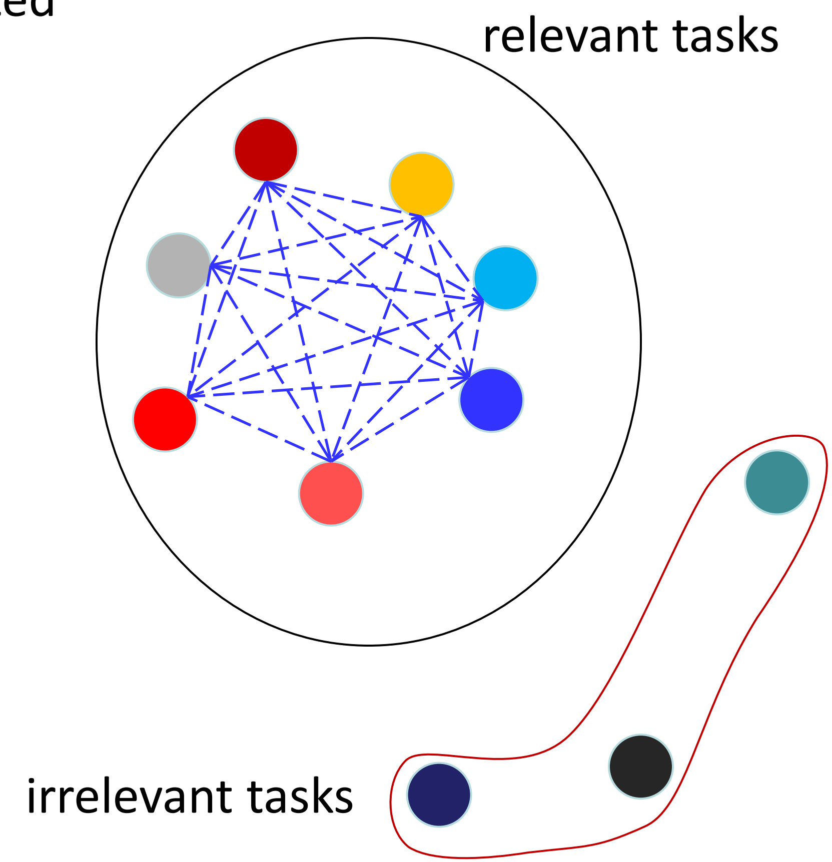
- **Lasso**: each phenotype represented as a circle is independently mapped to SNPs for association
- **Graph-constrained fused Lasso**: consider a QTN to search for an association between a SNP and a subnetwork of traits.
- **Graph-weighted fused Lasso**: consider a QTN with edge weights.



Robust Multi-Task Learning

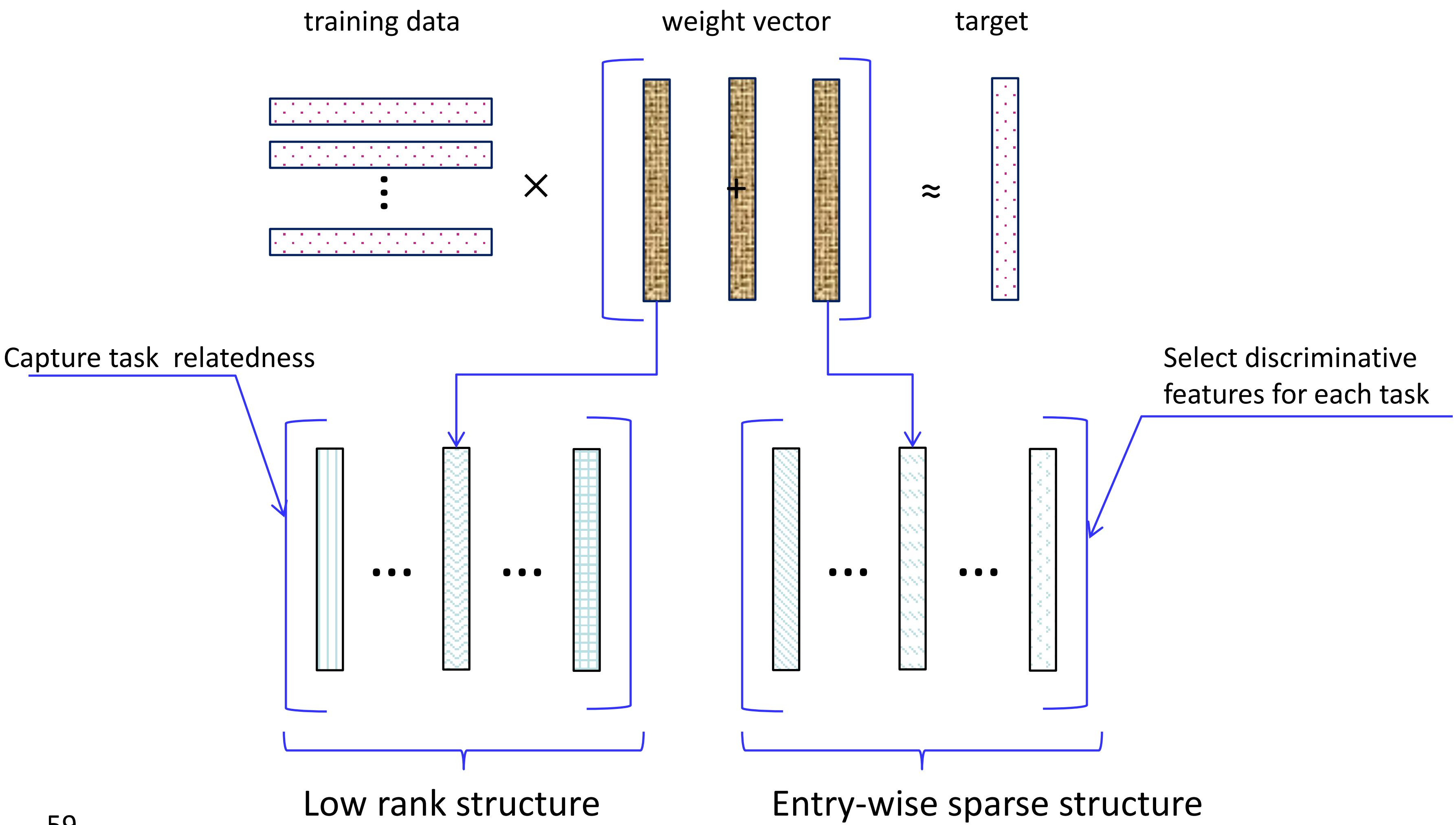
- Most Existing MTL Approaches
- Robust MTL Approaches





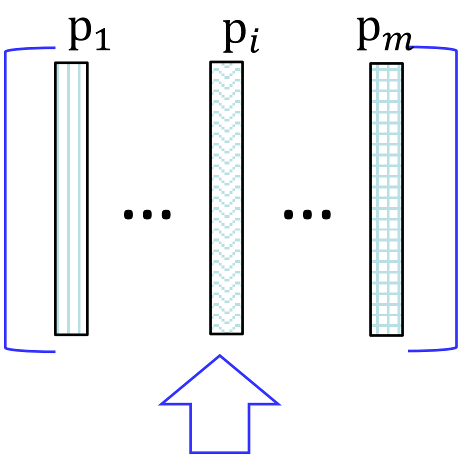
Incoherent Low-Rank and Sparse Structures

- Learning from the i-th task



Incoherent Low-Rank and Sparse Structures

Low-rank structure

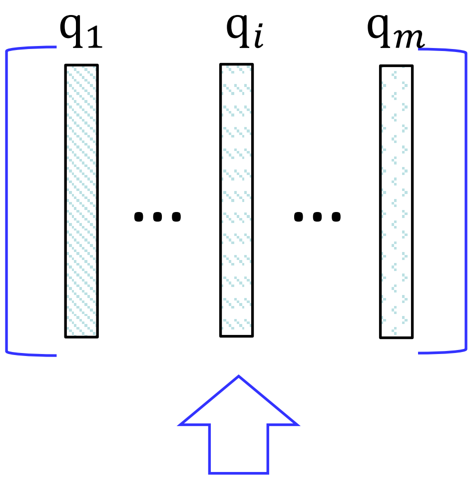


$$\|P\|_*$$

(Sum of singular values)

$$\|P\|_* \leqslant \eta$$

Entry-wise sparse structure



$$\|Q\|_1$$

(Sum of the absolute values of all entries)

$$\lambda\|Q\|_1$$

ISLR Formulation

- Empirical loss on the i-th task, e.g.,

$$\mathcal{L}_i(X_i(p_i + q_i), y_i) = \|X_i(p_i + q_i) - y_i\|^2$$

- Incoherent Sparse Low-Rank (ISLR) formulation

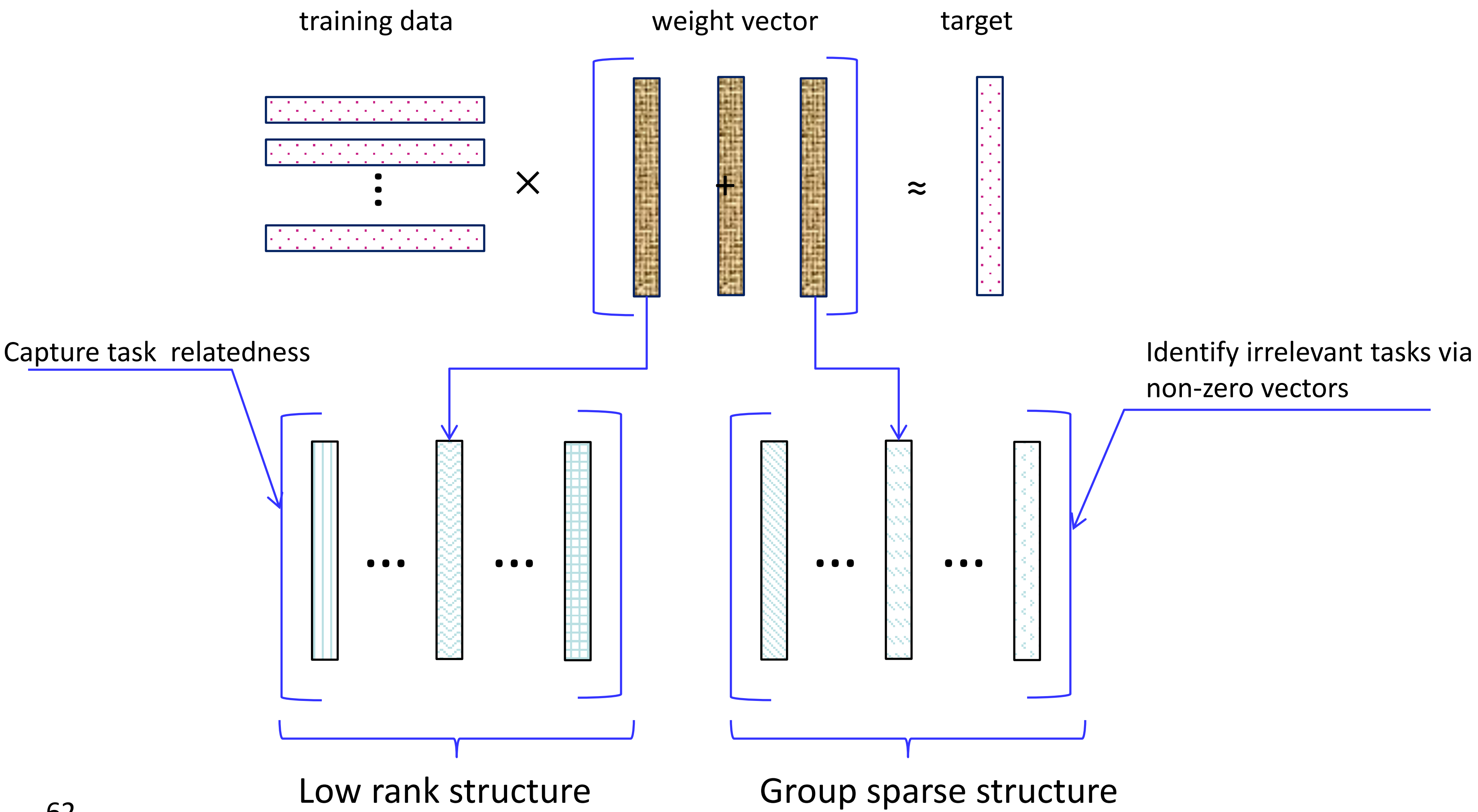
$$\underset{P,Q}{\text{minimize}} \quad \sum_{i=1}^{m} \mathcal{L}_i(X_i(p_i + q_i), y_i) + \lambda \|Q\|_1$$

$$\text{subject to} \quad \|P\|_* \leqslant \eta$$

  - Convex formulation
  - Decomposed sparse and low-rank structures

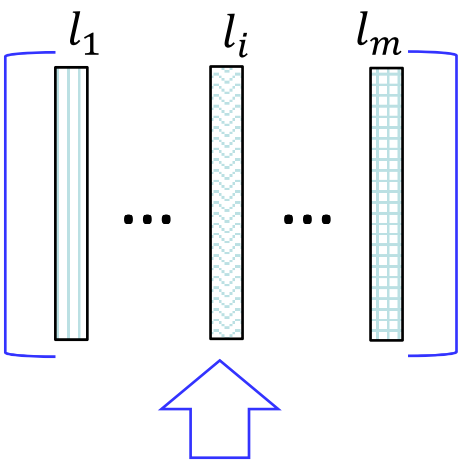
Low-Rank and Group Sparsity in MTL

- Learning from the i-th task



Low-Rank and Group Sparsity in MTL

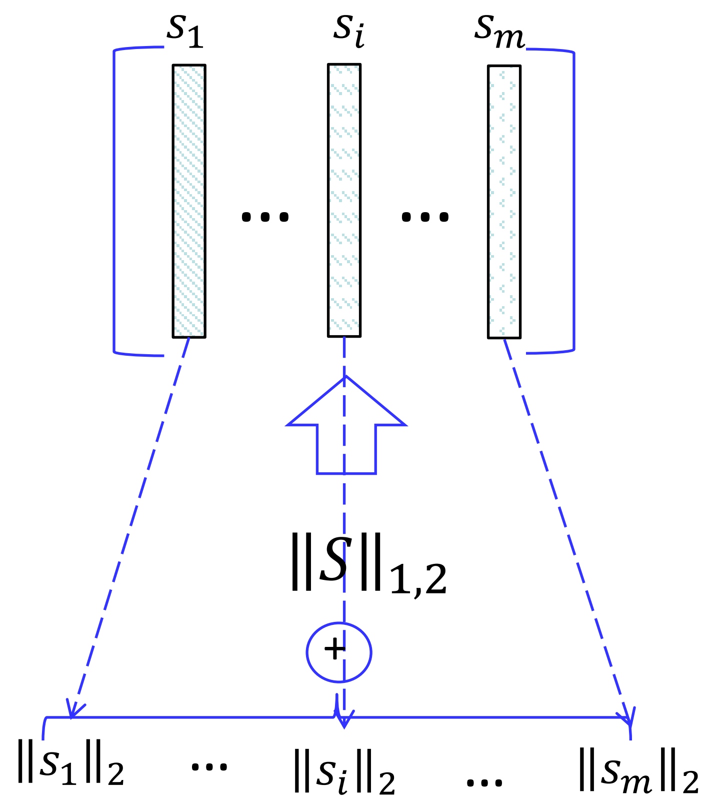
Low-rank structure



$\|L\|_*$

(Sum of singular values in L)

Group sparse structure



Robust MTL Formulation

- Empirical loss on the i-th task, e.g.,

$$\mathcal{L}_i(X_i(l_i + s), y_i) = \|X_i(l_i + s_i) - y_i\|^2$$

- Robust MTL Formulation

$$\underset{L,S}{\text{minimize}} \sum_{i=1}^{m} \mathcal{L}_i(X_i(l_i + s_i), y_i) + \alpha\|L\|_* + \beta\|S\|_{1,2}$$

  - Capture task relatedness via a low-rank structure
  - Identify irrelevant tasks via a group-sparse structure

Performance Bound

- Assumption on the existence of $\kappa_1(s)$ and $\kappa_2(q)$
  - Training data
  - Geometric structure of the coefficient matrices

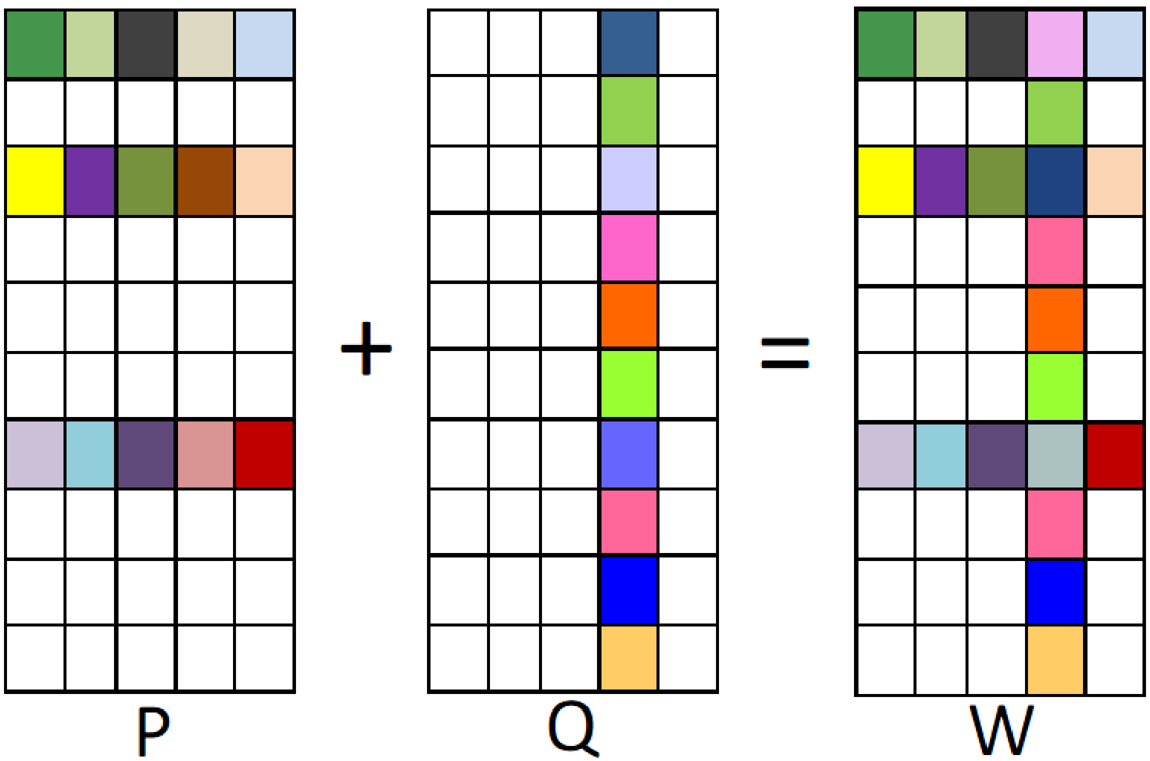
- Performance Bound

$$\frac{1}{T}\sum_{i=1}^{m}\left\|X_i^T(\hat{l}_i+\hat{s}_i)-\hat{f}_i\right\|_2^2 \leqslant (1+\varepsilon)\inf_{\{l_i,s_i\}}\frac{1}{T}\sum_{i=1}^{m}\left\|X_i^T(l_i+s_i)-\hat{f}_i\right\|_2^2+\Phi(\varepsilon)\left(\frac{\alpha^2}{\kappa_1^2(2r)}+\frac{\beta^2}{\kappa_2^2(c)}\right)$$

with the probability of at least $1-me^{-\frac{1}{2}\left(t-d\log\left(1+\frac{t}{d}\right)\right)}$.

Robust Multi-Task Feature Learning

- Simultaneously captures a common set of features among relevant tasks and identifies outlier tasks:



Robust Multi-Task Feature Learning

- Formulation:

$$\min_{W,P,Q} \sum_{i=1}^{m} \frac{1}{mn_i} \left\| X_i^T \mathbf{w}_i - \mathbf{y}_i \right\|^2 + \lambda_1 \|P\|_{1,2} + \lambda_2 \left\| Q^T \right\|_{1,2}$$

$$s.t.\ W = P + Q,$$

- Algorithm:
  - Accelerated Gradient Method
  - Proximal Operator problems:

$$P^k = \arg\min_P \frac{1}{2} \left\| P - \left(R^k - \frac{1}{\eta_k} \nabla_R l(R^k, S^k) \right) \right\|_F^2 + \frac{\lambda_1}{\eta_k} \|P\|_{1,2}$$

$$Q^k = \arg\min_Q \frac{1}{2} \left\| Q - \left(S^k - \frac{1}{\eta_k} \nabla_S l(R^k, S^k) \right) \right\|_F^2 + \frac{\lambda_2}{\eta_k} \left\| Q^T \right\|_{1,2}$$

Robust Multi-Task Feature Learning

- Theoretical Guarantees
  - With probability of at least $1 - \exp\left(-\frac{1}{2}\left(t - dm\log\left(1+\frac{t}{dm}\right)\right)\right)$

$$\sum_{i=1}^{m} \frac{1}{mn}\left\|X_i^T(\hat{\mathbf{p}}_i + \hat{\mathbf{q}}_i) - \mathbf{f}_i^\star\right\|^2 \le \sum_{i=1}^{m} \frac{1}{mn}\left\|X_i^T(\mathbf{p}_i + \mathbf{q}_i) - \mathbf{f}_i^\star\right\|^2$$

$$+ 2\lambda_1 \left\|(\hat{P} - P)^{\mathcal{J}(P)}\right\|_{1,2} + 2\lambda_2 \left\|(\hat{Q}^T - Q^T)^{\mathcal{J}(Q^T)}\right\|_{1,2}.$$

  - With probability of $1 - \exp\left(-\frac{1}{2}\left(t - dm\log\left(1+\frac{t}{dm}\right)\right)\right)$ $(t > 0)$

$$\frac{1}{mn}\left\|X^T \mathrm{vec}(\hat{P} + \hat{Q}) - \mathrm{vec}(F^\star)\right\|_F^2 \le \left(\frac{2\lambda_1\sqrt{r}}{\kappa_1(r)} + \frac{2\lambda_2\sqrt{c}}{\kappa_2(c)}\right)^2$$

Optimization Algorithms

Objective
$$\min f(x) = loss(x) + \lambda \times \Omega(x)$$

- Loss Function $loss(x)$
  - Least Squares Loss
  - Logistic Loss
- Convex and Smooth Penalty $\Omega(x)$
  - Regularized MTL
- Convex but Non-Smooth Penalty $\Omega(x)$
  - $\ell_{2,1}$ −Norm
  - Dirty MTL
  - Trace Norm
- Non-Convex Penalty $\Omega(x)$
  - Convex Relaxation
  - CMTL
  - ASO

Optimization Algorithms

Objective

$$\min f(x) = loss(x) + \lambda \times \Omega(x)$$

- Gradient Descent (GD)
- Accelerated Gradient Method (AGM)
  - Solving Proximal Operator

Gradient Descent

- Gradient descent is an algorithm to solve smooth optimization problems $\min_x f(x)$:
  - Repeat $x_{i+1} = x_i - \gamma_i f'(x_i)$ until convergence criterion is met.
  - $f(x)$ is continuously differentiable with Lipchitz continuous gradient L then if $\gamma_i \leq 1/L$ we can obtain the convergence rate of $O(1/N)$
- Most optimization problems in MTL are non-convex.
- Can we apply gradient descent to non-smooth problems?

Gradient Descent

Smooth Objective
$\min f(x)$

Repeat
$$x_{i+1} = x_i - \gamma_i f'(x_i)$$
until convergence

Equivalent

Repeat
$$x_{i+1} = \arg\min_x M(x_i, \gamma_i)$$
until convergence

Model

$$M(x_i, \gamma_i) = \underline{[f(x_i) + \langle f'(x_i), x - x_i \rangle]} + \frac{1}{2\gamma_i} \|x - x_i\|_2^2$$

1st order
Taylor expansion

Regularization

Gradient Descent

Objective

$$\min f(x) = \text{loss}(x) + \lambda \times \Omega(x)$$

Composite Model

$$M(x_i, \gamma_i) = \underbrace{[f(x_i) + \langle f'(x_i), x - x_i \rangle]}_{\text{1st order Taylor expansion}} + \underbrace{\frac{1}{2\gamma_i} \|x - x_i\|_2^2}_{\text{Regularization}} + \underbrace{\lambda \times \Omega(x)}_{\text{Non Smooth Penalty}}$$

Repeat

$$x_{i+1} = \arg\min_x M(x_i, \gamma_i)$$

until convergence

- Using the gradient descent with composite model to solve non-smooth optimization problems.
- Convergence Rate O(1/N)

Gradient Descent

Repeat

$$x_{i+1} = \arg\min_x M(x_i, \gamma_i)$$

until convergence

Composite Model

$$M(x_i, \gamma_i) = [f(x_i) + \langle f'(x_i), x - x_i \rangle] + \frac{1}{2\gamma_i}\|x - x_i\|_2^2 + \lambda \times \Omega(x)$$

Equivalent

Proximal Operator (Moreau, 1965)

$$x_{i+1} = \arg\min_x \frac{1}{2}\|x - v\|_2^2 + \rho \times \Omega(x)$$

$$v = x_i - \gamma_i loss'(x_i)$$

$$\rho = \gamma_i \lambda$$

Accelerated Gradient Method (AGM)

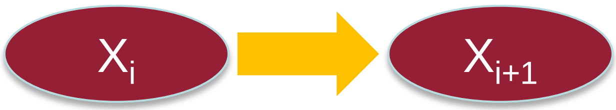
- A faster extension of gradient descent (Nesterov, 1983; Nemirovski, 1994; Nesterov, 2004)

Gradient Descent

Repeat

$$x_{i+1} = x_i - \gamma_i f'(x_i)$$

until convergence



Convergence: O(1/N)

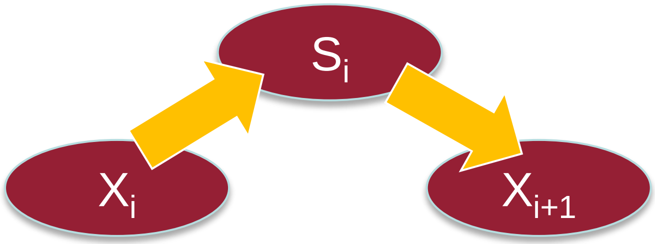
Accelerated Gradient Descent

Repeat

$$s_i = x_i + \alpha_i(x_i - x_{i-1})$$

$$x_{i+1} = x_i - \gamma_i f'(x_i)$$

until convergence



Convergence: O(1/N^2)

Accelerated Gradient Method (AGM)

Composite Model

$$M(x_i, \gamma_i) = [f(x_i) + \langle f'(x_i), x - x_i \rangle] + \frac{1}{2\gamma_i} \|x - x_i\|_2^2 + \lambda \times \Omega(x)$$

Gradient Descent

Repeat

$$x_{i+1} = \arg\min_x M(x_i, \gamma_i)$$

until convergence

Convergence: O(1/N)

Accelerated Gradient Descent

Repeat

$$s_i = x_i + \alpha_i (x_i - x_{i-1})$$

$$x_{i+1} = \arg\min_x M(s_i, \gamma_i)$$

until convergence

Convergence: O(1/N^2)

Can the proximal operator *M* be computed efficiently?

Optimization with Non-Convex Objectives

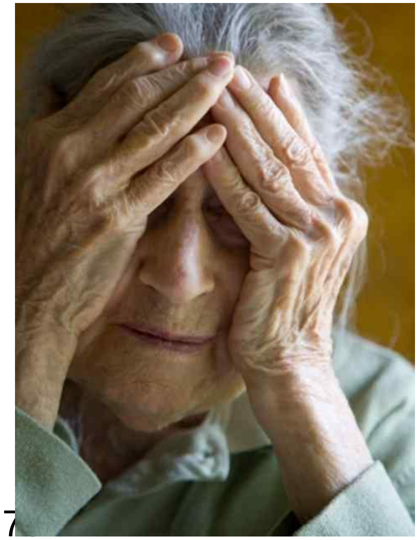
- In multi-task learning, optimization objectives involved may be non-convex (e.g. clustered multi-task learning).
- Directly applying convex optimization techniques may obtain suboptimal solutions.
- Convex Relaxation
  - General non-convex problem: find convex envelope
    - Rank minimization → Trace-norm minimization
  - Difference of convex (DC) problem: Convex-Concave Procedure (CCCP)[Yuille and Rangarajan NIPS 2001]
    - $\ell_1/\ell_{0.5}$-regularization → Reweighted group Lasso

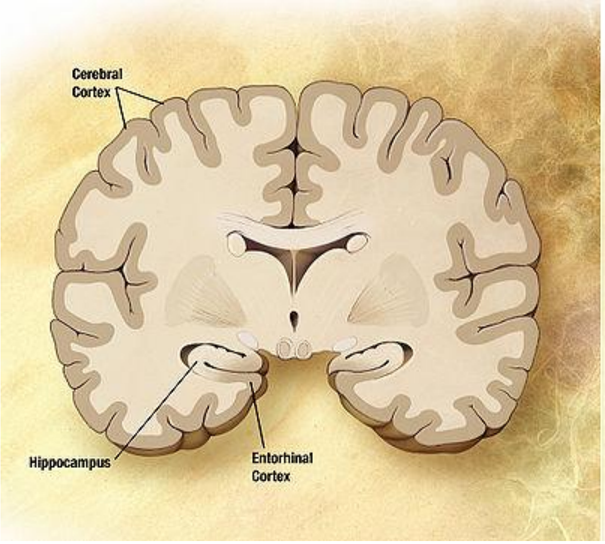
Difference of Convex (DC) Programming

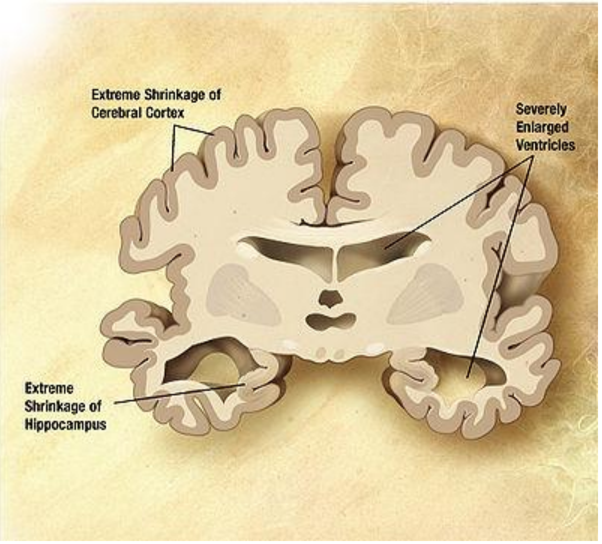
- The objective can be written in the form:
  - $\min_{x} f(x) - g(x)$
  - $f(x)$ and $g(x)$ are convex functions.
- We linearize $g(x)$ using the 1st order Taylor expansion at x':
  - $f(x) - g(x) = f(x) - g(x') - \langle \nabla g(x'), x - x' \rangle$
- In every iteration of CCCP, we minimize the upper bound:
  - $x_{k+1} = \mathrm{argmin}_x f(x) - \langle \nabla g(x_k), x \rangle$
- The objective function is guaranteed to decrease

Case Study: Disease Progression

- Alzheimer's Disease (AD) is
  - the most common type of dementia;
  - severe neurodegenerative disorder;
  - definitive diagnosed only through brain biopsy or autopsy;
  - clinically diagnosed by clinical/cognitive measures including MMSE and ADAS-Cog.

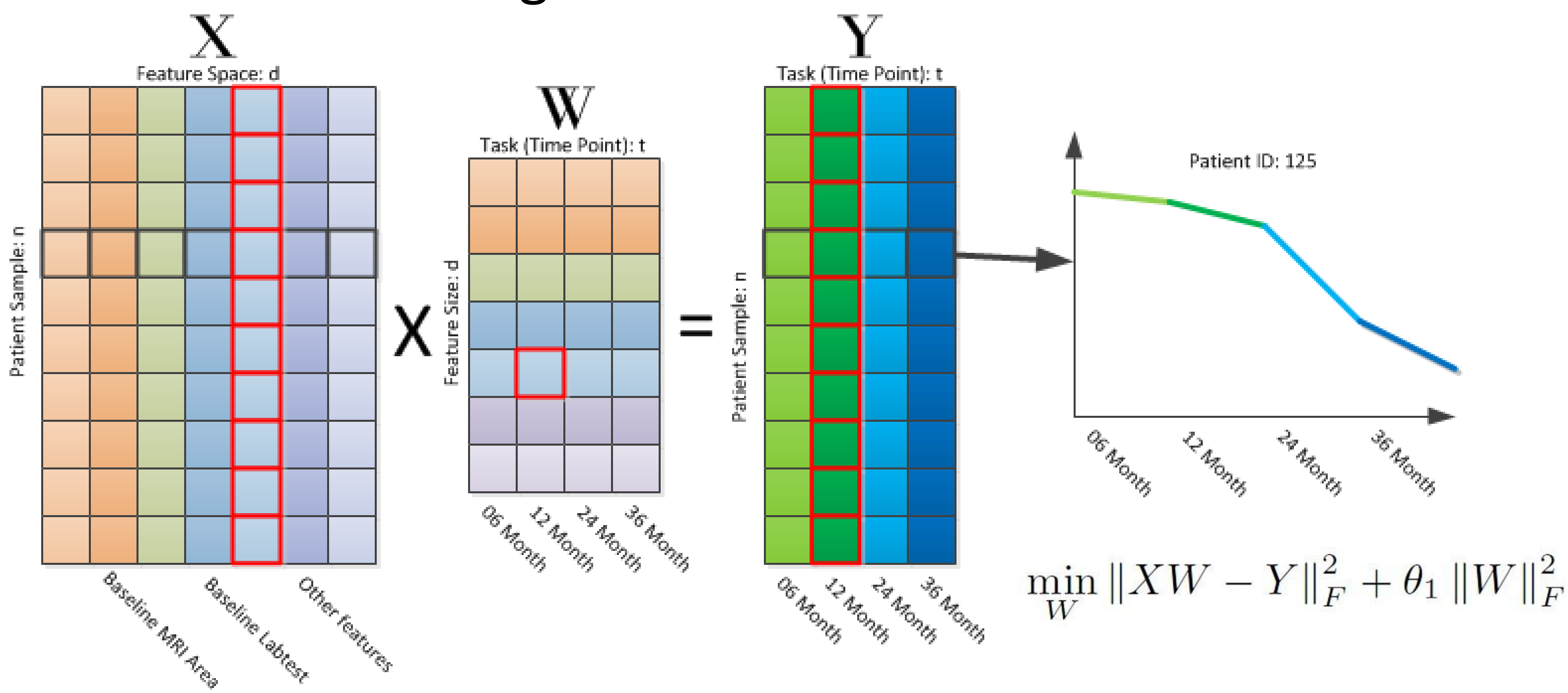






Modeling Disease Progression

- The prediction of cognitive scores at each time point can be modeled as a regression task.

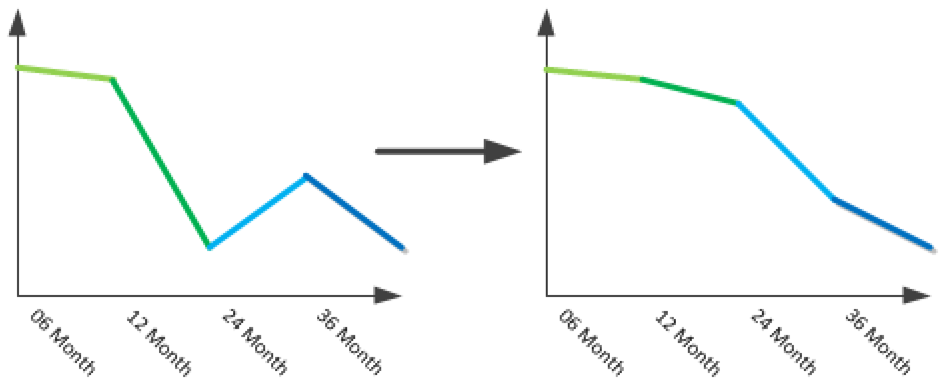


$$\min_{W} \|XW - Y\|_F^2 + \theta_1 \|W\|_F^2$$

- Motivation of using multi-task learning: the ability to explore inherent relationships among related tasks and enforce such knowledge using proper regularizations.

Temporal Smoothness

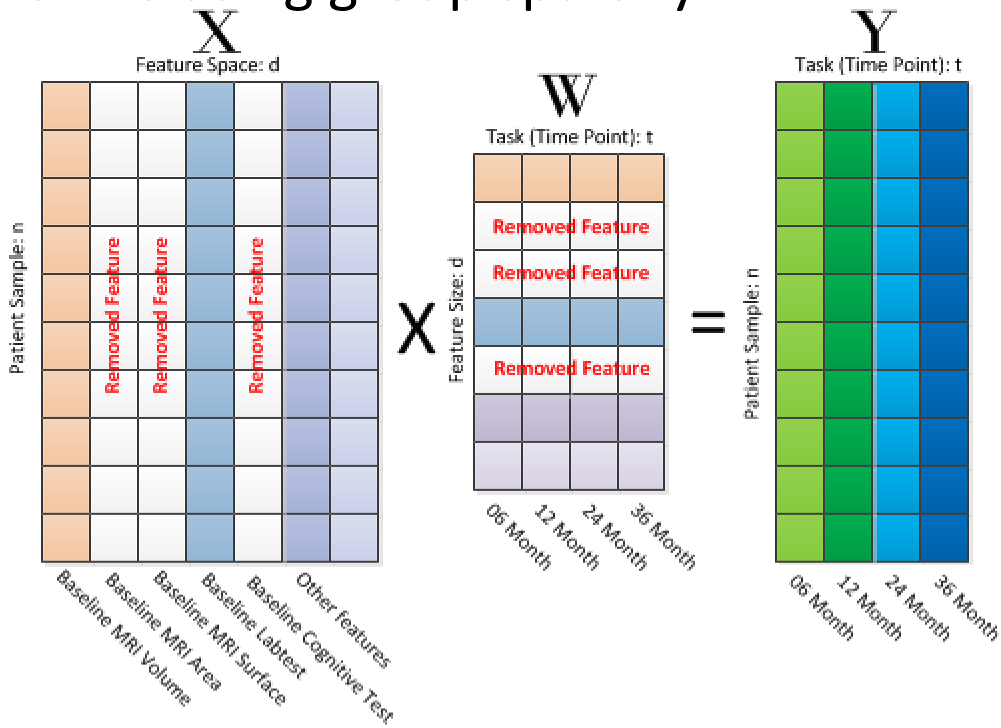
- Prior knowledge: the change of cognitive scores should be small for a patient. The scores should not fluctuate:



$$\min_{W} \|XW - Y\|_F^2 + \theta_1 \|W\|_F^2 + \theta_2 \sum_{i=1}^{t-1} \left\| w^i - w^{i+1} \right\|_2^2$$

Temporal Group Lasso

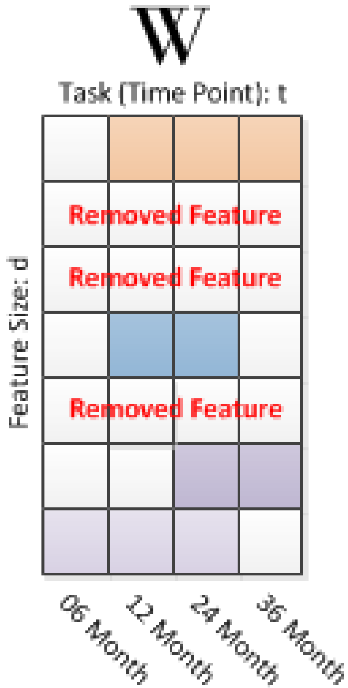
- Assumption: there is only a small subset of features related to disease progression, shared among tasks.
- Achieve this using group sparsity:



$$\min_{W} L(W) + \theta_1 \|W\|_F^2 + \theta_2 \left\| RW^T \right\|_F^2 + \theta_3 \|W\|_{2,1}$$

Fused Sparse Group Lasso

- Goal: find temporal patterns of the biomarkers in the disease progression.
- Simultaneous feature selection via Fused Lasso:
  - a common set of biomarkers for multiple time points
  - specific sets of biomarkers for different time points
- Incorporate the temporal smoothness via Group Lasso.



Fused Sparse Group Lasso

- The convex formulation:

$$\min_W L(W) + \lambda_1 \|W\|_1 + \lambda_2 \left\|RW^T\right\|_1 + \lambda_3 \|W\|_{2,1}$$

- Non-convex formulations:
  - Reduce shrinkage bias
  - Closer to the optimal l_0-norm
  - Fewer tuning parameters

$$\min_W L(W) + \lambda \sum_{i=1}^{d} \sqrt{\|\mathbf{w}_i\|_1} + \gamma \|RW^T\|_1$$

$$\min_W L(W) + \lambda \sum_{i=1}^{d} \sqrt{\|R\mathbf{w}_i^T\|_1 + \beta \|\mathbf{w}_i\|_1}$$

Performance

- MTL outperforms STL
- Fused sparse group Lasso formulations achieve better performance than Temporal group Lasso

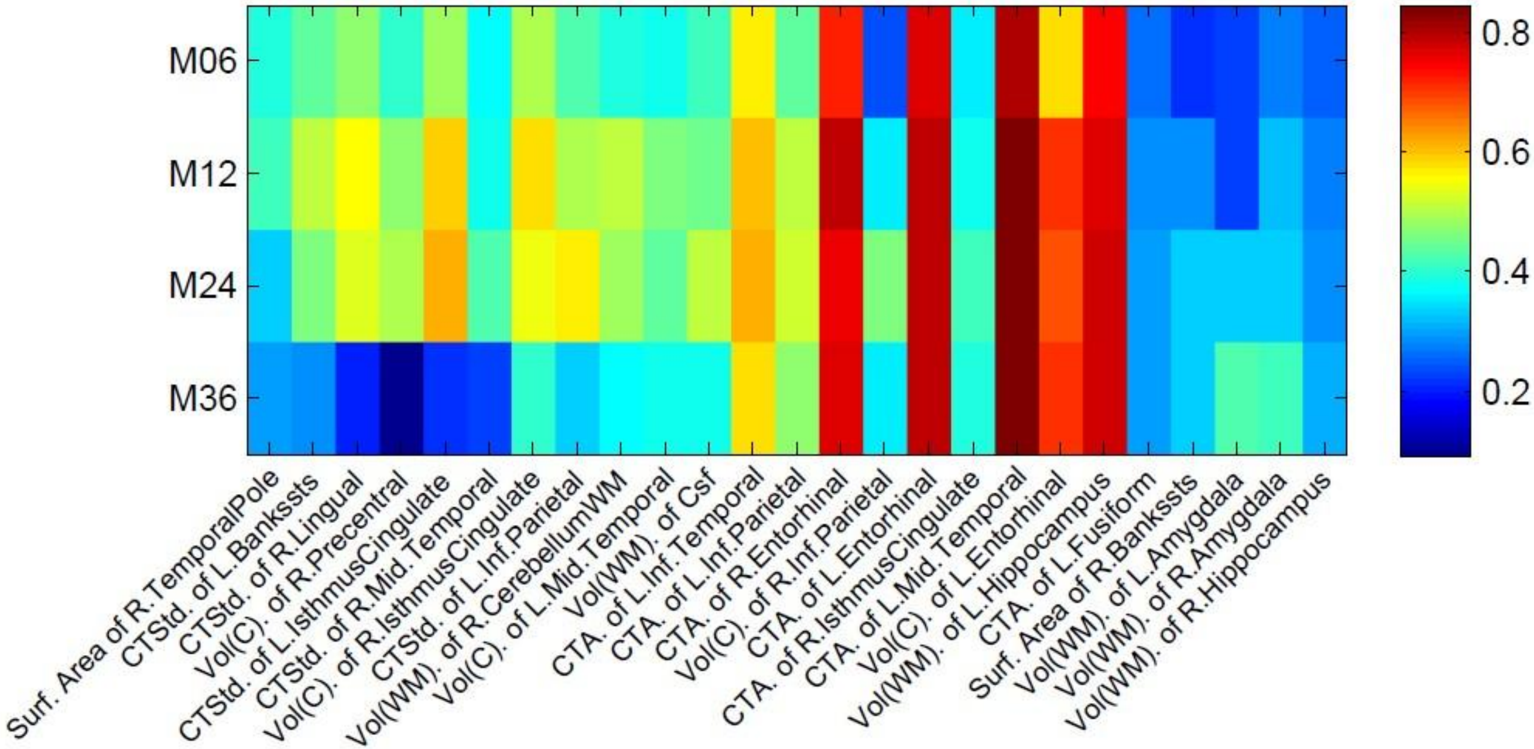
	Ridge	TGL	cFSGL	nFSGL1	nFSGL2
Target: MMSE					
nMSE	0.548 ± 0.057	0.449 ± 0.045	$\mathbf{0.400 \pm 0.053}$	0.412 ± 0.054	0.408 ± 0.056
R	0.689 ± 0.030	0.755 ± 0.029	0.790 ± 0.032	0.788 ± 0.031	$\mathbf{0.792 \pm 0.031}$
M06 MSE	2.269 ± 0.207	$\mathbf{2.038 \pm 0.262}$	2.069 ± 0.209	2.149 ± 0.194	2.181 ± 0.201
M12 MSE	3.266 ± 0.556	2.923 ± 0.643	2.803 ± 0.662	2.835 ± 0.662	$\mathbf{2.793 \pm 0.659}$
M24 MSE	3.494 ± 0.599	3.363 ± 0.733	3.016 ± 0.624	3.031 ± 0.604	$\mathbf{2.979 \pm 0.546}$
M36 MSE	4.003 ± 0.853	3.768 ± 0.962	3.302 ± 0.781	3.263 ± 0.785	$\mathbf{3.211 \pm 0.786}$
M48 MSE	4.328 ± 1.310	3.631 ± 1.226	2.787 ± 0.871	2.780 ± 0.855	$\mathbf{2.766 \pm 0.826}$
Target: ADAS-Cog					
nMSE	0.532 ± 0.095	0.464 ± 0.067	0.404 ± 0.055	0.386 ± 0.060	$\mathbf{0.381 \pm 0.057}$
R	0.705 ± 0.043	0.747 ± 0.033	0.791 ± 0.026	$\mathbf{0.809 \pm 0.023}$	$\mathbf{0.809 \pm 0.023}$
M06 MSE	5.213 ± 0.522	4.820 ± 0.489	4.543 ± 0.374	4.458 ± 0.354	$\mathbf{4.428 \pm 0.351}$
M12 MSE	6.079 ± 0.775	5.813 ± 0.697	5.363 ± 0.595	5.183 ± 0.597	$\mathbf{5.136 \pm 0.617}$
M24 MSE	7.409 ± 1.154	6.835 ± 1.052	6.456 ± 0.974	6.174 ± 0.943	$\mathbf{6.153 \pm 0.911}$
M36 MSE	7.143 ± 1.351	6.938 ± 1.363	6.101 ± 1.071	$\mathbf{5.819 \pm 0.945}$	5.879 ± 0.972
M48 MSE	6.644 ± 2.750	6.000 ± 2.738	5.751 ± 2.081	5.889 ± 1.848	$\mathbf{5.837 \pm 2.160}$

Performance

	Ridge	TGL	cFSGL	nFSGL1	nFSGL2
Target: MMSE					
nMSE	0.404 ± 0.056	0.320 ± 0.044	0.311 ± 0.042	0.308 ± 0.046	$\mathbf{0.303 \pm 0.046}$
R	0.788 ± 0.032	0.839 ± 0.027	0.841 ± 0.026	0.839 ± 0.027	$\mathbf{0.843 \pm 0.027}$
M06 MSE	2.188 ± 0.194	1.943 ± 0.161	1.912 ± 0.153	1.935 ± 0.150	$\mathbf{1.906 \pm 0.149}$
M12 MSE	2.744 ± 0.638	2.366 ± 0.722	2.356 ± 0.713	2.374 ± 0.696	$\mathbf{2.326 \pm 0.707}$
M24 MSE	3.113 ± 0.560	2.821 ± 0.664	2.823 ± 0.656	2.766 ± 0.601	$\mathbf{2.730 \pm 0.604}$
M36 MSE	3.150 ± 0.517	2.933 ± 0.657	2.878 ± 0.640	$\mathbf{2.755 \pm 0.550}$	2.792 ± 0.523
M48 MSE	3.639 ± 0.959	3.544 ± 1.136	3.098 ± 1.013	$\mathbf{2.942 \pm 0.928}$	2.961 ± 0.969
Target: ADAS-Cog					
nMSE	0.314 ± 0.036	0.278 ± 0.034	$\mathbf{0.233 \pm 0.035}$	0.238 ± 0.035	0.243 ± 0.035
R	0.840 ± 0.015	0.868 ± 0.016	$\mathbf{0.886 \pm 0.014}$	0.884 ± 0.015	0.880 ± 0.013
M06 MSE	3.972 ± 0.415	3.560 ± 0.469	3.553 ± 0.375	3.659 ± 0.356	$\mathbf{3.535 \pm 0.403}$
M12 MSE	4.365 ± 0.469	4.080 ± 0.598	$\mathbf{3.678 \pm 0.389}$	3.739 ± 0.367	3.742 ± 0.430
M24 MSE	6.028 ± 1.128	5.888 ± 1.641	5.115 ± 1.277	$\mathbf{5.111 \pm 1.222}$	5.257 ± 1.337
M36 MSE	5.824 ± 1.076	5.639 ± 1.339	4.747 ± 0.957	$\mathbf{4.737 \pm 0.917}$	5.055 ± 1.033
M48 MSE	6.192 ± 2.327	6.337 ± 2.487	5.065 ± 1.446	$\mathbf{4.968 \pm 1.339}$	5.404 ± 1.802

Longitudinal Stability Selection on ADAS-Cog

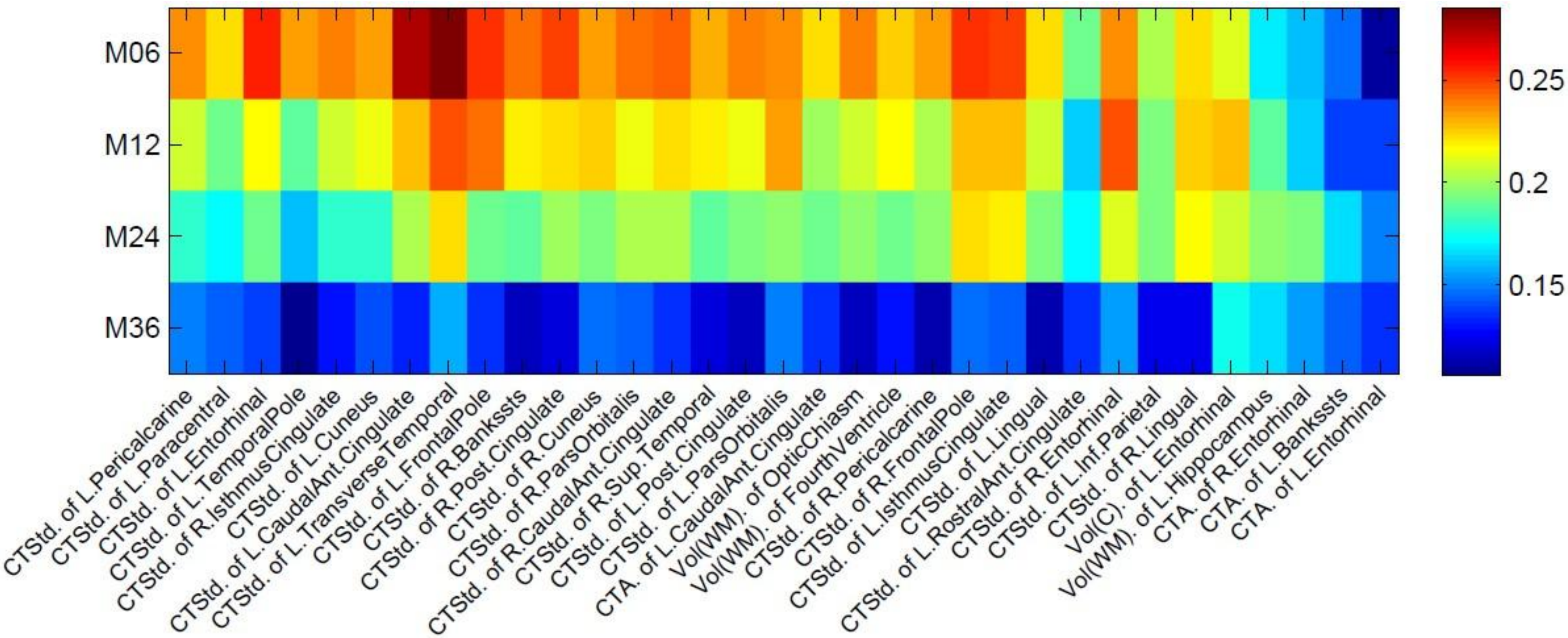
- Using FSGL
- From the distribution of stability scores, we can observe temporal patters of MRI biomarkers.



(a) Target: ADAS-Cog (25 stable features)

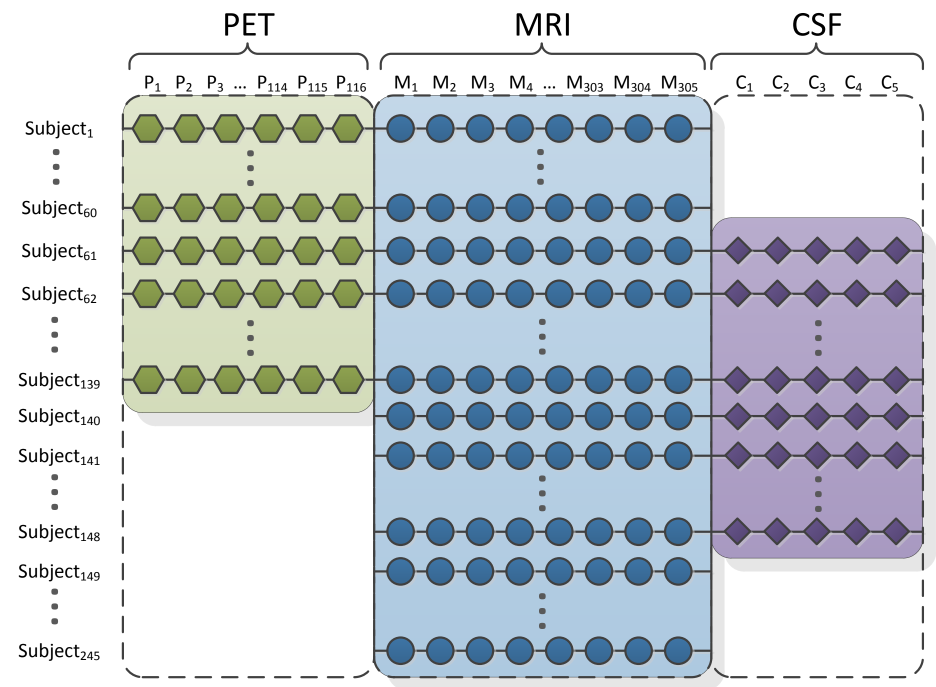
Longitudinal Stability Selection on MMSE

- From the distribution of stability scores, we can observe temporal patters of MRI biomarkers.



Case Study: Missing Data in Multi-Source Learning

- In many applications, multiple data sources may suffer from a considerable amount of missing data.
- In ADNI, over half of the subjects lack CSF measurements; an independent half of the subjects do not have FDG-PET.



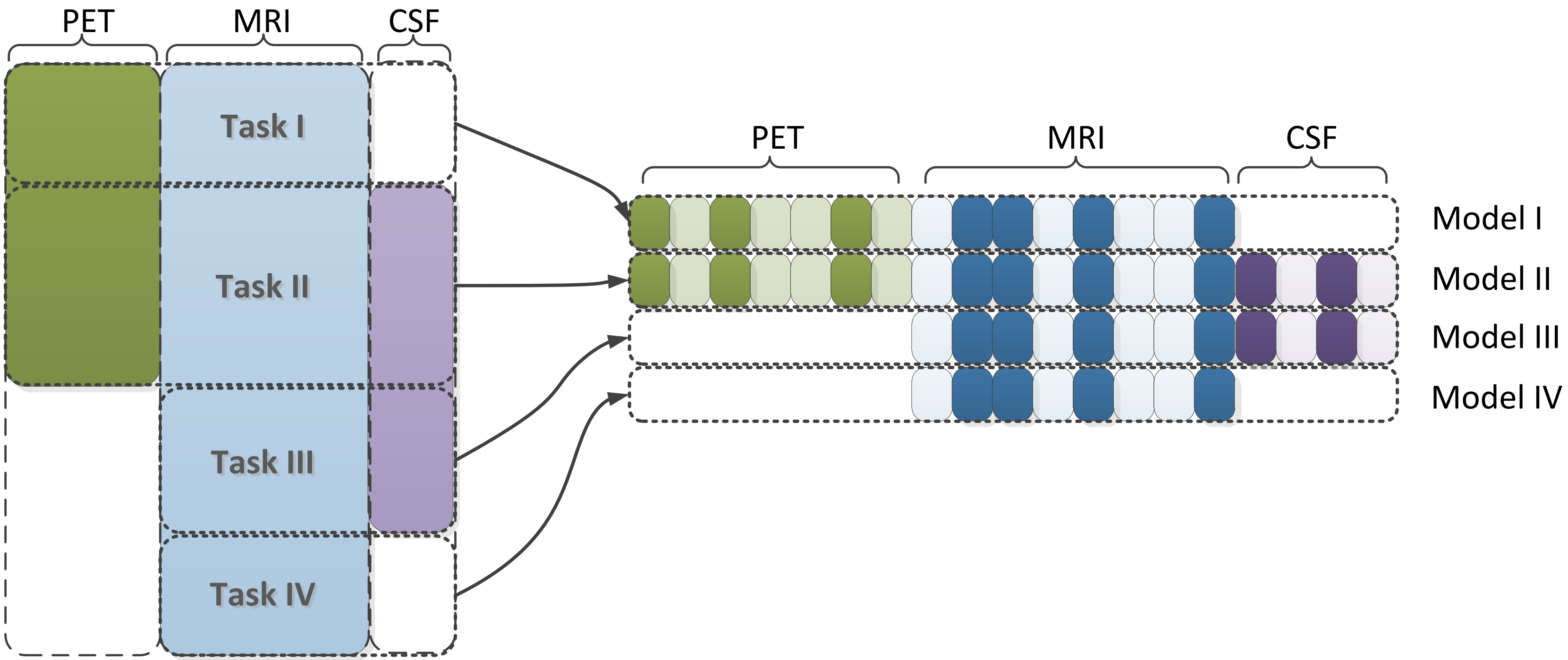
Challenges

- Simply removing samples with missing values will dramatically reduces the number of samples in the analysis.
- Plus, the resource and time devoted to those subjects with incomplete data are totally wasted.
- Estimating the entire chunk of missing values is very challenging.

Incomplete Multi Source Feature Learning (iMSF)

- A “row-wise” strategy
  - We first partition the samples into multiple blocks, one for each combination of data sources available
  - We then build one different model for each block of data
  - Using multi-task techniques, all models involving a specific source are constrained to select a common set of features for that particular source

Overview of iMSF

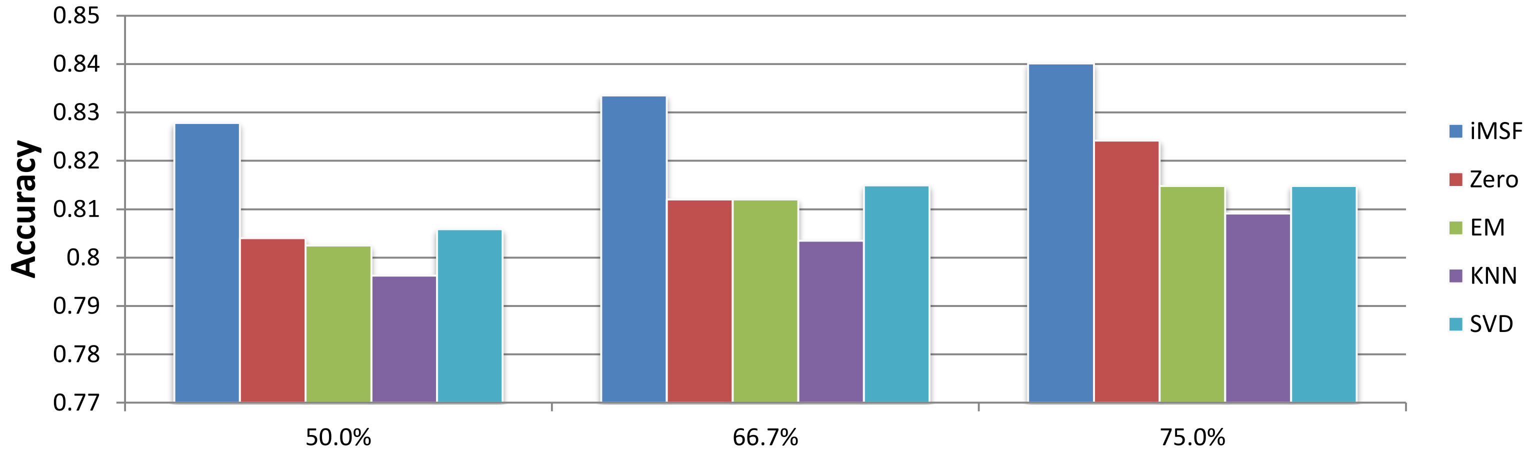


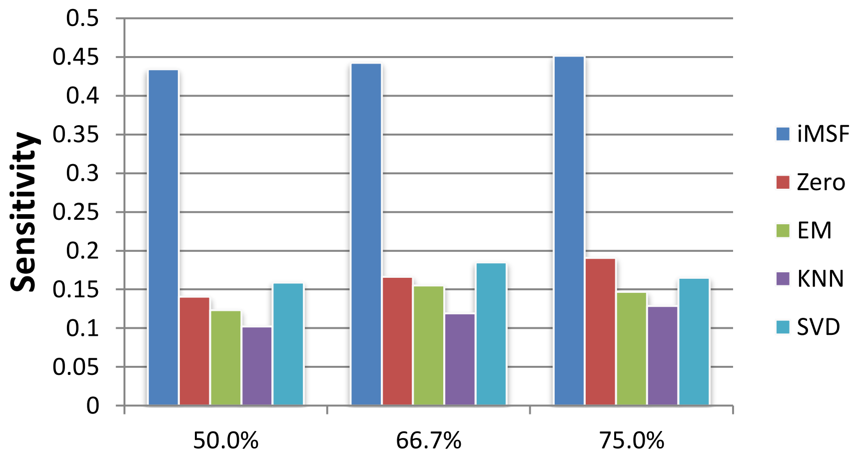
iMSF: the Formulation

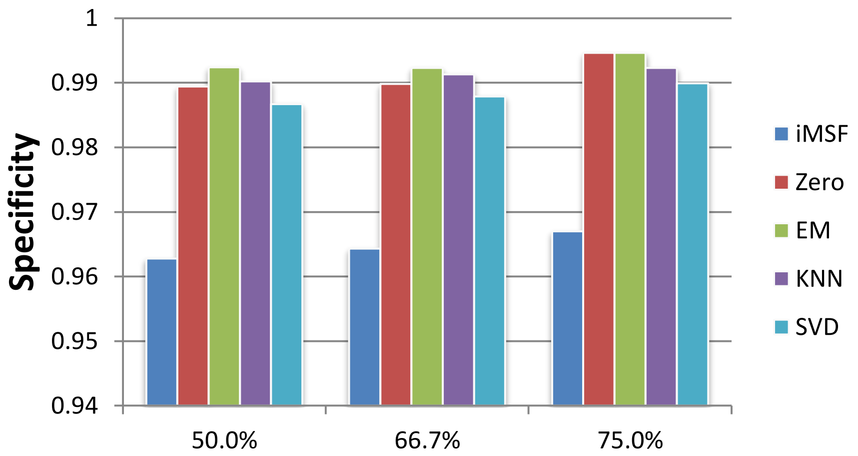
- Suppose the data set is divided into m tasks: $T^i = \{x_j^i, y_j^i\}$, $i = 1 \dots m, j = 1 \dots N_i$, where N_i is the number of subjects in the i-th task
- Denote β^i as the weight vector for the i-th task
- $\beta_{I(s,k)}$ denotes all the model parameters corresponding to the k-th feature in the s-th data source
- We Solve:

$$\min_{\beta} \frac{1}{m} \sum_{i=1}^{m} \frac{1}{N_i} \sum_{j=1}^{N_i} L\left(x_j^i, y_j^i, \beta^i\right) + \lambda \sum_{s=1}^{S} \sum_{k=1}^{p_s} \left\|\beta_{I(s,k)}\right\|_2$$

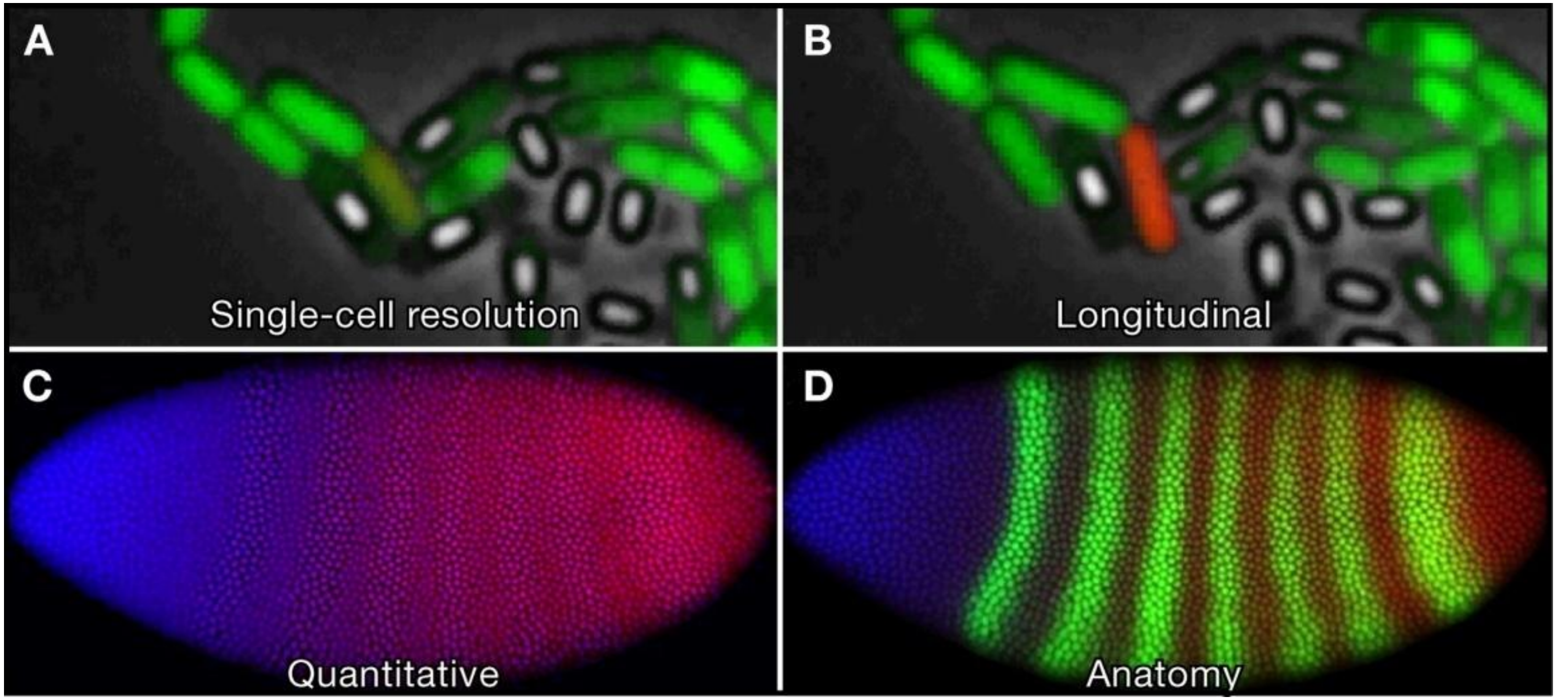
iMSF: Performance





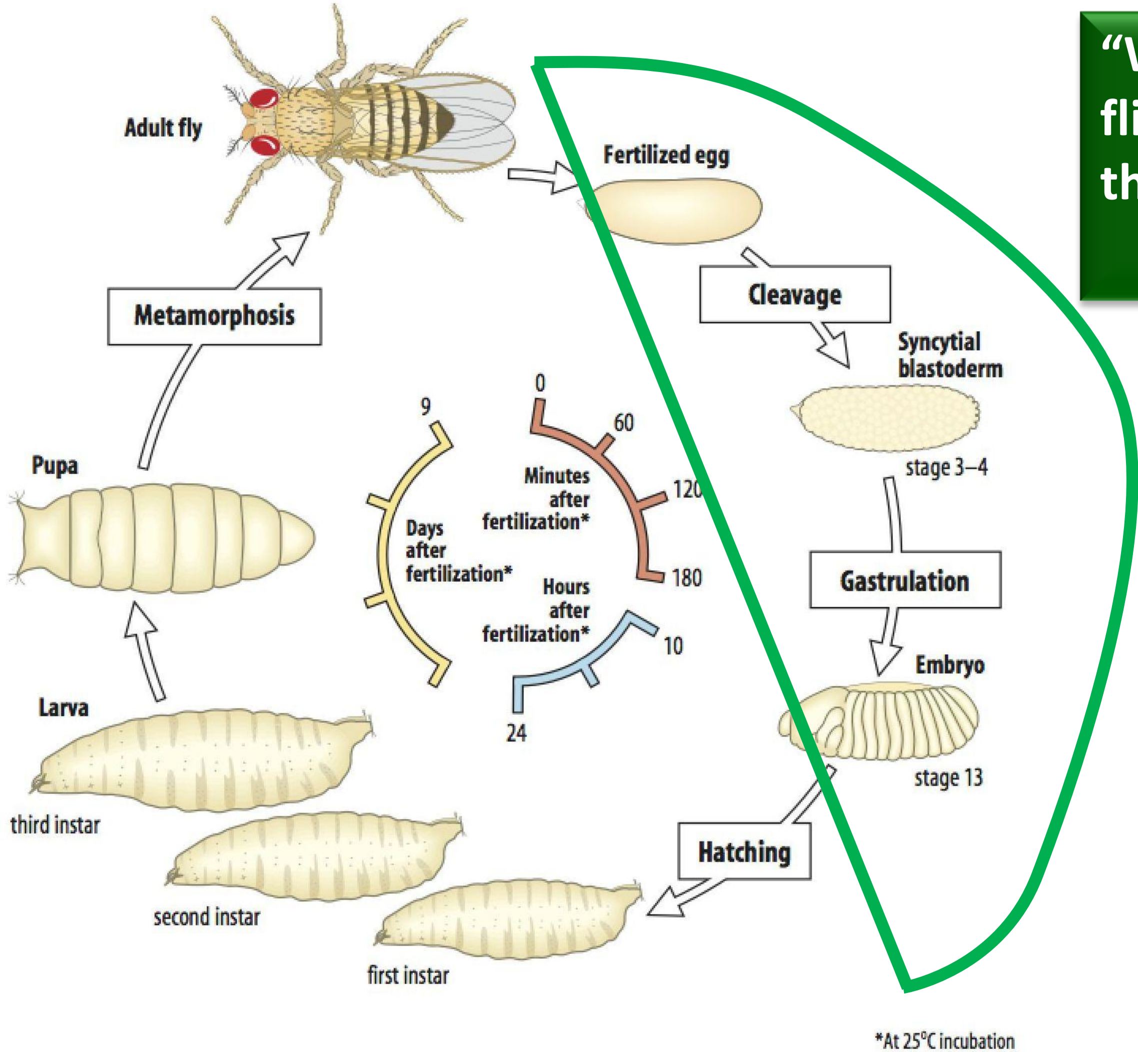


Case Study: *Drosophila* Gene Expression Image Analysis



[Megason and Fraser (2007) Cell]

Life cycle of fruit fly *Drosophila melanogaster*



"We are much more like flies in our development than you might think."
L. Wolpert

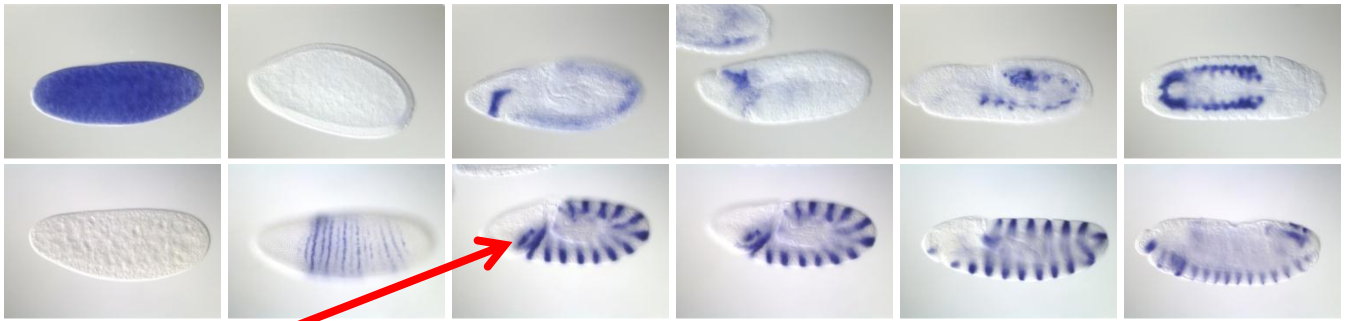
embryogenesis

[Wolpert *et al*. (2006) Principles of Development]

Drosophila gene expression pattern images

bgm

en



Stage 1-3 Stage 4-6 Stage 7-8 Stage 9-10 Stage 11-12 Stage 13-

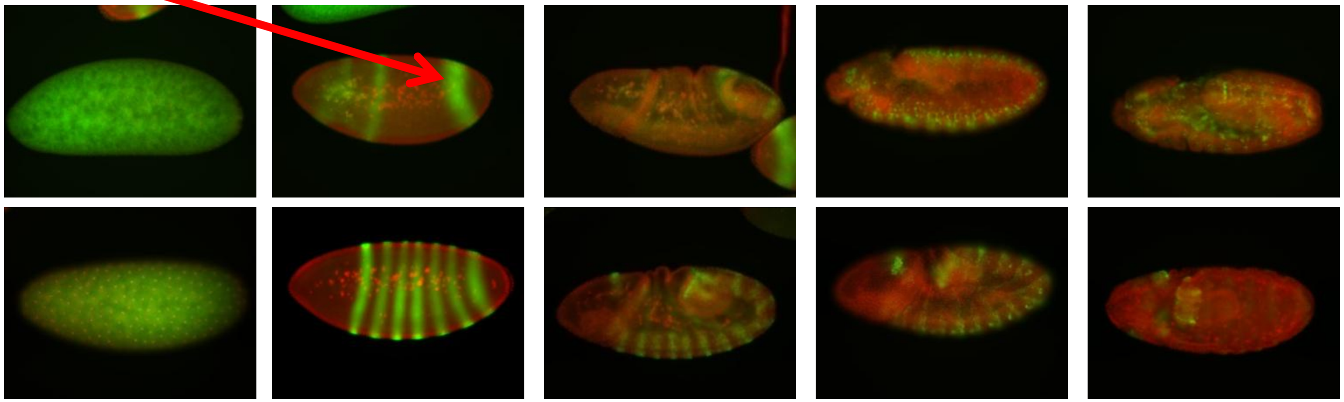
Berkeley *Drosophila* Genome Project (BDGP)

http://www.fruitfly.org/

Expressions

hb

runt



Stage 1-3 Stage 4-5 Stage 6-7 Stage 8-9 Stage 10-

Fly-FISH

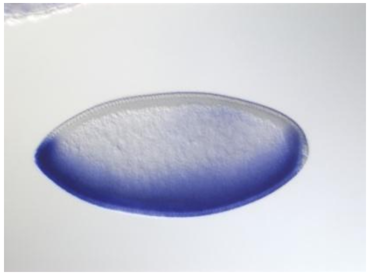
http://fly-fish.ccbr.utoronto.ca/

[Tomancak *et al*. (2002) Genome Biology; Lécuyer *et al*. (2007) Cell]

Comparative image analysis

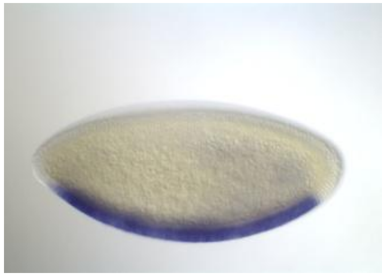
	Twist	*heartless*	*stumps*

stage 4-6



anterior endoderm AISN
trunk mesoderm AISN
subset
cellular blastoderm
mesoderm AISN

dorsal ectoderm AISN
procephalic ectoderm AISN
subset
cellular blastoderm
mesoderm AISN



anterior endoderm AISN
trunk mesoderm AISN
head mesoderm AISN

stage 7-8

trunk mesoderm PR
head mesoderm PR
anterior endoderm anlage

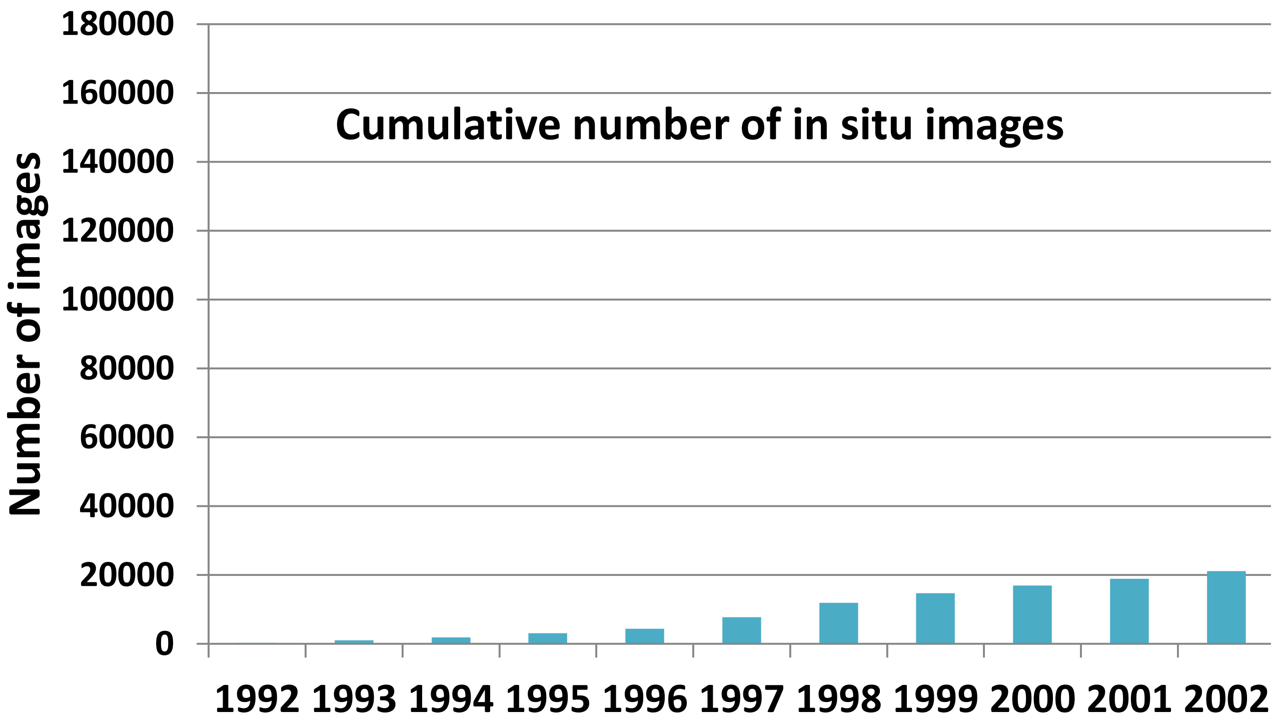
trunk mesoderm PR
head mesoderm PR

yolk nuclei
trunk mesoderm PR
head mesoderm PR
anterior endoderm anlage

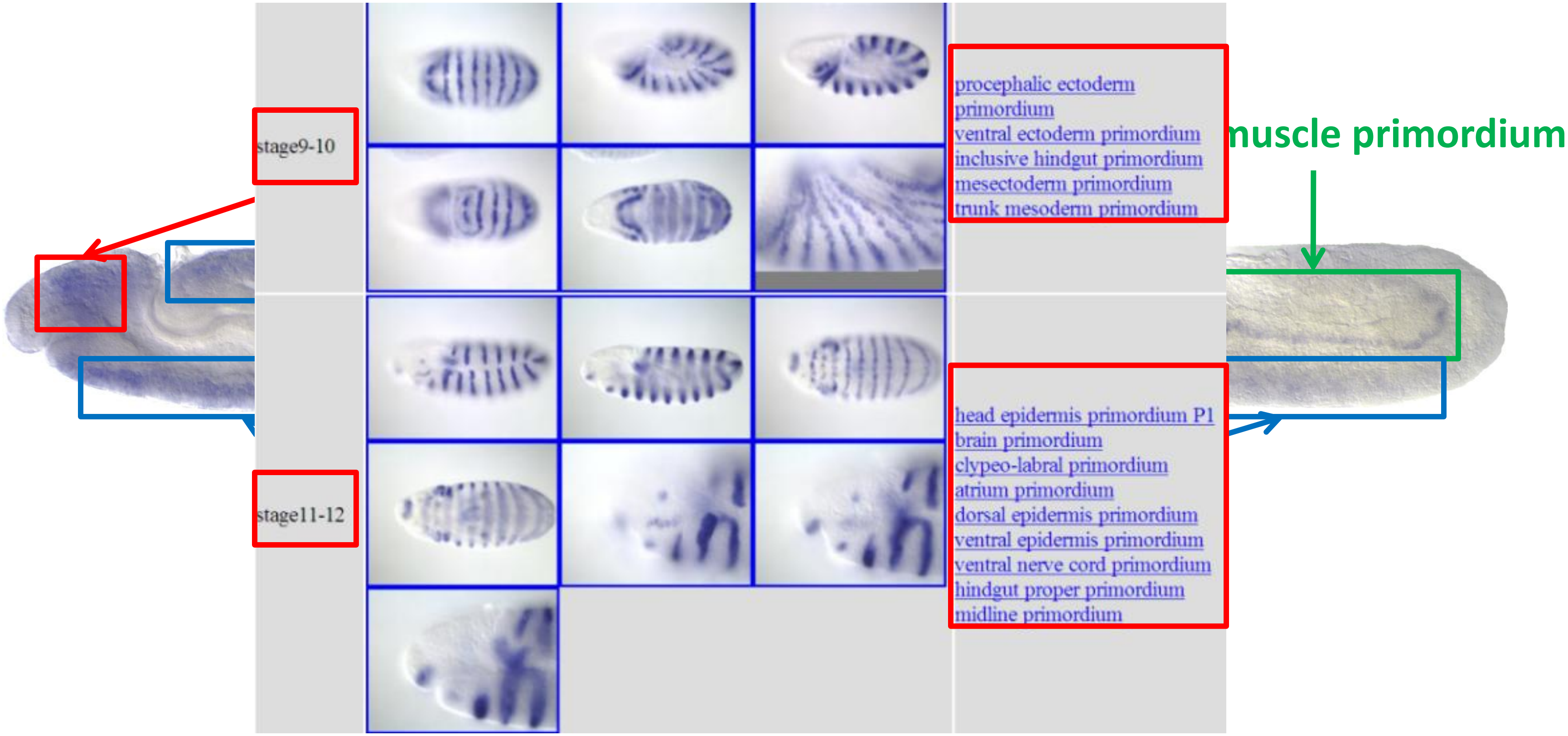
We need the spatial and temporal annotations of expressions

[Tomancak *et al*. (2002) Genome Biology; Sandmann *et al*. (2007) Genes & Dev.]

Challenges of manual annotation



Spatial keywords annotation



Multiple keywords are associated with multiple images

Exact correspondences among keywords and images are NOT given

- Prior approaches assume all keywords are associated with all images
  - Zhou and Peng (2007) Bioinformatics

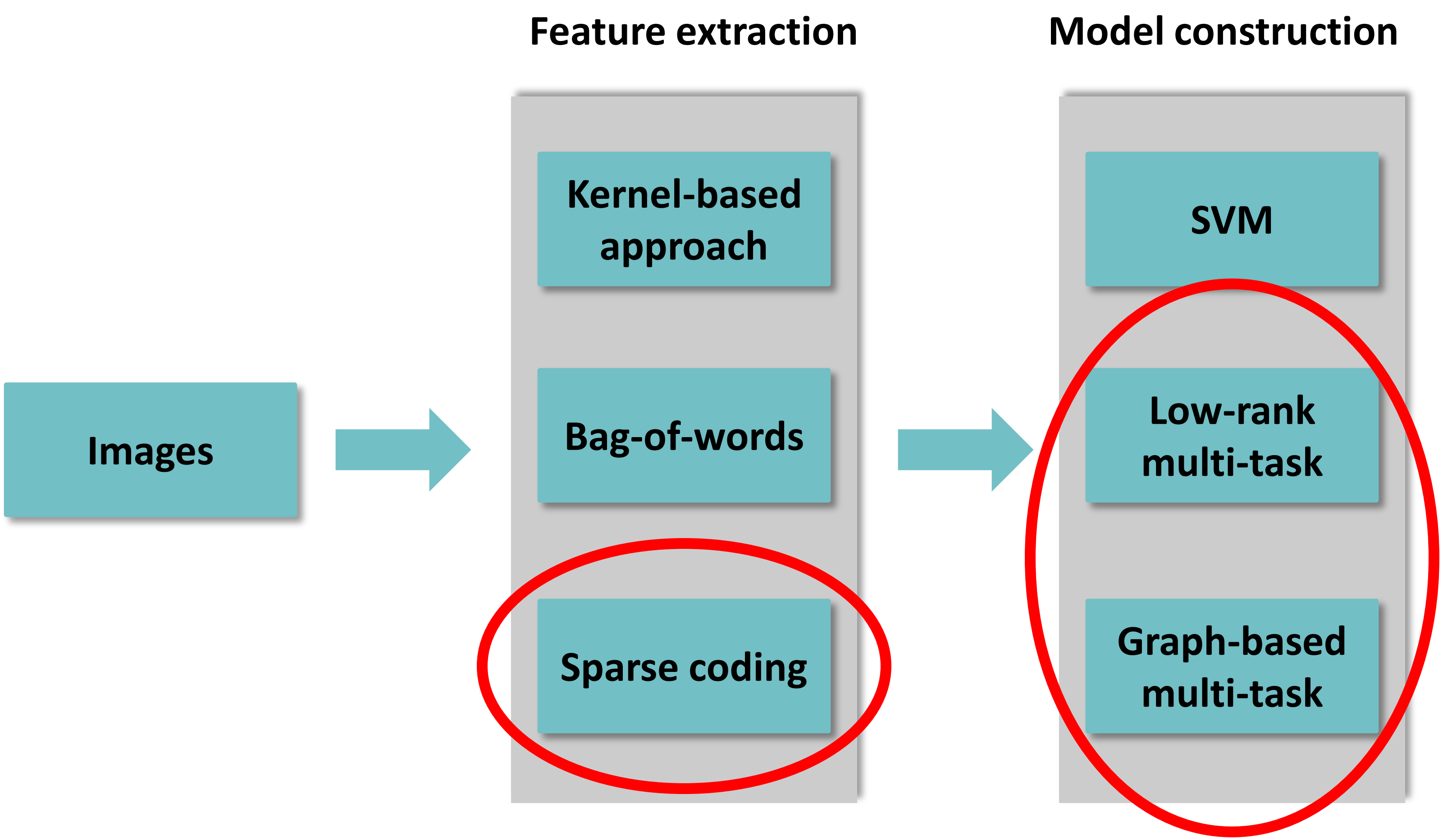
What are the challenges?

"We used human annotation, rather than automated approaches based on pattern recognition algorithms, because of the overwhelming complexity of annotation. Variation in morphology and incomplete knowledge of the shape and position of various embryonic structures make computational approaches impracticable at present."

P. Tomancak *et al.* (2002) Genome Biology

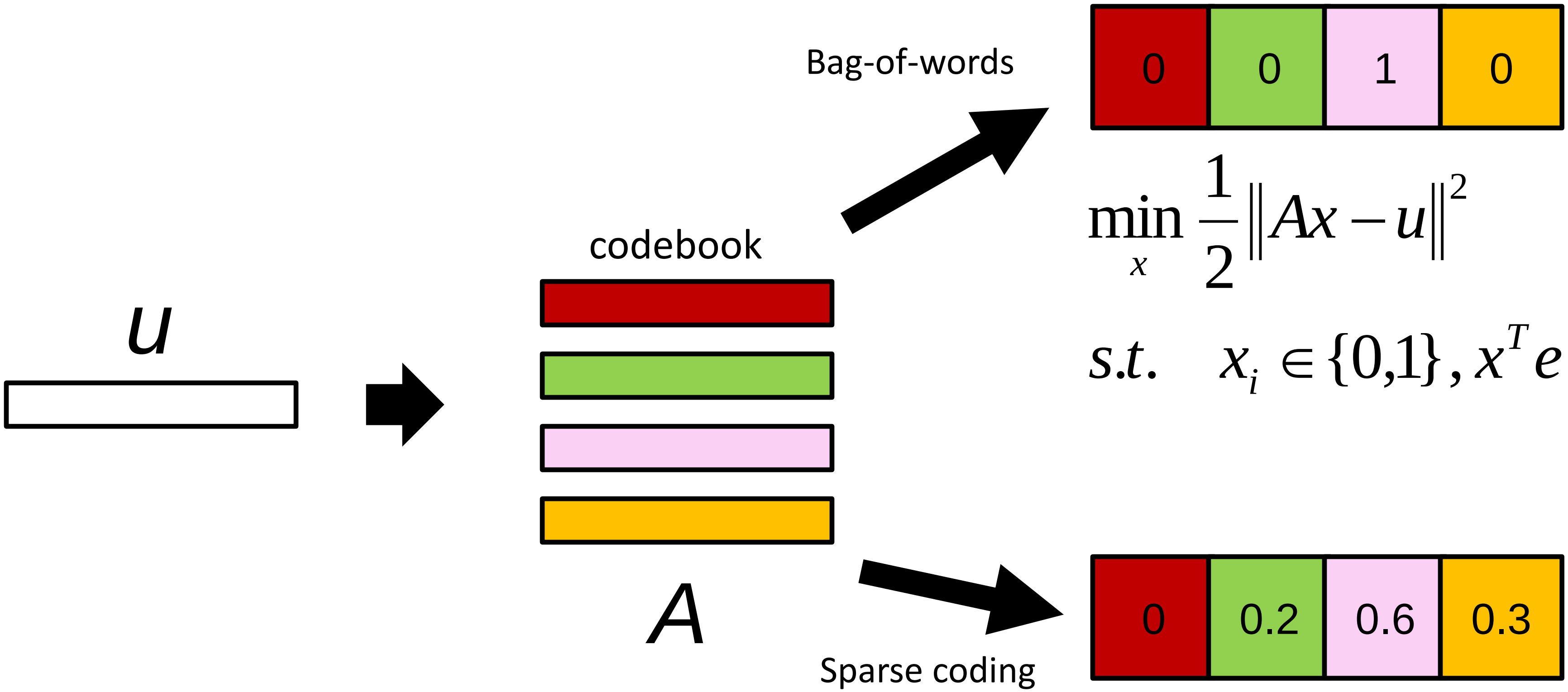
Local invariant features → High-level representations → Multi-task learning

Method outline



[Ji *et al*. (2008) Bioinformatics; Ji *et al*. (2009) BMC Bioinformatics; Ji *et al*. (2009) NIPS]

From bag-of-words to sparse coding



$$\min_x \frac{1}{2}\|Ax-u\|^2$$

$$s.t. \quad x_i \in \{0,1\}, x^T e = 1$$

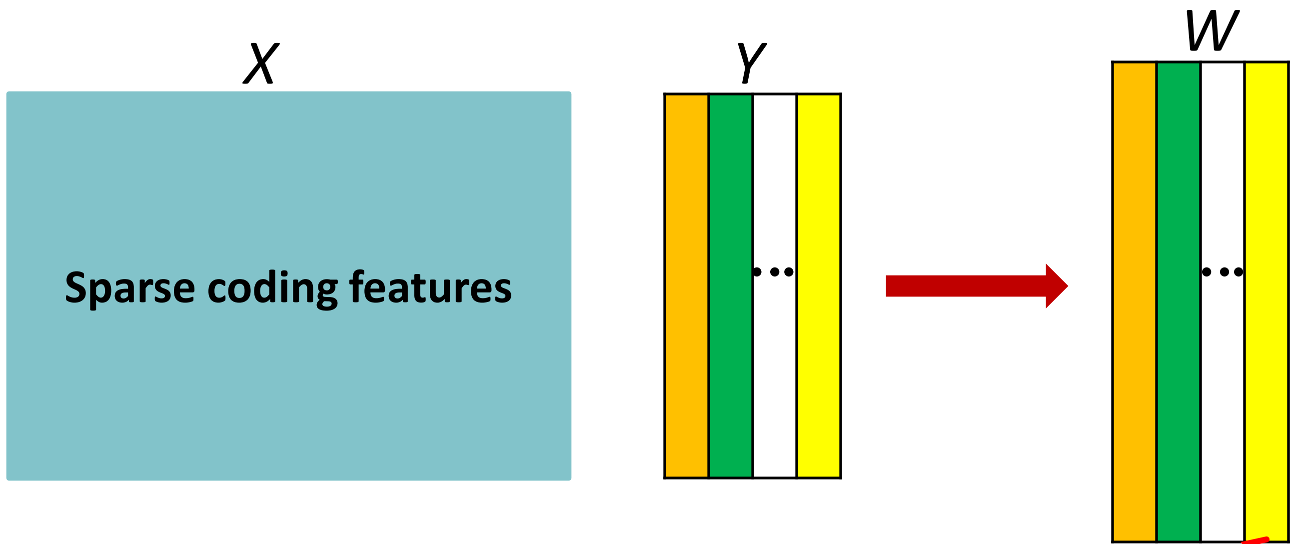
$$\min_x \frac{1}{2}\|Ax-u\|^2 + \lambda\|x\|_1$$

$$s.t. \quad x \geq 0$$

Both can be improved by incorporating the proximity information of local patches

[Ji *et al*. (2009) SIGKDD]

Low rank multi-task learning model



Low rank

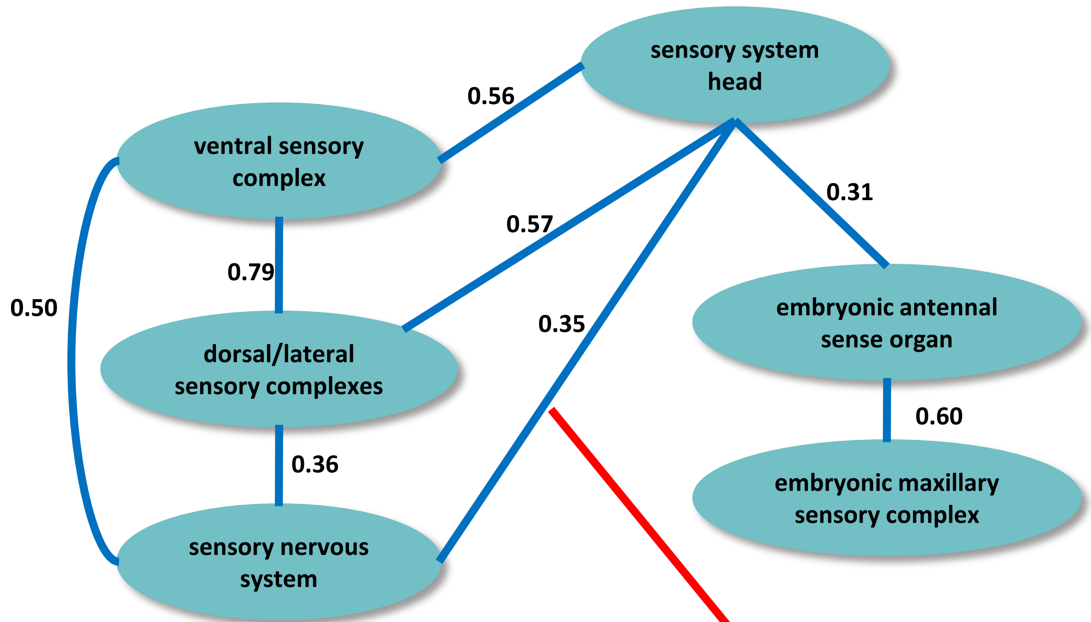
$$\sum_{i=1}^{k}\sum_{j=1}^{n}\boxed{L(w_i^T x_j, Y_{ji})}+\boxed{\lambda\|W\|_*}$$

trace norm = sum of singular values

Loss term

[Argyriou *et al*. (2008) Machine Learning]

Graph-based multi-task learning model



loss

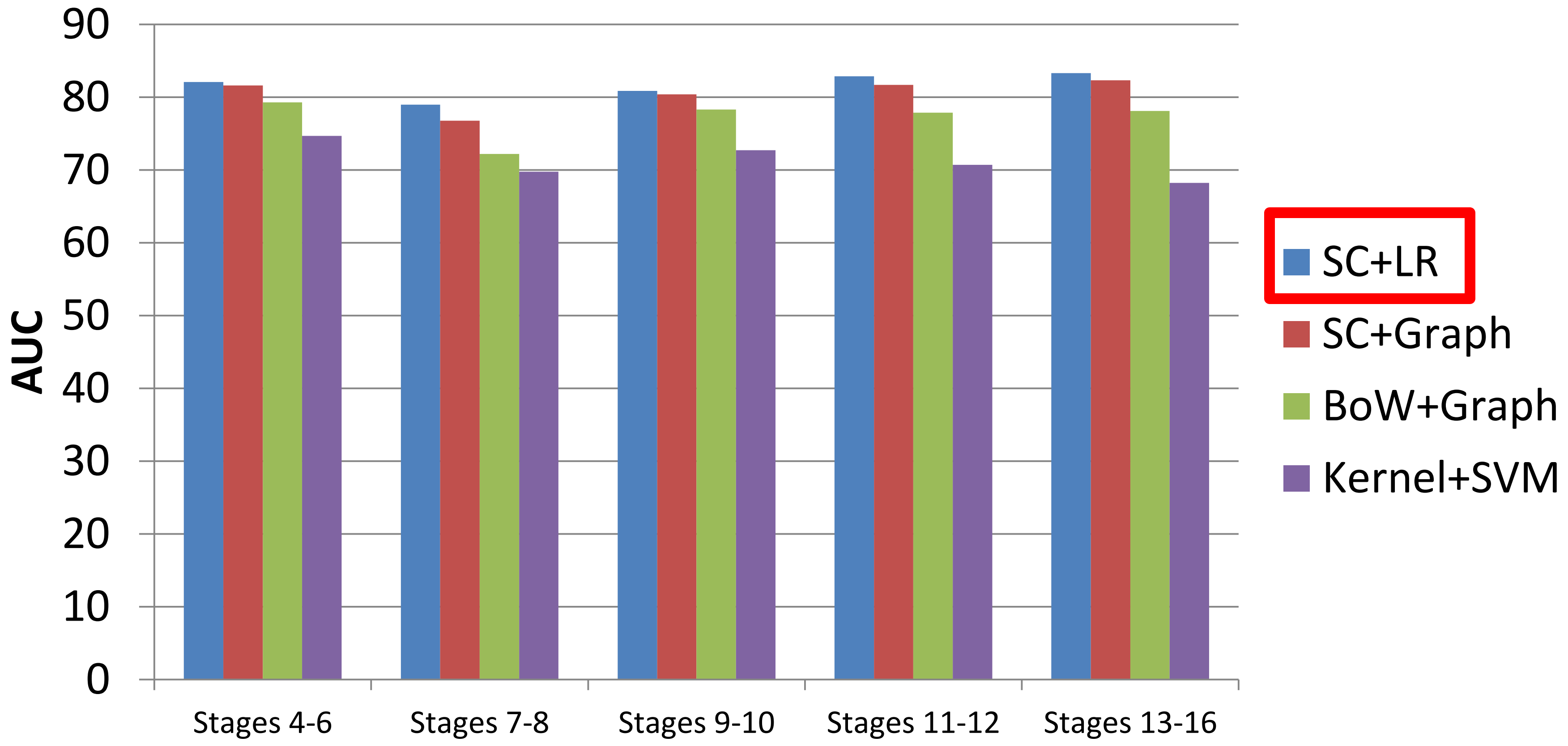
2-norm

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\boxed{L(w_i^T x_j, Y_{ji})}+\boxed{\lambda_1\|W\|_F^2}+\lambda_2\sum_{(p,q)\in G}\boxed{g(C_{pq})}\bullet\left\|w_p-\operatorname{sgn}(C_{pq})w_q\right\|^2$$

Closed-form solution

[Ji *et al*. (2009) SIGKDD]

Spatial annotation performance



- 50% data for training and 50% for testing and 30 random trials are generated
- Sparse coding with low rank multi-task learning achieves the best performance

MALSAR Package

Multi-T**A**sk **L**earning via
Structur**A**l **R**egularization
MALSAR package

- Jiayu Zhou, Jianhui Chen, Jieping Ye
- http://www.public.asu.edu/~jzhou29/Software/MALSAR/index.html

Functions in MALSAR Package

- Regularized Multi-Task Learning
- Joint Feature Learning
- Trace Norm Minimization
- ASO
- Clustered Multi-Task Learning
- Network Multi-Task Learning
- Robust Multi-Task Learning

Trends in Multi-Task Learning

- **Develop efficient algorithms** for large-scale multi-task learning. In many areas where multi-task learning is applied, such as bioinformatics, the dimensionality of data can be huge.
- Learn **task structures automatically** in MTL
- Most multi-task learning techniques deal with supervised learning problems. There is a high demand of developing new methods for **semi-supervised** and **unsupervised learning**.

Reference

Abernethy, J., Bach, F., Evgeniou, T., & Vert, J. (2006). Low-rank matrix factorization with attributes. *Arxiv preprint cs/0611124*.

Abernethy, J., Bach, F., Evgeniou, T., & Vert, J. (2009). A new approach to collaborative filtering: Operator estimation with spectral regularization. *The Journal of Machine Learning Research*, *10*, 803–826.

Agarwal, A., Daumé III, H., & Gerber, S. (2010). Learning multiple tasks using manifold regularization. .

Ando, R., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, *6*, 1817–1853.

Argyriou, A., Evgeniou, T., & Pontil, M. (2007). Multi-task feature learning. *Advances in neural information processing systems*, *19*, 41.

Argyriou, A., Evgeniou, T., & Pontil, M. (2008a). Convex multi-task feature learning. *Machine Learning*, *73*, 243–272.

Reference

Argyriou, A., Micchelli, C., Pontil, M., & Ying, Y. (2008b). A spectral regularization framework for multi-task structure learning. *Advances in Neural Information Processing Systems*, *20*, 25–32.

Bakker, B., & Heskes, T. (2003). Task clustering and gating for bayesian multitask learning. *The Journal of Machine Learning Research*, *4*, 83–99.

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*.

Bickel, S., Bogojeska, J., Lengauer, T., & Scheffer, T. (2008). Multi-task learning for hiv therapy screening. *Proceedings of the 25th international conference on Machine learning* (pp. 56–63).

Bonilla, E., Chai, K., & Williams, C. (2008). Multi-task gaussian process prediction. *Advances in Neural Information Processing Systems*, *20*, 153–160.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

Chen, J., Liu, J., & Ye, J. (2010). Learning incoherent sparse and low-rank patterns from multiple tasks. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1179–1188).

Reference

Chen, J., Tang, L., Liu, J., & Ye, J. (2009). A convex formulation for learning shared structures from multiple tasks. *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 137–144).

Evgeniou, T., Micchelli, C., & Pontil, M. (2006). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, *6*, 615.

Evgeniou, T., & Pontil, M. (2004). Regularized multi–task learning. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 109–117).

Gu, Q., Li, Z., & Han, J. (2011). Learning a kernel for multi-task clustering. *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Gu, Q., & Zhou, J. (2009). Learning the shared subspace for multi-task clustering and transductive transfer classification. *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on* (pp. 159–168).

Jacob, L., Bach, F., & Vert, J. (2008). Clustered multi-task learning: A convex formulation. *Arxiv preprint arXiv:0809.2085*.

Reference

Jebara, T. (2004). Multi-task feature and kernel selection for svms. *Proceedings of the twenty-first international conference on Machine learning* (p. 55).

Ji, S., & Ye, J. (2009). An accelerated gradient method for trace norm minimization. *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 457–464).

Lawrence, N., & Platt, J. (2004). Learning to learn with the informative vector machine. *Proceedings of the twenty-first international conference on Machine learning* (p. 65).

Liu, J., Ji, S., & Ye, J. (2009). Multi-task feature learning via efficient l 2, 1-norm minimization. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 339–348).

Reference

Nie, F., Huang, H., Cai, X., & Ding, C. (2010). Efficient and robust feature selection via joint l21-norms minimization. .

Obozinski, G., Taskar, B., & Jordan, M. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, *20*, 231–252.

Thrun, S., & OSullivan, J. (1998). Clustering learning tasks and the selective cross-task transfer of knowledge. *Learning to learn*, 181–209.

Wang, F., Wang, X., & Li, T. (2009). Semi-supervised multi-task learning with task regularizations. *2009 Ninth IEEE International Conference on Data Mining* (pp. 562–568).

Xue, Y., Liao, X., Carin, L., & Krishnapuram, B. (2007). Multi-task learning for classification with dirichlet process priors. *The Journal of Machine Learning Research*, *8*, 35–63.

Yu, K., Schwaighofer, A., Tresp, V., Ma, W., & Zhang, H. (2003). Collaborative ensemble learning: Combining collaborative and content-based information filtering via hierarchical bayes. *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*.

Reference

Yu, K., Tresp, V., & Schwaighofer, A. (2005). Learning gaussian processes from multiple tasks. *Proceedings of the 22nd international conference on Machine learning* (pp. 1012–1019).

Yu, K., Tresp, V., & Yu, S. (2004). A nonparametric hierarchical bayesian framework for information filtering. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 353–360).

Zha, H., He, X., Ding, C., Gu, M., & Simon, H. (2002). Spectral relaxation for k-means clustering. *Advances in Neural Information Processing Systems*, 2, 1057–1064.

Zhang, J., Ghahramani, Z., & Yang, Y. (2006). Learning multiple related tasks using latent independent component analysis. *Advances in neural information processing systems*, *18*, 1585.

Zhang, Y., & Yeung, D. (2010). A convex formulation for learning task relationships in multi-task learning. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 733–742).

Reference

Zhou, J., Chen, J., & Ye, J. (2011a). Clustered multi-task learning via alternating structure optimization. *Advances in Neural Information Processing Systems*.

Zhou, J., Yuan, L., Liu, J., & Ye, J. (2011b). A multi-task learning formulation for predicting disease progression. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 814–822). New York, NY, USA: ACM.

Reference

Optimization Algorithms

Nemirovski, A. Efficient methods in convex programming. *Lecture Notes*.

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate o (1/k2). *Soviet Mathematics Doklady* (pp. 372–376).

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, *103*, 127–152.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. *ReCALL, 76*.

Nesterov, Y., & Nesterov, I. (2004). *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer.

Thank You!