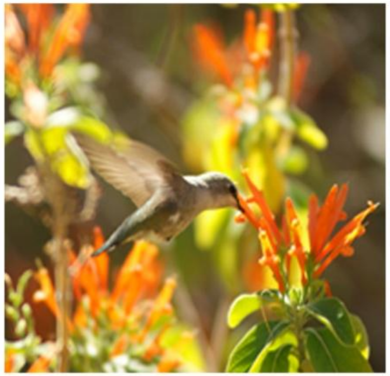
Multi-Task Learning: Theory, Algorithms, and Applications

Jiayu Zhou[1,2], Jianhui Chen[3], Jieping Ye[1,2]
[1]Computer Science and Engineering, Arizona State University, AZ
[2]The Biodesign Institute, Arizona State University, AZ
[3]GE Global Research, NY





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 multi-task learning applications
- 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
  - Incomplete Multi-Source Fusion
  - *Drosophila* Gene Expression Image Analysis
- Part IV: An MTL Package (MALSAR)
- Current and future directions

Tutorial Road Map

- **Part I: Multi-task Learning (MTL) background and motivations**
- Part II: MTL formulations
- Part III: Case study of real-world applications
  - Incomplete Multi-Source Fusion
  - *Drosophila* Gene Expression Image Analysis
- Part IV: An MTL Package (MALSAR)
- Current and 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 (Birth year … Previous score); school-dependent (School ranking …)

→

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 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

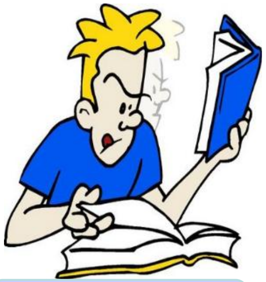
School 139 - Rosemead High School

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%	…		?

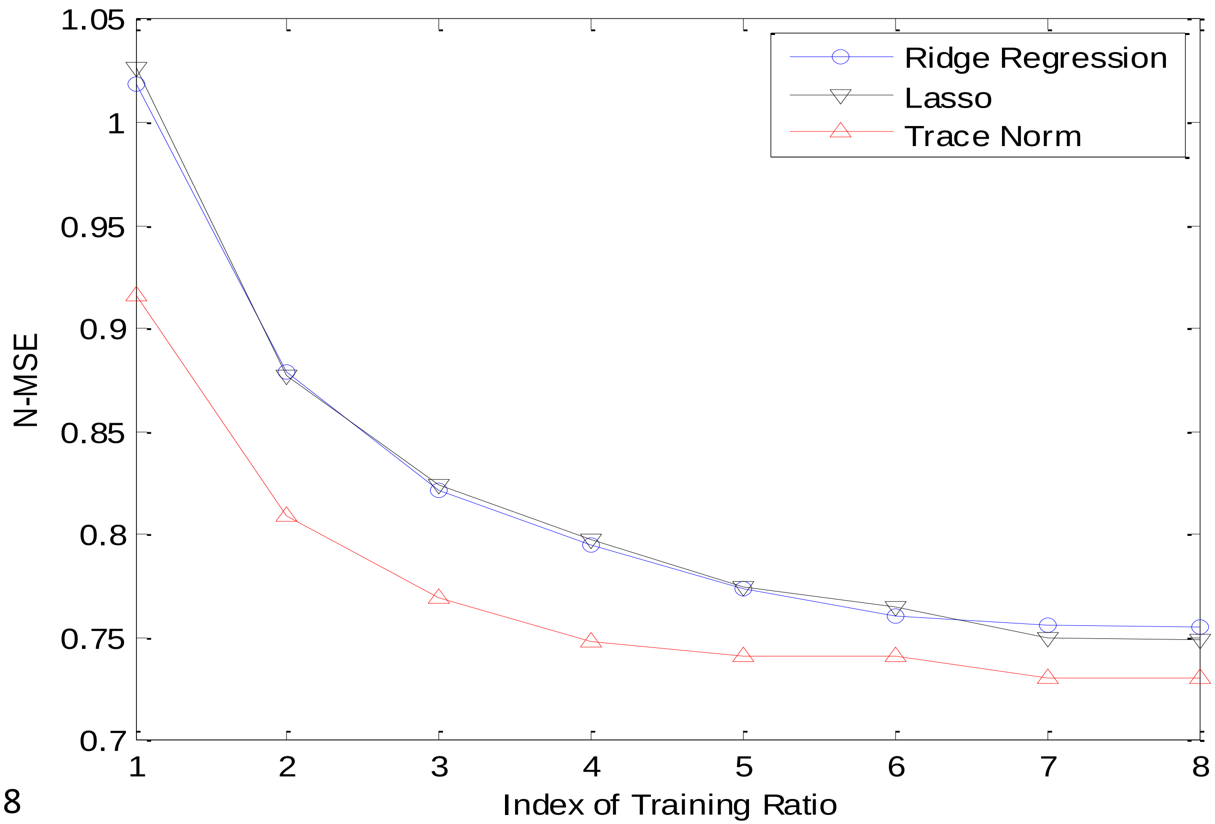
⇨ task 139th

Learn tasks simultaneously
Model the tasks relationship

⇨ ……

Performance of MTL

- Evaluation on the *School* data:
  - Predict exam scores for 15362 students from 139 schools
  - Describe each student by 27 attributes
  - Compare single task learning approaches (Ridge Regression, Lasso) and one multi-task learning approach (trace-norm regularized learning)

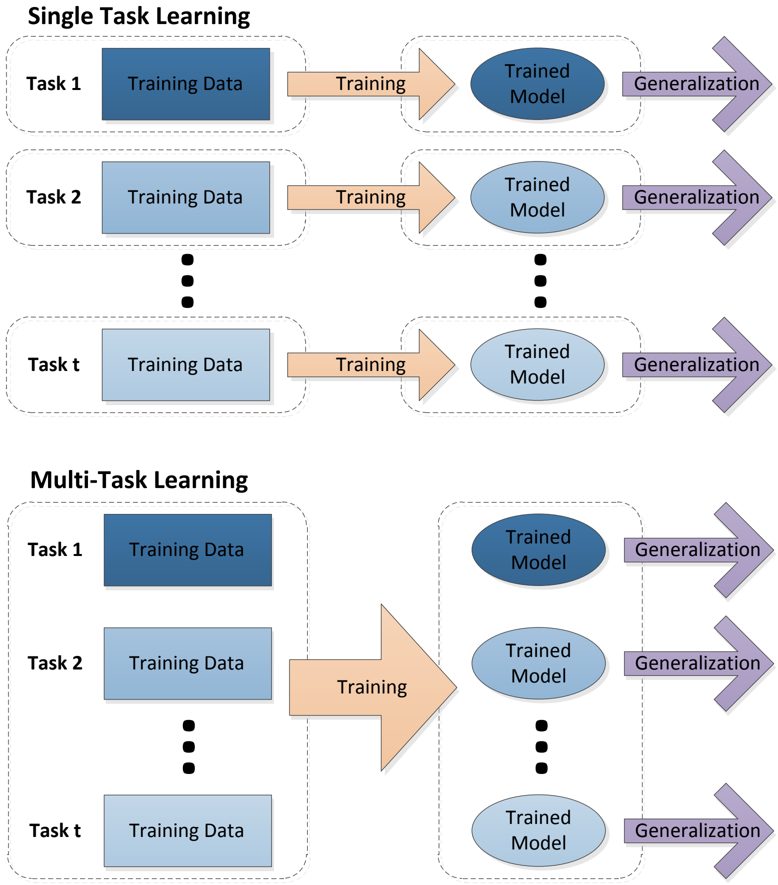


Performance measure:

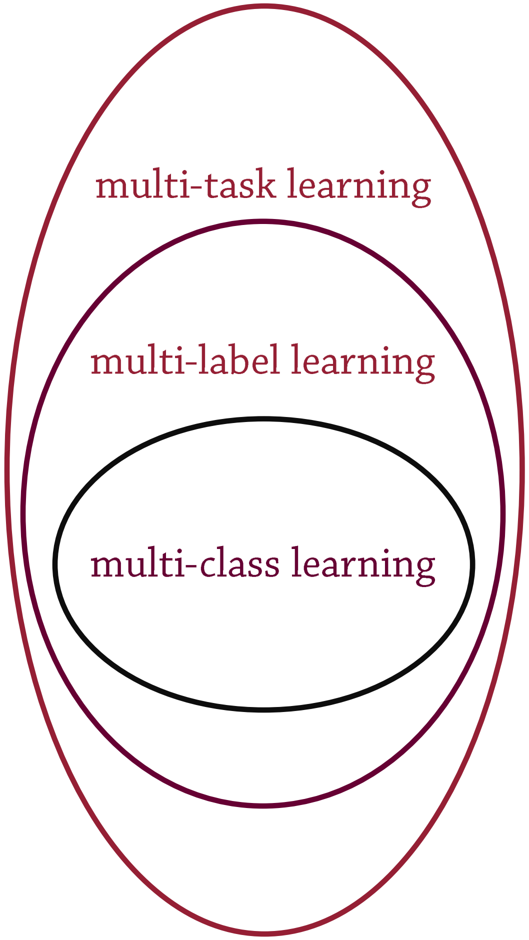
$$\text{N-MSE} = \frac{\text{mean squared error}}{\text{variance (target)}}$$

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



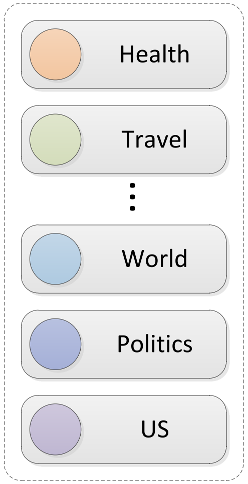
- Multi-task Learning
  - Model the task relatedness
  - Learn all tasks simultaneously
  - Tasks may have different data/features
- Multi-label Learning
  - Model the label relatedness
  - Learn all labels simultaneously
  - Labels share the same data/features
- Multi-class Learning
  - Learn the classes independently
  - All classes are exclusive

Web Pages Categorization

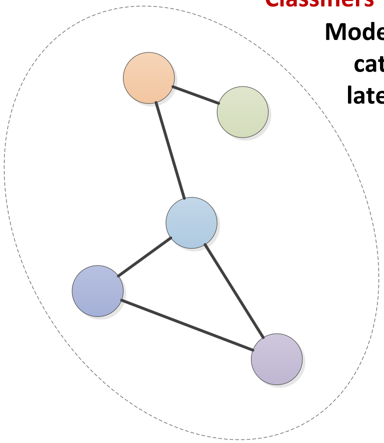
Chen *et. al*. 2009 ICML

- Classify documents into categories
- The classification of each category is a task
- The tasks of predicting different categories may be latently related

Category Classifiers



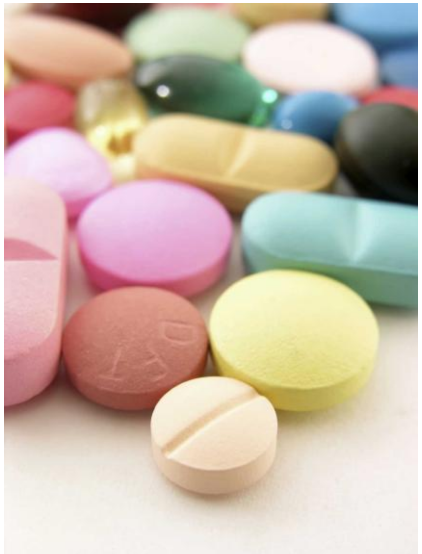
Classifiers' Parameters
Models of different categories are latently related

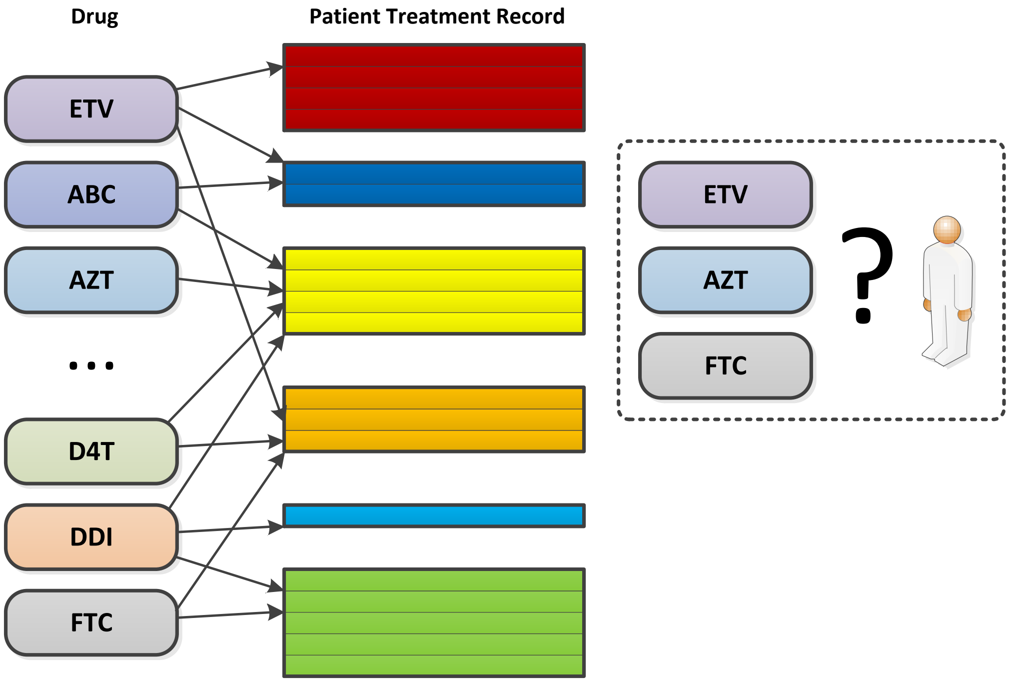


MTL for HIV Therapy Screening

Bickel *et. al.* ICML 08

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





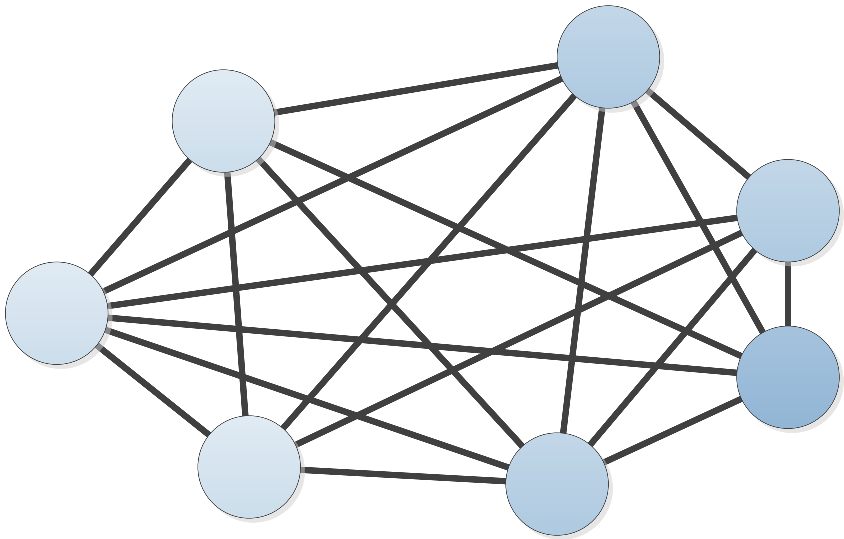
Other Applications

- Portfolio selection [Ghosn and Bengio, NIPS'97]
- Collaborative ordinal regression [Yu *et. al.* NIPS'06]
- Web image and video search [Wang *et. al.* CVPR'09]
- Disease progression modeling [Zhou *et. al.* KDD'11]
- Disease prediction [Zhang *et. al.* NeuroImage 12]

Tutorial Road Map

- Part I: Multi-task Learning (MTL) background and motivations
- **Part II: MTL formulations**
- Part III: Case study of real-world applications
  - Incomplete Multi-Source Fusion
  - *Drosophila* Gene Expression Image Analysis
- Part IV: An MTL Package (MALSAR)
- Current and future directions

How Tasks Are Related

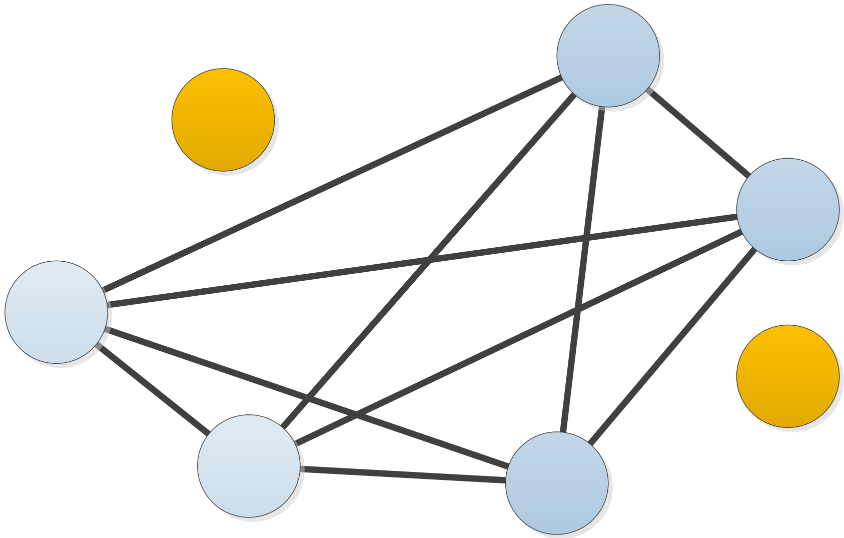


Assumption:
All tasks are related

Methods

- Mean-regularized MTL
- Joint feature learning
- Trace-Norm regularized MTL
- Alternating structural optimization (ASO)
- Shared Parameter Gaussian Process

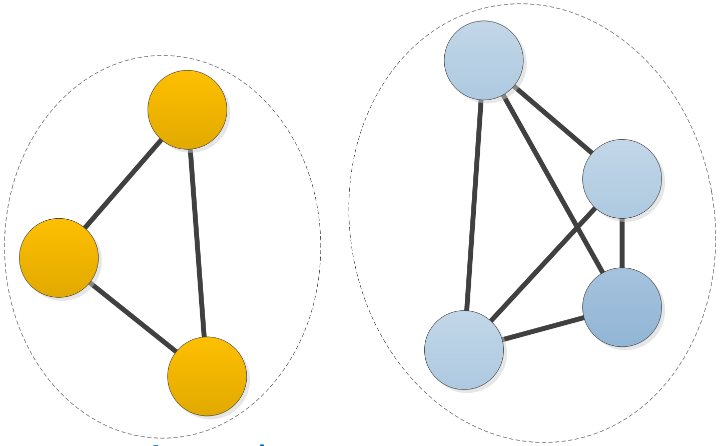
How Tasks Are Related



Assumption:
There are outlier tasks

- Assume all tasks are related may be too strong for practical applications.
- There are some irrelevant (outlier) tasks.

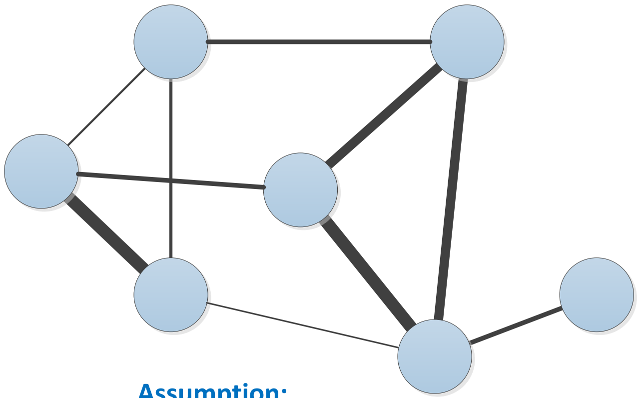
How Tasks Are Related



Assumption:
Tasks have group structures

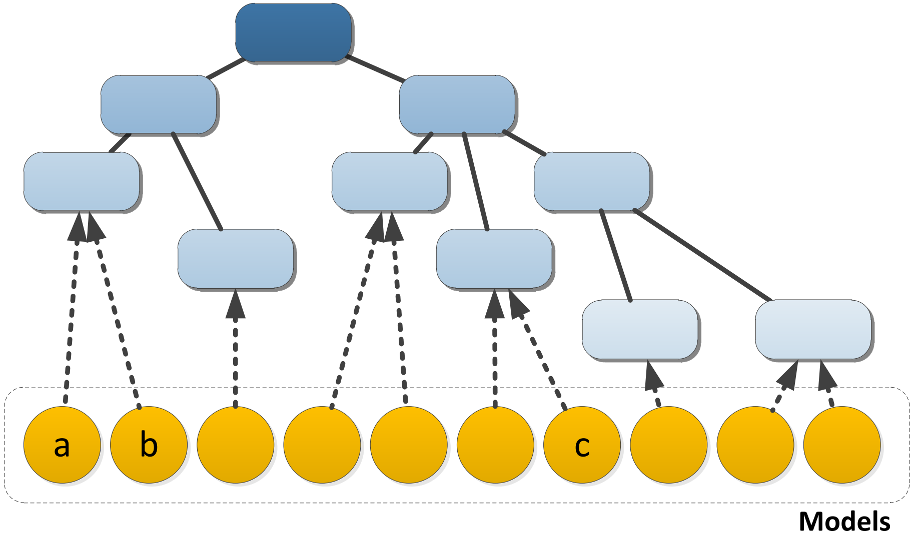
Methods

- Clustered MTL
- Tree MTL
- Network MTL



Assumption:
Tasks have graph/network structures

Assumption:
Tasks have tree structures



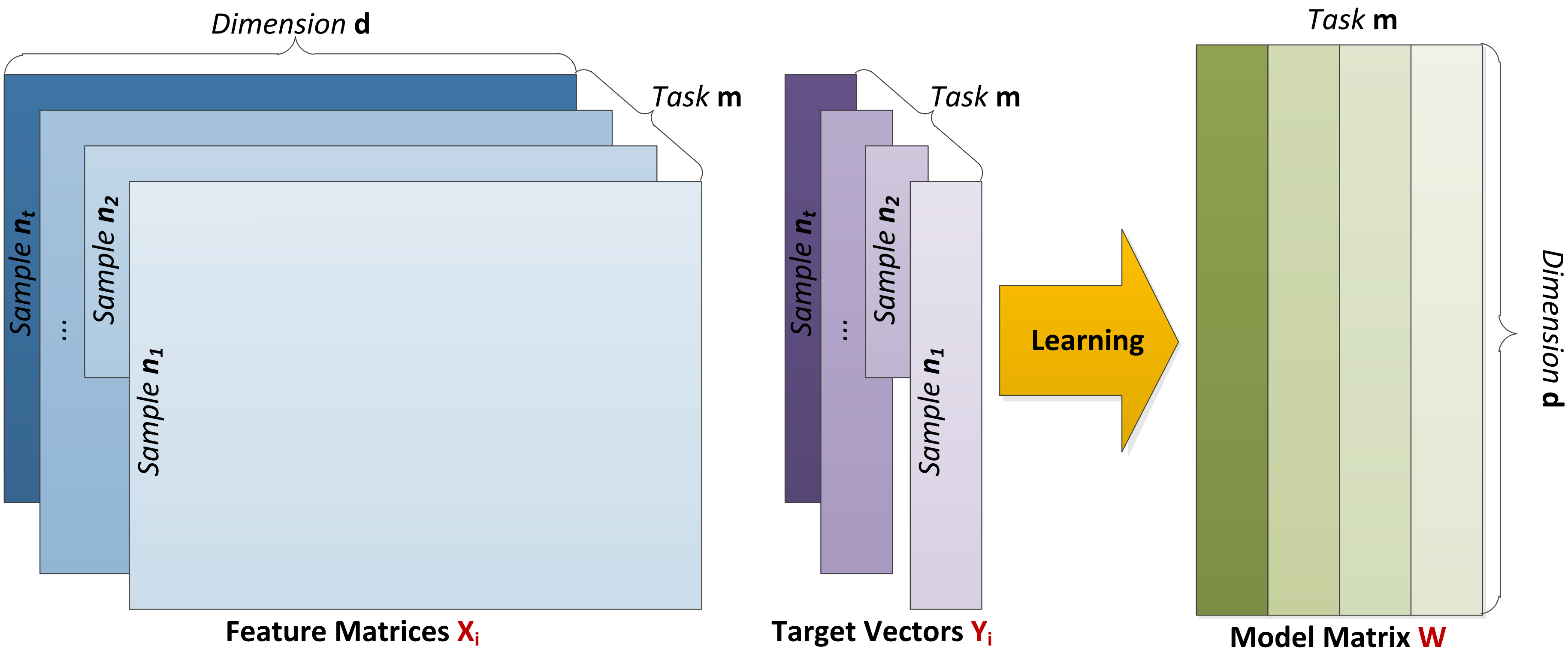
Multi-Task Learning Methods

- Regularization-based MTL
  - All tasks are related
    - regularized MTL, joint feature learning, low rank MTL, ASO
  - Learning with outlier tasks: robust MTL
  - Tasks form groups/graphs/trees
    - clustered MTL, network MTL, tree MTL
- Other Methods
  - Shared Hidden Nodes in Neural Network
  - Shared Parameter Gaussian Process

Regularization-based Multi-Task Learning

- All tasks are related
  - **Mean-Regularized MTL**
  - MTL in high dimensional feature space
    - Embedded Feature Selection
    - Low-Rank Subspace Learning
- Clustered MTL
- MTL with Tree/Graph structure

Notation



- We focus on linear models: $\mathrm{Y_i} = X_i \times W_i$

$$X_i \in \mathbb{R}^{n_i \times d}, Y_i \in \mathbb{R}^{n_i \times 1}, W = [W_1, W_2, \dots, W_m]$$

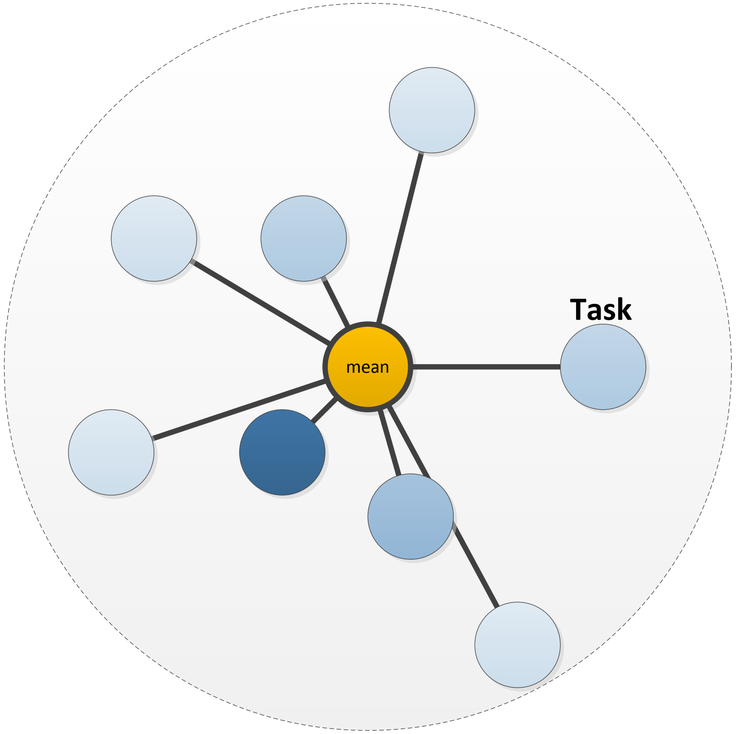
Mean-Regularized Multi-Task Learning

Evgeniou & Pontil, 2004 KDD

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

Regularization
penalizes the deviation of each task from the mean

$$\min_{W} \frac{1}{2}\|XW - Y\|_F^2 + \lambda \sum_{i=1}^{m} \left\| W_i - \frac{1}{m}\sum_{s=1}^{m} W_s \right\|_2^2$$



Regularization-based Multi-Task Learning

- All tasks are related
  - Mean-Regularized MTL
  - **MTL in high dimensional feature space**
    - Embedded Feature Selection
    - Low-Rank Subspace Learning
- Clustered MTL
- MTL with Tree/Graph structure

Multi-Task Learning with High Dimensional Data

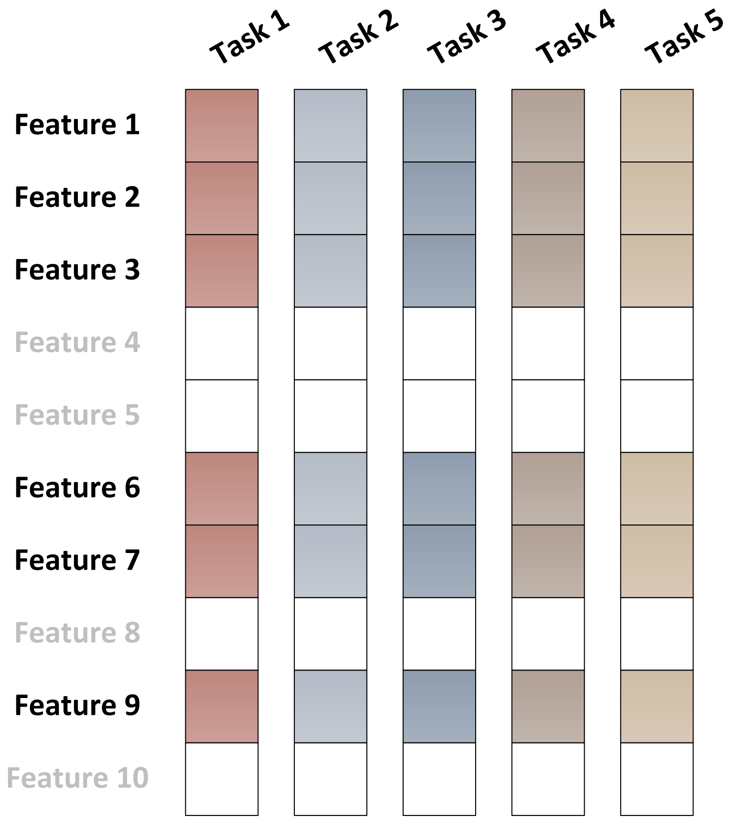
- In practical applications, we may deal with high dimensional data.
  - Gene expression data, biomedical image data
- Curse of Dimensionality
- Dealing with high dimensional data in multi-task learning
  - Embedded feature selection: L_1/L_q - Group Lasso
  - Low-rank subspace learning: low-rank assumption – ASO, Trace-norm regularization

Regularization-based Multi-Task Learning

- All tasks are related
  - Mean-Regularized MTL
  - MTL in high dimensional feature space
    - **Embedded Feature Selection**
    - Low-Rank Subspace Learning
- Clustered MTL
- MTL with Tree/Graph structure

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.

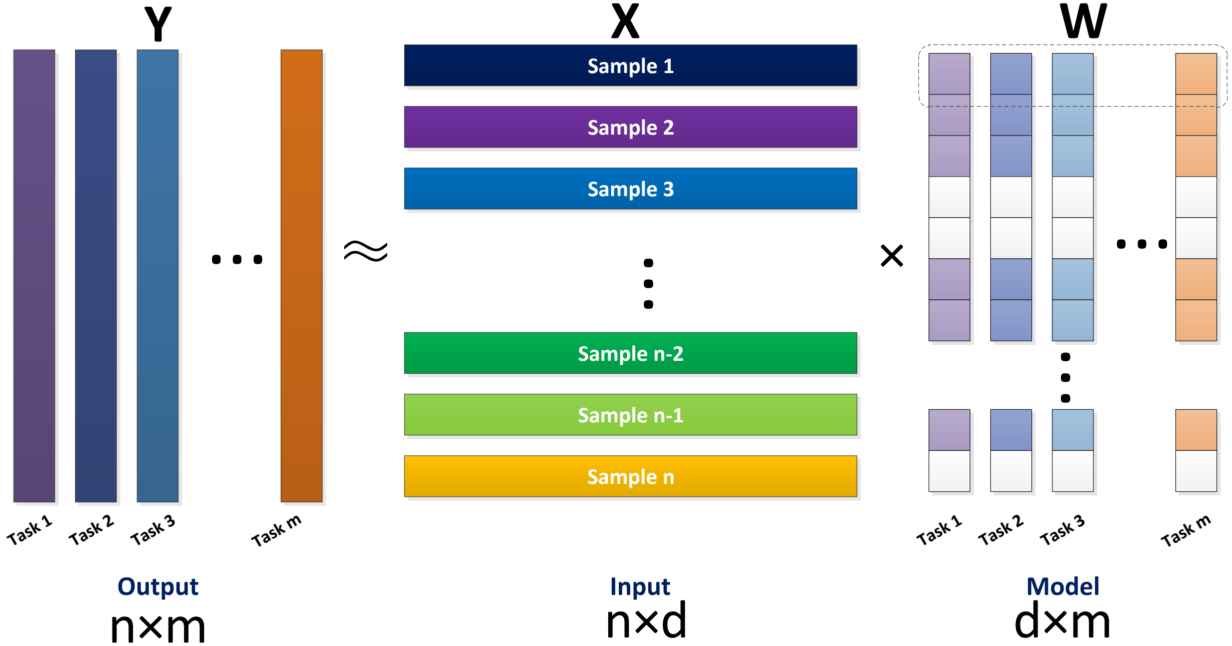


Multi-Task Learning with Joint Feature Learning

Obozinski *et. al.* 2009 Stat Comput, Liu *et. al.* 2010 Technical Report

- Using group sparsity: ℓ_1/ℓ_q-norm regularization
- When q>1 we have group sparsity.

$$\|W\|_{1,q} = \sum_{i=1}^{d} \|\boldsymbol{w}_i\|_q$$

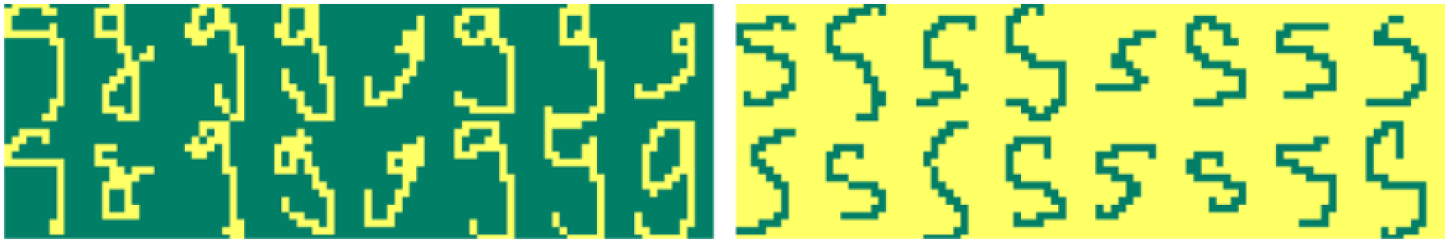


$$\min_W \frac{1}{2}\|XW - Y\|_F^2 + \lambda\|W\|_{1,q}$$

Writer-Specific Character Recognition

Obozinski, Taskar, and Jordan, 2006

- Each task is a classification between two letters for one writer.

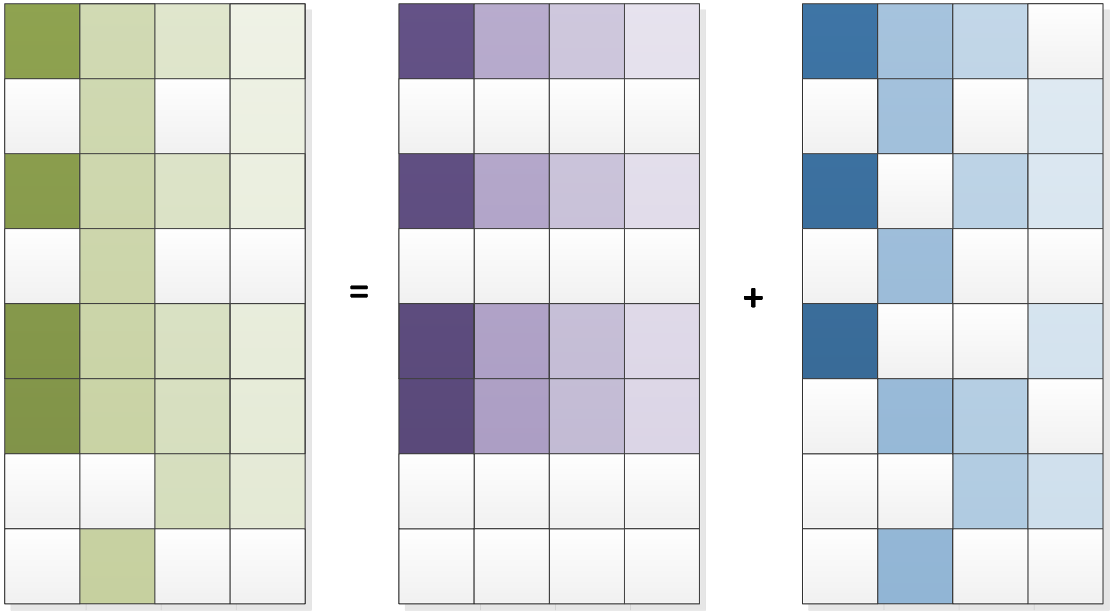


	pixels: error (%)			
Task	ℓ_1/ℓ_2	ℓ_1/ℓ_1	id.ℓ_1	pool
c/e	**4.0**	8.5	9.0	4.5
g/y	**11.4**	16.1	17.2	18.6
g/s	**4.4**	10.0	10.3	6.9
m/n	**2.5**	6.3	6.9	4.1
a/g	**1.3**	3.6	4.1	3.6
i/j	12.0	14.0	14.0	**11.3**
a/o	**2.8**	4.8	5.2	4.2
f/t	**5.0**	6.7	6.1	8.2
h/n	**3.2**	14.3	18.6	5.0

Dirty Model for Multi-Task Learning

Jalali *et. al.* 2010 NIPS

- In practical applications, it is too restrictive to constrain all tasks to share a single shared structure.



Model
W

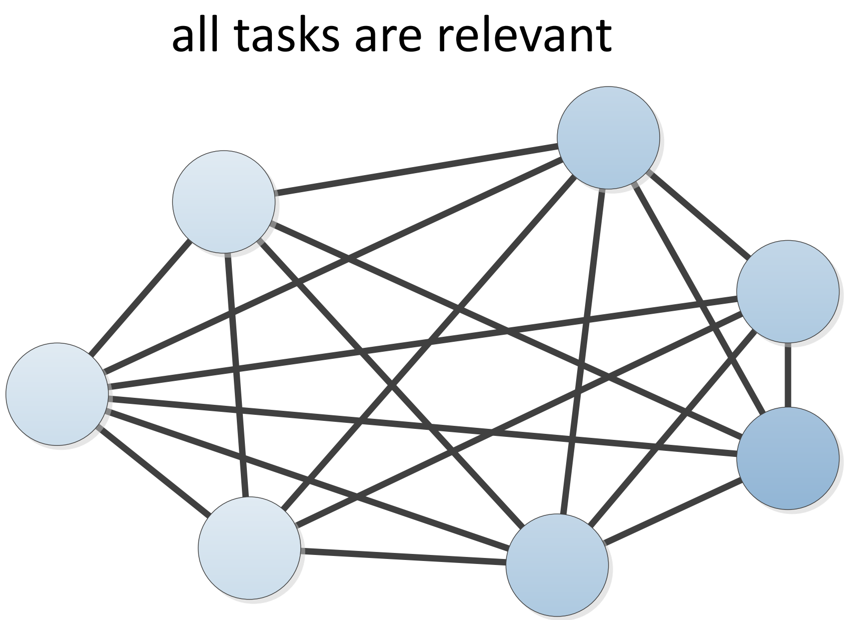
Group Sparse Component
P

Sparse Component
Q

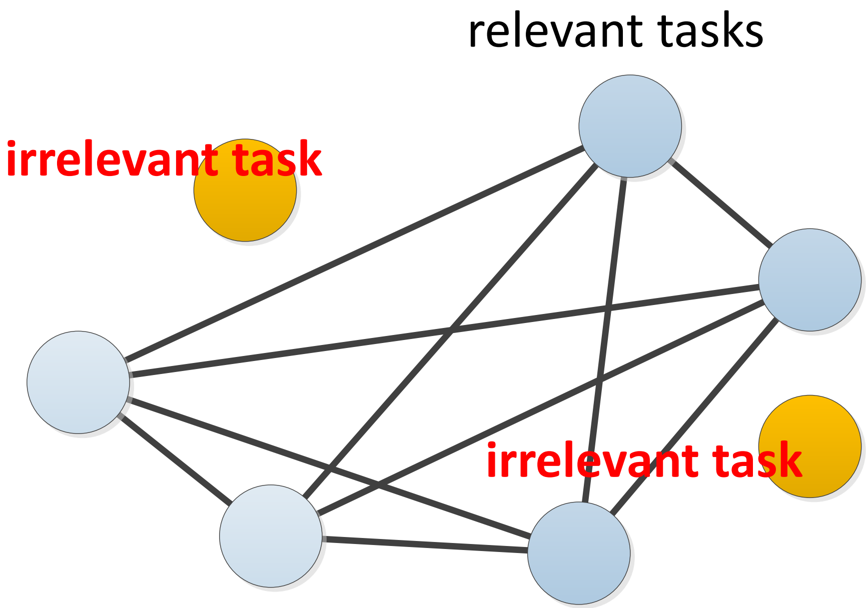
$$\min_{P,Q} \|Y - X(P+Q)\|_F^2 + \lambda_1 \|P\|_{1,q} + \lambda_2 \|Q\|_1$$

Robust Multi-Task Learning

- Most Existing MTL Approaches
- Robust MTL Approaches



Assumption:
All tasks are related

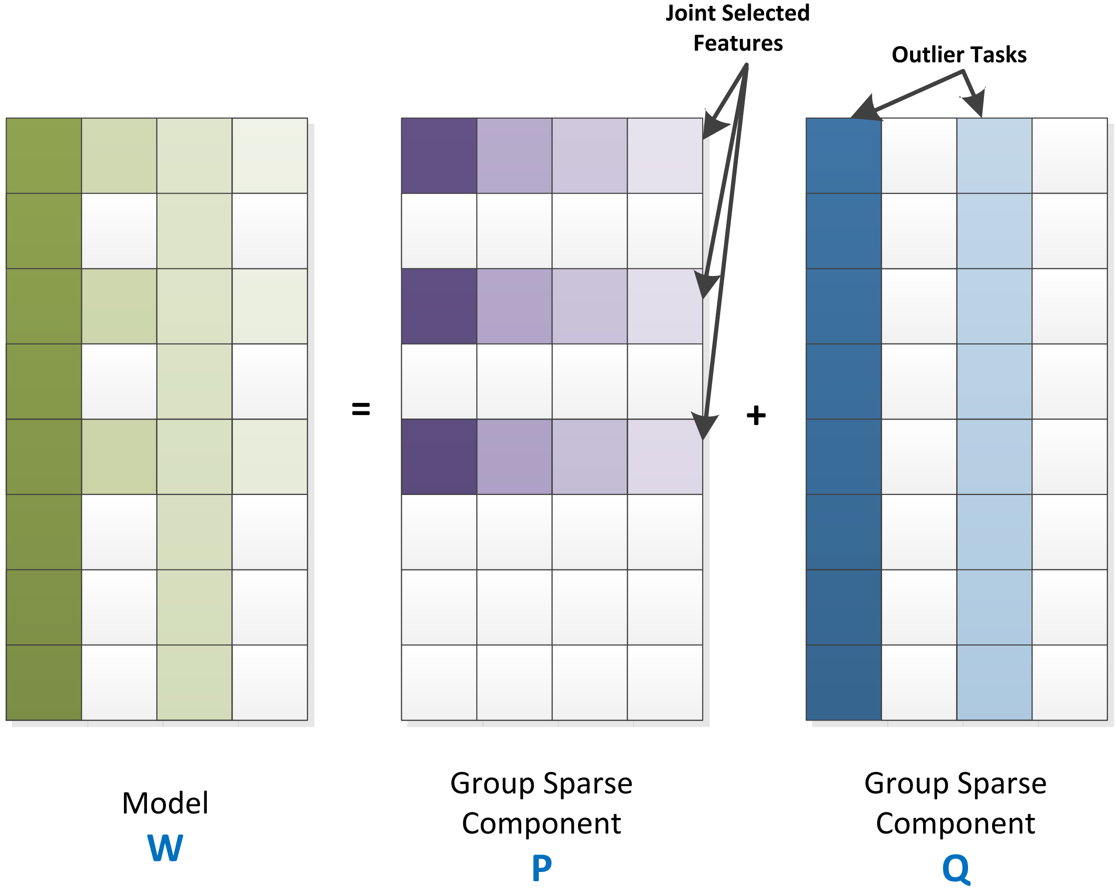


Assumption:
There are outlier tasks

Robust Multi-Task Feature Learning

Gong *et. al.* 2012 Submitted

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



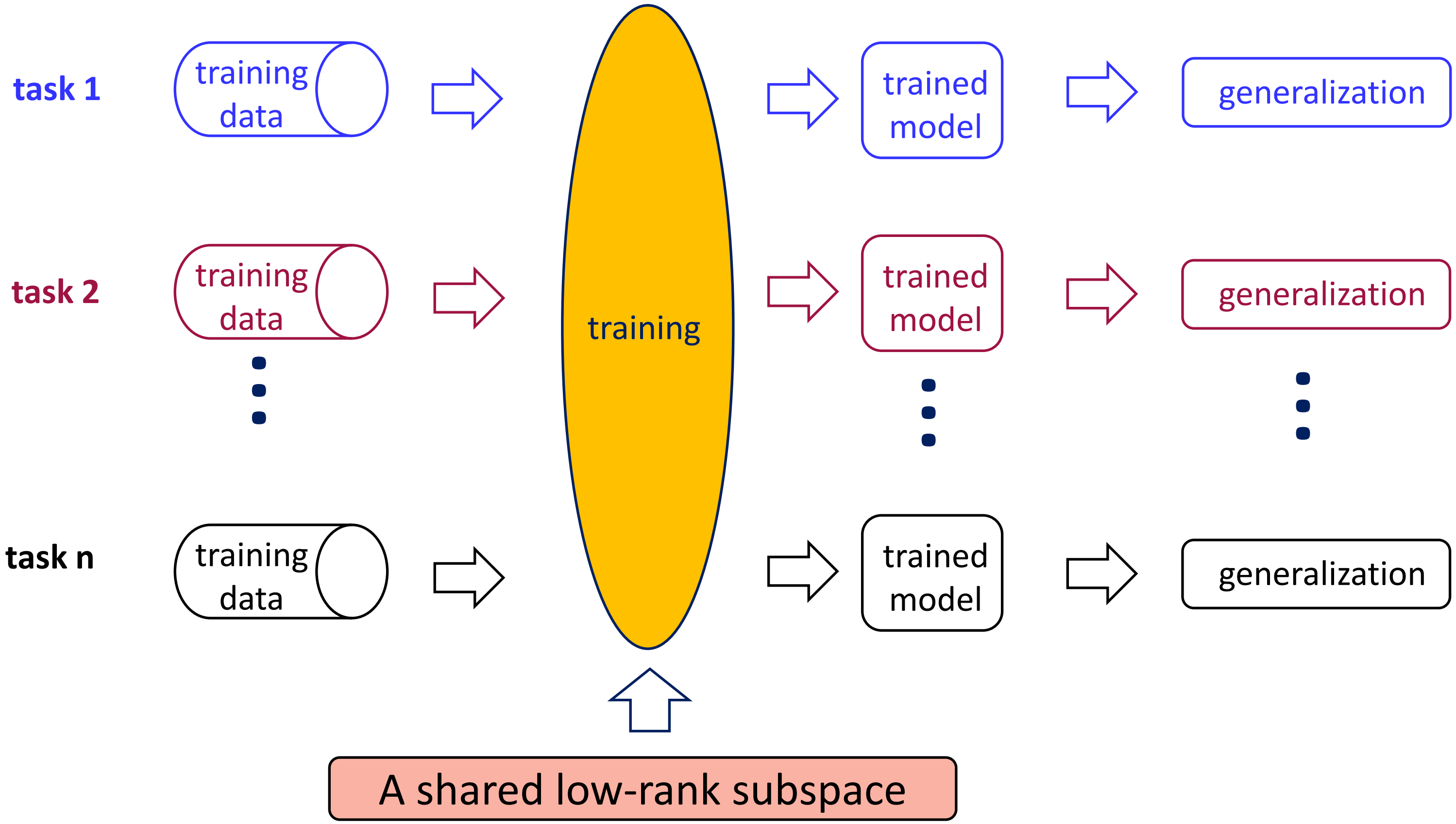
$$\min_{P,Q} \|Y - X(P+Q)\|_F^2 + \lambda_1 \|P\|_{1,q} + \lambda_2 \|Q^T\|_{1,q}$$

Regularization-based Multi-Task Learning

- All tasks are related
  - Mean-Regularized MTL
  - MTL in high dimensional feature space
    - Embedded Feature Selection
    - **Low-Rank Subspace Learning**
- Clustered MTL
- MTL with Tree/Graph structure

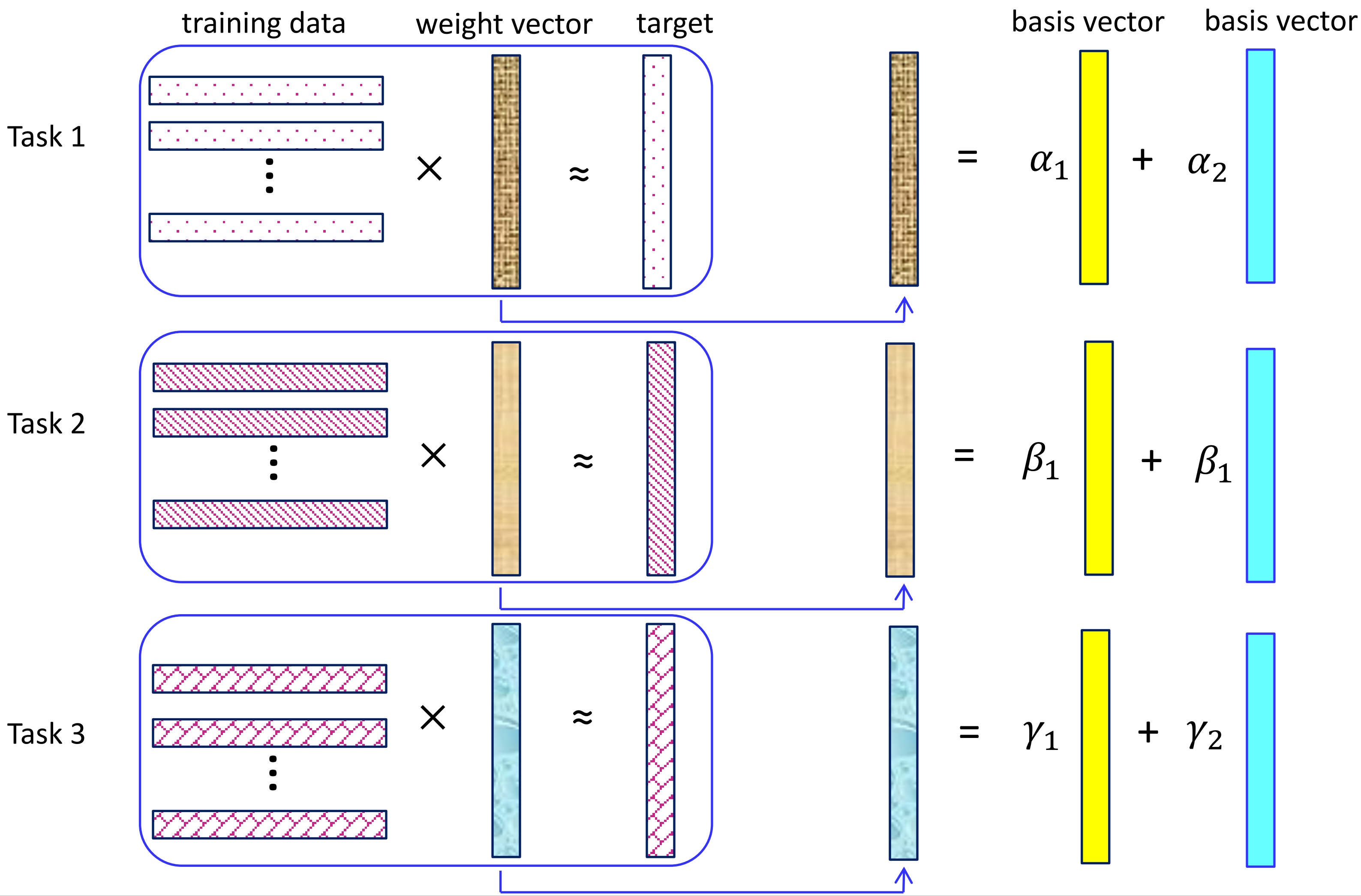
Trace-Norm Regularized MTL

- Capture task relatedness via a shared low-rank structure

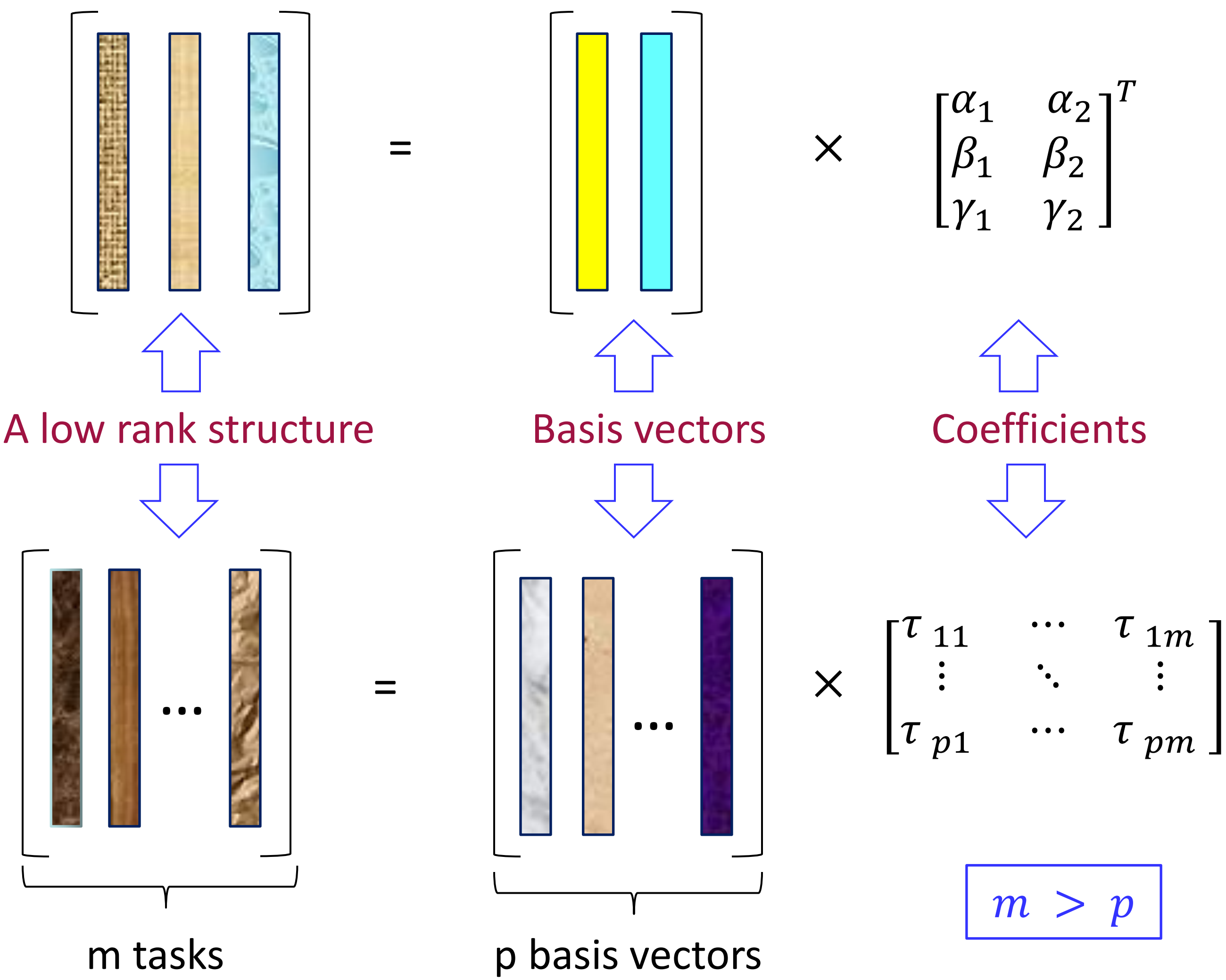


Low-Rank Structure for MTL

- Assume we have a rank 2 model matrix:



Low-Rank Structure for MTL



Low-Rank Structure for MTL

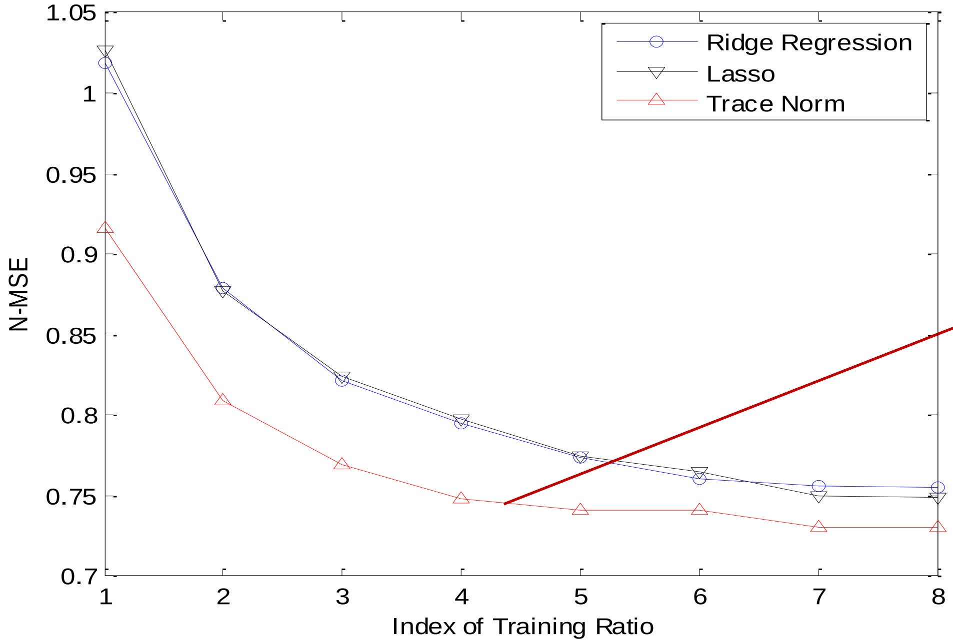
Ji *et. al.* 2009 ICML

- Rank minimization formulation
  - $\min_{W} \text{Loss}(W) + \lambda \times \text{Rank}(W)$
  - Rank minimization is *NP-Hard* for general loss functions

- Convex relaxation: trace norm minimization
  - $\min_{W} \text{Loss}(W) + \lambda \times \|W\|_*$ $\|W\|_*$: sum of singular values of W
  - The trace norm is theoretically shown to be a good approximation for rank function (Fazel et al., 2001).

Low-Rank Structure for MTL

- Evaluation on the *School* data:
  - 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:

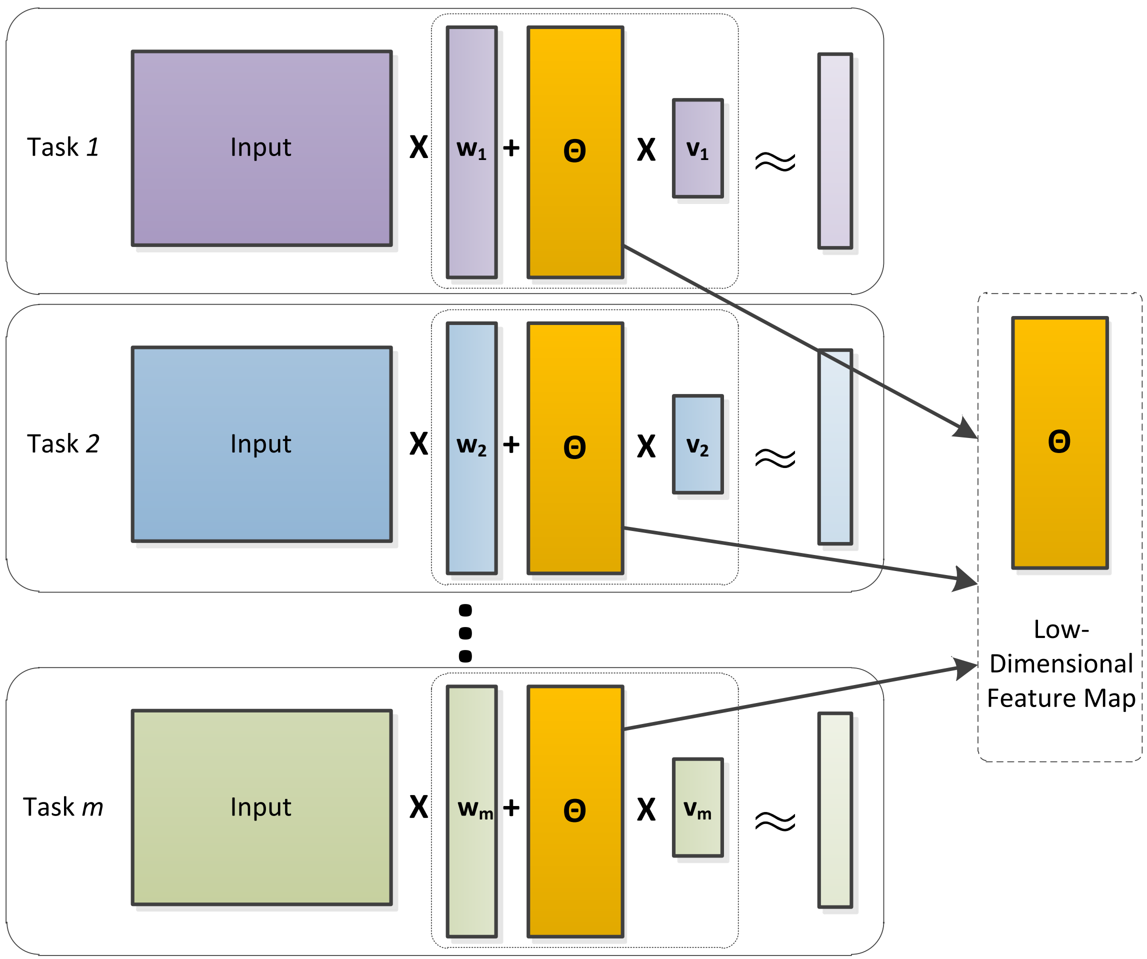
$$\text{N-MSE} = \frac{\text{mean squared error}}{\text{variance (target)}}$$

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

Alternating Structure Optimization (ASO)

Ando and Zhang, 2005 JMRL

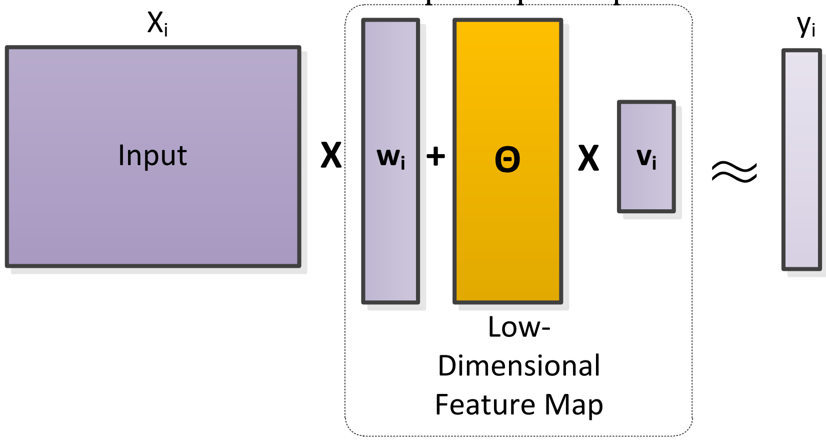
- ASO assumes that the model is the sum of two components: a task specific one and a shared low dimensional subspace.



Alternating Structure Optimization (ASO)

Ando and Zhang, 2005 JMRL

- Learning from the i-th task $u_i = \Theta v_i + w_i$



- Empirical loss function for i-th task

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

- ASO simultaneously learns *models* and the *shared structure*:

$$\min_{\Theta,\{v_i,w_i\}} \sum_{i=1}^{m} \{\mathcal{L}_i(X_i(\Theta v_i + w_i), y_i) + \alpha\|w_i\|^2\}$$

subject to $\Theta^T\Theta = I$

iASO Formulation

Chen et al., 2009 ICML

- iASO formulation

$$\min_{\Theta,\{v_i,w_i\}} \sum_{i=1}^{m} \{\mathcal{L}_i(X_i(\Theta v_i + w_i), y_i) + \alpha \|\Theta v_i + w_i\|^2 + \beta\, \|w_i\|^2\}$$

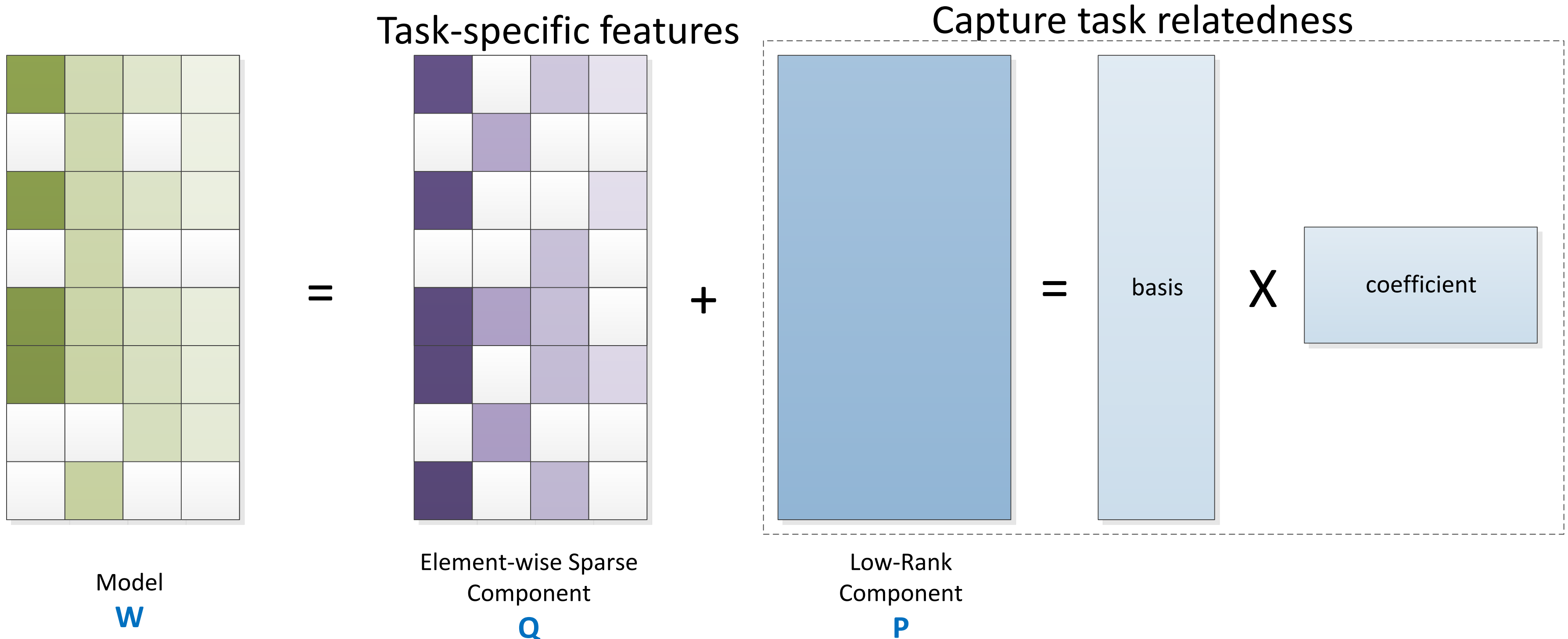
subject to $\Theta^T\Theta = I$

- control both *model complexity* and task relatedness
- subsume ASO (Ando et al.'05) as a special case
- naturally lead to a convex relaxation (Chen et al., 09, ICML)
- Convex relaxed ASO is equivalent to iASO under certain mild conditions

Incoherent Low-Rank and Sparse Structures

Chen *et. al.* 2010 KDD

- ASO uses L2-norm on task-specific component, we can also use L1-norm to learn task-specific features.



$$\min_{P,Q} \sum_{i=1}^{m} \mathcal{L}_i(X_i(P_i + Q_i), y_i) + \lambda \|Q\|_1$$

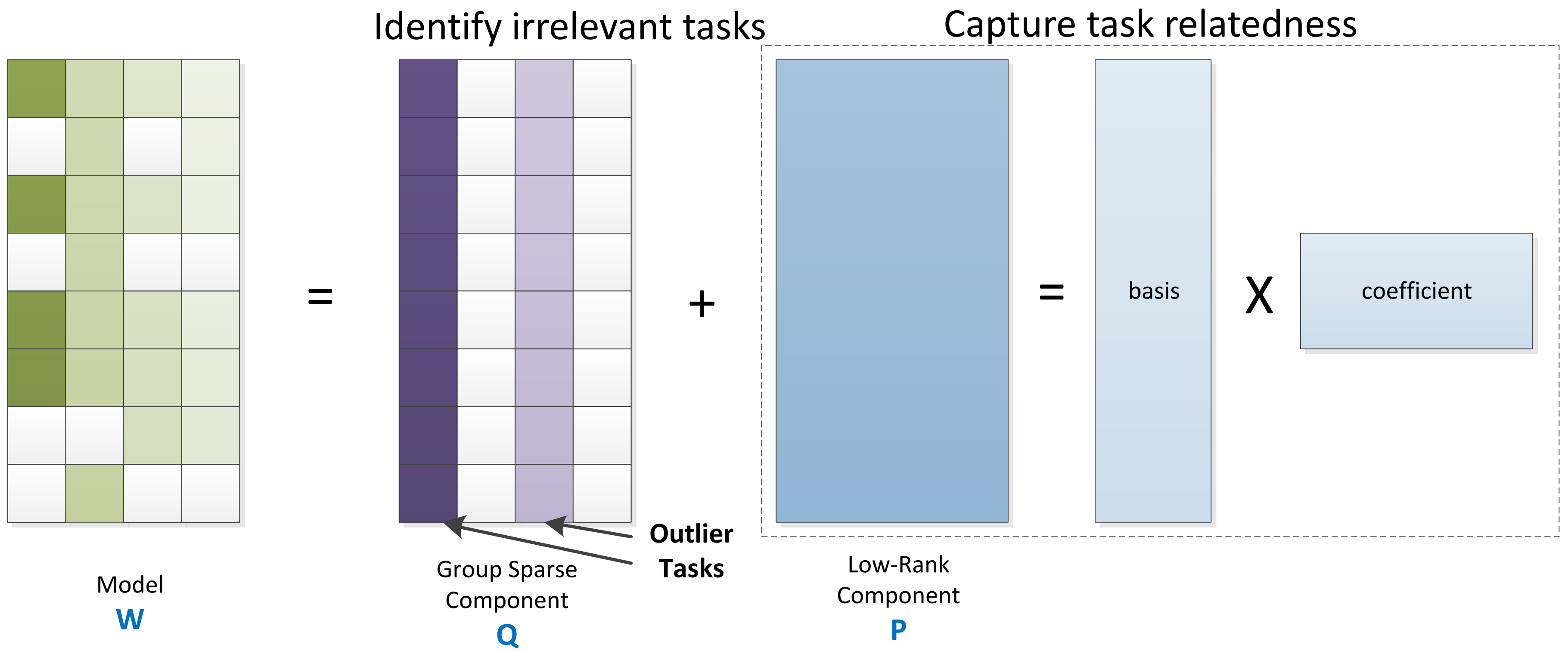
subject to $\|P\|_* \leqslant \eta$

Convex formulation

Robust Low-Rank in MTL

Chen *et. al.* 2011 KDD

- Simultaneously perform low-rank MTL and identify outlier tasks.



$$\min_{\mathrm{P,Q}} \sum_{i=1}^{m} \mathcal{L}_{\mathrm{i}}(\mathrm{X_i}(P_i + Q_i), \mathrm{y_i}) + \alpha\|\mathrm{P}\|_* + \beta\left\|\mathrm{Q}^{\mathrm{T}}\right\|_{1,q}$$

Summary

- All multi-task learning formulations discussed above can fit into the **W=P+Q** schema.
  - Component **P**: shared structure
  - Component **Q**: information not captured by the shared structure

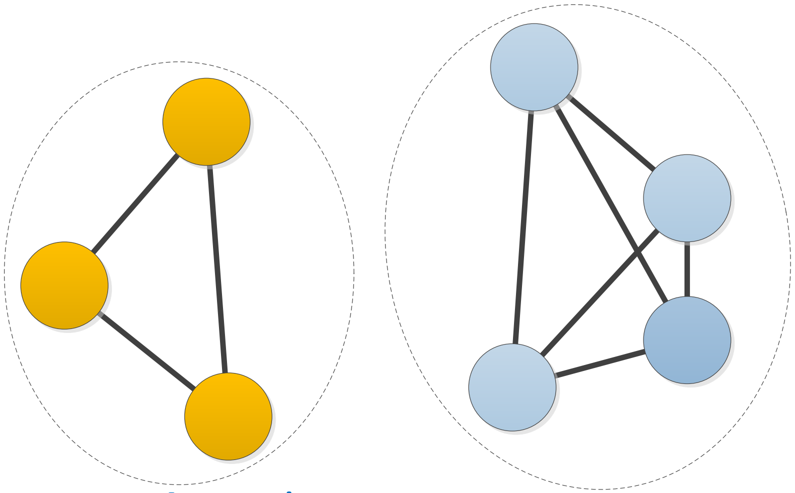
Embedded Feature Selection	Shared Structure P	Component Q
L1/Lq	Feature Selection (L1/Lq Norm)	0
Dirty	Feature Selection (L1/Lq Norm)	L1-norm
rMTFL	Feature Selection (L1/Lq Norm)	Outlier (column-wise L1/Lq Norm)
Low-Rank Subspace Learning		
Trace Norm	Low-Rank (Trace Norm)	0
ISLR	Low-Rank (Trace Norm)	L1-norm
ASO	Low-Rank (Shared Subspace)	L2-norm on independent comp.
RMTL	Low-Rank (Trace Norm)	Outlier (column-wise L1/Lq Norm)

Regularization-based Multi-Task Learning

- All tasks are related
  - Mean-Regularized MTL
  - MTL in high dimensional feature space
    - Embedded Feature Selection
    - Low-Rank Subspace Learning
- **Clustered MTL**
- MTL with Tree/Graph structure

Multi-Task Learning with Clustered Structures

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



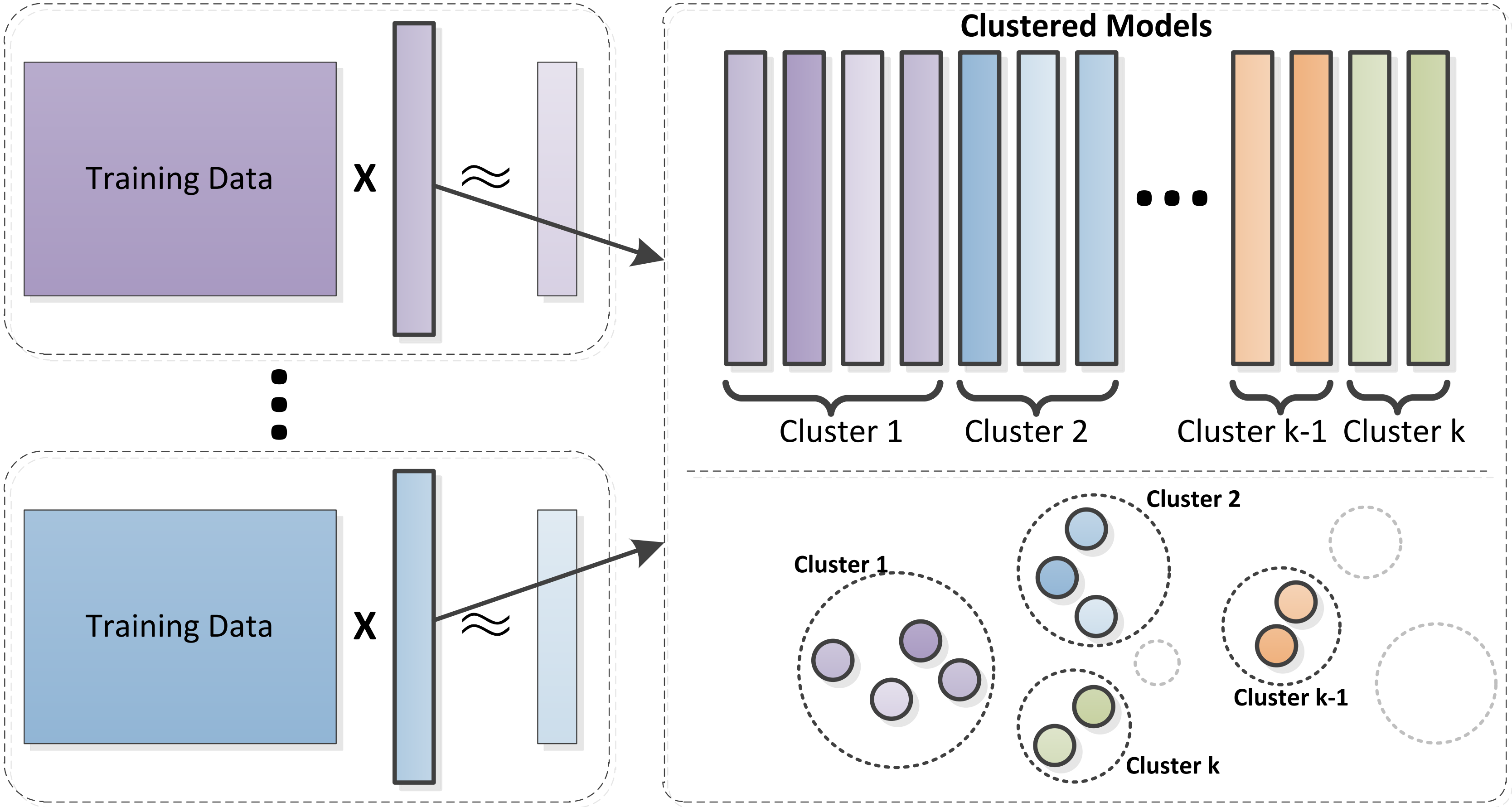
Assumption:
Tasks have group structures

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

Clustered Multi-Task Learning

Jacob *et. al.* 2008 NIPS, Zhou *et. al.* 2011 NIPS

- Use regularization to capture clustered structures.



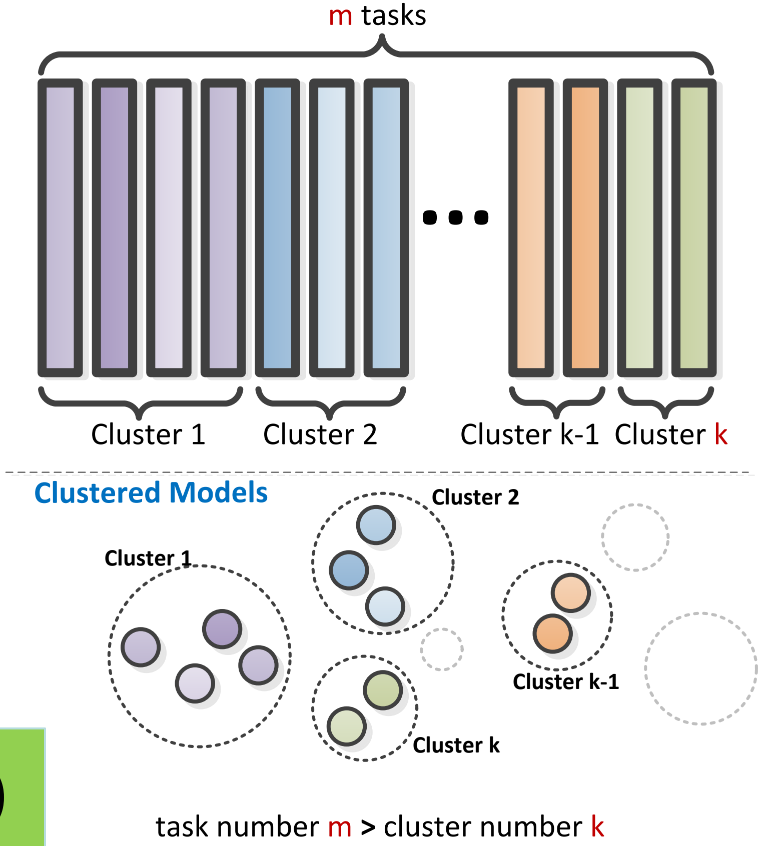
Clustered Multi-Task Learning

Zhou *et. al.* 2011 NIPS

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

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

I_j index set of j^{th} cluster



Equivalent

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

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

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

Clustered Multi-Task Learning

Zhou *et. al.* 2011 NIPS

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

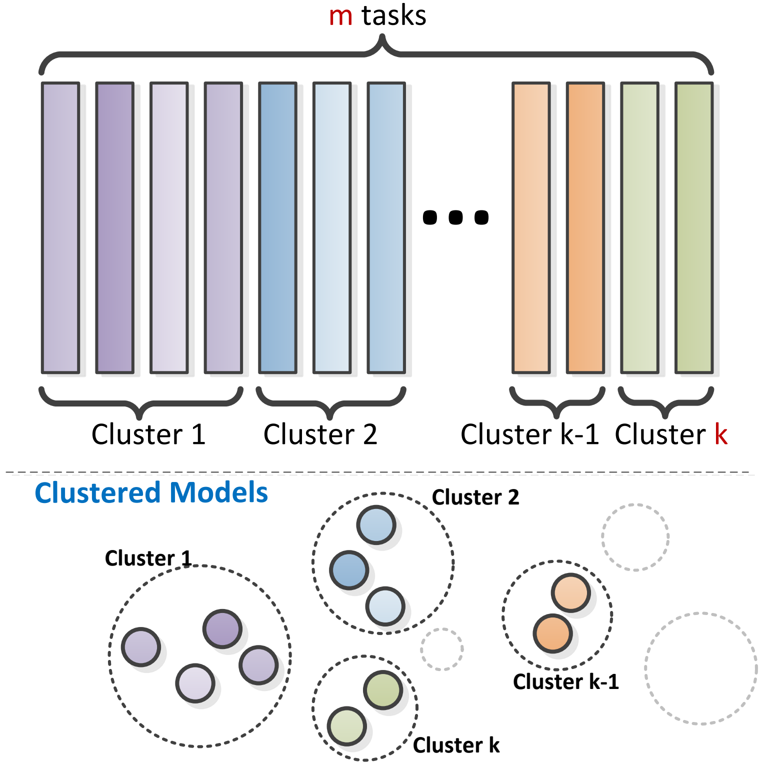
$$\min_{F} \mathrm{tr}(W^T W) - \mathrm{tr}(F^T W^T W F)$$

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

Spectral Relaxation

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

Zha et. al. 2001 NIPS



task number m > cluster number k

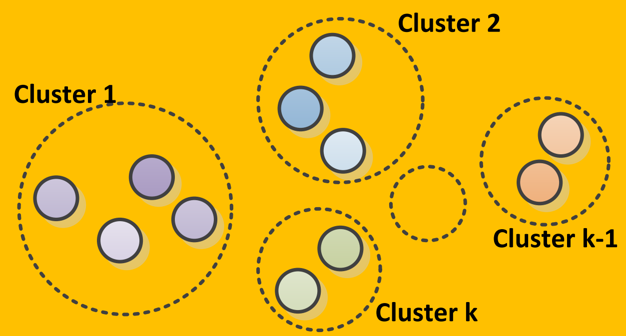
Clustered Multi-Task Learning

Zhou *et. al.* 2011 NIPS

- Clustered multi-task learning (CMTL) formulation

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

capture cluster structures

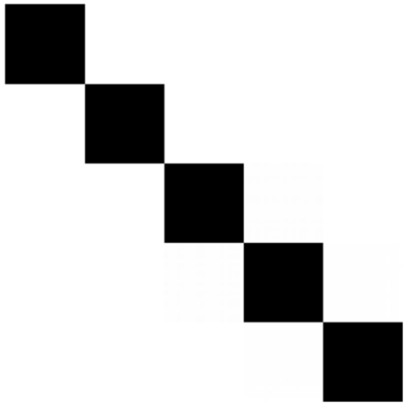


Improves generalization performance

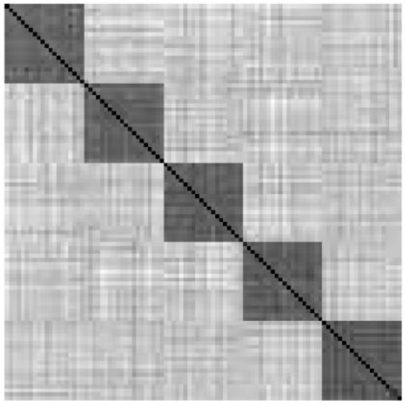
- CMTL has been shown to be equivalent to ASO
  - Given the dimension of the shared low-rank subspace in ASO and the cluster number in clustered multi-task learning (CMTL) are the same.

Convex Clustered Multi-Task Learning

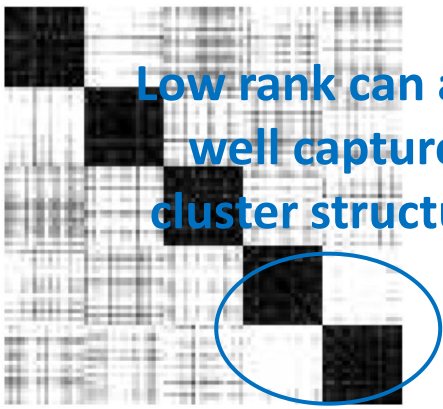
Zhou *et. al.* 2011 NIPS



Ground Truth

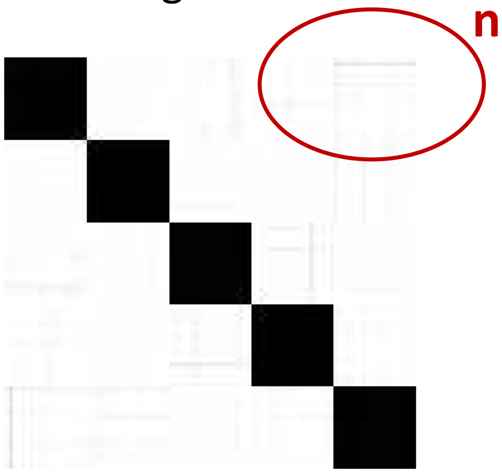


Mean Regularized MTL



Trace Norm Regularized MTL

noise introduced by relaxations



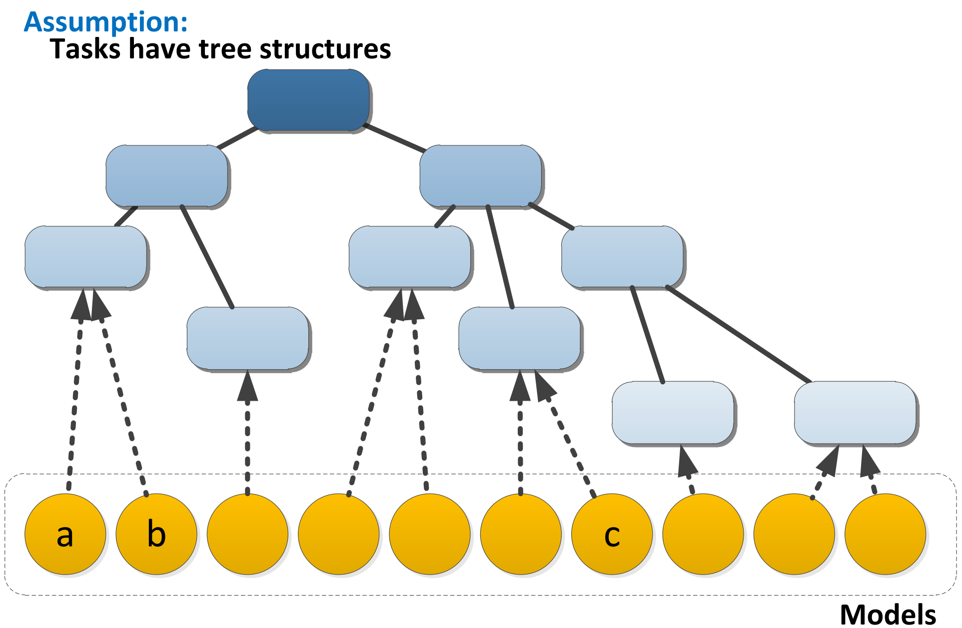
Convex Relaxed CMTL

Regularization-based Multi-Task Learning

- All tasks are related
  - Mean-Regularized MTL
  - MTL in high dimensional feature space
    - Embedded Feature Selection
    - Low-Rank Subspace Learning
- Clustered MTL
- **MTL with Tree/Graph structure**

Multi-Task Learning with Tree Structures

- In some applications, the tasks may be equipped with a tree structure:
  - The tasks belonging to the same node are similar to each other
  - The similarity between two nodes is related to the depth of the 'common' tree node



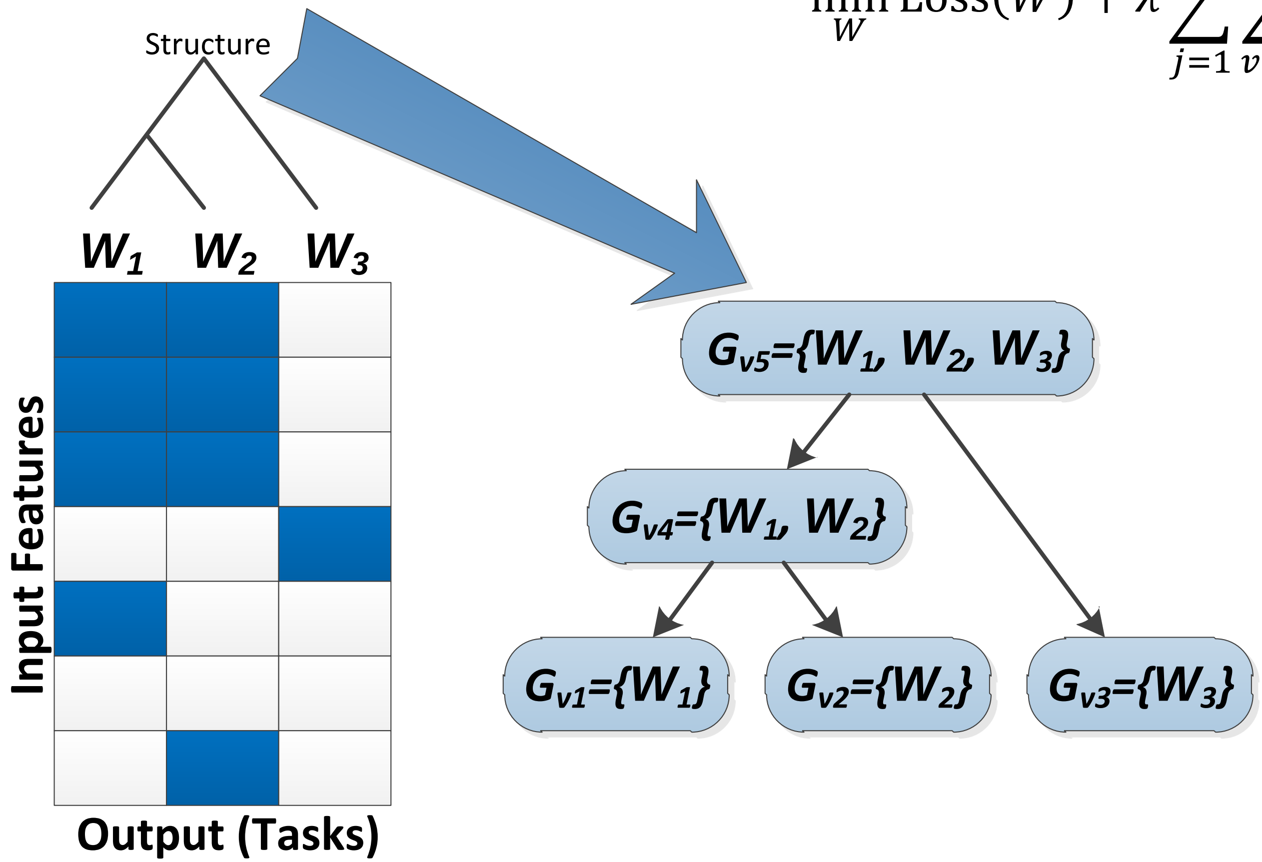
Task a is more similar with b, compared to c

Multi-Task Learning with Tree Structures

Kim and Xing 2010 ICML

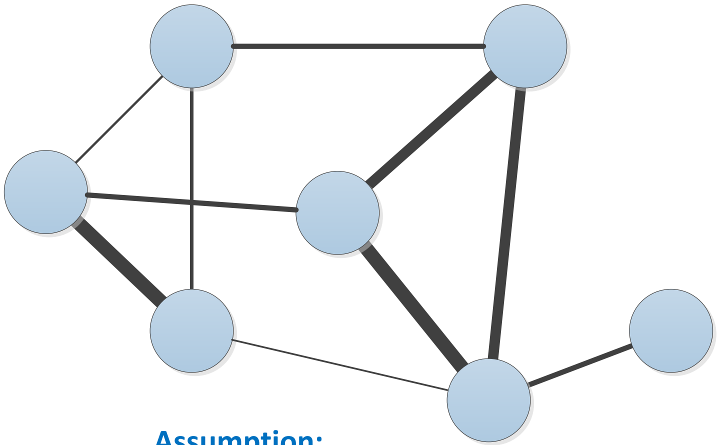
- Tree-Guided Group Lasso

$$\min_{W} \mathrm{Loss}(W) + \lambda \sum_{j=1}^{d} \sum_{v \in V} w_v \left\| W_{G_v}^{j} \right\|_2$$



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

- A simple way to encode graph structure is to penalize the difference of two tasks that have an edge between them
- Given a set of edges E, we thus penalize:

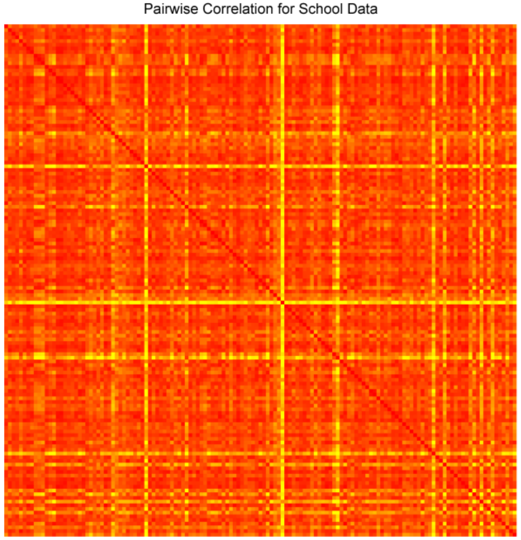
$$\sum_{i=1}^{|E|} \left\| W_{e_{\{i,1\}}} - W_{e_{\{i,2\}}} \right\|_2^2 = \|W R^T\|_F^2 \quad R \in \mathbb{R}^{|E|\times m}$$

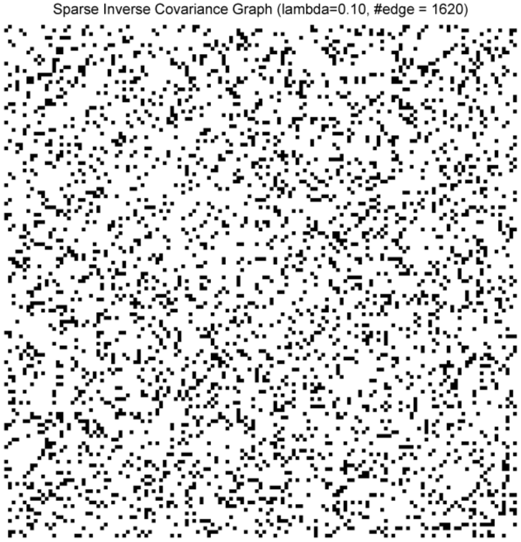
- The graph regularization term can also be represented in the form of Laplacian term

$$\|W R^T\|_F^2 = tr\left((W R^T)^T W R^T\right) = tr(W R^T R W^T) = tr(W \mathcal{L} W^T)$$

Multi-Task Learning with Graph Structures

- How to obtain graph information
  - External domain knowledge
    - protein-protein interaction (PPI) for microarray
  - Discover task relatedness from data
    - Pairwise correlation coefficient
    - Sparse inverse covariance (Friedman *et. al.* 2008 Biostatistics)

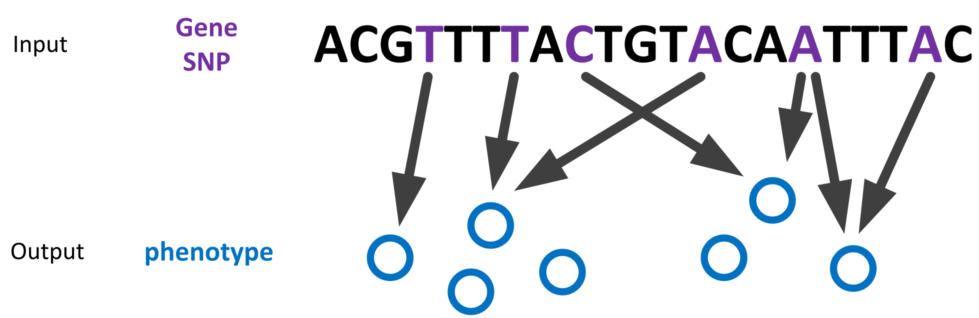




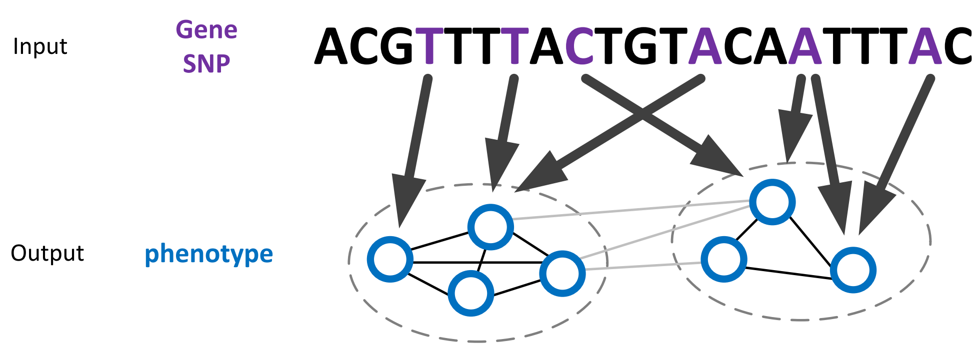
Multi-Task Learning with Graph Structures

Chen *et. al.* 2011 UAI, Kim *et. al.* 2009 Bioinformatics

- Graph-guided Fused Lasso



Lasso



Graph-Guided Fused Lasso

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

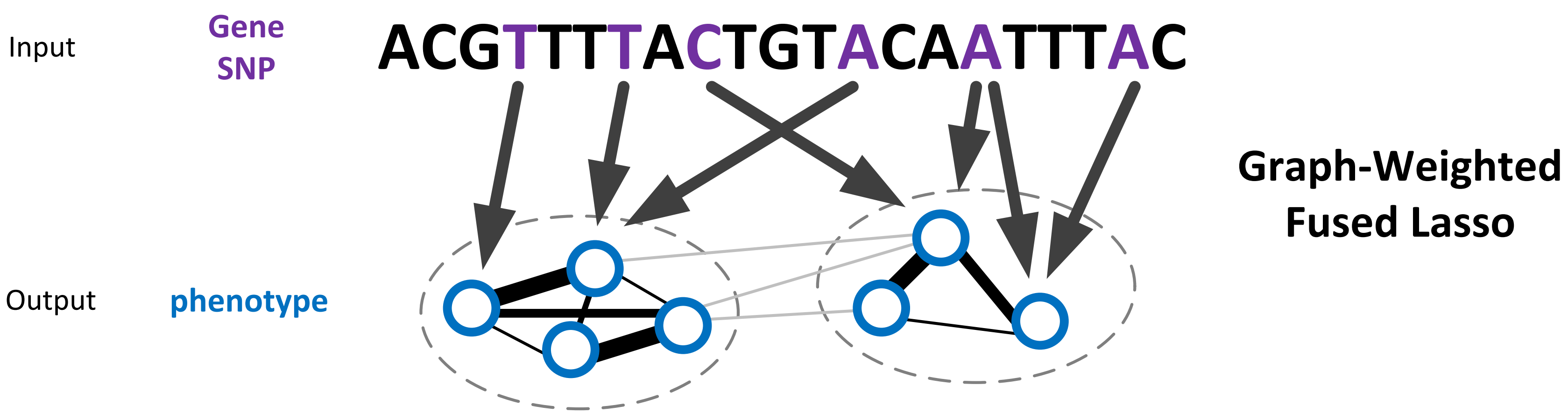
Graph-guided Fusion Penalty

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

Multi-Task Learning with Graph Structures

Kim *et. al.* 2009 Bioinformatics

- In some applications, we know not only which pairs are related, but also **how** they are related.
- Graph-Weighted Fused Lasso.



$$\min_{W} \mathrm{Loss}(W) + \lambda \|W\|_1 + \gamma \sum_{e=(m,l)\in E} \tau(r_{ml}) \sum_{j=1}^{J} \left| W_{jm} - sign(r_{ml}) W_{jl} \right|$$

Added weight information!

Practical Guideline

- **MTL versus STL**
  - MTL is preferred when dealing with multiple related tasks with small number of training samples
- **Shared features versus shared subspace**
  - Identifying shared features is preferred When the data dimensionality is large
  - Identifying a shared subspace is preferred when the number of tasks is large

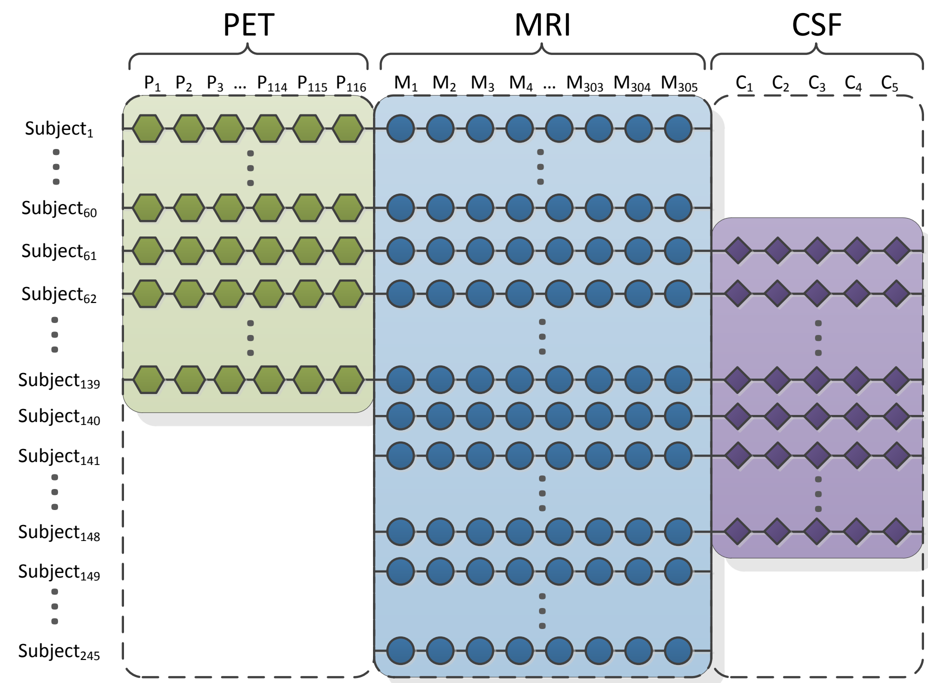
Tutorial Road Map

- Part I: Multi-task Learning (MTL) background and motivations
- Part II: MTL formulations
- **Part III: Case study of real-world applications**
  - Incomplete Multi-Source Fusion
  - *Drosophila* Gene Expression Image Analysis
- Part IV: An MTL Package (MALSAR)
- Current and future directions

Case Study I: Incomplete Multi-Source Data Fusion

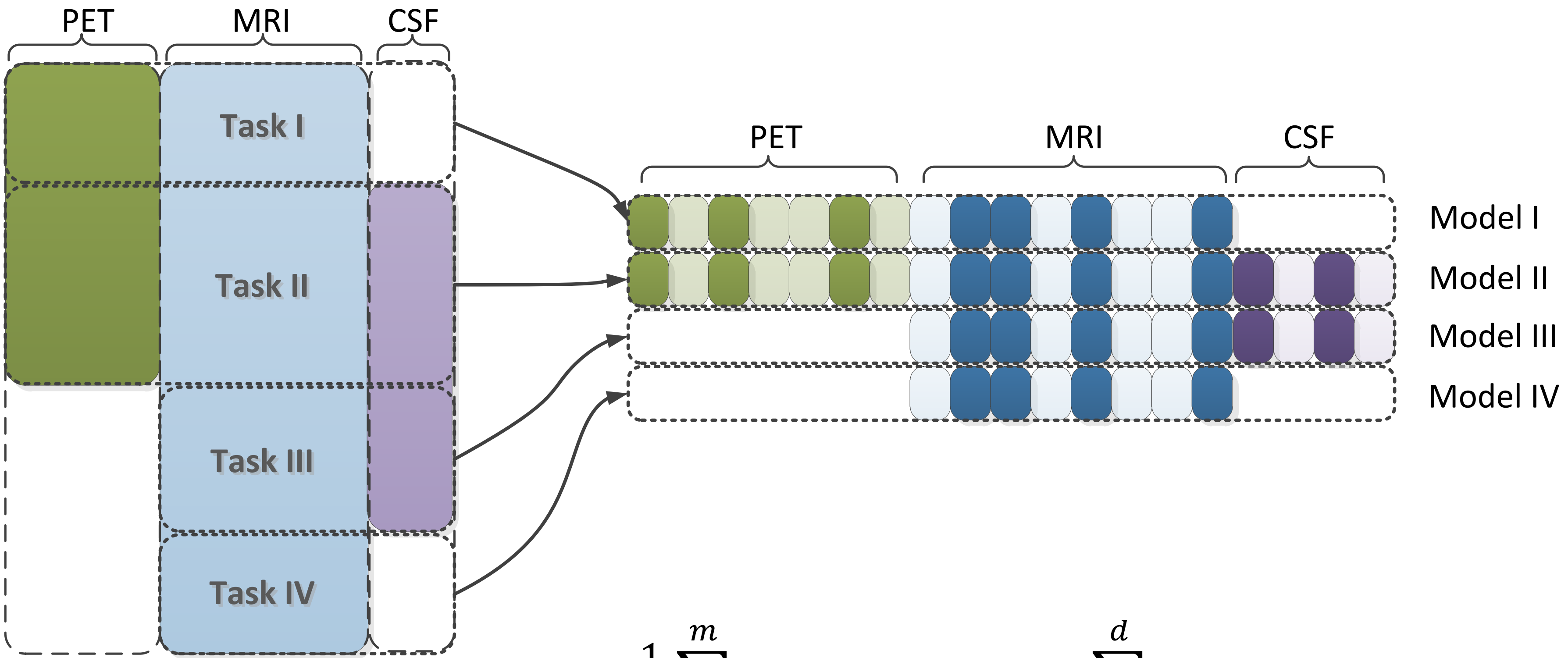
Yuan *et. al.* 2012 NeuroImage

- In many applications, multiple data sources may contain 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.



Overview of iMSF

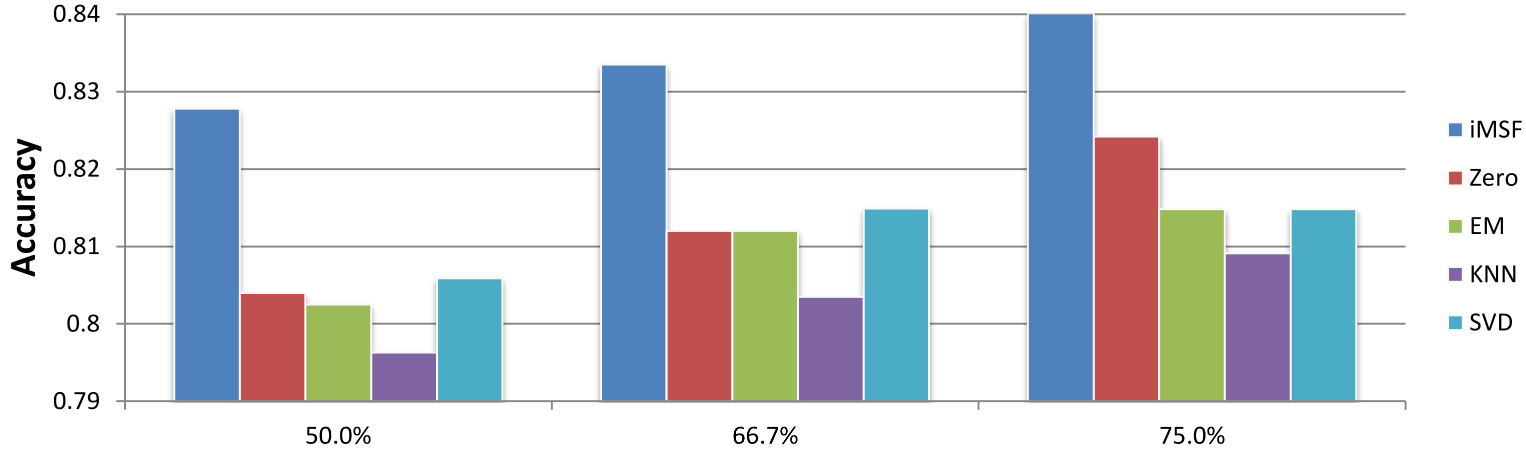
Yuan *et. al.* 2012 NeuroImage

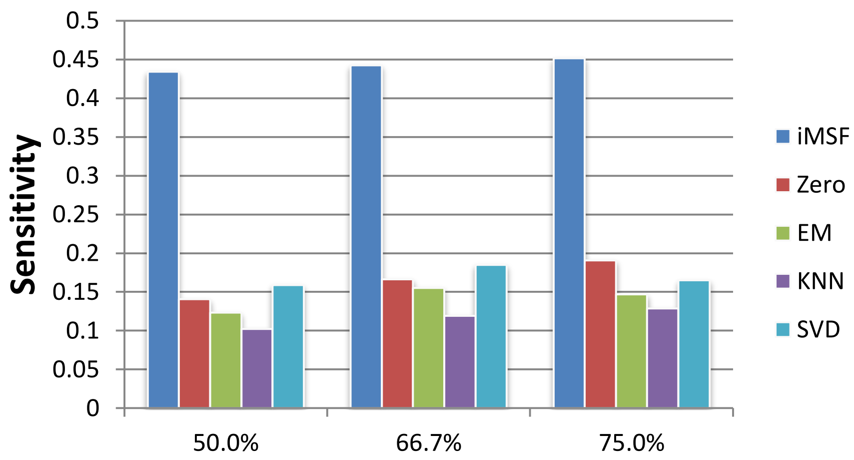


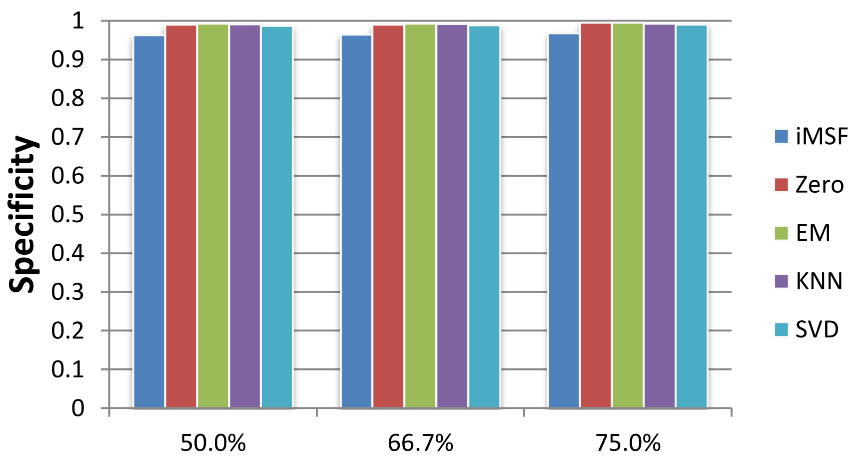
$$\min_{W} \frac{1}{m} \sum_{i=1}^{m} Loss(X_i, Y_i, W_i) + \lambda \sum_{k=1}^{d} \left\| W_{G_k} \right\|_2$$

iMSF: Performance

Yuan *et. al.* 2012 NeuroImage

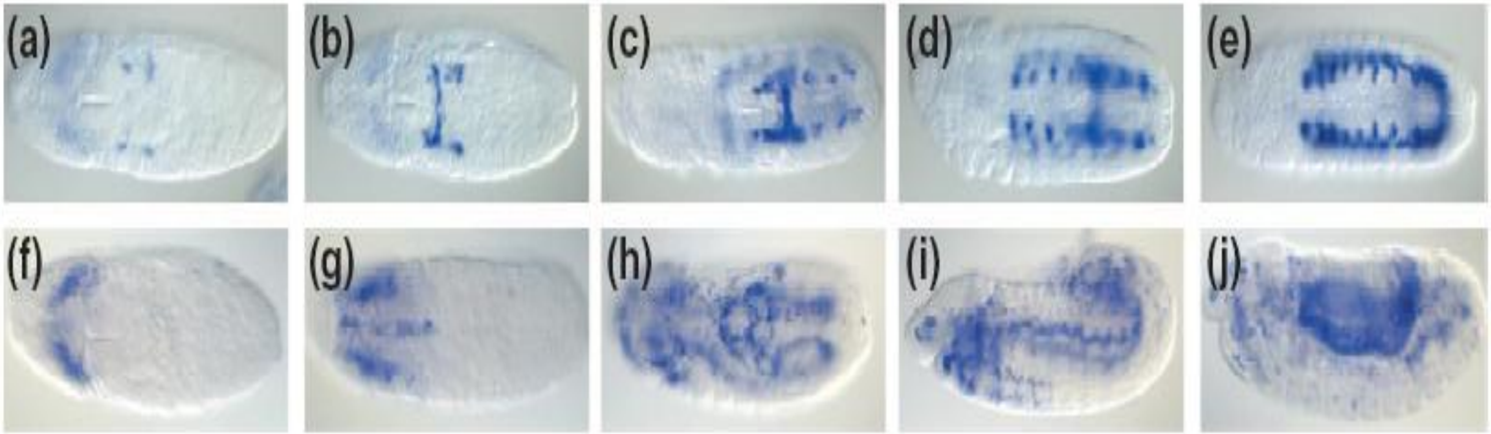






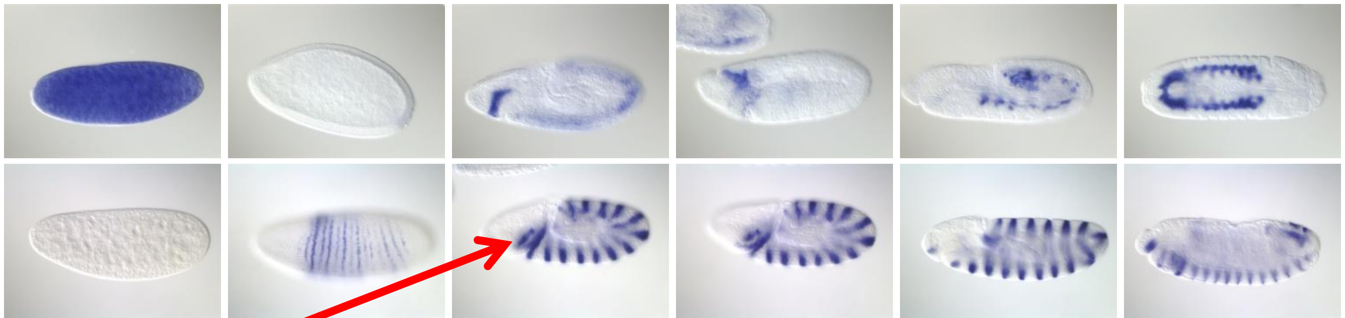
Case Study II: *Drosophila* Gene Expression Image Analysis

- *Drosophila* (fruit fly) is a favorite model system for geneticists and developmental biologists studying embryogenesis.
  - The small size and short generation time make it ideal for genetic studies.
- *In situ hybridization* allows us to generate images showing when and where individual genes were active.
  - The analysis of such images can potentially reveal gene functions and gene-gene interactions.



Drosophila gene expression pattern images

bgm
en

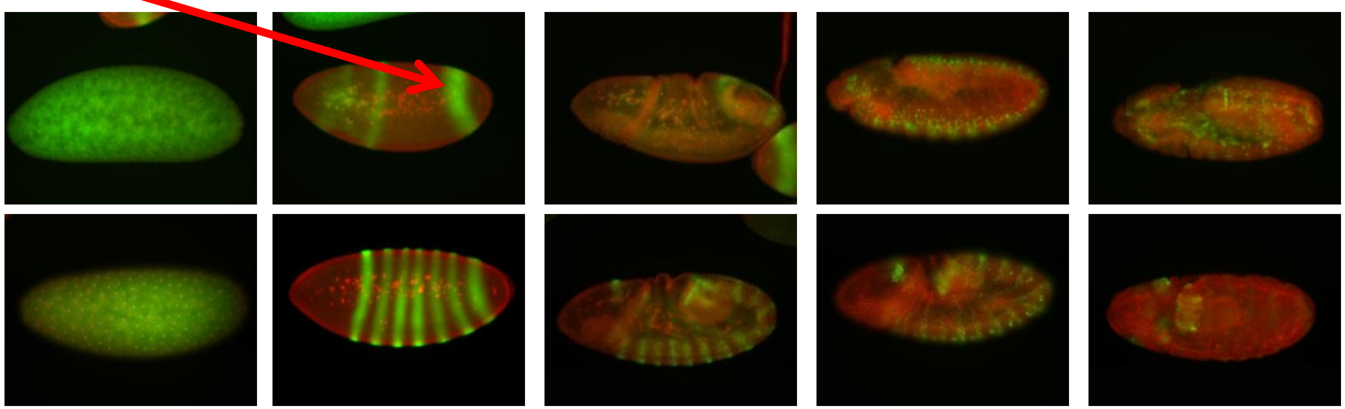


Stage 1-3 Stage 4-6 Stage 7-8 Stage 9-10 Stage 11-12 Stage 13-

Expressions

Berkeley *Drosophila* Genome Project (BDGP)
http://www.fruitfly.org/

hb
runt

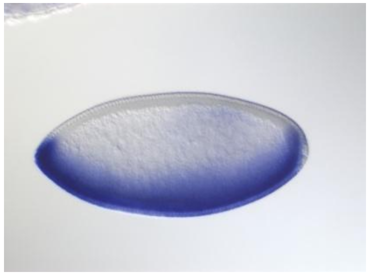
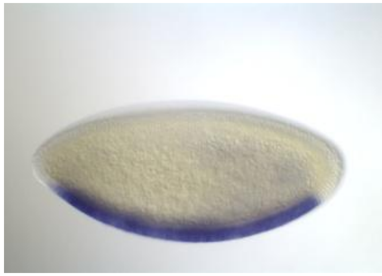


Stage 1-3 Stage 4-5 Stage 6-7 Stage 8-9 Stage 10-

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

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

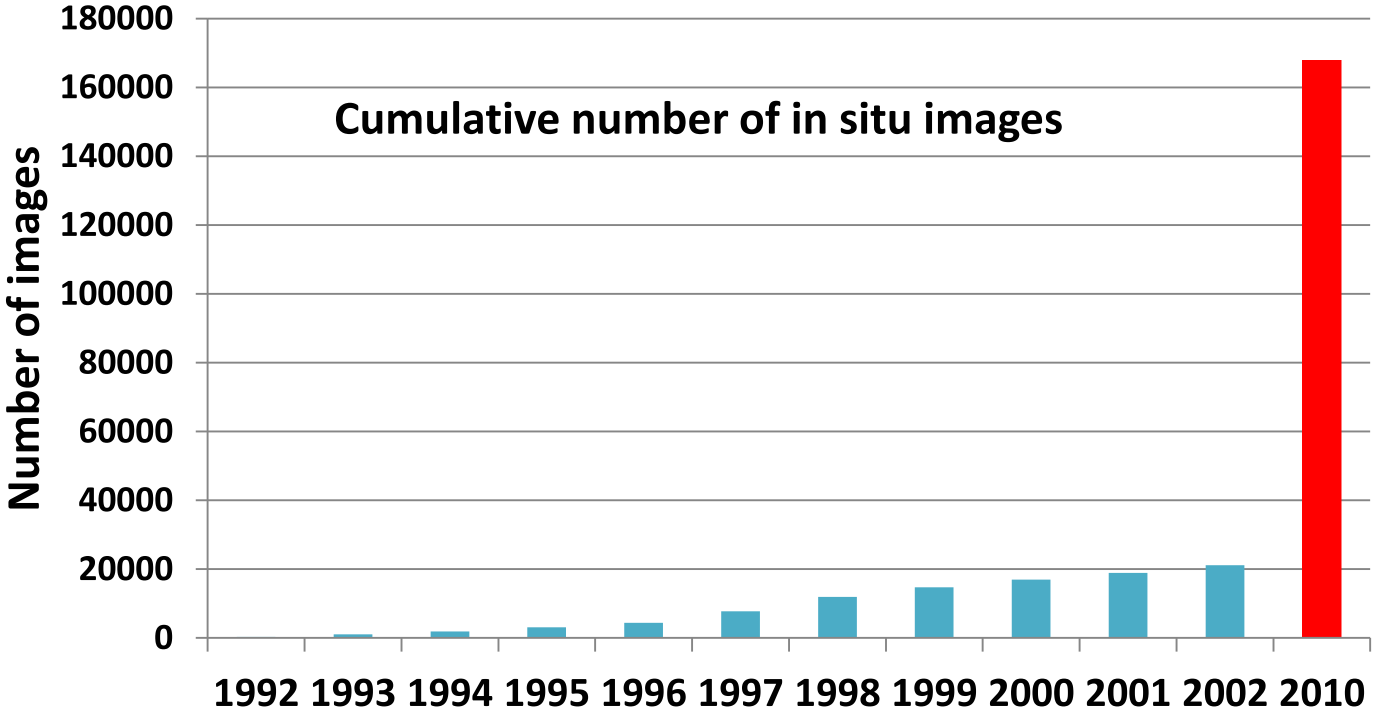
Comparative image analysis

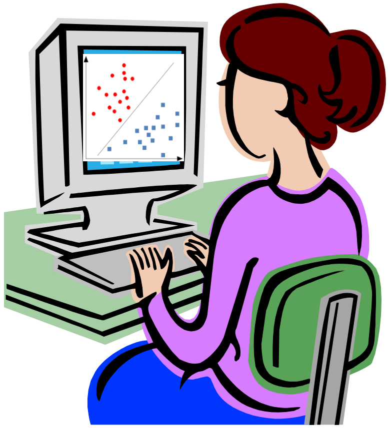
	Twist	*heartless*	*stumps*
stage 4-6	 anterior endoderm AISN trunk mesoderm AISN subset cellular blastoderm mesoderm AISN	 dorsal ectoderm AISN procephalic ectoderm AISN subset cellular blastoderm mesoderm AISN	 anterior endoderm AISN trunk mesoderm AISN head mesoderm AISN
stage 7-8	 trunk mesoderm PR 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.]

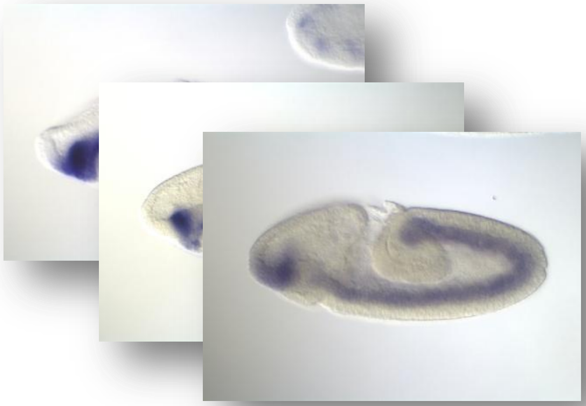
Challenges of manual annotation





Method outline

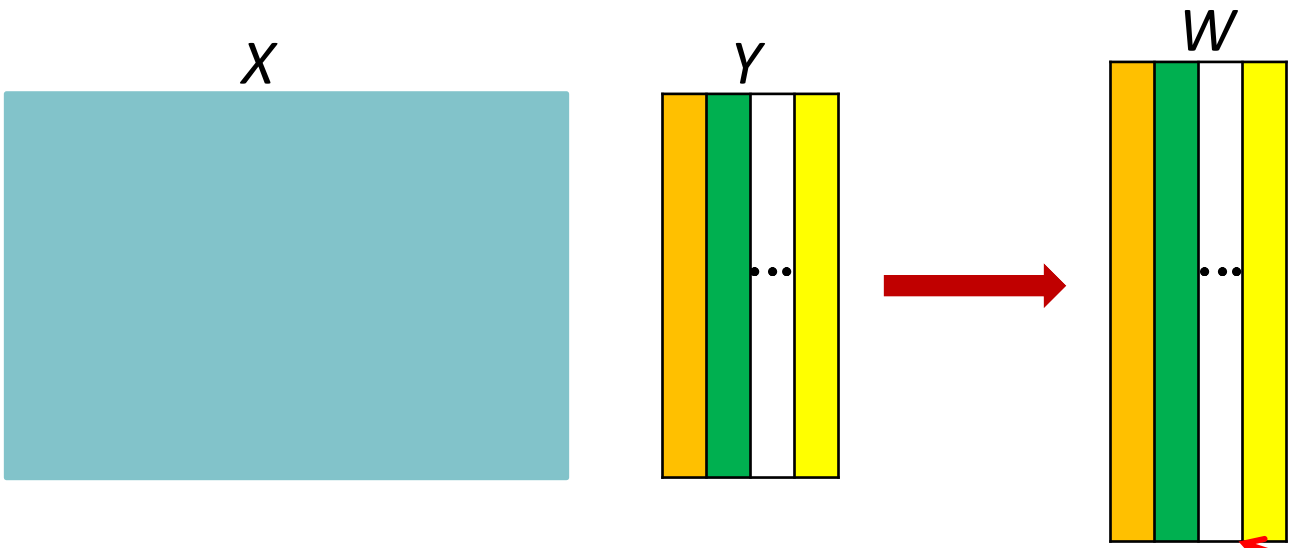
Ji *et. al*. 2008 Bioinformatics; Ji *et. al*. 2009 BMC Bioinformatics; Ji *et. al*. 2009 NIPS



Images → Feature extraction: Sparse coding → Model construction: Low-rank multi-task; Graph-based multi-task

Low rank multi-task learning model

Ji *et. al*. 2009 BMC Bioinformatics



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

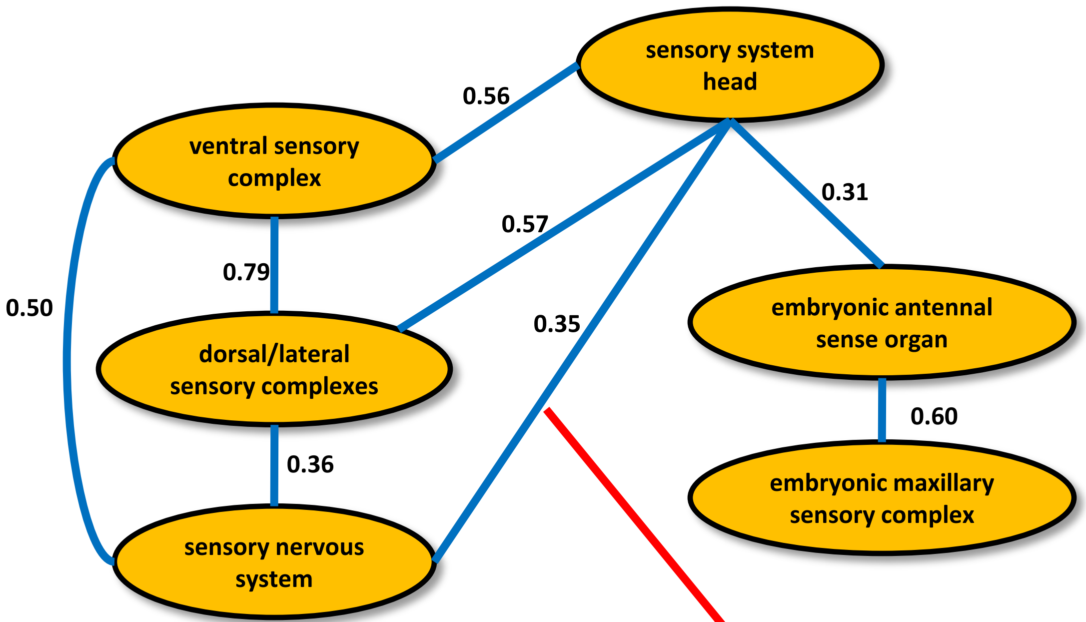
Low rank

trace norm = sum of singular values

Loss term

Graph-based multi-task learning model

Ji *et. al*. 2009 SIGKDD

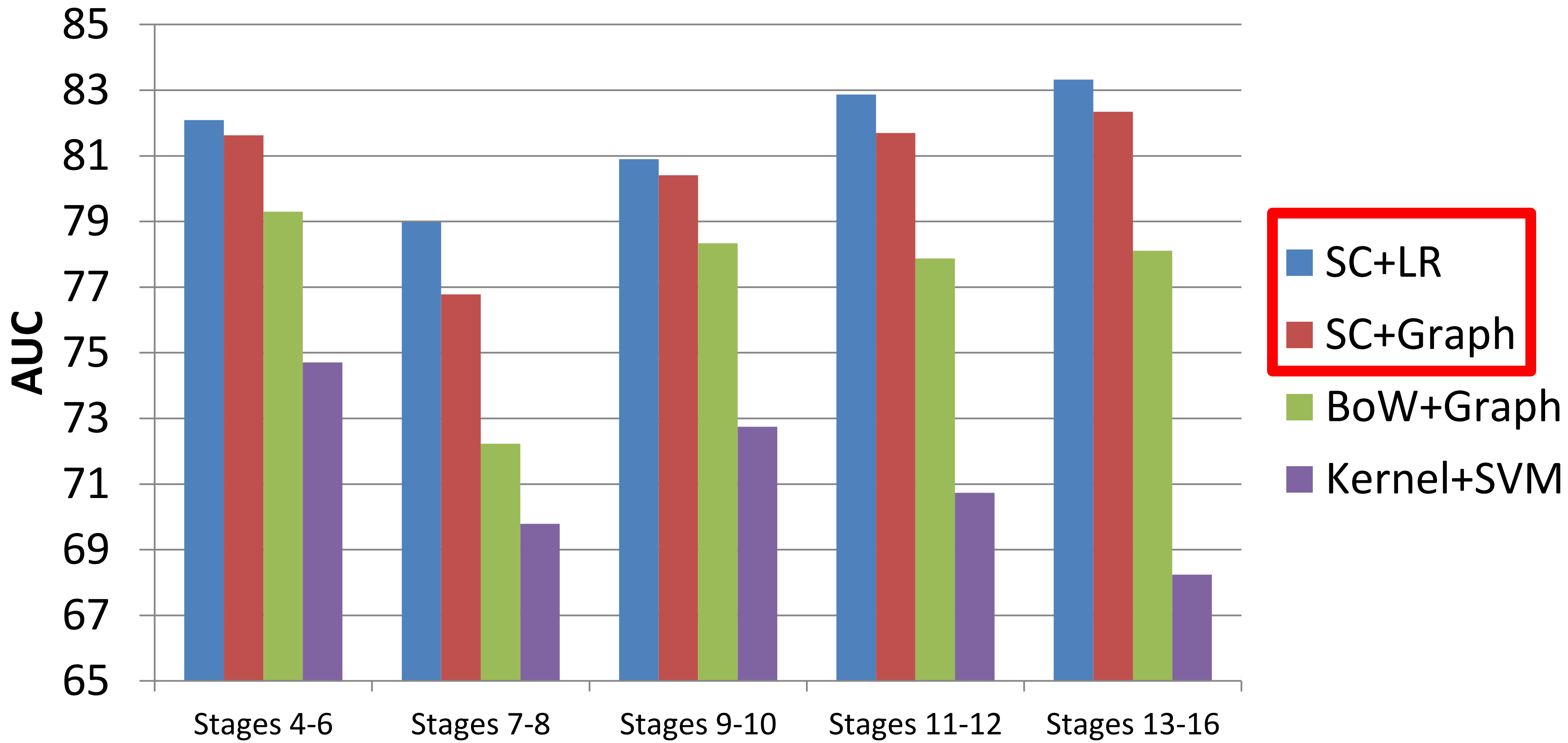


loss 2-norm

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

Closed-form solution

Spatial annotation performance



- 50% data for training and 50% for testing and 30 random trials are generated
- Multi-task approaches outperform single-task approaches

Tutorial Road Map

- Part I: Multi-task Learning (MTL) background and motivations
- Part II: MTL formulations
- Part III: Case study of real-world applications
  - Incomplete Multi-Source Fusion
  - *Drosophila* Gene Expression Image Analysis
- **Part IV: An MTL Package (MALSAR)**
- Current and future directions

MALSAR

Multi-Task Learning via Structural Regularization

Jiayu Zhou, Jianhui Chen, Jieping Ye

- A multi-task learning package
- Encode task relationship via structural regularization
- www.public.asu.edu/~jye02/Software/MALSAR/

MTL Algorithms in MALSAR 1.0

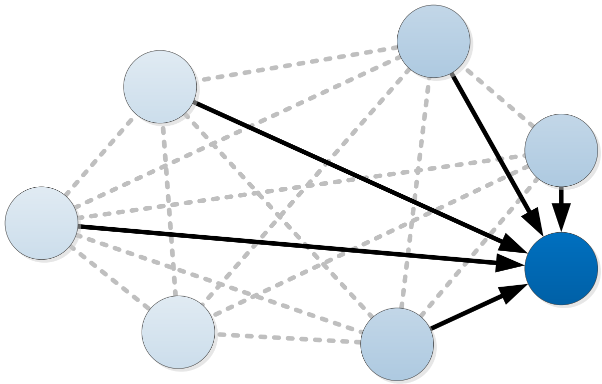
- Mean-Regularized Multi-Task Learning
- MTL with Embedded Feature Selection
  - Joint Feature Learning
  - Dirty Multi-Task Learning
  - Robust Multi-Task Feature Learning
- MTL with Low-Rank Subspace Learning
  - Trace Norm Regularized Learning
  - Alternating Structure Optimization
  - Incoherent Sparse and Low Rank Learning
  - Robust Low-Rank Multi-Task Learning
- Clustered Multi-Task Learning
- Graph Regularized Multi-Task Learning

Tutorial Road Map

- Part I: Multi-task Learning (MTL) background and motivations
- Part II: MTL formulations
- Part III: Case study of real-world applications
  - Incomplete Multi-Source Fusion
  - *Drosophila* Gene Expression Image Analysis
- Part IV: An MTL Package (MALSAR)
- **Current and future directions**

Trends in Multi-Task Learning

- **Develop efficient algorithms** for large-scale multi-task learning.
- **Semi-supervised** and **unsupervised MTL**
- Learn **task structures automatically** in MTL
- Asymmetric MTL
- Cross-Domain MTL
  - The features may be different



The relationship is not mutual

Acknowledgement

- National Science Foundation
- National Institute of Health

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.