Multi-Source Visual Analytics

Jieping Ye
Arizona State University

Co-PIs: Anshuman Razdan, Peter Wonka

Multi-Source Data Transformation and Visualization

- Multiple sources for
  - Neuroimages and ge
    subjects
  - Texts and images for a collection of web pages

- Multiple sources for t
  - Biomedical images
  - GWAS data from mu

❑ Heterogeneous

❑ Incomplete

Different data distributions

❑ Marginal

❑ Conditional

Technical Achievements

- Methods and theories for multi-source domain adaptation (KDD'11, NIPS'11, NIPS'12, TKDD'13)
  - Distribution differences
- Methods for incomplete multi-source data fusion (KDD'12, NeuroImage'12)
  - Heterogeneous, incomplete
- Supervised and unsupervised dimensionality reduction algorithms
  - Part of FODAVA Testbed Software

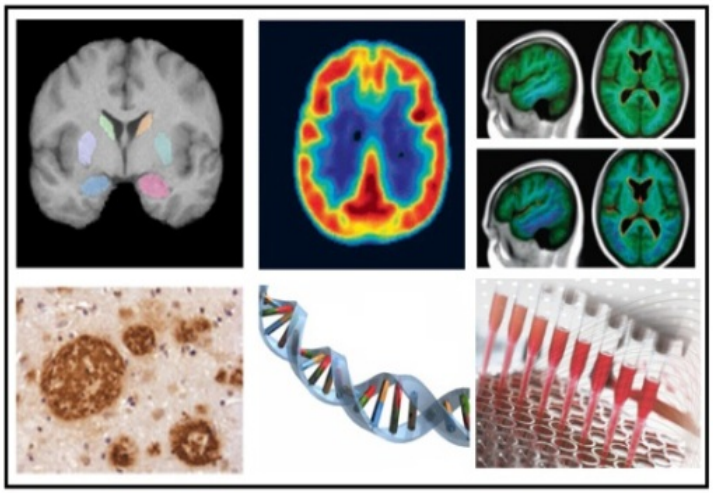
Technical Achievements

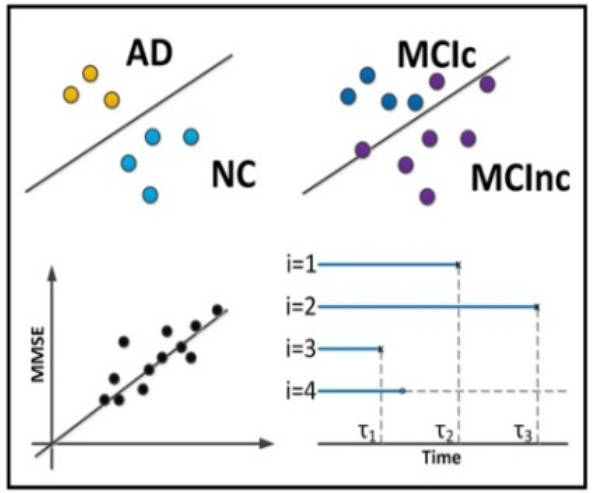
- Methods and theories for multi-source domain adaptation (KDD'11, NIPS'11, NIPS'12, TKDD'13) (**FODAVA Annual Review'11**)

- **Methods for incomplete multi-source data fusion (KDD'12, NeuroImage'12)**

- Supervised and unsupervised dimensionality reduction algorithms (**FODAVA Annual Review'10**)
  - Part of FODAVA Testbed Software

Alzheimer's Disease Neuroimaging Initiative





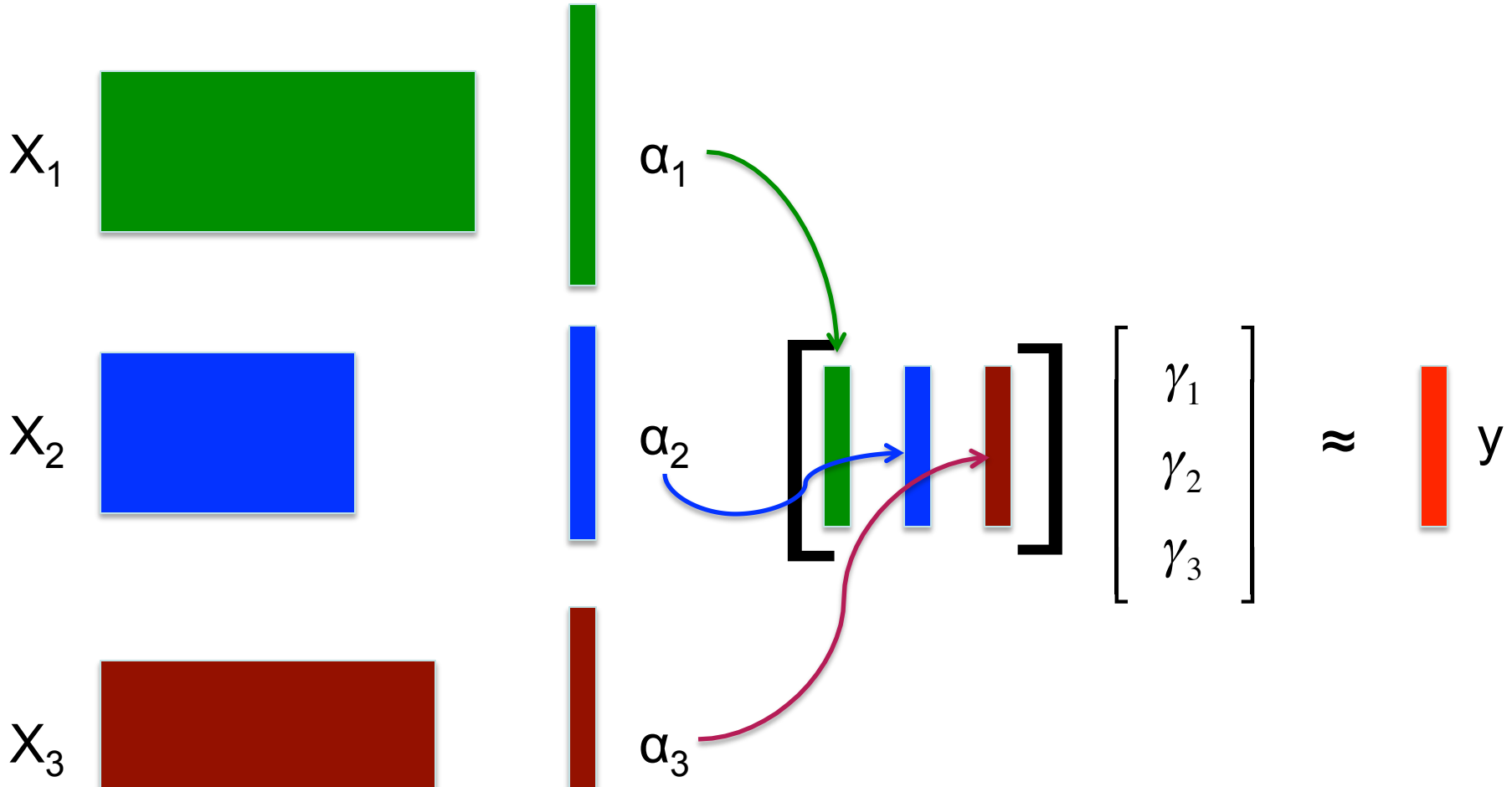
200 NC (normal controls)
400 MCI (mild cognitive impairment)
200 AD (Alzheimer's disease patient)

MRI, PET, Proteomics, GWAS, CSF

The primary goal of ADNI has been to test whether **serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined** to measure the progression of MCI and early AD.

ADNI Alzheimer's Disease Neuroimaging Initiative

Multi-Source Feature Learning: A Unified Framework



$$\frac{1}{2}\left\| y - \sum_{i=1}^{S} \gamma_i \cdot X_i \alpha_i \right\|_2^2 + \sum_{i=1}^{S} \lambda_i \left\| \alpha_i \right\|_p^p + \sum_{i=1}^{S} \eta_i \left| \gamma_i \right|^q$$

Multi-Source Feature Learning: A Unified Framework

$$\min_{\alpha,\gamma} \frac{1}{2}\left\| y - \sum_{i=1}^{S} \gamma_i \cdot X_i \alpha_i \right\|_2^2 + \sum_{i=1}^{S} \lambda_i \left\| \alpha_i \right\|_p^p + \sum_{i=1}^{S} \eta_i \left| \gamma_i \right|^q$$

- p=1, q=∞: Lasso **(feature selection)**
- p=2, q=2: Group Lasso via $L_{2,1}$ norm (multiple kernel learning)
- p=∞, q=1: Group Lasso via $L_{\infty,1}$ norm **(source selection)**
- p=1, q=2: Sparse group Lasso **(source and feature selection)**

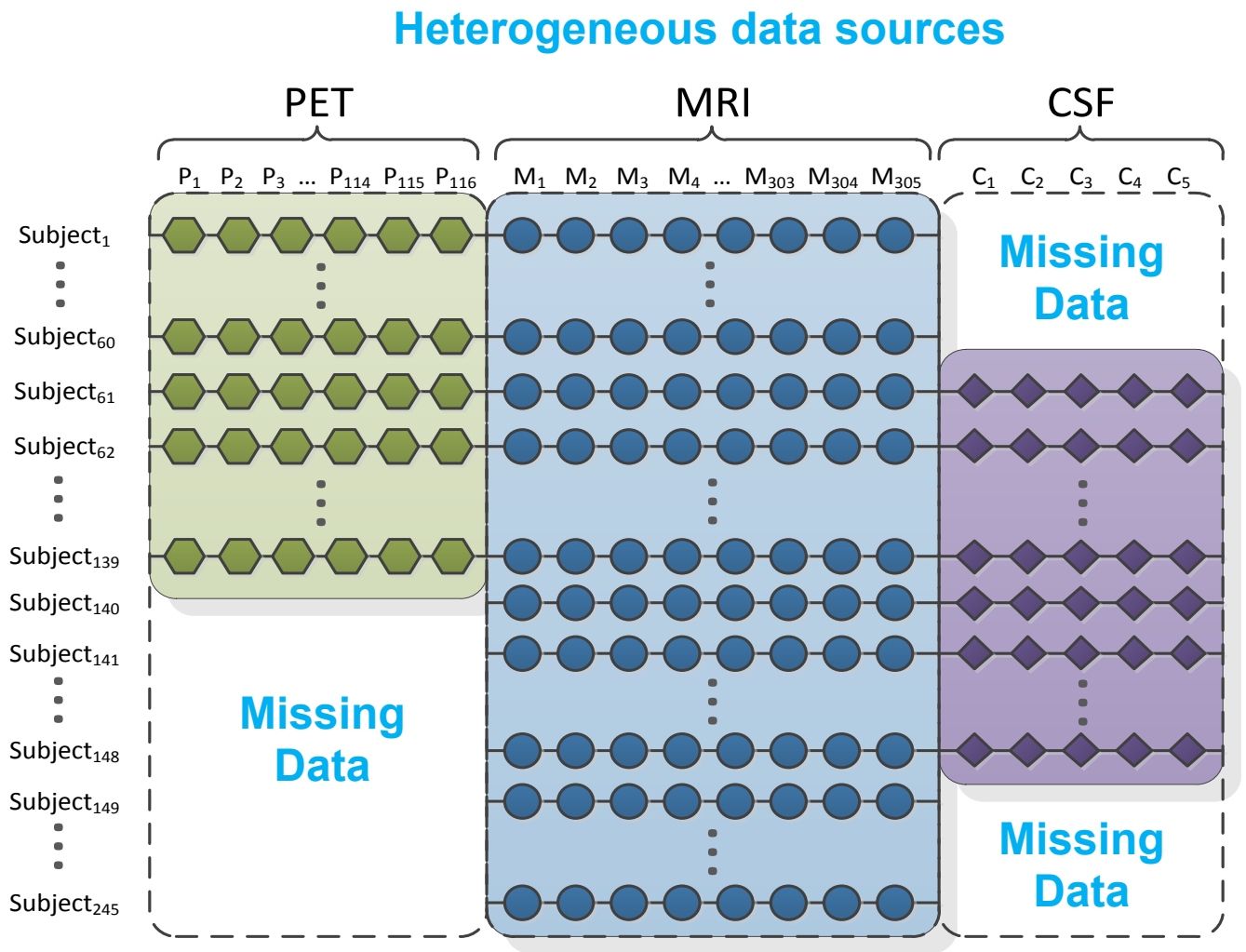
Challenges: Blockwise Missing Data

- Missing data may be due to
  - High cost of certain measures (e.g., PET scans)
  - Poor data quality
  - Dropout of the patients from the study

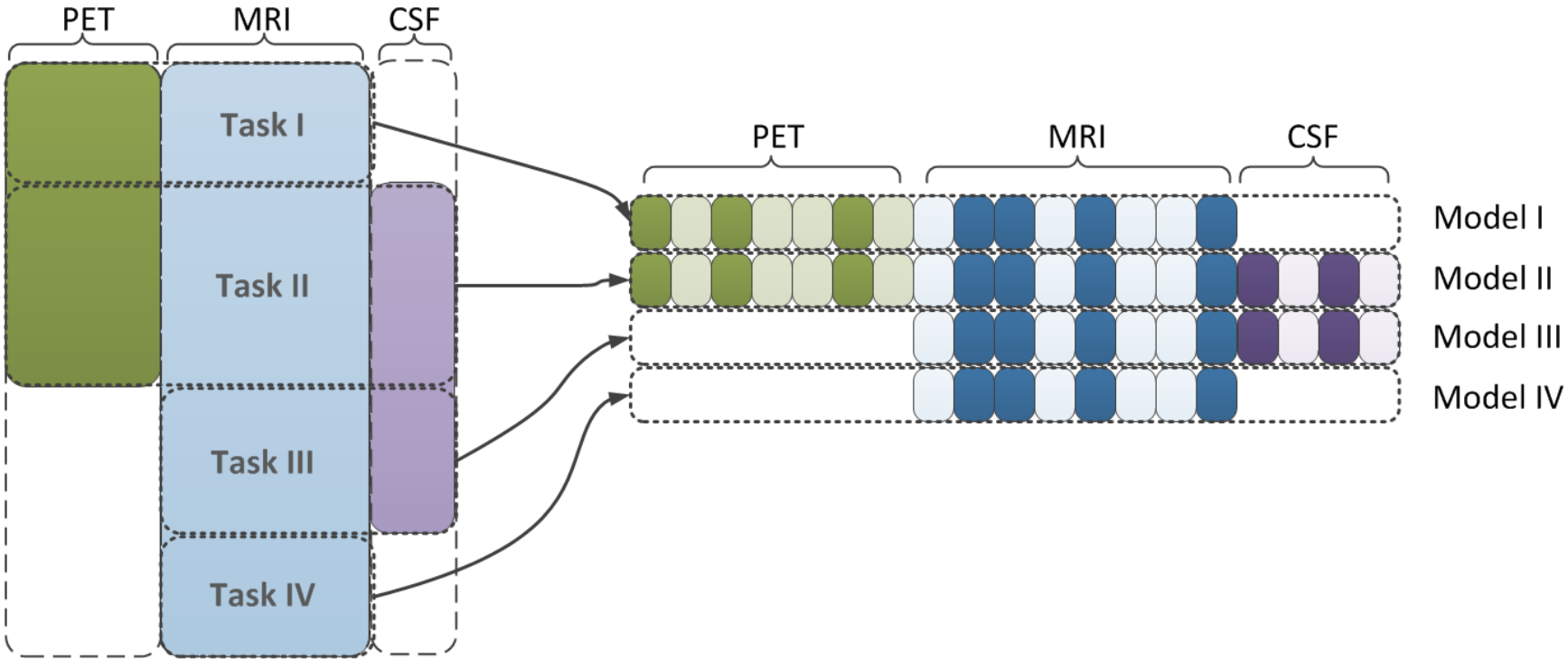
- Some measures require more invasive procedures which not all study participants are willing to consent to.

- Some subjects in a longitudinal study may miss at least one of the regular assessments.

Example: The ADNI Database



Incomplete Multi-Source Feature Learning Model (iMSF)



L. Yuan, Y. Wang, P. Thompson, V. Narayan, and J. Ye. **NeuroImage**, 2012.

iMSF: Formulation

- Partition the problem into multiple tasks according to the availability of data sources
- A linear model is learned for each task
- Features are selected jointly among all tasks
- $\ell_{2,1}$-norm regularization is used for joint feature learning (consistent feature selection)

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

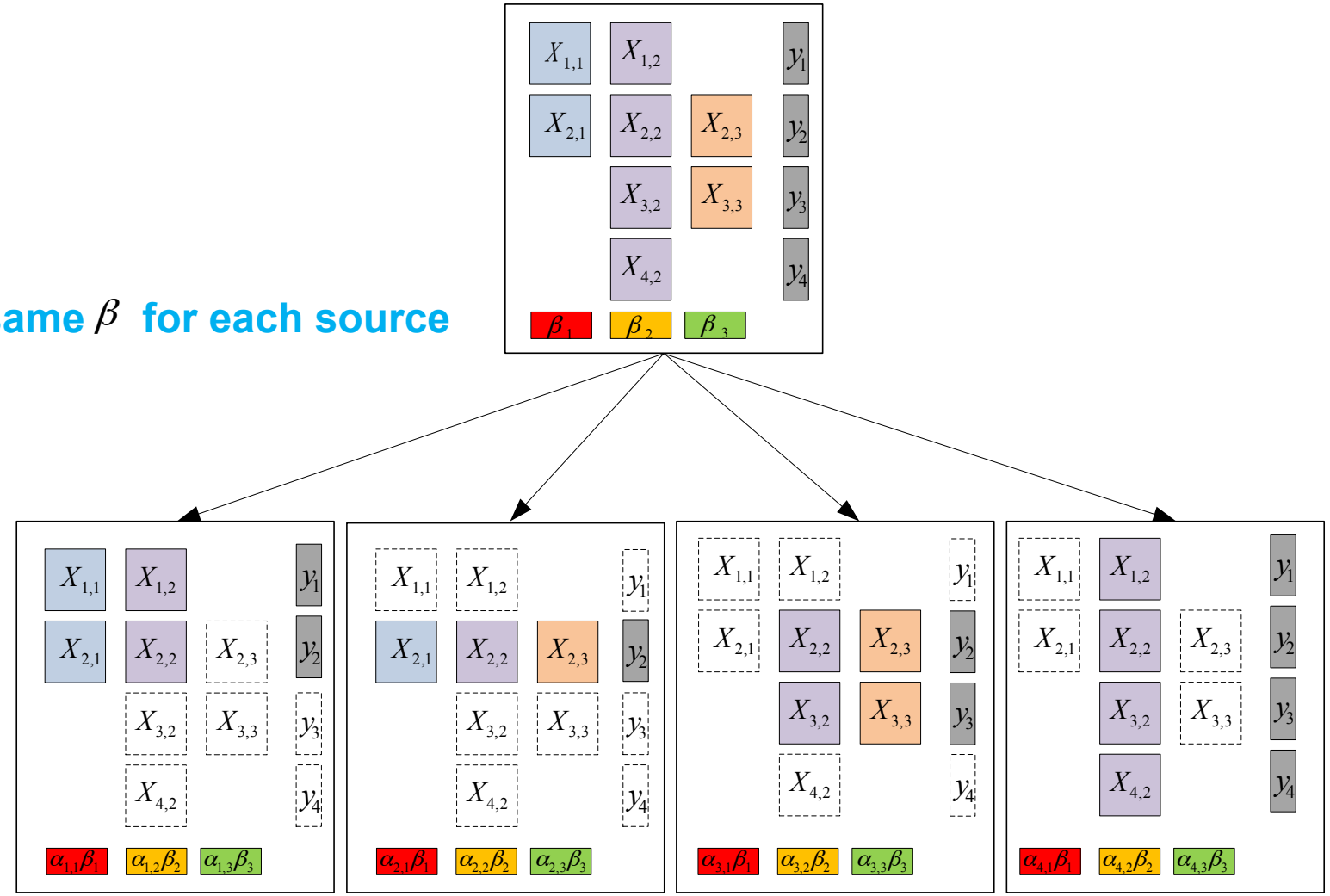
iMSF: Pros and Cons

- Pros
  - Every sample can be used as long as at least one of the data source is available
  - No imputation needed
  - Interpretable feature learning results due to sparsity
  - Efficient convex optimization
- Cons
  - The data source combination grows exponentially
  - Different models for the same source seems unnatural
  - Unable to perform source-level selection

Can we do better ?

The Alpha-Beta Model

same β for each source



Learn the weight α for each combination

The Alpha-Beta Model: Formulation

- Partition the problem into multiple tasks according to the availability of data sources
- Learn single model for each data source
- Assign different weights for data sources in different source combinations
- Perform feature-level and source-level selection via regularization on β and α

$$\min_{\alpha,\beta} \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_i}\sum_{j=1}^{N_i} L(x_j^i, y_j^i, \alpha_j^i, \beta_j) + \lambda\|\beta\|_1$$

$$s.t. \quad \|\alpha\|_1 \le 1$$

The Alpha-Beta Model: Optimization

- Jointly non-convex model

- Alternating Minimization Method:
  - Fix β, compute α: constrained Lasso problem

source-level selection

  - Fix α, compute β: Lasso problem

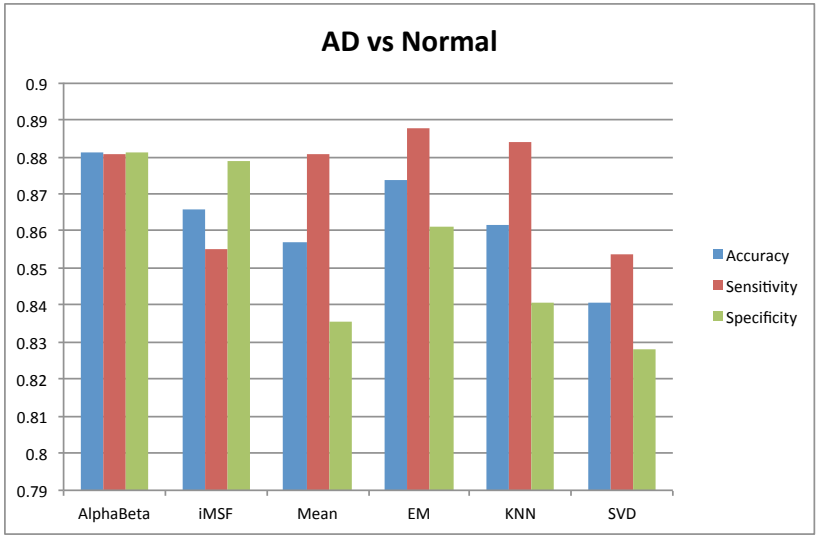
feature-level selection

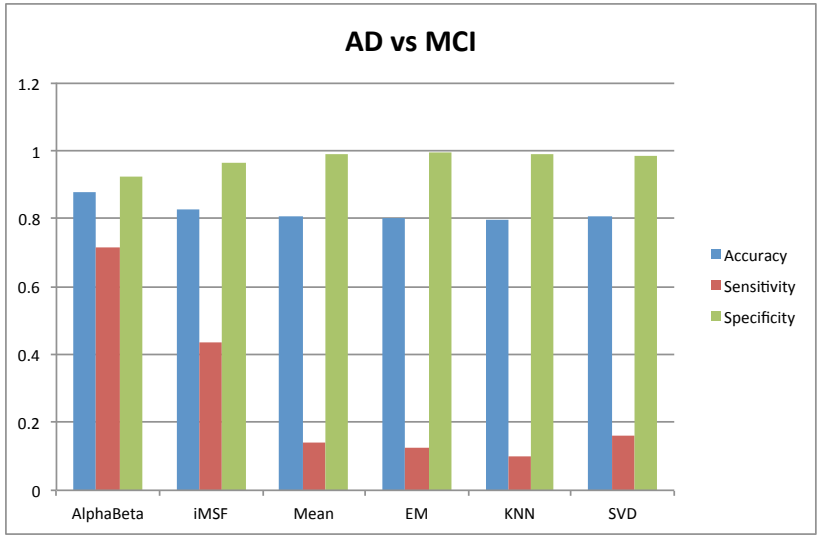
The Alpha-Beta Model: Pros and Cons

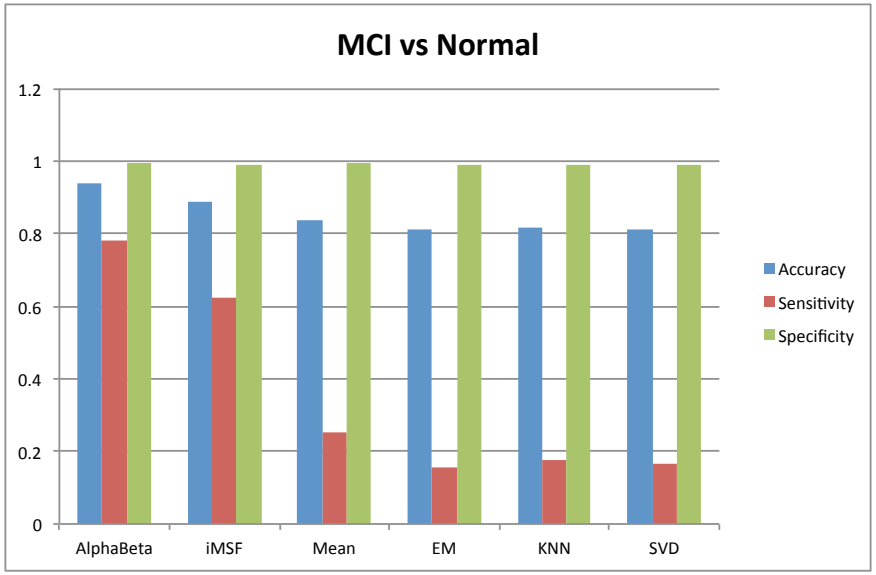
- Pros
  - Consistent model for each data source.
  - Easy generalization to more complex structure via different regularization on β , e.g., fused and tree structure.
  - Source selection via a regularization term on α.

- Cons
  - Jointly non-convex model. Alternating minimization provides efficient but locally-optimal solutions.

Experimental Results (ADNI)







Extensions and Outreach

- Co-Organizer, Visual Analytics and Information Fusion Workshop (In conjunction with KDD 2011)
- Co-Organizer, Mini-symposium on Data Mining for Biomedical Informatics (In conjunction with SDM 2011)
- Tutorial Speaker, Multi-Task Learning: Theory, Algorithms, and Applications, SDM 2012
- A book titled “Multi-label Dimensionality Reduction” to be published in 2013

Selected Publications

- Lei Yuan, Yalin Wang, Paul Thompson, Vaibhav Narayan, and Jieping Ye. Multi-Source Feature Learning for Joint Analysis of Incomplete Multiple Heterogeneous Neuroimaging Data. **NeuroImage**, 2012.
- Chao Zhang, Lei Zhang, and Jieping Ye. Generalization bounds for domain adaptation. **NIPS**, 2012.
- Rita Chattopadhyay, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Batch Mode Active Sampling based on Marginal Probability Distribution Matching. **KDD** 2012. **KDD Best Research Paper Award nomination**
- Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A Two-Stage Weighting Framework for Multi-Source Domain Adaptation. **NIPS**, 2011.
- Rita Chattopadhyay, Sethuraman Panchanathan, Ian Davidson, Wei Fan, and Jieping Ye. Multi-Source Domain Adaptation and Its Application to Early Detection of Fatigue. **KDD**, 2011. **KDD Best Research Paper Award nomination**

Thank you!