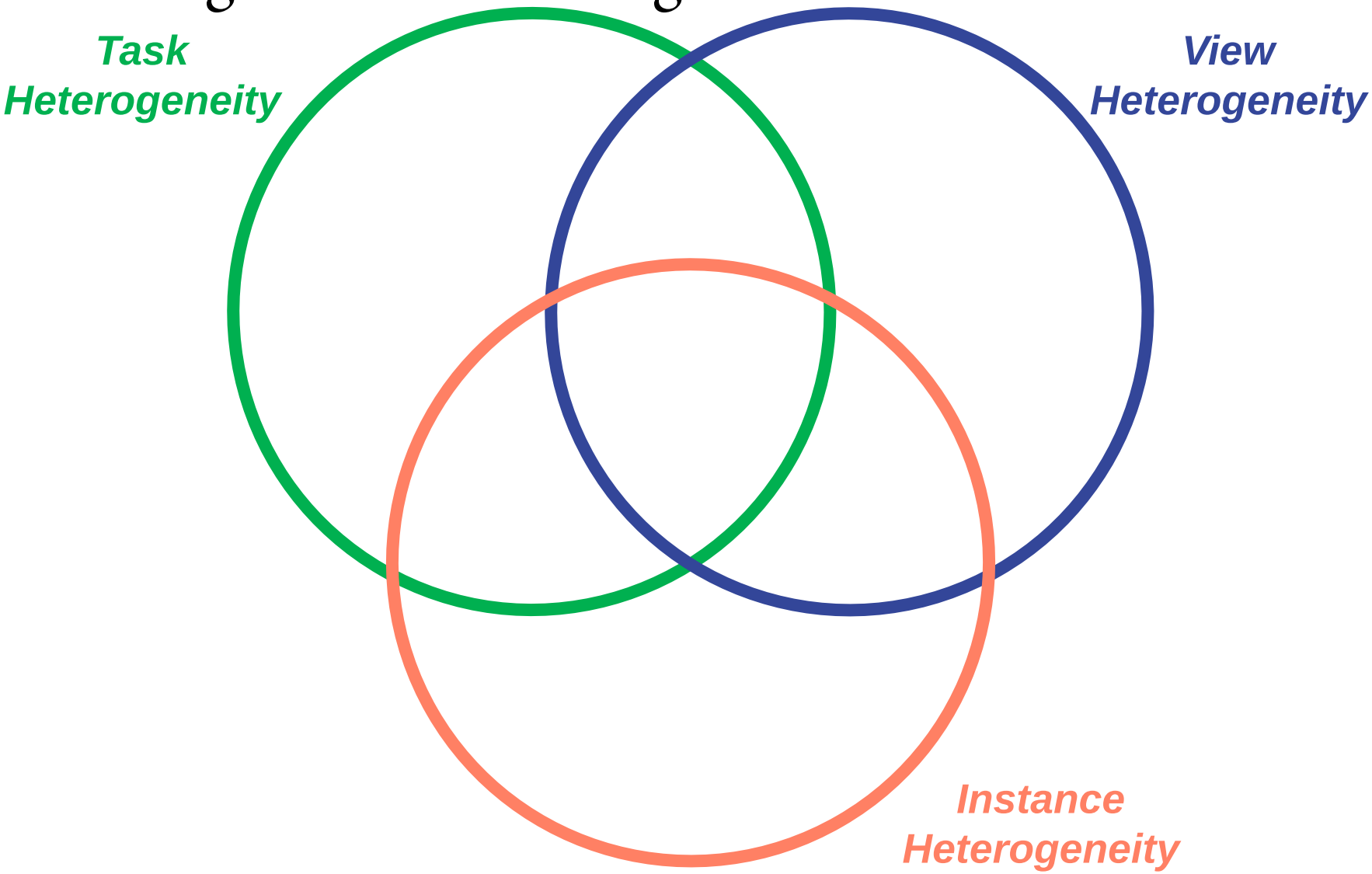
Learning with Dual Heterogeneity: A Nonparametric Bayes Model

Hongxia Yang (IBM Research) and Jingrui He (Arizona State)

What is ***Heterogeneity***?

- Definition
  - ***Inhomogeneous*** property of a target application
- Types of ***Heterogeneity***
  - Task Heterogeneity (multi-task learning)
  - View Heterogeneity (multi-view learning)
  - Instance Heterogeneity (multi-instance learning)
  - Label Heterogeneity (multi-label learning)
  - Oracle Heterogeneity (crowd sourcing)
  - ……

Heterogeneous Learning: Overview



Applications

20 News Groups

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

WebKB Dataset

WebKB			
Class	**# train docs**	**# test docs**	**Total # docs**
project	336	168	504
course	620	310	930
faculty	750	374	1124
student	1097	544	1641
Total	**2803**	**1396**	**4199**

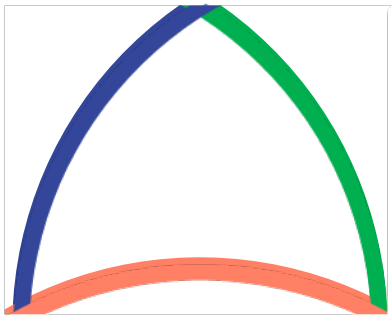
Email Spam

Roadmap

- Introduction
- Multi-Task Multi-View Learning
  - Overview
  - Nonparametric Bayes Learning with Dual Heterogeneity
    - Motivation
    - Model Formulation
    - Updating Algorithm
    - Experimental Results
- Conclusions

Multi-task Multi-view Learning

Task Heterogeneity



View Heterogeneity

Instance Heterogeneity

An Example: Good Guy vs. Bad Guy

Queen

Snow White

An Example: Features

Kind-hearted

Queen

Wesley

鬼祟

Funny

Snow White

Beauty

和善

Sneaky

Doc

Features

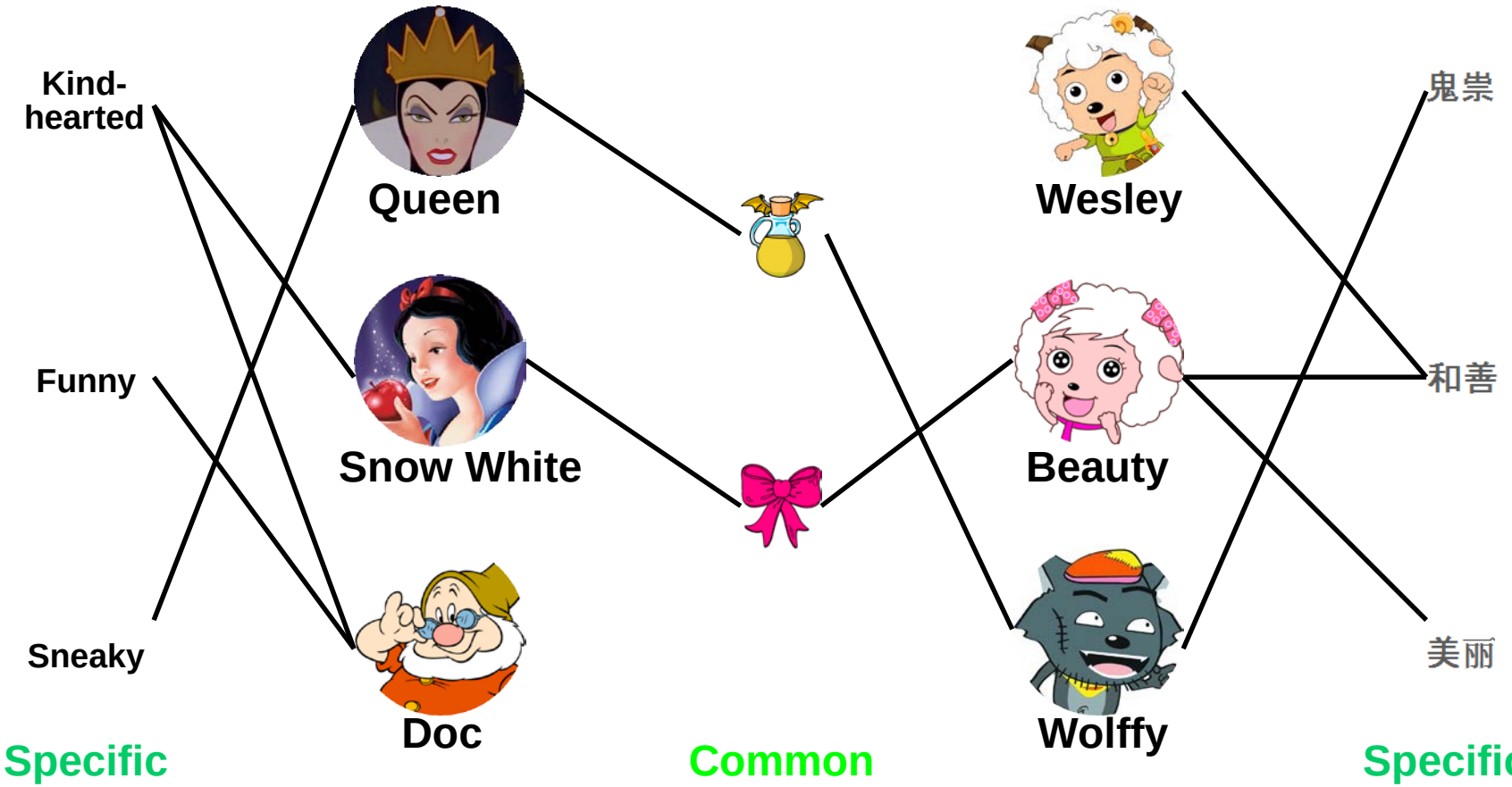
Features

Wolffy

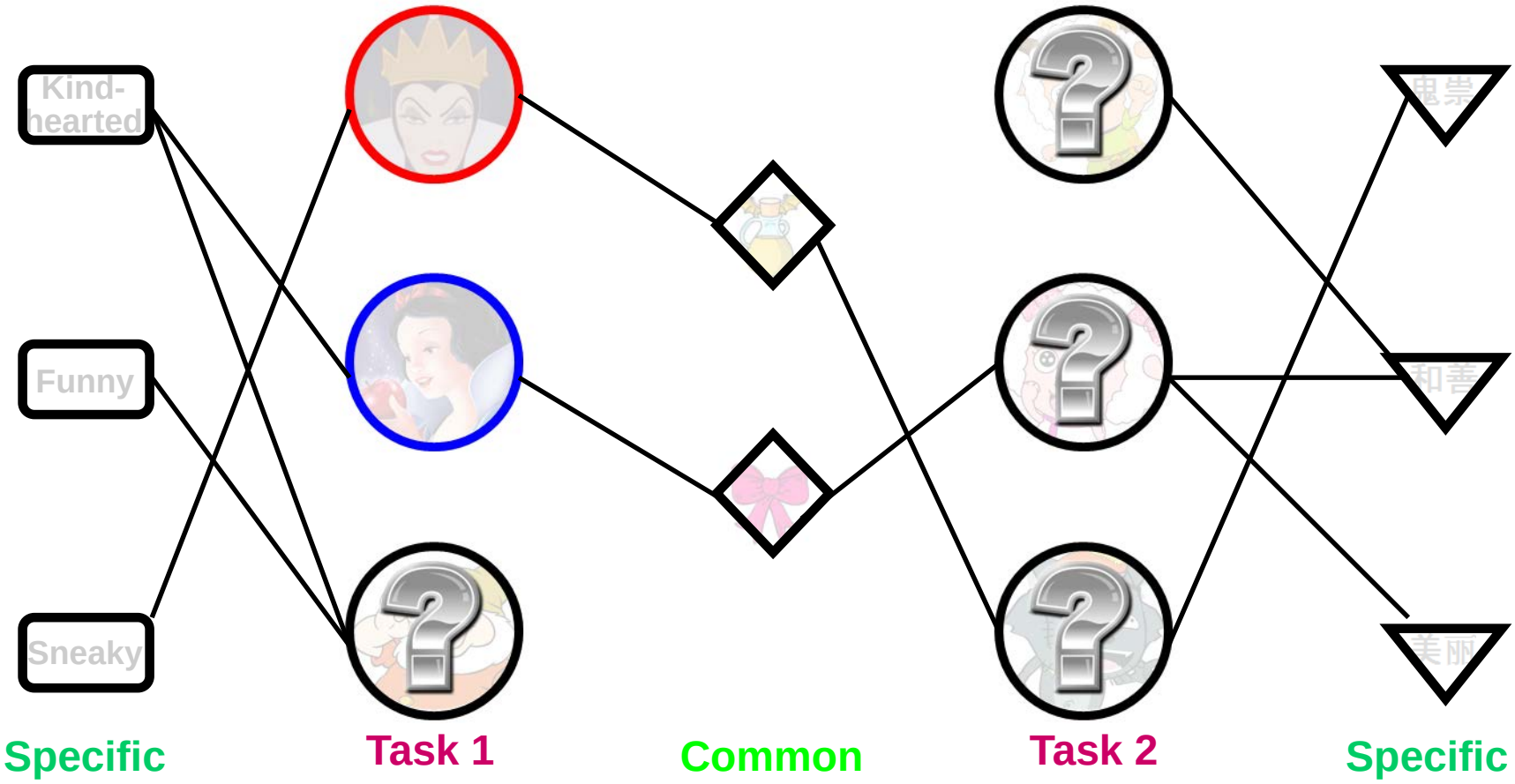
Features

美丽

An Example: View Heterogeneity



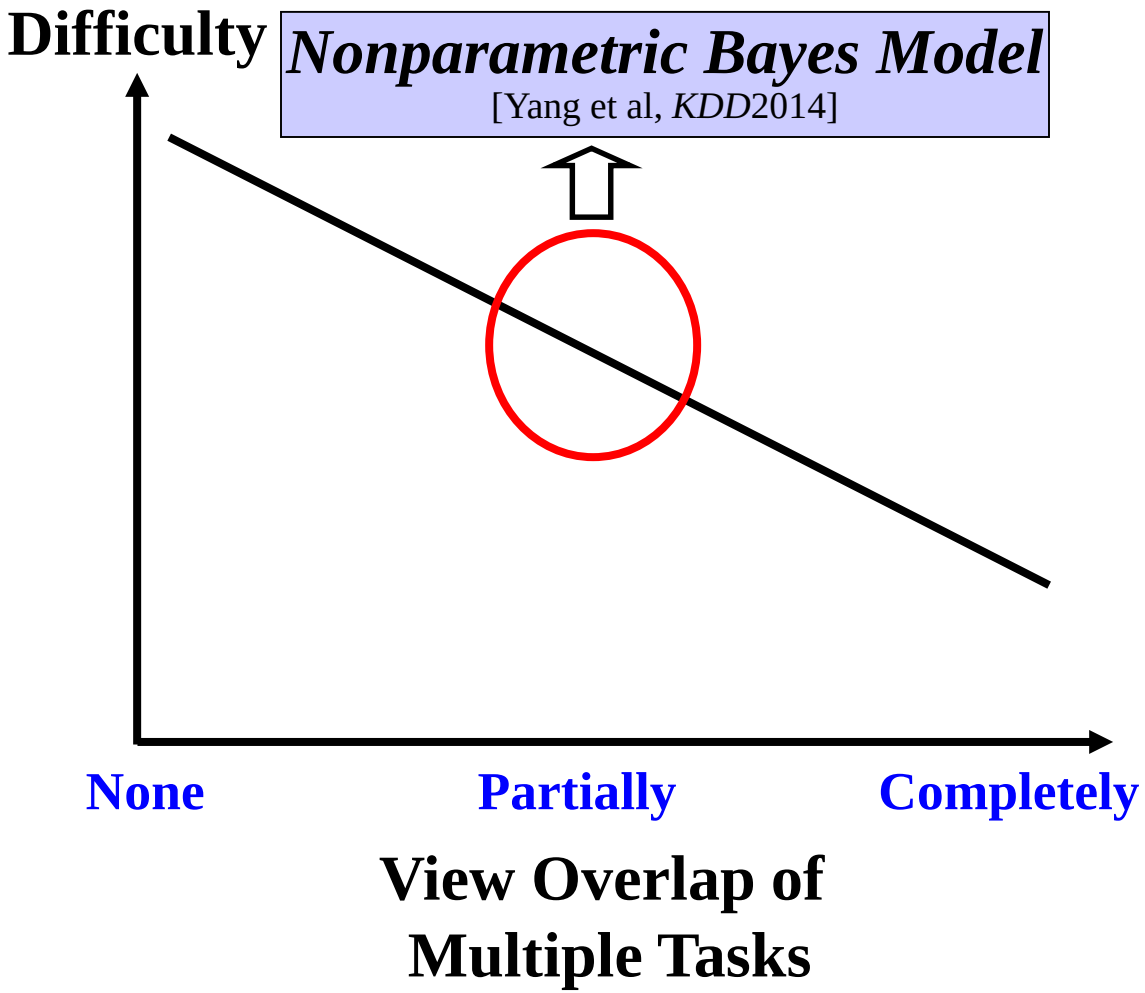
An Example: Task Heterogeneity



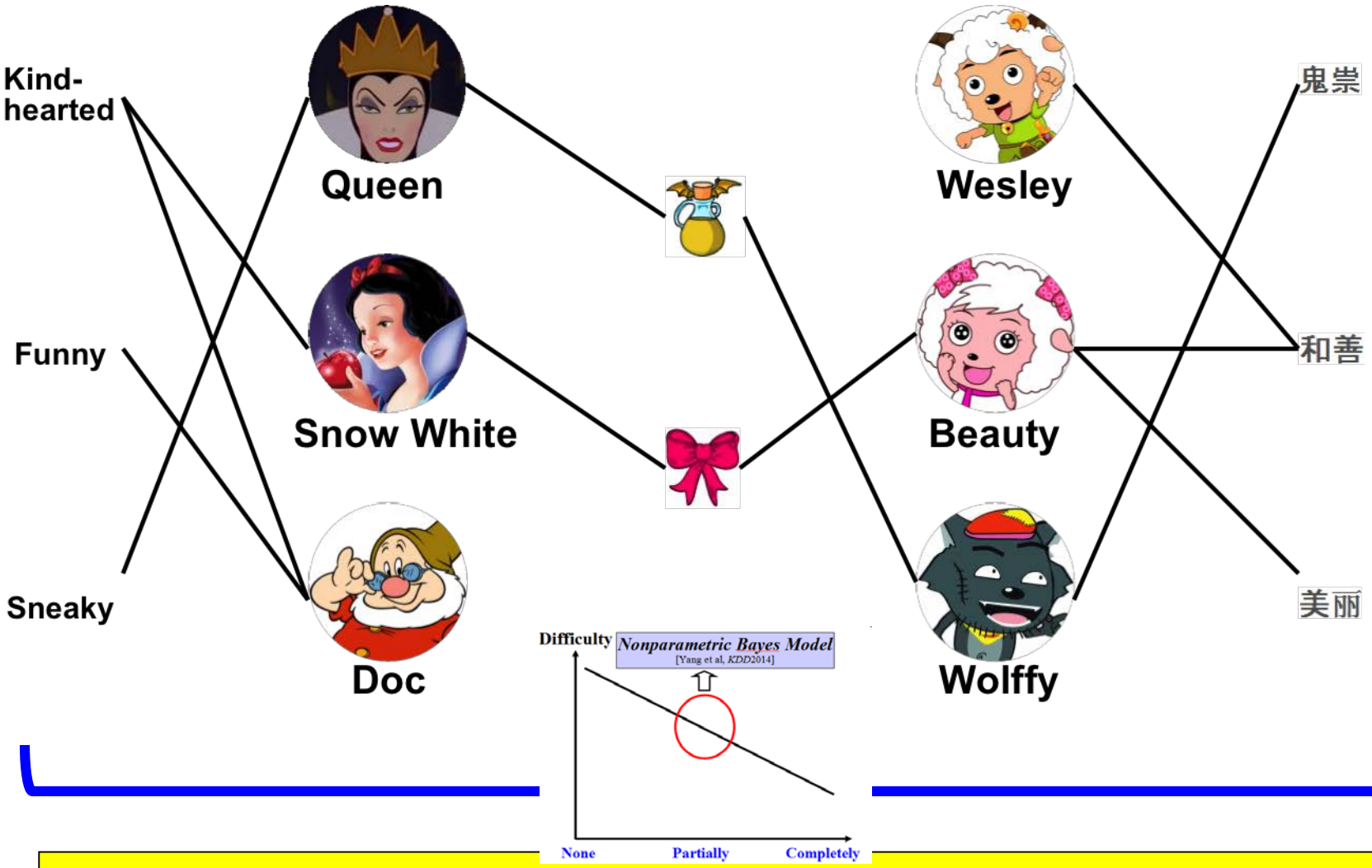
Roadmap

- Introduction
- Multi-Task Multi-View Learning
  - Overview
  - **Nonparametric Bayes Learning with Dual Heterogeneity**
    - **Motivation**
    - **Model Formulation**
    - **Updating Algorithm**
    - **Experimental Results**
- Conclusions

Partially Overlapping Views



Partially Overlapping Views



Learning with Dual Heterogeneity: A Nonparametric Bayes Model

[Yang and He, KDD 2014]

Dirichlet Process (DP) Mixture Models

- DP models uncertainty about the prior density *P*
  - Can be analytically integrated out of the conditional distribution of $\theta_T|\theta_{1:(T-1)}$
  - Specifically, the random variable θ_T has Polya urn distribution: $\theta_T|\theta_{1:(T-1)} \sim \frac{1}{\alpha+T-1}\sum_{t=1}^{T-1}\delta_{\theta_t} + \frac{\alpha}{\alpha+T-1}G_0.$

- In a DP mixture, θ is a latent parameter to an observed data point *y*

$$P \sim \mathrm{DP}(\alpha G_0),\ \theta_t \sim P, y_t|\theta_t \sim f(\cdot|\theta_t).$$

  - "Infinite clustering" model: observations are grouped by their shared parameters
  - The DP prior does not allow local clustering of tasks/views.

Notations

- Suppose that we have *T* tasks and *V* views in total.
  - For the v^{th} view, there are d_v features
  - For the t^{th} task, there are n_t examples with label $\hat{y}_{ts}$

$$\boldsymbol{x}_{ts} = [(\boldsymbol{x}_{ts1})', \ldots, (\boldsymbol{x}_{tsV})']'$$

- WOLG, suppose that we know the output of the first m_t examples
- Our goal is to leverage both the label information from all the related tasks
- As well as the consistency among different views of a single task to predict the output of the remaining unknown tasks

Model Formulation

- We first decompose each task into multiple single-view models

$$\hat{y}_{ts} = \sum_{v=1}^{V} (\boldsymbol{x}_{tsv})' \boldsymbol{f}_{tv} + \epsilon_{ts},$$

  - $\boldsymbol{f}_{tv} \in \mathcal{R}^{d_v}$ is the coefficient vector.
  - $\epsilon_{tsv} \in \mathcal{R}$ is the observational error.
- Based on the above model, we estimate the task relatedness and the view consistency through the nonparametric Bayes framework.

Task Relatedness

- We assume that $\epsilon_{\mathbf{s}} = \{\epsilon_{\mathbf{ts}}\}_{\mathbf{t}=\mathbf{1},\ldots,\mathbf{T}} \sim \mathrm{N}(\mathbf{0}, \mathbf{K})$
  - $K \in \mathcal{R}^{T \times T}$ is the kernel matrix of the Gaussian process, and it is the key to determining the various task relatedness.
  - We fully leverage the multi-view property to estimate *K* in a more reliable way.
  - We define a task graph as follows: the graph consists of *T* nodes with each node representing a single task.
  - Let *B* denote the adjacency matrix of the graph

$$B_{tt'} = \frac{1}{n_t n_{t'}} \sum_{s=1}^{n_t} \sum_{s'=1}^{n_{t'}} < \boldsymbol{x}_{ts}, \boldsymbol{x}_{t's'} >$$

  - We can compute the Laplacian $\Delta = D - B$
  - We obtain *K* as follows:

$$K = \left[\beta(\Delta + \frac{1}{\sigma^2}\mathbf{I})\right]^{-1}$$

To Be Continued…

- Global Relatedness
  - The kernel matrix *K*, whose elements indicate the similarity among various tasks, depends on the inverse of the regularized graph Laplacian Δ .
- Robust to noise
  - All the unlabeled data are used to define the adjacency matrix B (since it does not require label information)
- More Reliable
  - The adjacency matrix *B* depends on the features from all the views through x_{ts}

View Consistency

- To estimate the various view consistency, we jointly model the coefficient vectors $\boldsymbol{f}_{tv}\ (v = 1, \ldots, V)$

$$\begin{pmatrix} \boldsymbol{f}_{t1} \\ \vdots \\ \boldsymbol{f}_{tV} \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0}\ , \begin{bmatrix} \Psi_{11} & \Psi_{12} & \cdots & \Psi_{1V} \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{V1} & \Psi_{V2} & \cdots & \Psi_{VV} \end{bmatrix} \right)$$

- We extend the matrix DP to define the covariance matrix which encourages cross-view sharing of data.

To Be Continued…

- Assuming that $\Psi_{vv'} \overset{\text{ind}}{\sim} F_{vv'}, \quad \mathcal{F} \sim \mathcal{P}$
- Next, our focus is on the specification of $\mathcal{P}$

$$F_{vv'} = \sum_{h=1}^{\infty} \{W_{vv',h} \prod_{l<h} (1 - W_{vv',l})\} \delta_{\Theta_h}, \Theta_h \overset{\text{ind}}{\sim} G,$$

- For the stick-breaking weights, we decompose them as follows

$$W_{vv',h} = \gamma_{vh}\gamma_{v'h}, \gamma_{vh} \sim \text{Beta}(1, \alpha), \alpha \overset{\text{ind}}{\sim} \text{Ga}(1, \alpha_0)$$

  - The definition of γ_{vh} ensures that $\sum_{h=1}^{\infty} \{W_{vv',h} \prod_{l<h} (1 - W_{vv',l}) = 1\}$
  - We can verify that

$$\Pr(\Theta_{V_1V_2} = \Theta_{V_1V_3}) \leq \Pr(\Theta_{V_1V_2} = \Theta_{V_1V_3} | \Theta_{V_4V_2} = \Theta_{V_4V_3})$$

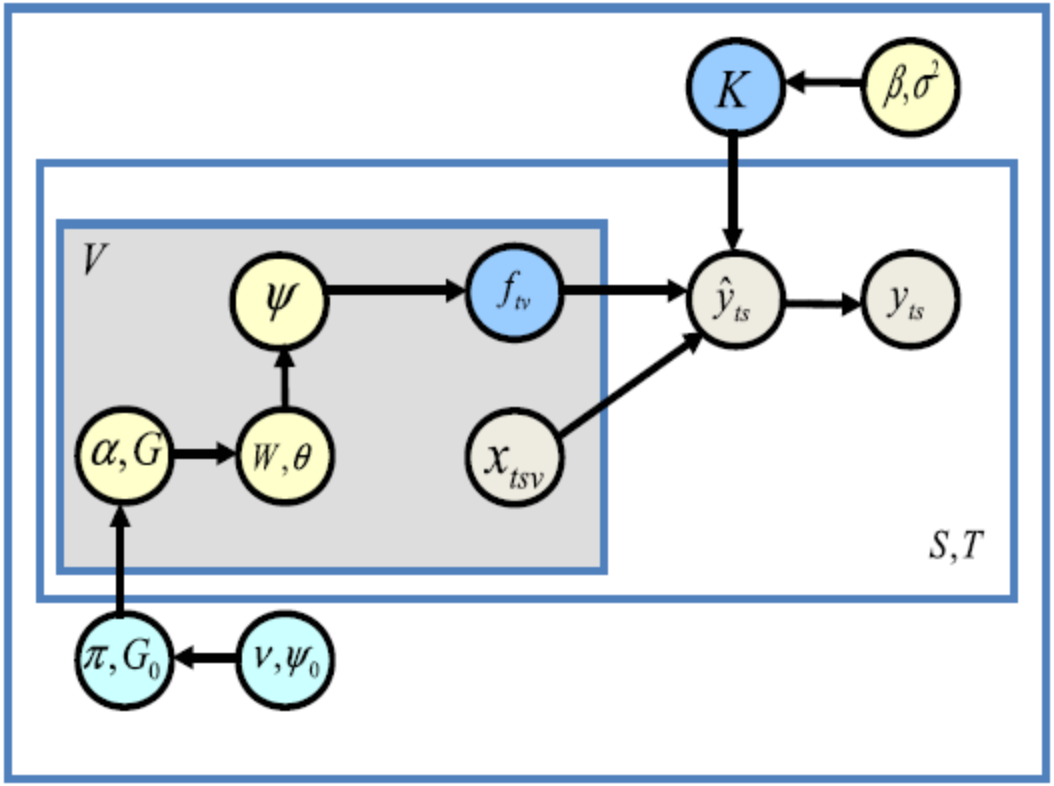
To Be Continued…

- For the base measure *G* of the covariance matrix

$$G = \pi \mathrm{I}_0 + (1-\pi)G_0, G_0 \sim \mathrm{IW}(\nu, \Psi_0)$$

  - IW: Inverse-Wishart distribution with df ν and scale matrix Ψ_0
  - When the covariance matrix falls into the null cluster, the corresponding covariance matrix will be a zero matrix, and the non-significant

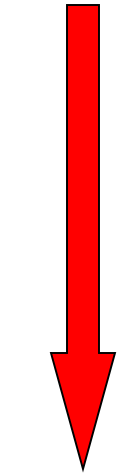
Graphical Representation



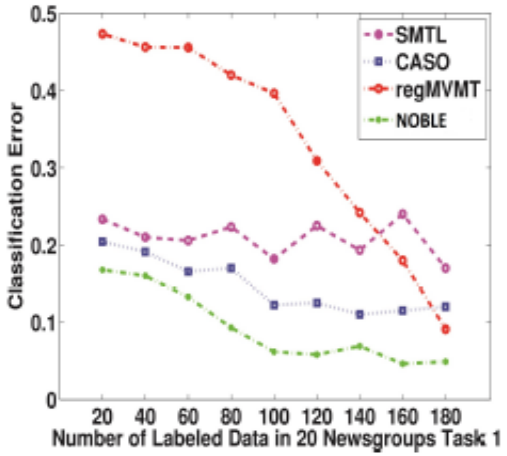
Experimental Results: Multiple Tasks, Non-identical Views

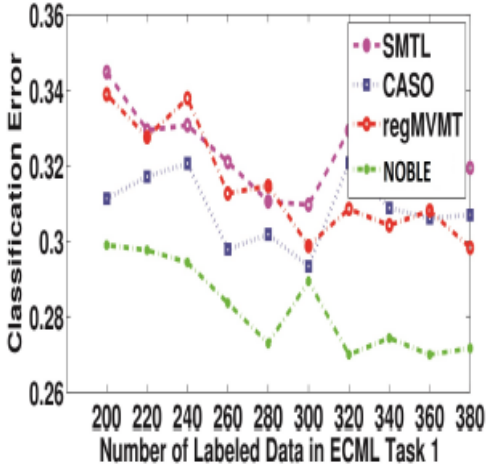
- Data set:
  - ECML 2006 discovery challenge data set
  - 20 Newsgroups data set
- Common view: common vocabulary
- Task-specific view: task-specific vocabulary

worse



better

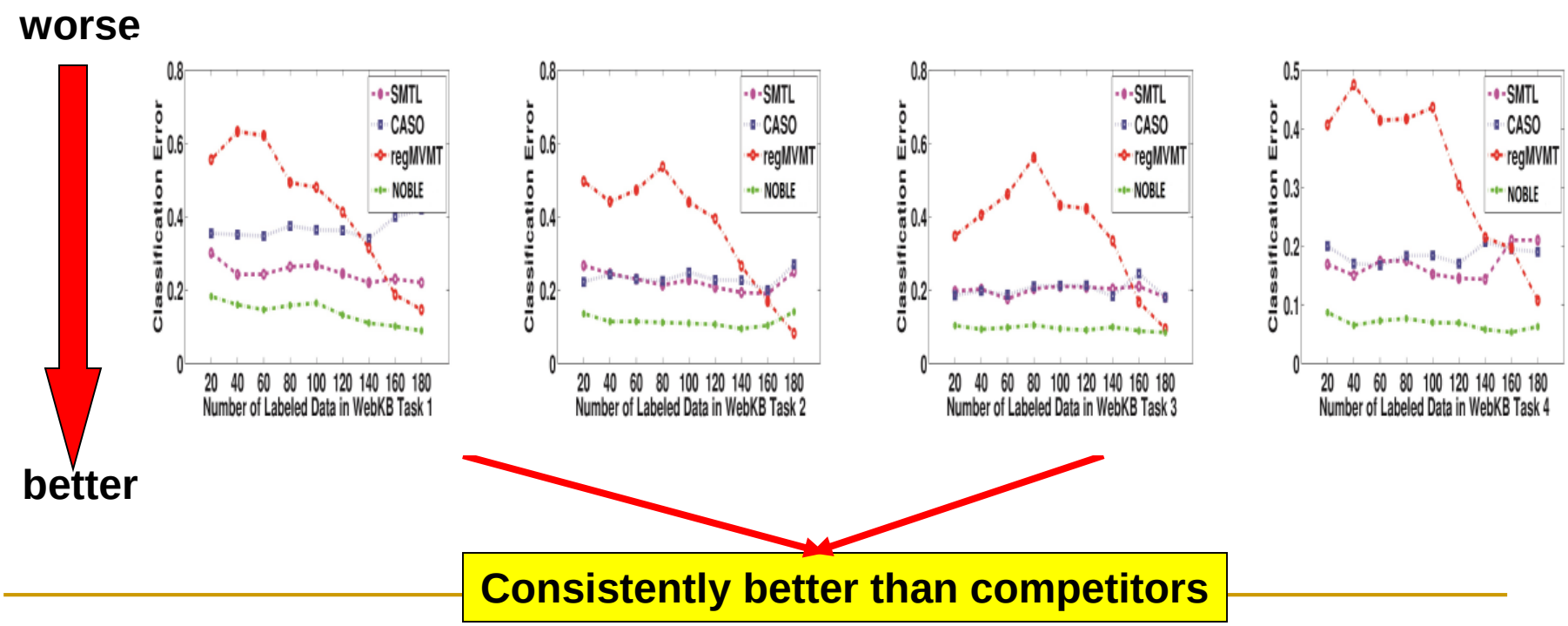




Consistently better than competitors

Experimental Results: Multiple Tasks, Identical Views

- Data set:
  - WebKB data set
- Common view: common vocabulary



Roadmap

- Introduction
- Multi-Task Multi-View Learning
  - Overview
  - Nonparametric Bayes Learning with Dual Heterogeneity
    - Motivation
    - Model Formulation
    - Updating Algorithm
    - Experimental Results
- **Conclusions**

Conclusions

- Propose a nonparametric Bayes model for addressing problems with dual-heterogeneity,
  - Tasks equally related? Views equally consistent?
  - To what extent for the relatedness and consistency?
- Gaussian Process Prior and Matrix DP Extensions
- Efficient Gibbs Sampler
- Competitive Results

Thank You!