Multivariate Models for Relational Learning

Version 1.0

http://www.dbs.ifi.lmu.de/~tresp/ILPTUT2010/WebPageILPTutorial.html

Volker Tresp

Siemens

Corporate Technology

Overview

I. Introduction: Multivariate Models for Relational Learning

II. Hierarchical Bayes - Mixed Models
III. Projection Methods

IV. Multivariate Models: Unstructured
V. Multivariate Models: Structured

VI. Conclusions
VII. Literature

I. Introduction: Multivariate Models for Relational Learning

Statistical Relational Learning (SRL)

- Advantages of SRL approaches
  - can handle complex structured representations
  - can connect both to machine learning and knowledge representation
  - connects to first order logic (FOL) via inductive logic programming (ILP)
- Object-to-object relationships can be modeled

Graph Representation

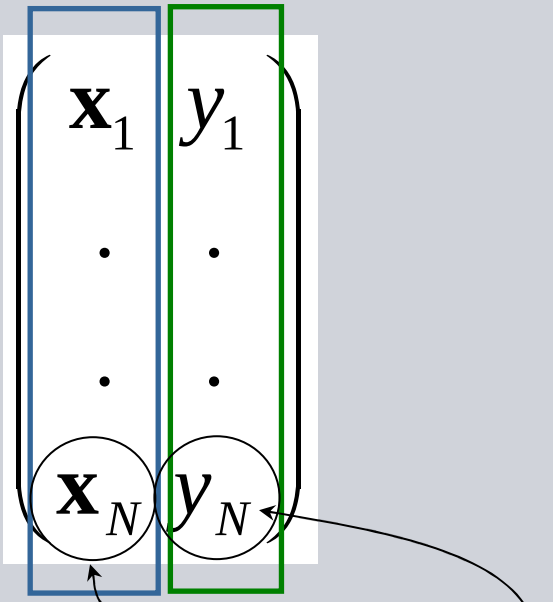
- Representation as directed graph
  - Example: RDF (Resource Description Framework) graph of the Semantic Web
- Thus: connections to
  - Learning on graphs

SRL and Multivariate Predictive Models

- Multivariate models can be quite useful for SRL
- We assume a quite general definition of a multivariate model
  - The model contains (maybe as a building block) a representation where several outputs are predicted
  - In particular our definition includes Hierarchical Bayesian modeling

A Classical Generic Supervised Learning Task

Data matrix

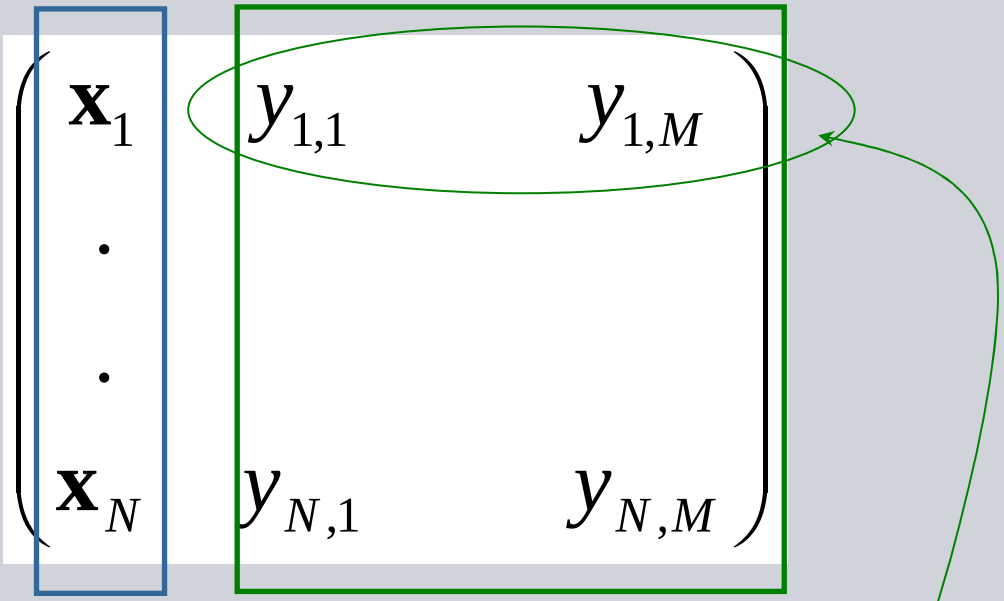


$$y_i = f_{\mathbf{w}}(\mathbf{x}_i) + \varepsilon_i$$

$$P(y_i = 1) = \sigma\left(f_{\mathbf{w}}(\mathbf{x}_i)\right)$$

- Rows: data points
- Columns:
  - Input vector ***x***
  - Output scalar *y*

Prediction of Several Output Variables



- Rows: data points
- Columns:
  - Input vector ***x***
  - Output **vector** ***y***

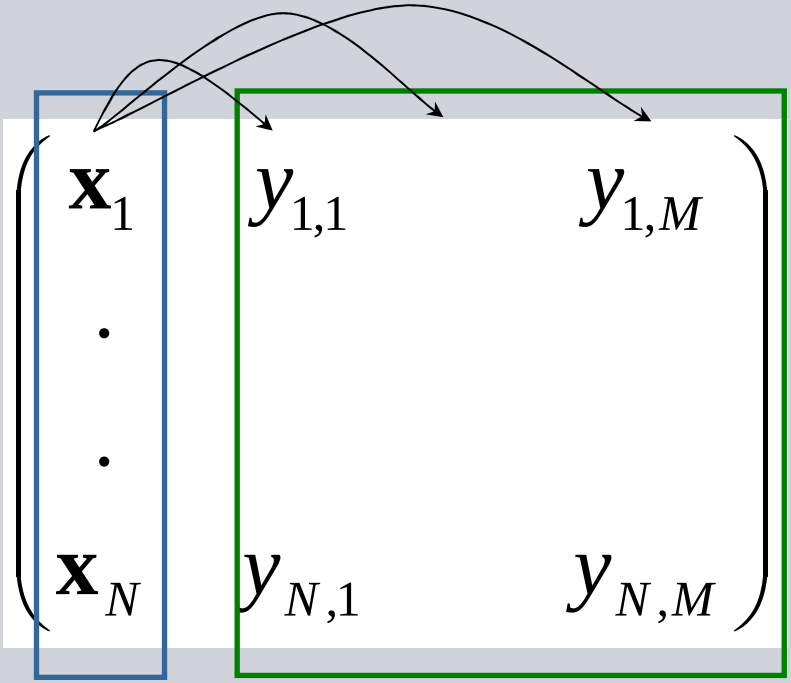
Multiple Outputs

“Multiple outputs do not affect each others least squares estimates”

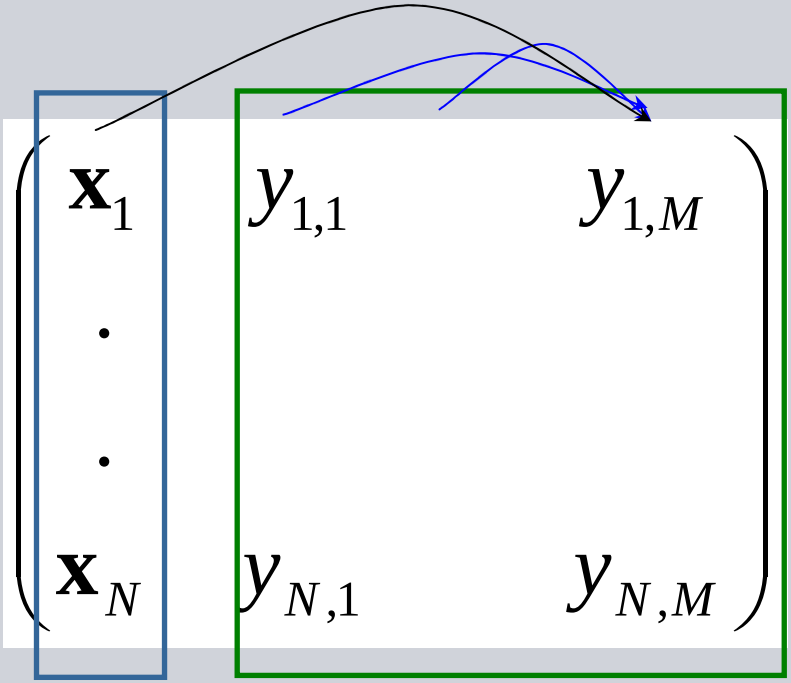
Hastie, Tibshirani, Friedman (2001)

We will study cases, where this statement is not applicable!

Model Each Output Separately?



Condition on All Other Variables?



- Dependency networks

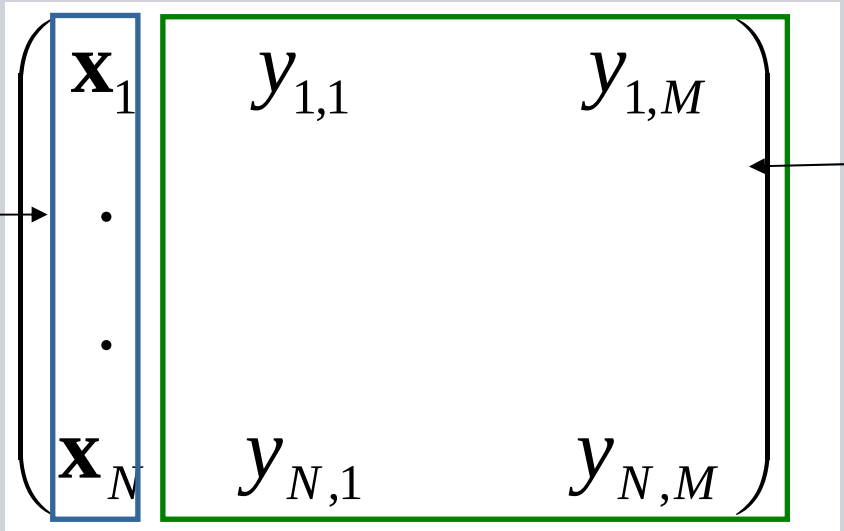
- Let’s consider some generic examples …

1: Entity-Centered Prediction

- Let's assume a relational data model
- Definition of a statistical unit: typically entities of some type (*person*)
- Definition of a population: typically entities of some type with some properties (*students in Munich*)
- Definition of random variables associated with entities: attributes, relationships
- Definition of inputs: typically aggregated information; could also be attributes
- We can derive a regular data matrix (*matrixification*)
- Advantage: formulated as standard learning problem where powerful learning algorithms are available

(known) Inputs

- aggregated information
- attributes



Tresp et al., 2009

Outputs

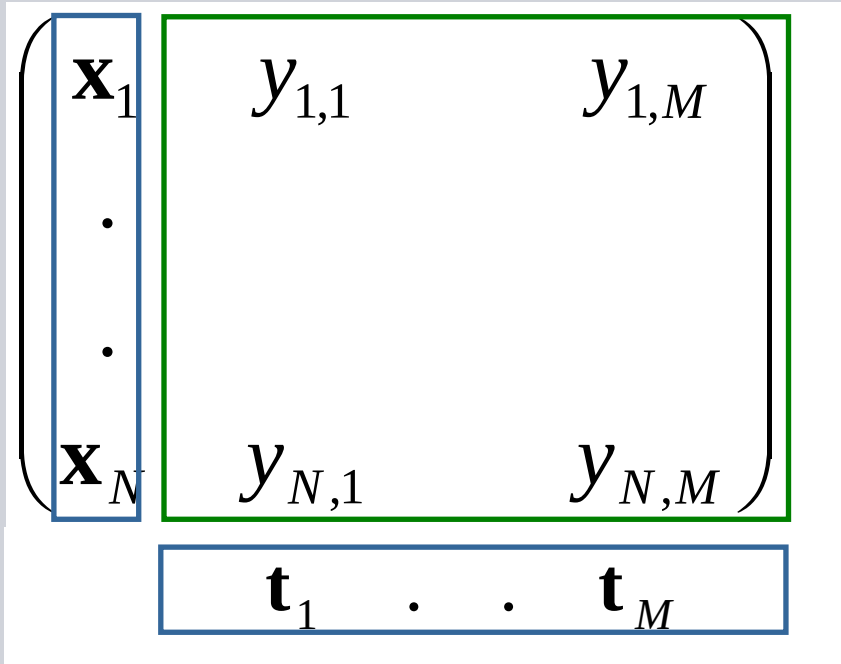
- (one or several) relationships
- attributes

"The preference of a user for a movie depends on preference patterns and on movie attributes"

2: Relationship-Centered Prediction

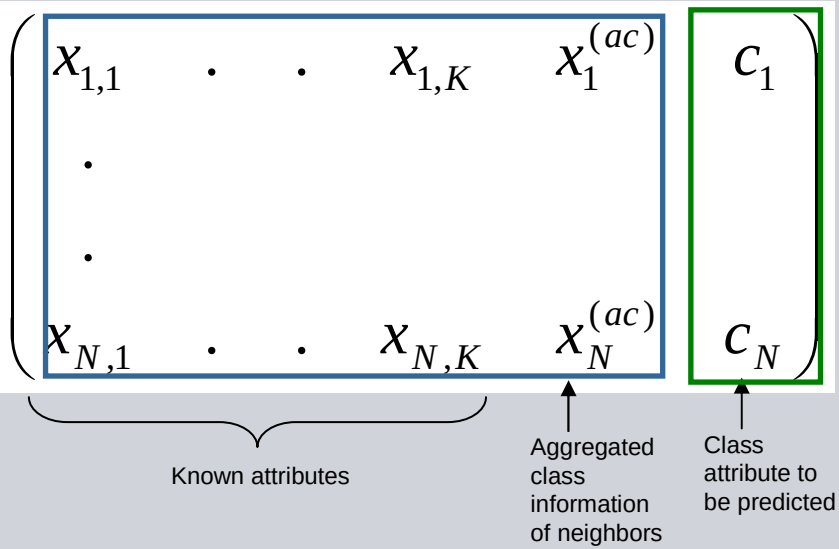
"young males like action movies"

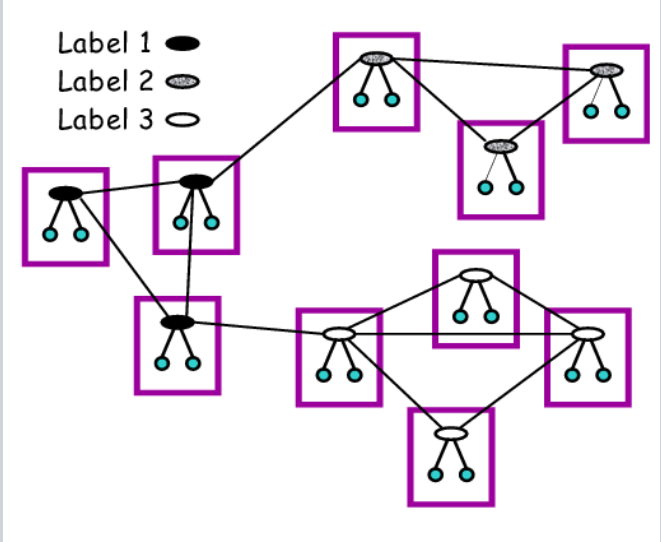
- Symmetrical representation: attributes of two entity types are included



3: Joint Models as Coupled Multivariate Models

- The aggregated information might be calculated based in information that is incomplete
- One obtains a global coupling of variables or a joint model

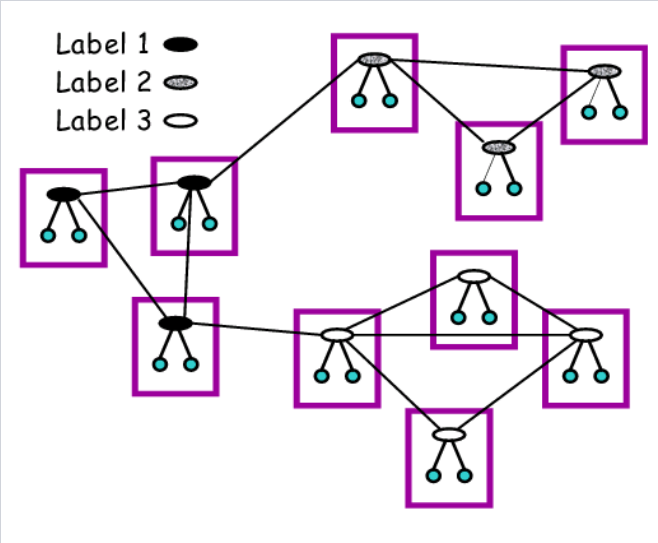


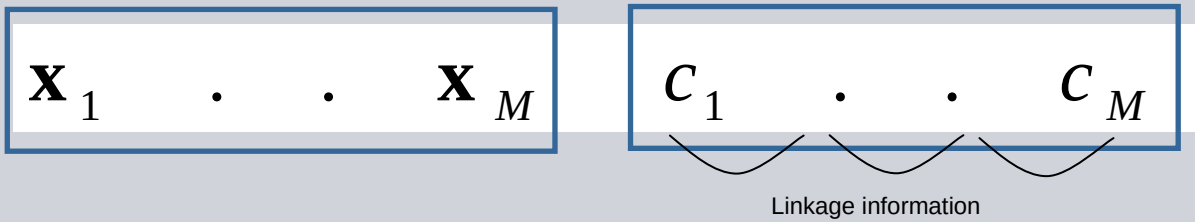


"a class label for a document depends on document properties and depends on the class label of cited documents (which might partially be unknown)"

4: Joint Models as Structured Multivariate Models

- The network represents one data point
- All known attributed form the input of one data point
- The target variables have a structured form
- Conditional Random Fields (CRFs)





The network represents one data point

In the remaining part we will discuss different approaches to multivariate models and illustrate their application to SRL

Overview

II. Hierarchical Bayes - Mixed Models
III. Projection Methods

IV. Multivariate Models: Unstructured
V. Multivariate Models: Structured

VI. Conclusions
VII. Literature

II. Hierarchical Bayes

- Predicting the same thing (patient's length of stay) but in different situations (different hospitals)
- *Relational setting: many users like many movies*

II.A.
Problem Settings and Simple Solutions

Problem Setting

- Data is collected for *M* different situations (entities/sites/tasks) and the goal is to learn predictive models $f_j(\mathbf{x}), \quad j = 1, \cdots, M$
- Can data from other situations help to improve the prediction of both $f_j(\mathbf{x})$ and for a new situation $f_{new}(\mathbf{x})$?
- For simplicity, we consider models linear in the parameters of the form

$$f_j(\mathbf{x}) = \sum_{l=1}^{L} w_{j,l} \phi_l(\mathbf{x})$$

Typically we only have access to $y_j(\mathbf{x}) = f_j(\mathbf{x}) + \varepsilon_j(\mathbf{x})$

Problem Setting: SRL

Entity-Centered Prediction

- $f_j(\mathbf{x}_i)$ describes the relationship between object *i* with attributes $\mathbf{x}_i$ and object *j*
- Example: movie *i* and user *j*

Simple Solution: One Global Model

$$f(\mathbf{x}) = \sum_{l=1}^{L} w_l \phi_l(\mathbf{x})$$

- We learn one model with all data: Fruits, not apple and oranges
- Data efficient solution

- Problem: ignores differences in different situations

- *Dependent on its attributes, a movie is either liked or disliked (independent of the user)*

Simple Solution: Separate Models

- A model for each situation is trained solely on its own data

$$f_j(\mathbf{x}) = \sum_{l=1}^{L} w_{j,l}\phi_l(\mathbf{x})$$

- Problem: no sharing of statistical strength
  (but sometimes the correct solution)
  - Only one output dimension contributes to parameter estimates

- *A preference of user j1 tells me nothing about the preference of user j2*

Simple Solution: Situation as Input

- The situation is just another set of inputs to the model, e.g., in form of indicator variables

$$f(\mathbf{x}, \mathbf{u}_j)$$

$$\mathbf{u}_j = (0,0,\cdots,u_{j,j}=1,\cdots 0,0,0)^T$$

- Data efficient
- Problem: sometimes suitable but the influence of the situation might be quite complex in which case this approach might fail

- *Dependent on its attributes, a movie is either liked or disliked; in addition users have a general preference for movies*

From Multi-categorical Inputs to Hierarchical Modeling

Another view

- Consider a problem with a multi-categorical input variable

$$f(\mathbf{x}, \mathbf{u}_j)$$

$$\mathbf{u}_j = (0,0,\cdots,u_{j,j}=1,\cdots 0,0,0)^T$$

- This is exactly the case where one might think of applying Hierarchical Bayesian modeling!
  - The multi-categorical input is treated as situation

II.B.
Hierarchical Bayes / Mixed Models

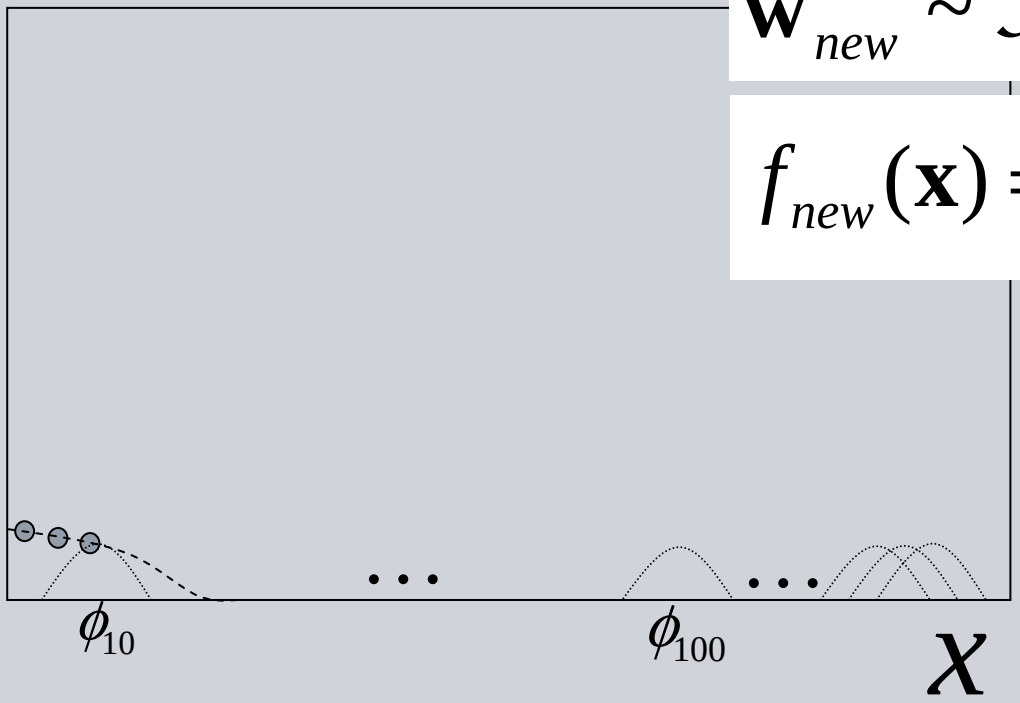
New Situation with Few Data Points

- Assume a few data points local in input space

$$\hat{f}_{new}(x)$$

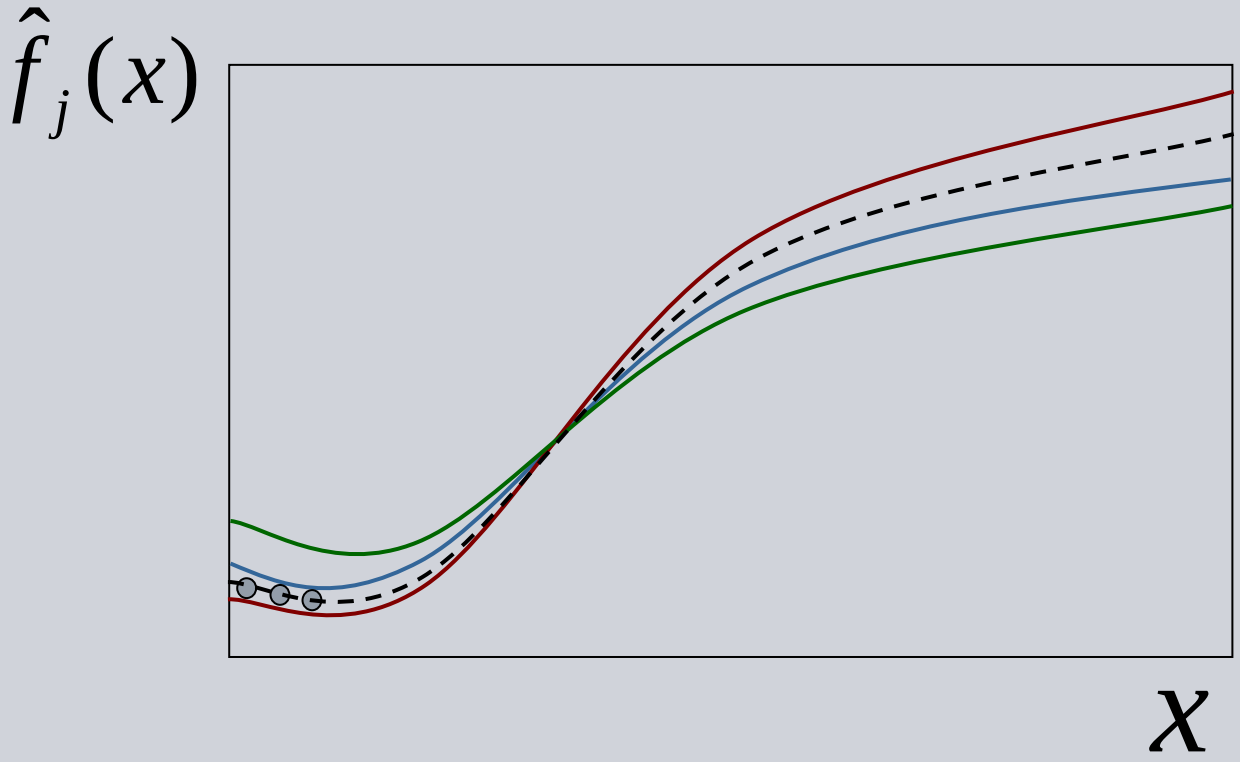
$$\mathbf{w}_{new} \sim \mathcal{N}(0, \alpha^2 I)$$

$$f_{new}(\mathbf{x}) = \sum_{l=1}^{L} w_{new,l} \phi_l(\mathbf{x})$$

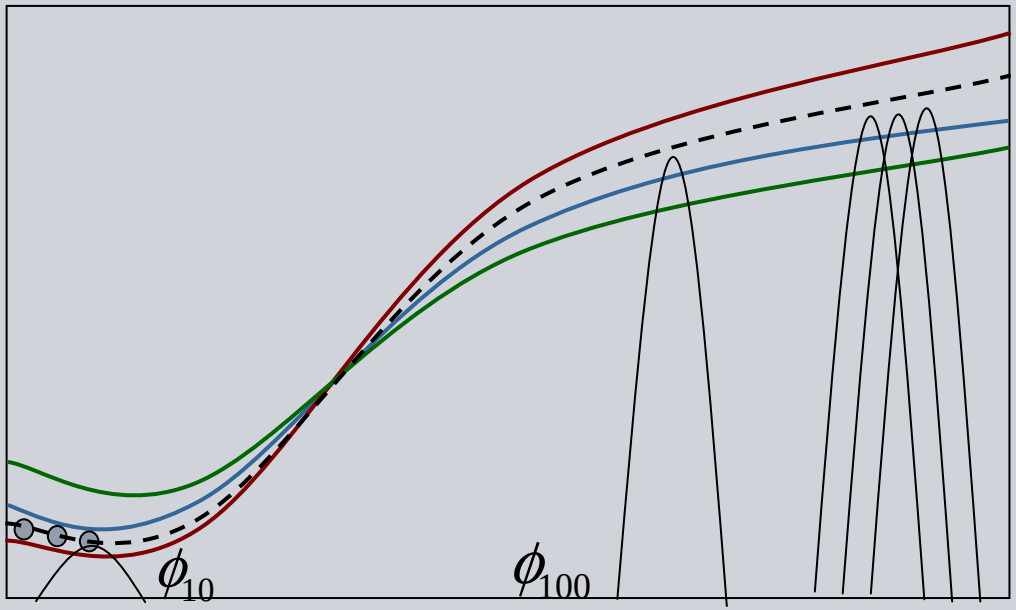


Motivation for Hierarchical Bayes

- Looking at other models $\hat{f}_j(x)$
  another solution becomes more likely

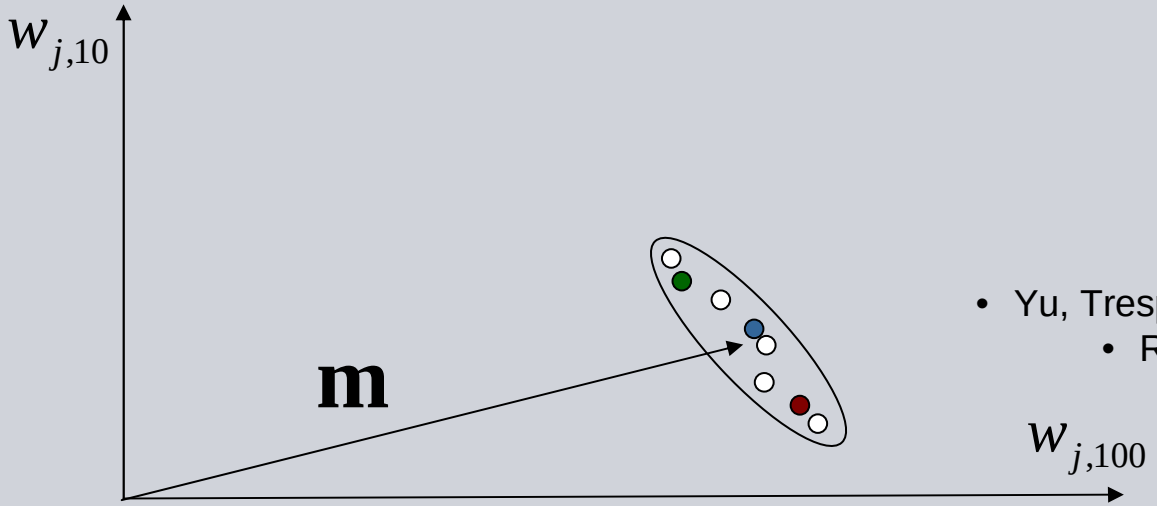


Parameter Distributions



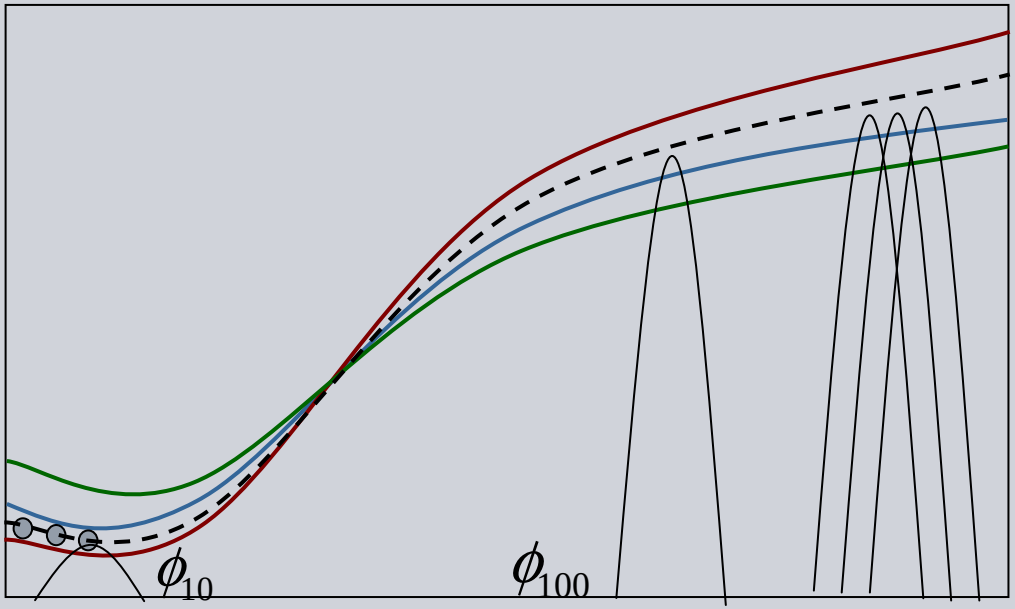
- The parameters for the different models might form again a Gaussian distribution

$$\mathbf{w}_{new} \mid \mathcal{D} \sim \mathcal{N}(\mathbf{m}, \Sigma)$$



- Yu, Tresp, Schwaighofer (2005)
- Raina, Ng, Koller (2006)

Parameter Distributions: *Entity-Centered Prediction*



$$\mathbf{w}_{new} \mid \mathcal{D} \sim \mathcal{N}(\mathbf{m}, \Sigma)$$

If a person prefers action movies, this person dislikes romantic movies

Learned Prior

- A new model sees the “learned” prior

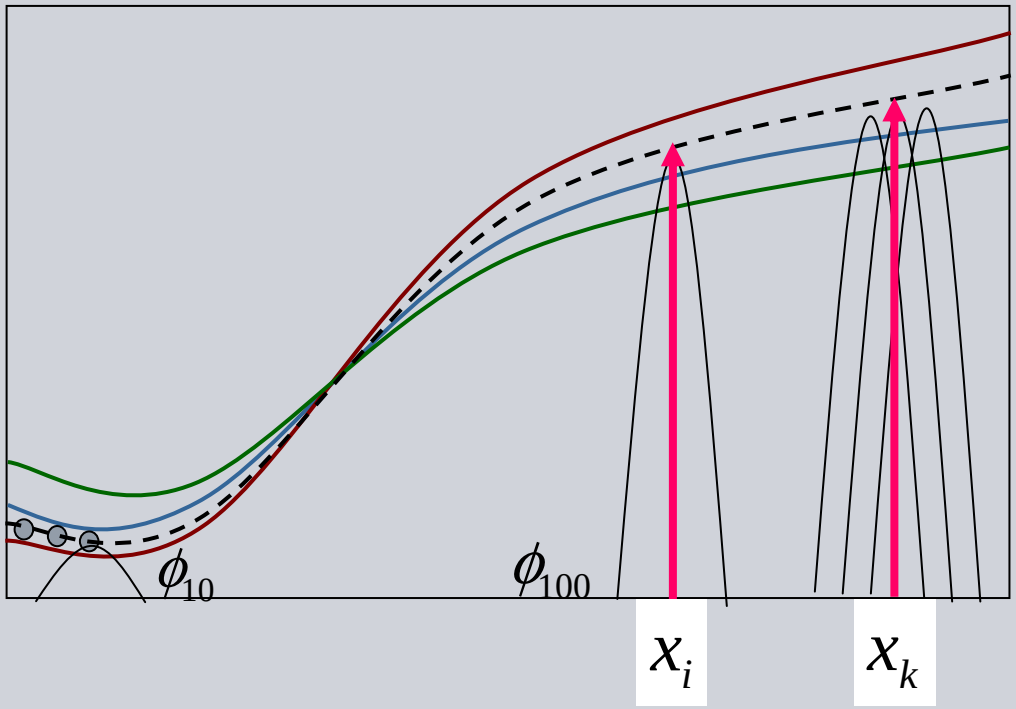
$$\mathbf{w}_{new} \mid \mathcal{D} \sim \mathcal{N}(\mathbf{m}, \Sigma)$$

- With a Gaussian (learned) prior we obtain a Gaussian process with mean function and covariance kernel given by

$$E(f_{new}(\mathbf{x}) \mid \mathcal{D}) = \sum_{l=1}^{L} m_l \phi_l(\mathbf{x})$$

$$\text{cov}(f_{new}(\mathbf{x}_i), f_{new}(\mathbf{x}_j) \mid \mathcal{D}) = \phi^T(\mathbf{x}_i)\Sigma\phi(\mathbf{x}_j)$$

Learned Prior in Function Space



$$E(f_{new}(\mathbf{x}) \mid \mathcal{D}) = \sum_{l=1}^{L} m_l \phi_l(\mathbf{x})$$

$$\text{cov}(f_{new}(\mathbf{x}_i), f_{new}(\mathbf{x}_j) \mid \mathcal{D}) = \phi^T(\mathbf{x}_i)\Sigma\phi(\mathbf{x}_j)$$

- So we got what we wanted: the new function is guided by the previously learned functions

Covariance and Basis Functions

- We can decompose using a Principal Component Analysis (PCA):

$$\Sigma = VDD^T V^T$$

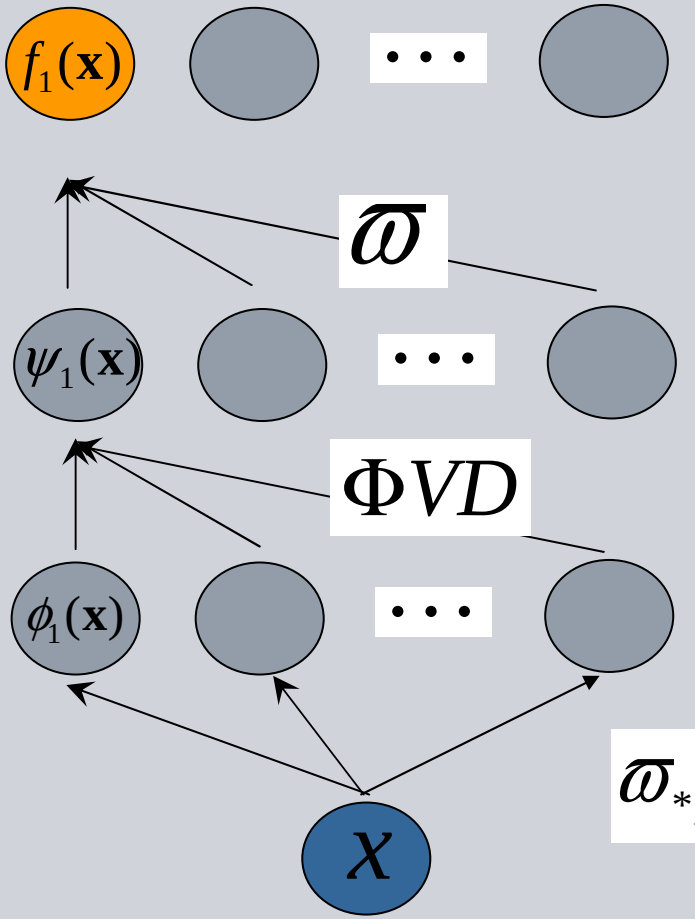
- And obtain:

$$\phi^T(\mathbf{x}_i)\Sigma\phi(\mathbf{x}_j) = \left(D^T V^T \phi(\mathbf{x}_i)\right)^T \left(D^T V^T \phi(\mathbf{x}_j)\right)$$

- From this view point the new model has a Gaussian parameter distribution with identity covariance matrix and with *new learned basis functions formed as linear combinations of the original basis functions:*

$$\psi_k(\mathbf{x}) = d_{k,k} \sum_l v_{l,k} \phi_l(\mathbf{x})$$

Architecture: Hierarchical Bayesian Modeling



$$f_j(\mathbf{x}) = \sum_{l=1}^{L} m_l \phi_l(\mathbf{x}) + \sum_{k=1}^{K} \varpi_{k,j} \psi_k(\mathbf{x})$$

$$\psi_k(\mathbf{x}) = d_{k,k} \sum_{l} v_{l,k} \phi_l(\mathbf{x})$$

Bottleneck!

$$\phi_l(\mathbf{x})$$

$\varpi_{*,j}$ is trained using solely the data for output j with penalty $\text{cost} + \lambda \sum_k \varpi_{k,j}^2$

Compare: Neural networks with hidden layers and multiple outputs

Technical Details: EM Updates

- In typical applications noisy measurements for the different situations are available. The design matrix for situation *j*: Φ_j inverse Wishart: $\mathcal{IW}$
- Complete data likelihood

$$\mathcal{N}(\mathbf{m} \mid \mu, \eta^{-1}\Sigma)\mathcal{IW}_L(\Sigma \mid \delta, \kappa)\prod_{j=1}^{M}\mathcal{N}(y_{*,j} \mid \Phi_j \mathbf{w}_j, \sigma^2 I)\mathcal{N}(\mathbf{w}_j \mid \mathbf{m}, \Sigma)$$

- E-step $P(\mathbf{w}_j \mid y_{*,j}, \mathbf{m}, \Sigma) = \mathcal{N}(\mathbf{w}_j \mid \mathbf{r}_j, V_j)$

$$V_j = (\Sigma^{-1} + \frac{1}{\sigma^2}\Phi_j{}^T\Phi_j)^{-1} \text{ and } \quad \mathbf{r}_j = V_j(\frac{1}{\sigma^2}\Phi_j{}^T y_{*,j} + \Sigma^{-1}\mathbf{m})$$

- M-Step

$$\mathbf{m} = \frac{1}{M+\eta}(\sum_{j=1}^{M}\mathbf{r}_j + \eta\mu)$$

$$\Sigma = \frac{1}{M+\eta+\delta+2L}(\eta(\mathbf{m}-\mu)^T(\mathbf{m}-\mu) + \sum_{j=1}^{M}(\mathbf{m}-\mathbf{r}_j)^T(\mathbf{m}-\mathbf{r}_j) + \sum_{j=1}^{M}V_j + \kappa)$$

Definition of Inverse Wishart

$$\mathcal{IW}_L(\Sigma \mid \delta, \kappa) \propto \left(\det\Sigma\right)^{-(\delta+2L)/2} \exp[-\frac{1}{2}\mathrm{tr}(\kappa\Sigma^{-1})]$$

This definition has the advantage that it is marginalization consistent

A. P. Dawid. *Some matrix-variate distribution theory: Notational considerations and a Bayesian application.* Biometrika, 68(1), 1981

As Matrix Factorization

- If we look at the estimation of the training targets we get the matrix factorization

$$E(Y) = F = \Phi R$$

$$R = \left(\mathbf{r}_1, \ldots \mathbf{r}_M\right)$$

Learned Basis Functions

- The key benefit in Hierarchical Bayesian modeling for linear systems is that common basis functions are learned that are used for all outputs

- Kernel before and after Training

$$\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_k) \rightarrow \phi^T(\mathbf{x}_i)\Sigma\phi(\mathbf{x}_k) = \psi^T(\mathbf{x}_i)\psi(\mathbf{x}_k)$$

Comments

- Advantages of Hierarchical Bayes:
  - Inclusion of prior knowledge by defining the basis functions
  - Generalization to new inputs/rows and new outputs/columns
  - No problems with missing outputs

- Alternatively: in Hierarchical Bayes inference is often performed via Gibbs sampling or other approximate methods such as variational learning (see, e.g., Latent Dirichlet Allocation, LDA)

(Blei, Ng, Jordan, 2003)

- Naturally Hierarchical Bayes is also applicable beyond linear models

- Gelman, Carlin, Stern and Rubin (2003) provide a thorough discussion of Hierarchical Bayesian models

Three Phases in HB modeling

- **First Phase:** With no data yet available the model for a new situation follows the prior (the mean function)
- **Second Phase:** With some data available for a new situation, a model follows more closely a previous model that fits those data well
- **Finally:** With increasing data available, the model becomes independent of the learned prior

- Dimensional reduction: Derived basis functions

$$\psi_k(\mathbf{x}) = d_{k,k} \sum_l \nu_{l,k} \phi_l(\mathbf{x})$$

with a small $d_{k,k}$ are ignored

When Hastie's Statement is Applicable

- If the hyper parameters (in our case: $\mathbf{m}, \Sigma$) are known a priori, i.e., they represent the empirical parameter distribution, then all output functions are independent
  - Or: if output functions have no common prior distribution (predicting apples and oranges)

- In contrast, if the prior is learned then all measurements influence all predictions!

Frequentist Equivalent: Mixed Models

$$y_{*,j} = \Phi_j \mathbf{m} + Z_j \mathbf{b}_j + \varepsilon_j$$

$$\mathbf{b}_j \propto \mathcal{N}_K(0, \Sigma)$$
$$\varepsilon_j \propto \mathcal{N}_{N_j}(0, \sigma^2 \Lambda_j)$$

- Known: Φ_j, Z_j
- (unknown but) Fixed effect: $\mathbf{m}$
- Random effect: $\mathbf{b}_j$
- Special case: $Z_j = \Phi_j$
  - *regression model with random coefficients*
- Relationship to HB-model: $\mathbf{w}_j = \mathbf{m} + \mathbf{b}_j$

- **New: correlated contributions that cannot be explained by the inputs ("noise")**
- **Collaborative effect!**

- MM: As Bayesian as a frequentist will ever get
- HB: as frequentist as a Bayesian will ever get

Gaussian Process Hierarchical Bayes (GP-HB)

- As already discussed, a system of fixed basis functions and Gaussian weight prior

$$f(\mathbf{x}_i) = \sum_l^M w_l \phi_l(\mathbf{x}_i) \qquad \mathbf{w} \propto \mathcal{N}(\mathbf{m}, \Sigma)$$

- … is technically equivalent to a Gaussian process with covariance $k(\mathbf{x}_i, \mathbf{x}_k) = \phi^T(\mathbf{x}_i)\Sigma\phi(\mathbf{x}_k)$

and mean function $m(\mathbf{x}) = \sum_l^M m_l \phi_l(\mathbf{x}_i)$

- Thus:
  - as parametric HB boils down to learning $\mathbf{m}, \Sigma$
  - GP-HB, boils down to learning $m(\mathbf{x}), k(\mathbf{x}_i, \mathbf{x}_k)$

GP-HB: Learning in Function Space

- Now we consider GP-HB in *function* space
- A prior for mean and covariance kernel is defined for a finite set of L points (typically the training data and some test points))

$$\mathcal{N}(\mathbf{m} \mid \mu, \eta^{-1}K)\mathcal{IW}_L(K \mid \delta, \kappa)$$

- MAP estimates for kernel and mean are calculated using EM equations
- $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ is the base kernel and can be used to represent prior knowledge about the kernel shape

. κ is the respective Gram matrix.

EM Learning for GP-HB

- In typical applications noisy measurements for the different situations are available (for missing data: simply set noise variance to infinity)
- Complete data likelihood

$$\mathcal{N}(\mathbf{m} \mid \mu, \eta^{-1}K)\mathcal{IW}_L(K \mid \delta, \kappa)\prod_{j=1}^{M}\mathcal{N}(y_{*,j} \mid f_{*,j}, S_j)\mathcal{N}(f_{*,j} \mid \mathbf{m}, K)$$

- E-step

$$P(f_{*,j} \mid y_{*,j}, \mathbf{m}, K) = \mathcal{N}(f_{*,j} \mid \mathbf{r}_j, V_j) \qquad S_j = diag(\sigma_{i,j}^2)$$

$$V_j = (K^{-1} + S_j^{-1})^{-1} \quad \text{and} \quad \mathbf{r}_j = V_j(S_j^{-1}y_{*,j} + K^{-1}\mathbf{m})$$

- M-Step

$$\mathbf{m} = \frac{1}{M+\eta}(\sum_{j=1}^{M}\mathbf{r}_j + \eta\mu)$$

$$K = \frac{1}{M+\eta+\delta+2L}(\eta(\mathbf{m}-\mu)^T(\mathbf{m}-\mu) + \sum_{j=1}^{M}(\mathbf{m}-\mathbf{r}_j)^T(\mathbf{m}-\mathbf{r}_j) + \sum_{j=1}^{M}V_j + \kappa)$$

Reconstruction

- If we look at the estimation of the training targets we get

$$E(Y) = F = R$$

Induction: Generalizing to New Inputs

- GP-HB does not distinguish between the content-based effect and the collaborative effect
- Only the content-based effect can be generalized to new inputs (movies)

- To generalize to new inputs (induction) one can use different approximations. Schwaighofer, Tresp, Yu (2004) propose

$$k(x_i, x_k) = \kappa^T(\cdot, x_i)(\kappa + \lambda I)^{-1} K (\kappa + \lambda I)^{-1} \kappa(\cdot, x_k)$$

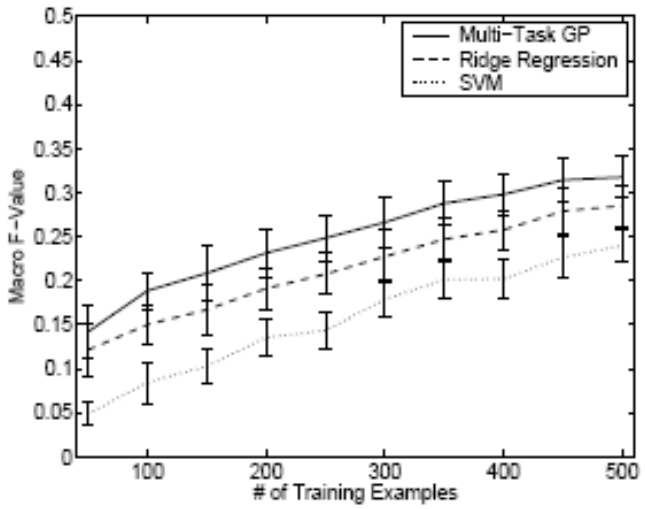
- Schwaighofer, Tresp, Yu (2004)
- Yu, Tresp, Schwaighofer (2005)
- Lawrence and Platt (2004): similar approach but without priors on mean and kernel

Predicting Reuter's labels

- 10000 documents with a total of 81 labels (situations) with TFIDF features; on average each document has 3.96 labels.
- The test set contains 9700 examples; All: evaluation on all the test points. Partially Labeled: each test document with at least one label in some category.

Table 1. Comparison of four algorithms for text categorization on RCV1

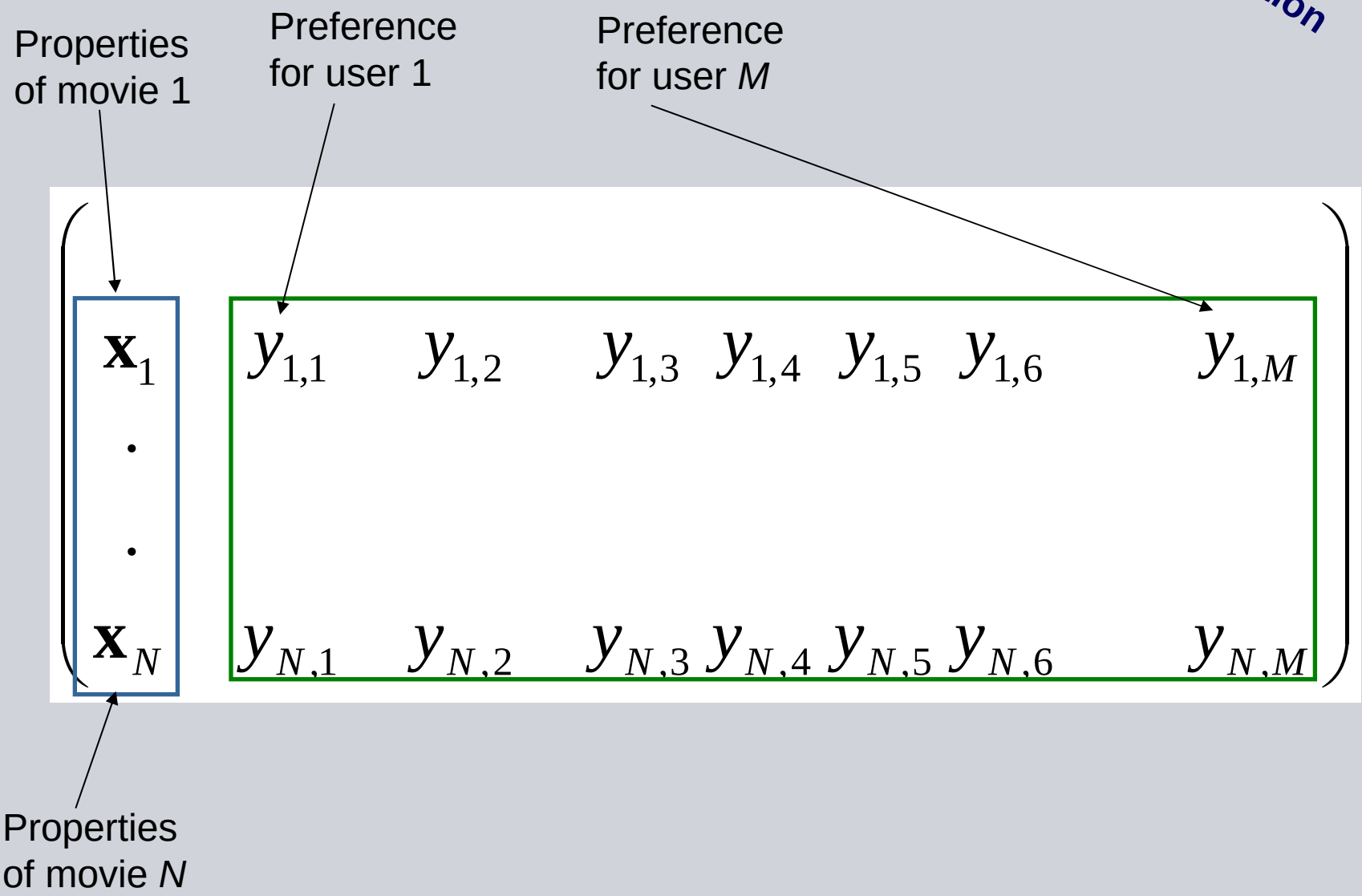
	ALL			PARTIALLY LABELED		
	AUC	F-MICRO	F-MACRO	AUC	F-MICRO	F-MACRO
MULTI-TASK GP	0.773	0.605	0.260	0.826	0.623	0.281
REGULARIZED MULTI-TASK LEARNING	0.701	0.571	0.232	0.709	0.545	0.216
RIDGE REGRESSION	0.756	0.584	0.245	0.771	0.564	0.240
SVM	0.697	0.573	0.221	0.716	0.547	0.212



- SVM, RR: separate models for each task
- RMTL: only learns common mean function

Predicting Relation between Users and Items : Recommendation System

Entity-Centered Prediction



Fast Implementation of GP-HB

Table 5: RMSE of various matrix factorization methods on the Netflix test set

Method	RMSE
Baseline	0.9514
VB [6]	0.9141
SVD [5]	0.920
BPMF [10]	0.8954
NSVD	0.9216
NPCA	**0.8926**

- Straightforward of the EM approach on Netflix will take thousands of hours per iteration

- Fast implementation plus model simplification leads to 5h/iterations
- VB: variational Bayes matrix factorization. SVD: SVD for sparse matrices. BPMF: Bayesian Probabilistic Matrix Factorization. NSVD: Max Margin Matrix Factorization. NPCA: nonparametric PCA (GP-HB)

  - Yu, Zhu, Lafferty, Gong (2009)

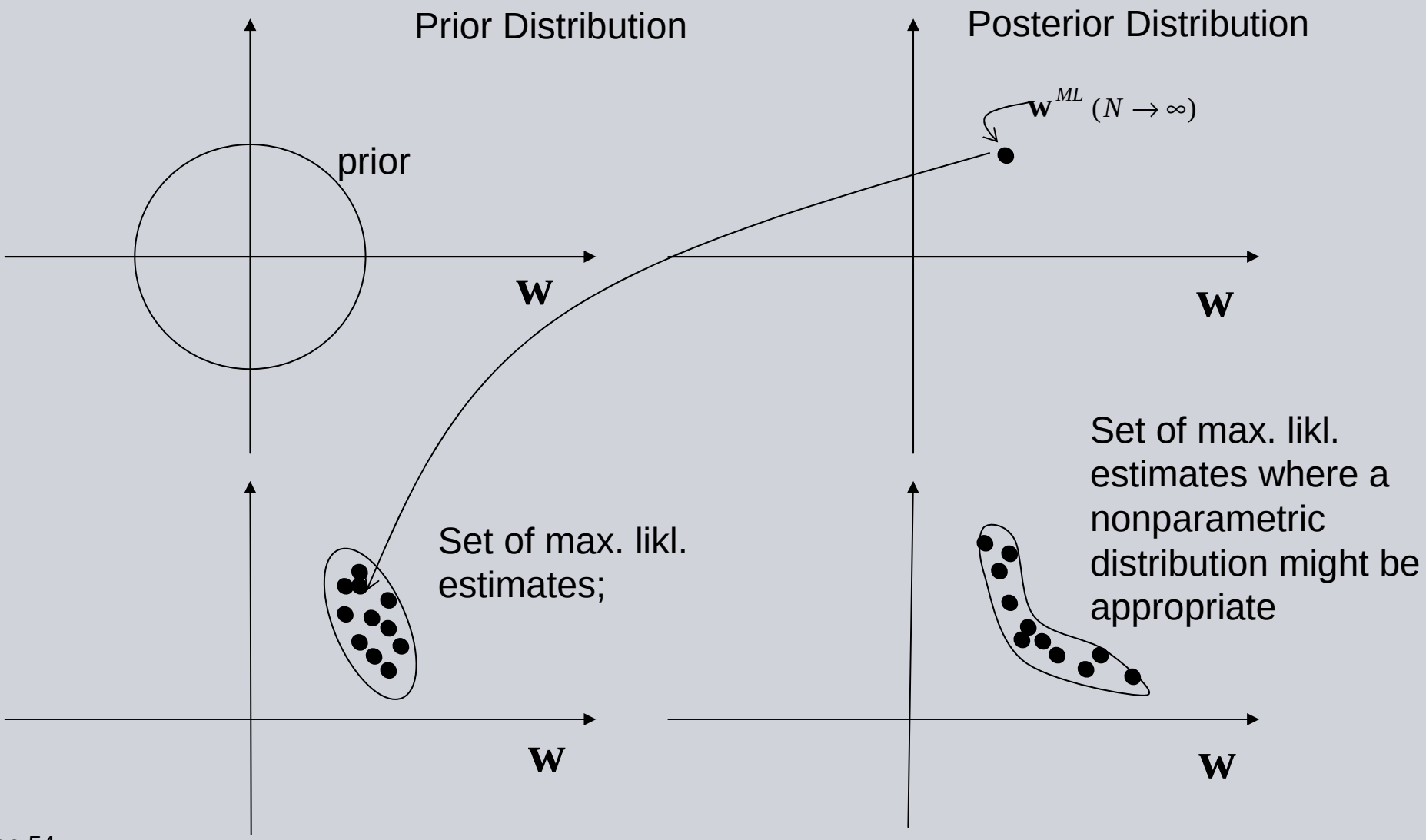
Summary Hierarchical Bayes

- **Main benefit:** data for a given situation is supported by data from other situations
- **Training:**
  - Inputs (objects) can be arbitrary in different situations
    (from another view: ***no problems with missing outputs***)
- **Generalization**
  - to new objects (inputs) is possible
  - to new situations (output dimensions) is possible
- **Output driven regularization / dimensionality reduction!**
- **Not limited to models that are linear in the parameters**
- More helpful references:
- Caruana (1995), Thrun (1996): early work
  - Zhang, Ghahramani and Yang (2005): find latent independent components (not just uncorrelated components)
  - Barutcuoglu, Schapire and Troyanskaya (2006): application to gene function prediction
  - Krishnapuram, Yu, Yakhnenko, Rao, Carin (2008): recent NIPS workshop

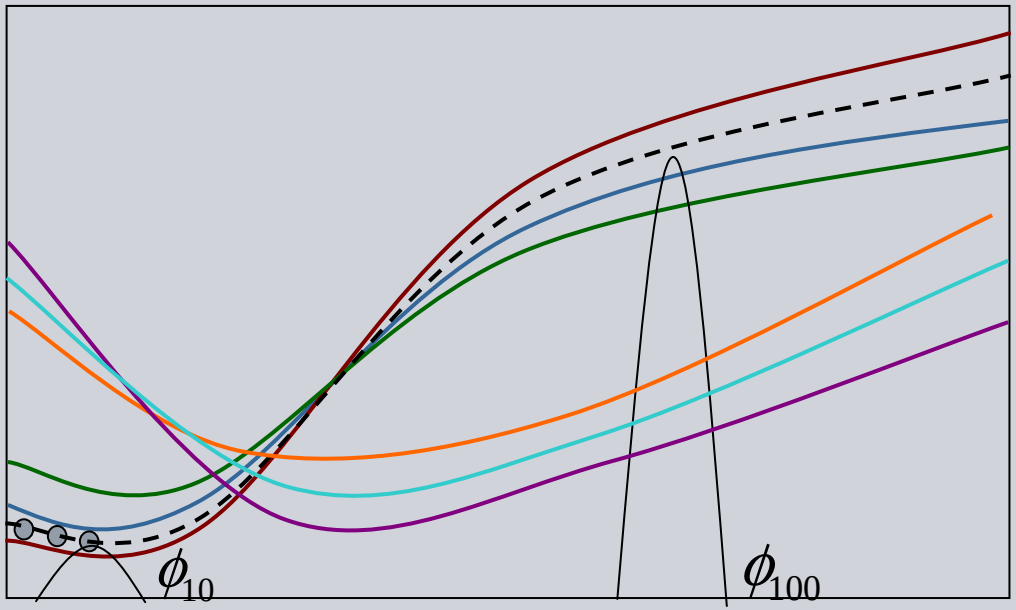
II.C.
Nonparametric Hierarchical Bayes

- The prior parameterization needs to be quite expressive!

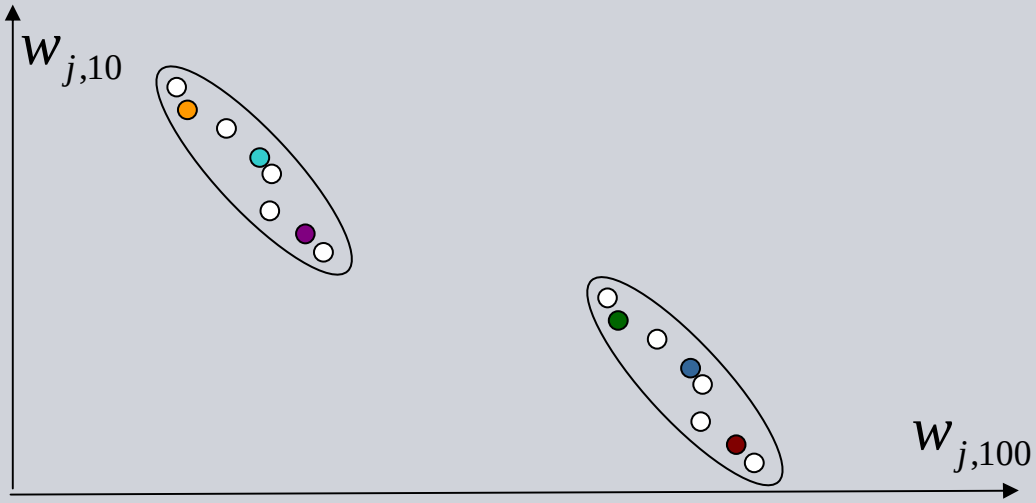
A Problem with Low-dimensional HB Approaches



Another View

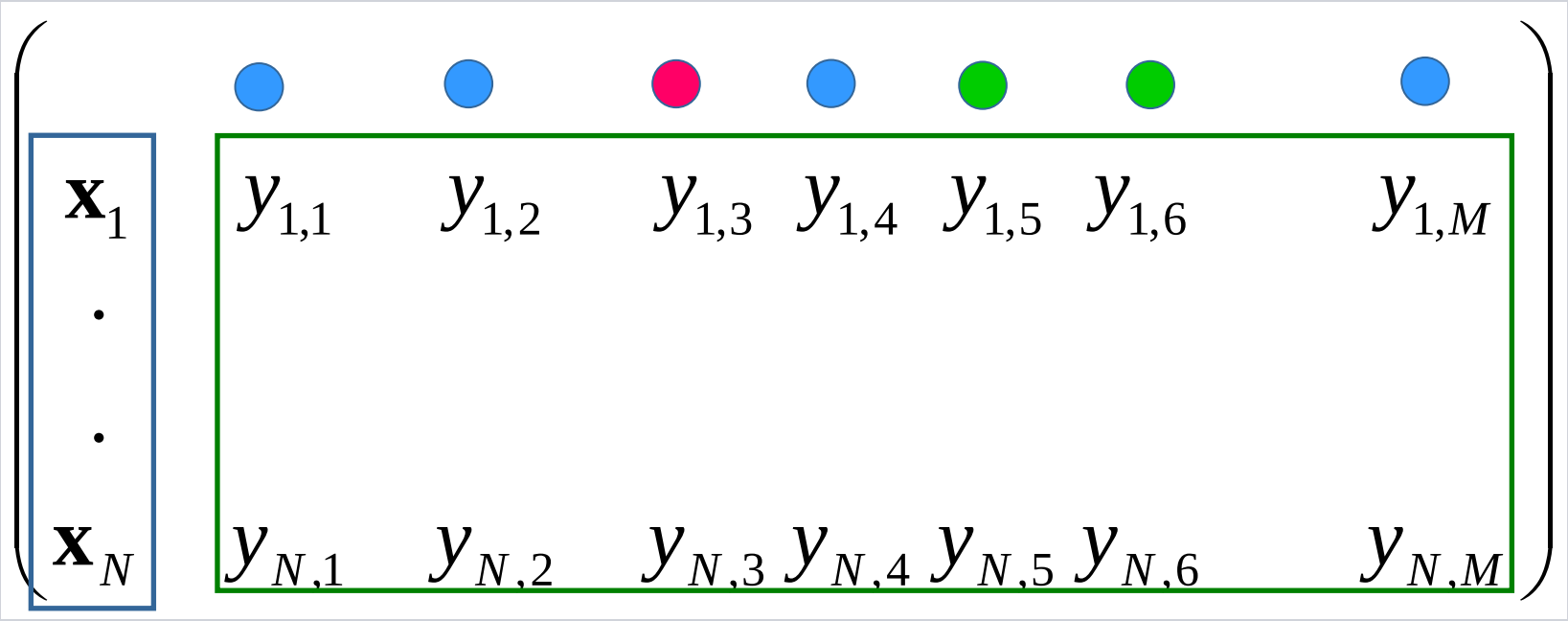


- A latent mixture model for the distribution of the parameters
- Latent variable (clustering) model of functions, not data points!
- Multi-modal learned prior distribution



(Soft) Grouping of Variables or Functions

- Colors: cluster assignment (grouping of outputs/functions, not data points)
- In each cluster, parameters are shared



Finite Models:
A Particular Mixtures of Experts Models (Regression)

- After training, let parameter vector $\mathbf{w}_l$ be assigned to cluster *l*
- As a prediction for situation *j*, based its past data $\mathcal{D}_j$ one obtains

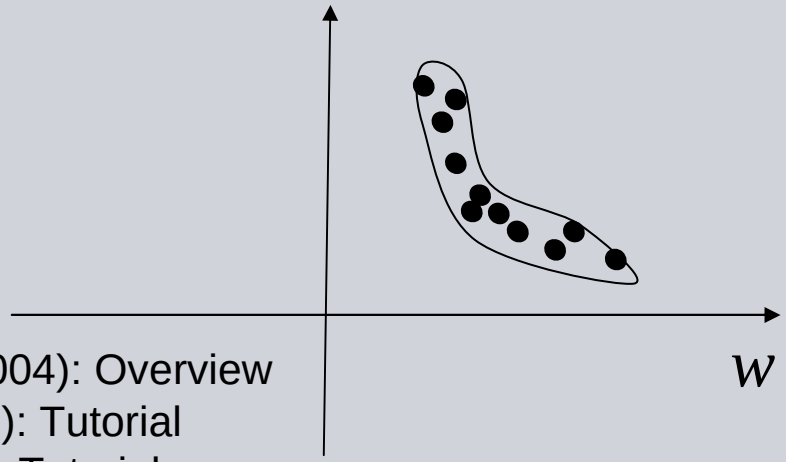
$$E(f_j(\mathbf{x})) = \sum_{l=1}^{L} f(\mathbf{x}, \mathbf{w}_l) P(Z_j = l \mid \mathcal{D}_j)$$

- Can be interpreted as a mixture of expert approach with experts $f(\mathbf{x}, \mathbf{w}_l)$ and weight $P(Z_{j=l} \mid \mathcal{D}_j)$
- Note that in contrast to the typical mixture of expert approach, we assign a whole function (i.e., situation) to a component

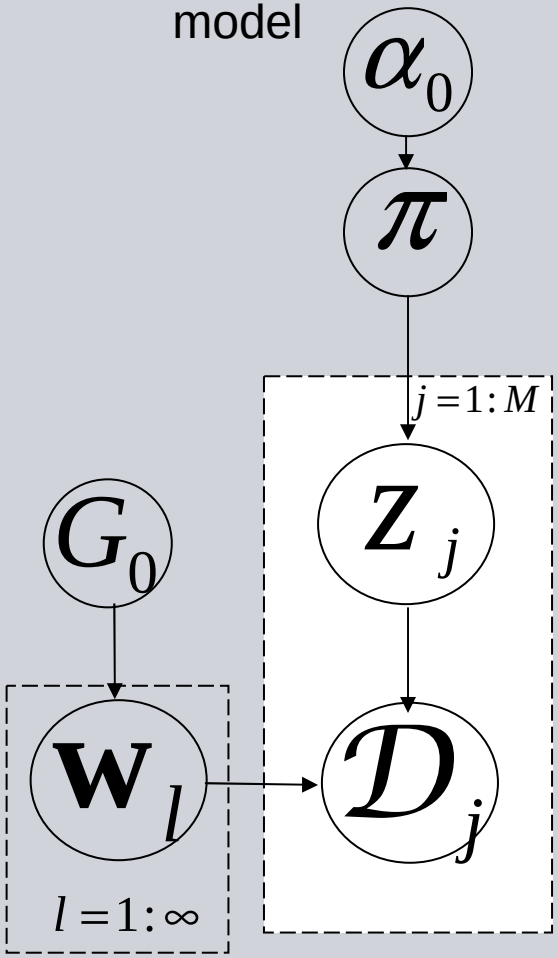
- Tresp and Yu (2004)

Dirichlet Process Mixture Models for Multitask Learning

- If, in a Bayesian approach, we let the *number of components go to infinity*, we obtain a *Dirichlet process mixture model*
- Automatic model selection: in the sampling procedure only a finite number of states is being used
- *This is equivalent to a nonparametric hierarchical Bayesian approach*



Stick breaking representation of a Dirichlet process mixture model



- Tresp, Yu (2004): Overview
- Jordan (2005): Tutorial
- Tresp (2006): Tutorial
- Xue, Liao, Carin, Krishnapuram (2007)

II.D
Nonparametric Hierarchical Bayes for Relational Learning

- Dirichlet Process MixtureModels
- Gaussian Processes

Generalization of Dirichlet Process Mixture Models (Nonparametric HB)

Predicting a Single Relationship Type

- We will first be concerned with the situation where only one relationship type is concerned
- In this case a matrix representation is appropriate where $y_{i,j}$

  describes the relationship between row entity *i* and column entity *j*

- A new aspect: attributes for both input entities and output entities are available!
- Symmetrical representation
- Note that, as before, the whole network of interlinked entities should be considered to represent a single data point, thus the matrix does not represent i.i.d samples

- In the spirit of the previous discussion we will focus on generalizations of nonparametric models

Hierarchical Bayesian versus Multivariate Mixture Models

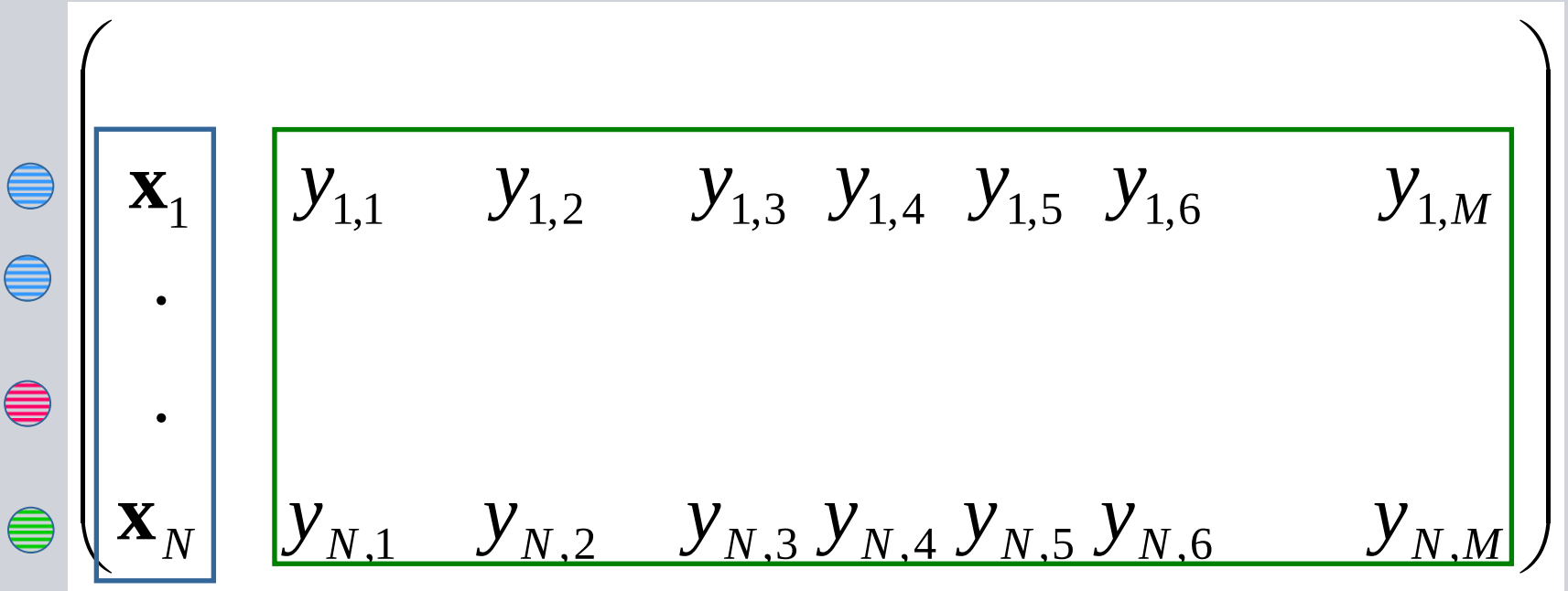
- Hierarchical Bayes:
  - In a mixture model: **columns are grouped** and share parameters
    - A common parameter vector is assigned to several output dimensions or columns (in the same cluster)

- In a multivariate analysis
  - In a mixture model: **rows are grouped** and share parameters
    - A common parameter vector is assigned to several data points (in the same cluster)

- Now
  - A mixture model for both rows and columns

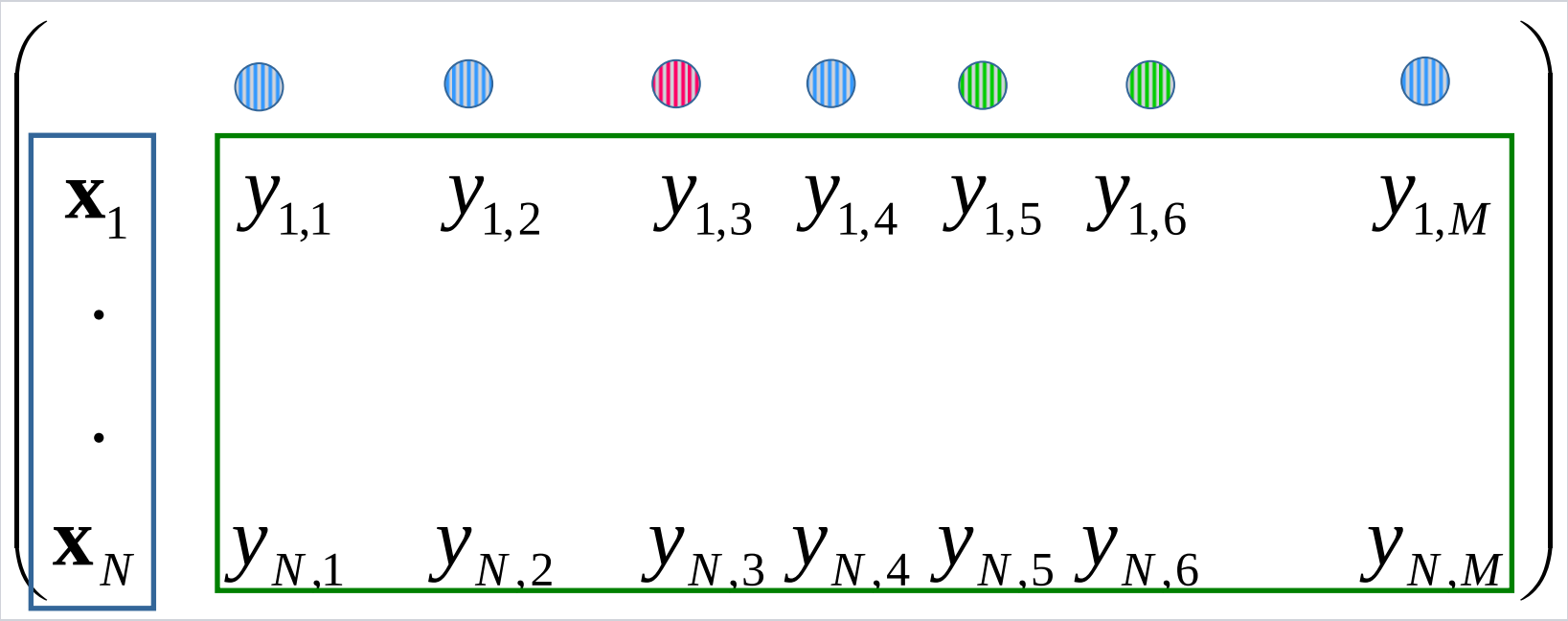
Mixture Analysis of Multivariate data

- Colors: cluster assignment (grouping of data points)



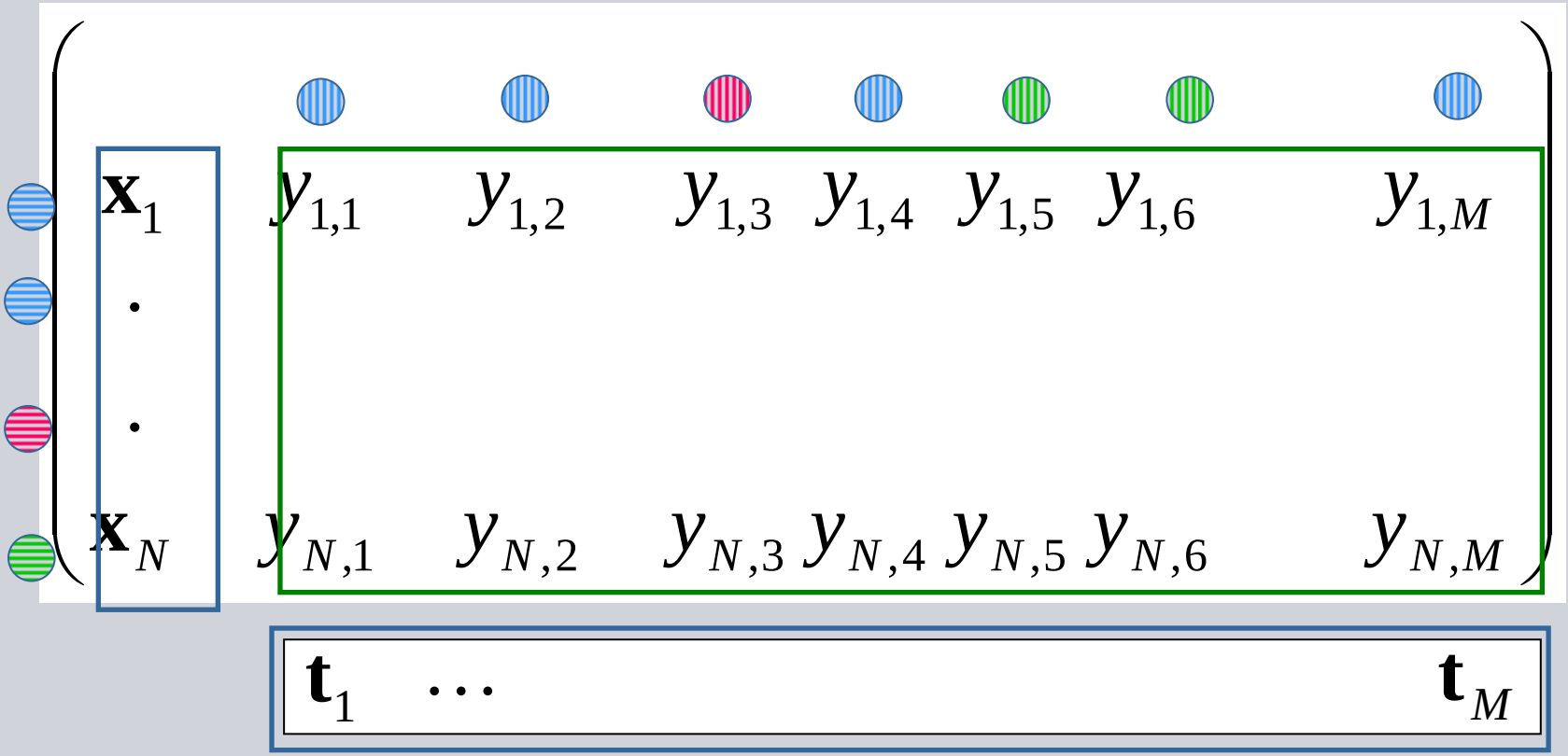
Recall: Mixture Analysis of Outputs

- Dirichlet process mixture models (Nonparametric Hierarchical Bayes)
- Colors: cluster assignment (grouping of outputs/functions, not data points)



Mixture Analysis of Input Objects and Output Objects

- Colors: cluster assignment (grouping of outputs/functions, not data points)
- ***t***: attributes of output objects
- Infinite Hidden Relational Model (IHRM, Xu et al. 2006, Kemp et al. 2006)



- Note: not really one matrix anymore: a relational data base would require at least two tables

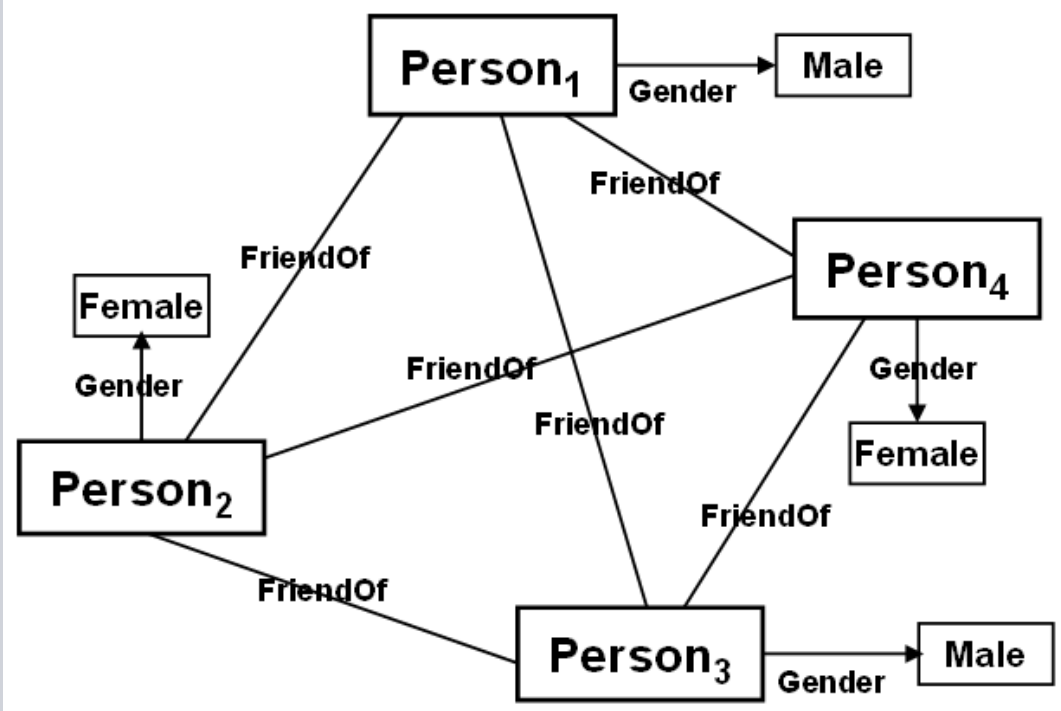
Example: Social Network

To introduce the IHRM we use a social network example

- Some persons are known to be friends
- Persons can either be male or female
- Can we predict friendship?

Graphical representation:

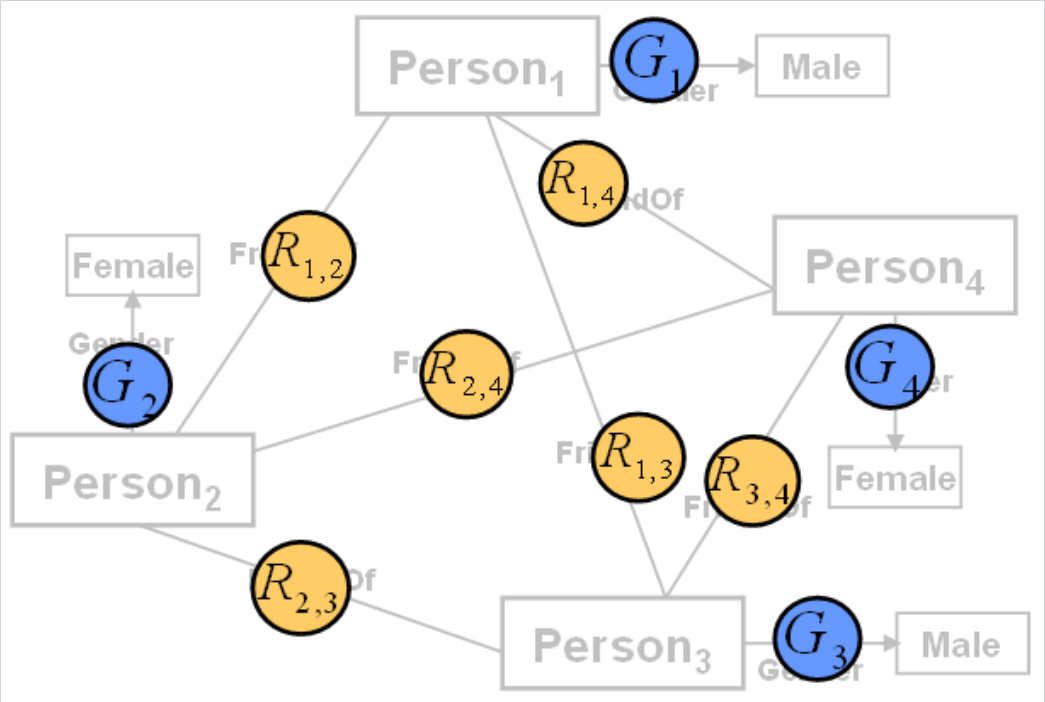
- Sociogram
- Entity-relationship graph
- RDF-Graph



- Xu, Tresp, Yu, Yu (2008)

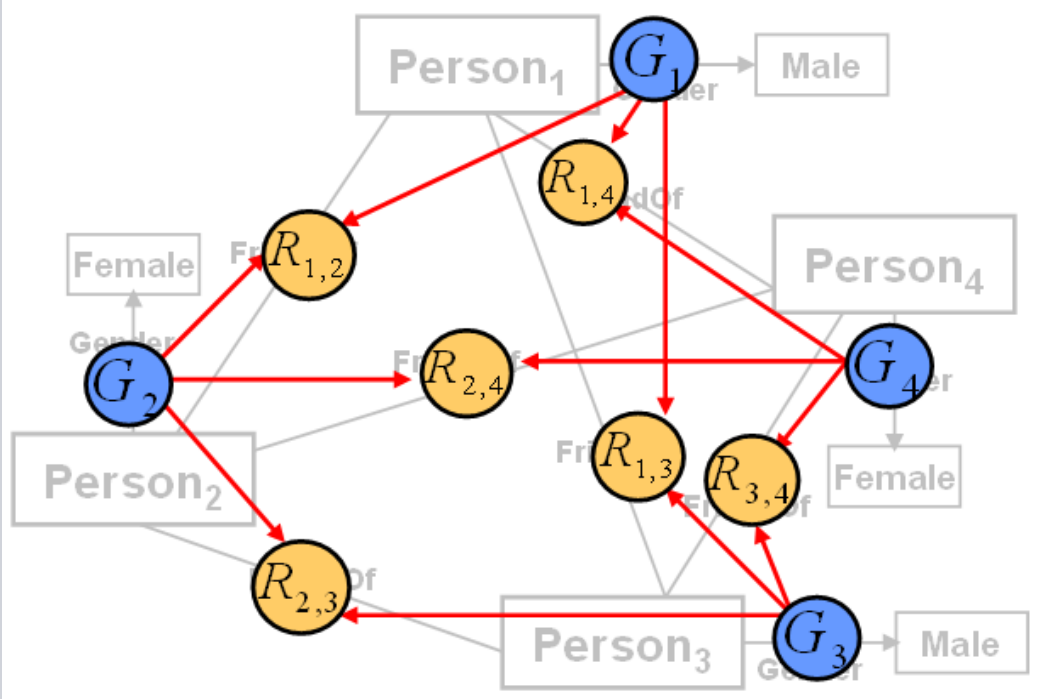
Relational Graph and Random Variables

- Each random variable stands for the truth value of a statement



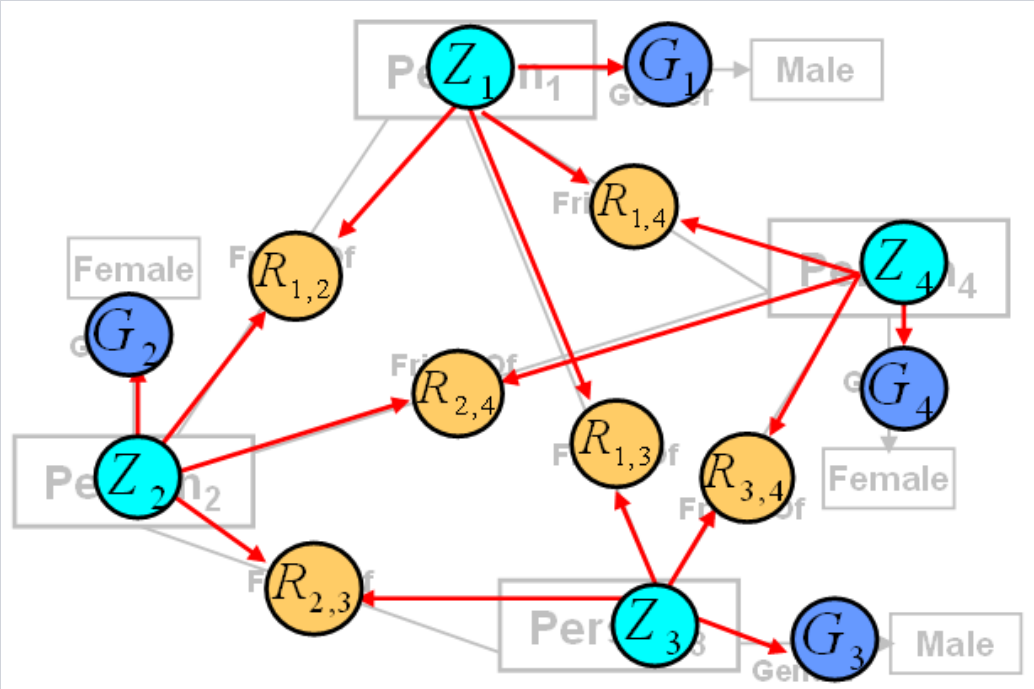
A Possible Ground Bayesian Networks

- The red directed arcs indicate direct probabilistic dependencies
- Here we assume that friendship can be predicted by the attributes (gender)
- We obtain a ground Bayesian network
- Problems:
  - Only local dependencies; no global propagation of information
  - No collaborative effect (exploiting friendship patterns)



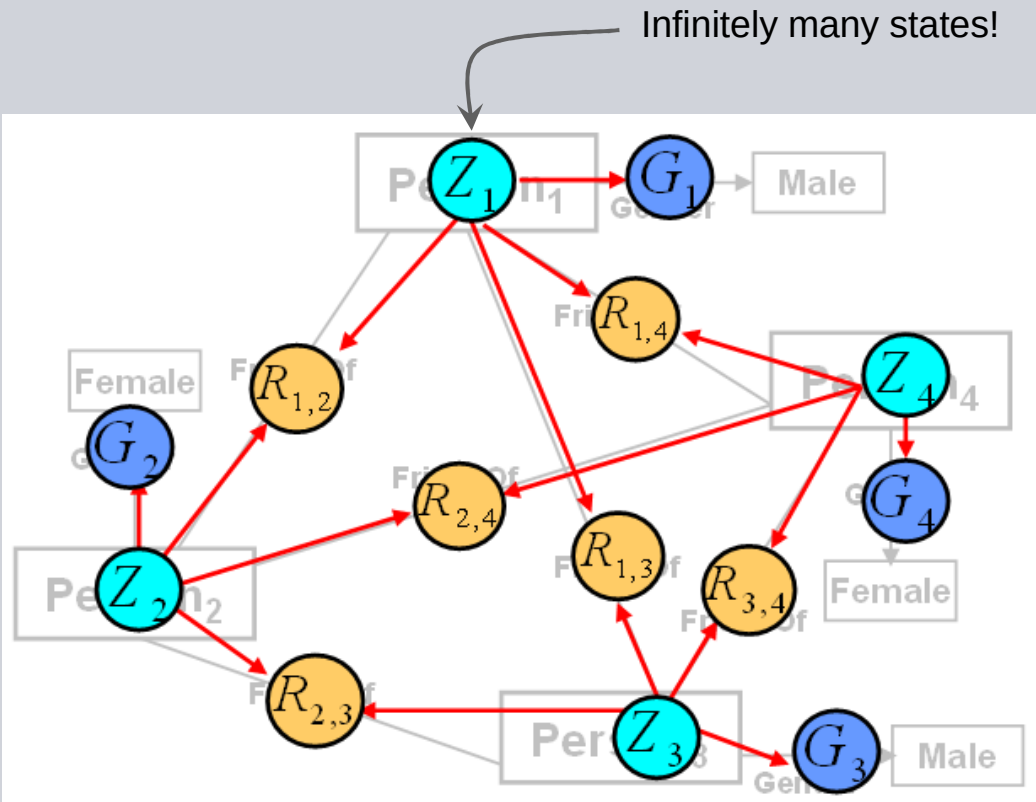
Hidden Relational Model (HRM)

- In the HRM we introduce a latent (cluster) variable for each object
- The latent variable is the parent of all nodes involving statements that include the object
- The latent variable represents the unknown information that would be sufficient to predict links (latent attributes)
- The state of the latent variable depends on
  - The attributes (gender)
  - The links an object is involved in and the states of the latent variables of the objects involved in the link.
- Identification of *roles of actors*



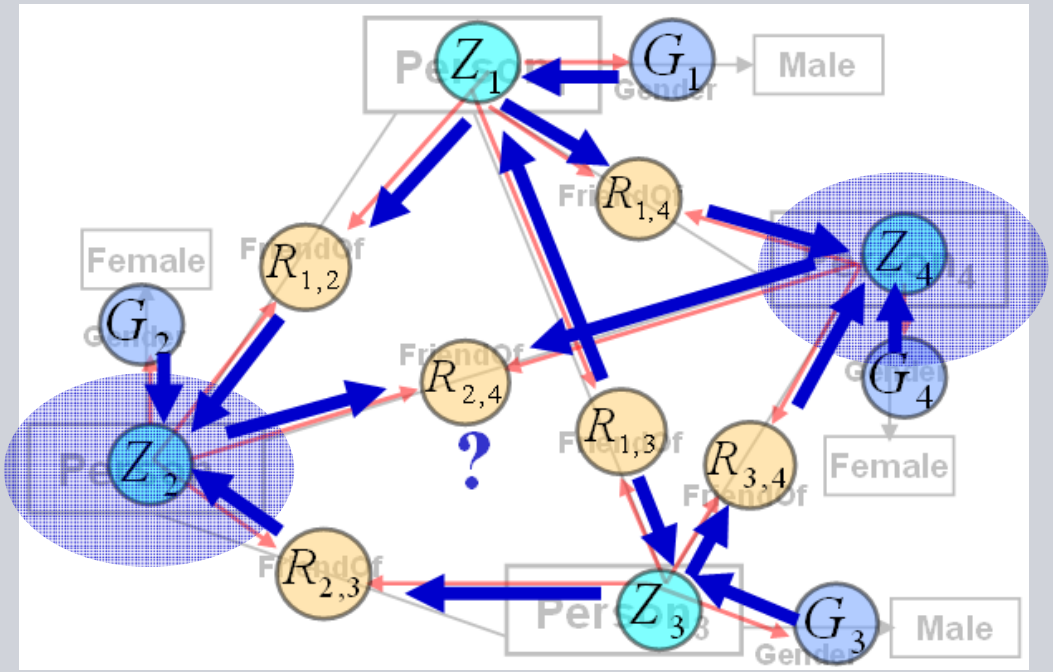
Infinite Hidden Relational Model (IHRM)

- In the IHRM the number of states in each latent variable is infinite
- We achieve a nonparametric *hierarchical Bayesian model* in form of a *Dirichlet process mixture model*
- A property of the Dirichlet process mixture models: During inference, the number of hidden states is adapted to the data in a self organized way
  - Important if different object types are involved



Information Propagation in IHRM

- Information propagates along “relational paths”
- All known information propagates to the relation of interest via hidden variables of the involved objects



Advantages of the IHRM

- Easy to apply without any extensive structural learning
  - Structural learning in Statistical Relational Learning can be quite demanding
- Information can flow through the network of latent variables and have a global effect
  - Collaborative effect (exploiting friendship patterns)
- The ground network is guaranteed to have no directed loops
- Clustering in relational domain (multi-relational clustering)
  - Analysis of clustering structure based on relational information
  - Each entity class can learn its optimal number of clusters
- No computationally-expensive feature construction (aggregation) and no global normalization

Inference/Learning in the IHRM

- A full Bayesian approach for learning and inference in the IHRM is feasible (and even practical) using Gibbs sampling
- Mean-field approximations
- Gibbs sampling simulates the model (i.e., samples from parameters and variables) conditioned on the observations

Generalization

Joint Models as Coupled Multivariate Models

- N-ary relations

$$P(R_{i,j,k} = 1 \mid H_i^{(1)}, H_j^{(2)}, H_k^{(3)})$$

- Arbitrary number of entity types

- Multiple kinds of relations

$$P(R_{i,j,k}^{type=l} = 1 \mid H_i^{(1)}, H_j^{(2)}, H_k^{(3)})$$

- General Relational Model!

Inference in the IHRM

We derived and compared various inference and learning approaches

- Gibbs sampler derived from the Chinese restaurant process representation (Kemp et al. 2004, 2006, Xu et al. 2006);
- Gibbs sampler derived finite approximations to the stick breaking representation
  - Dirichlet multinomial allocation (DMA)
  - Truncated Dirichlet process (TDP)
- Two mean field approximations based on those two approximations
- A memory-based empirical approximation (EA)

Experiment 1:
Experimental Analysis on Movie Recommendation

Task description

- To predict whether a user likes a movie given attributes of users and movies, as well as known ratings of users.
- Data set: MovieLens

MovieLens Attributes

User	Age (6)	>61; 60~46; 45~27; 26~19; 18~13; 12~4
	Gender (2)	Female; Male
	Occupation (21)	Administrator; Artist; Doctor; Educator; Engineer; Entertainment; Executive; Healthcare; Homemaker; Lawyer; Librarian; marketing; None; Other; Programmer; Retired; Salesman; Scientist; Student; Technician; Writer;
Movie	Genre (18)	Action; Adventure; Animation; Children's; Comedy; Crime; Documentary; Drama; Fantasy; Film-Noir; Horror; Musical; Mystery; Romance; Sci-Fi; Thriller; War; Western
	Year (4)	1998~1995; 1994~1990; 1989~1980; after 1979

Experimental Analysis on Movie Recommendation

Method	Prediction Accuracy (%)				Time (s)	#Comp^u	#Comp^m
	given5	given10	given15	given20			
GS-CRP	65.13	65.71	66.73	68.53	164993	47	77
GS-TDP	65.51	66.35	67.82	68.27	33770	59	44
GS-DMA	65.64	65.96	67.69	68.33	25295	52	34
MF-TDP	65.26	65.83	66.54	67.63	2892	9	6
MF-DMA	64.23	65.00	66.54	66.86	2893	8	12
EA	63.91	64.10	64.55	64.55	386	---	---

- Sampling based on the stick-breaking representation is faster than CRP-based Gibbs sampling since Z can be updated in a block; it also gave comparable performance
- Gibbs sampling finds many more components than mean field but only less than 10 have significant weight

Movie cluster analysis
Gibbs sampling with CRP

Cluster 1 (161/207) very new and popular	Cluster 2 (76/113) old, non US, drama	Cluster 3 (49/98) comedy	Cluster 4 (32/51) children
My Best Friend's Wedding (1997) **G.I. Jane (1997)** **The Truth About Cats & Dogs (1996)** **Phenomenon (1996)** **Up Close & Personal (1996)** **Tin Cup (1996)** **Bed of Roses (1996)** **Sabrina (1995)** **Clueless (1995)……**	**Big Night (1996)** **Antonia's Line (1995)** **Three Colors: Red (1994)** **Three Colors: White (1994)** **Cinema Paradiso(1989)** **Henry V (1989)** **Jean de Florette (1986)** **A Clockwork Orange (1971)** **Citizen Kane (1941)** **Mr. Smith Goes to Washington (1939) ……**	**Swingers (1996)** **Get Shorty (1995)** **Mighty Aphrodite (1995)** **Welcome to the Dollhouse (1995)** **Clerks (1994)** **Ed Wood (1994)** **The Hudsucker Proxy (1994)** **What's Eating Gilbert Grape (1993)** **Groundhog Day (1993)……**	**Event Horizon (1997)** **Batman & Robin (1997)** **Escape from L.A. (1996)** **Batman Forever (1995)** **Batman Returns (1992)** **101 Dalmatians (1996)** **The First Wives Club (1996)** **Nine Months (1995)** **Casper (1995)** **…… **
Cluster 5 (16/27) new action	Cluster 6 (9/15) old action	Cluster 7 (8/13) old drama	Cluster 8 (3/6) H. Ford, Star Wars
Conspiracy Theory (1997) **The Game (1997)** **Air Force One (1997)** **Ransom (1996)** **The Rock (1996)** **Primal Fear (1996)** **Crimson Tide (1995)** **In the Line of Fire (1993)** **The Abyss (1989)** **……**	**Brave Heart (1995)** **Forrest Gump (1994)** **Fugitive (1993)** **Terminator 2: Judgment Day (1991)** **Indiana Jones and the Last Crusade (1989)** **Die Hard (1988)** **Aliens (1986)** **Terminator (1984)** **Return of the Jedi (1983)**	**Shawshank Redemption (1994)** **Wrong Trousers (1993)** **Schindler's List (1993)** **Silence of the Lambs (1991)** **One Flew Over the Cuckoo's Nest (1975)** **Godfather (1972)** **Rear Window (1954)** **Casablanca (1942)**	**Star Wars (1977)** **Star Wars: The Empire Strikes Back (1980)** **Raiders of the Lost Ark (1981)**

Experiment 2:
Gene Interaction and Gene Function

Task

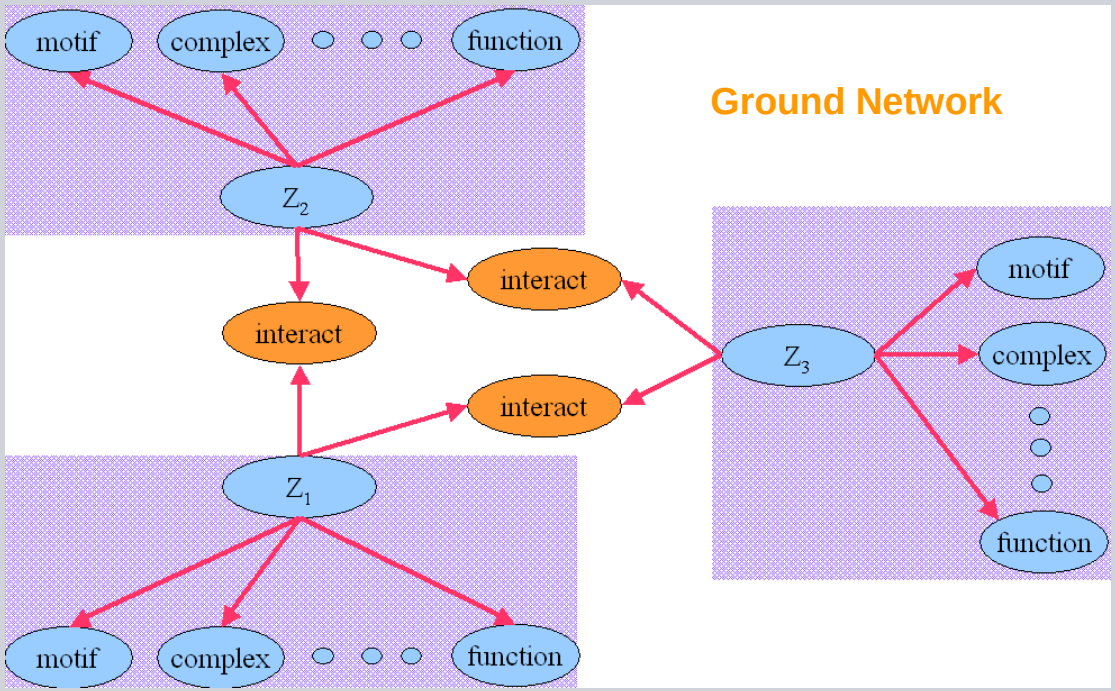
- Cluster analysis
- Prediction of gene functions given the information on the gene level and the protein level, as well as the interaction between the genes.

Attribute data: CYGD (Comprehensive Yeast Genome Database) from MIPS (Munich Information Center for Protein Sequences)

- 1000 Genes
- Attributes: Chromosome, Motif, Essential, Class, Phenotype, Complex, Function

Interaction data: DIP (data base of interacting proteins)

IHRM Model



Task: Genes (1243) have one or more **functions** (14)[1-4] (cell growth, cell organization, transport, …) to be predicted; 862 for genes for training, 381 for testing

Genes might **interact** with one another

For a gene one or more **phenotypes** (11)[1-6] are observed in the organism

How the expression of the gene can **complex** with others to form a larger protein (56)[1-3]

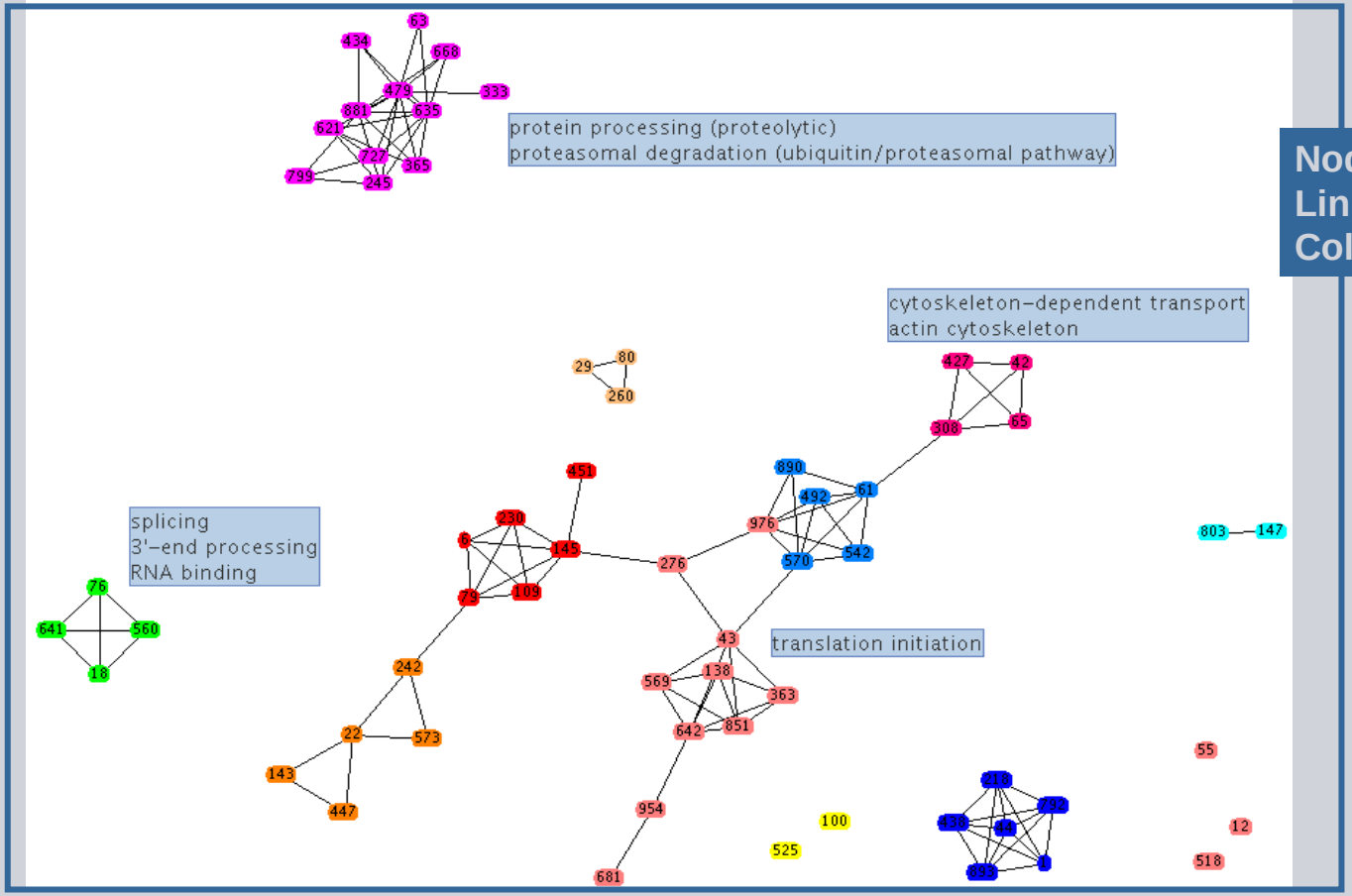
The protein coded by the gene might belong to one or more **structural categories** (24) [1-2]

A gene might contain one or more characteristic **motifs** (351) [1-6] (information about the amino acid sequence of the protein)

Gene **attributes** are: essential (an organism with a mutation can survive?), which chromosome

Cluster Structure

Some gene clusters: the genes in the same cluster have dense interactions; but the genes in the different clusters have rare interactions.



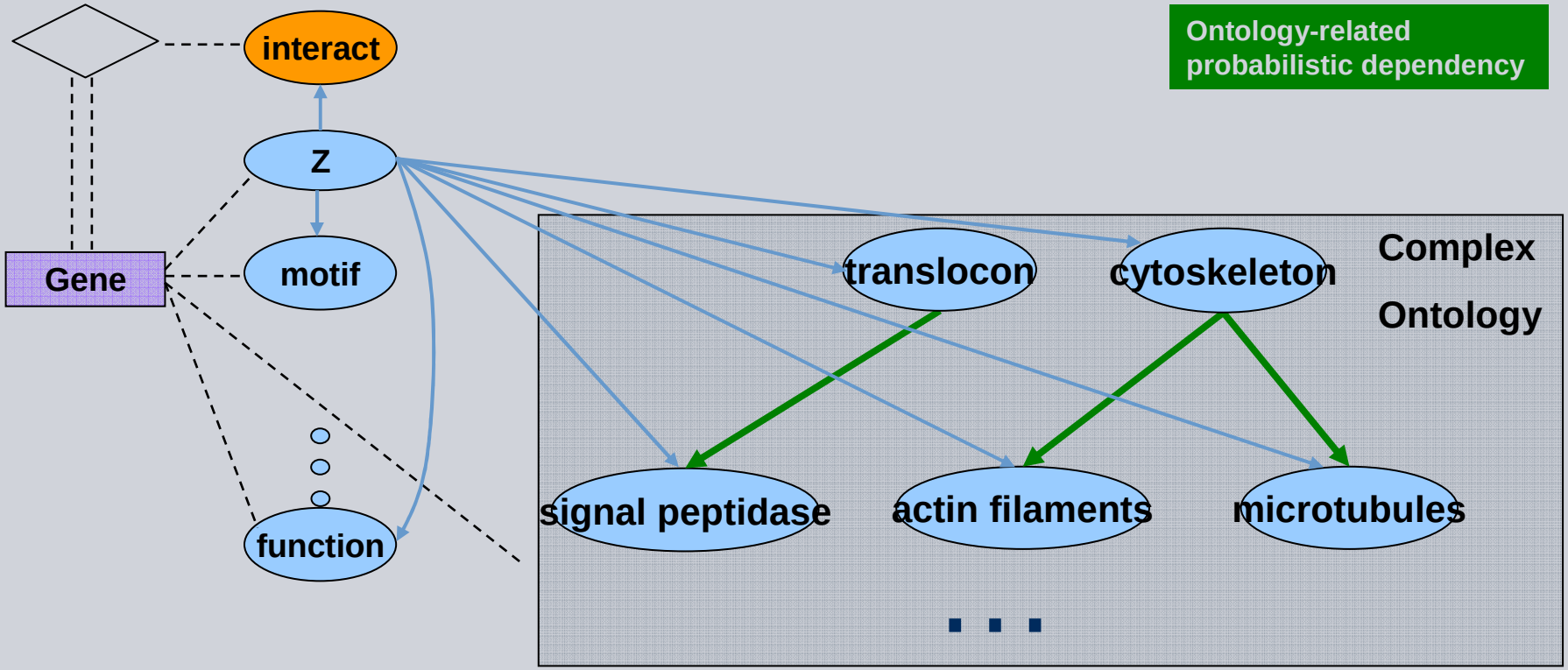
Relevance of Attributes and Relationships

The importance of a variety of relationships in function prediction of genes

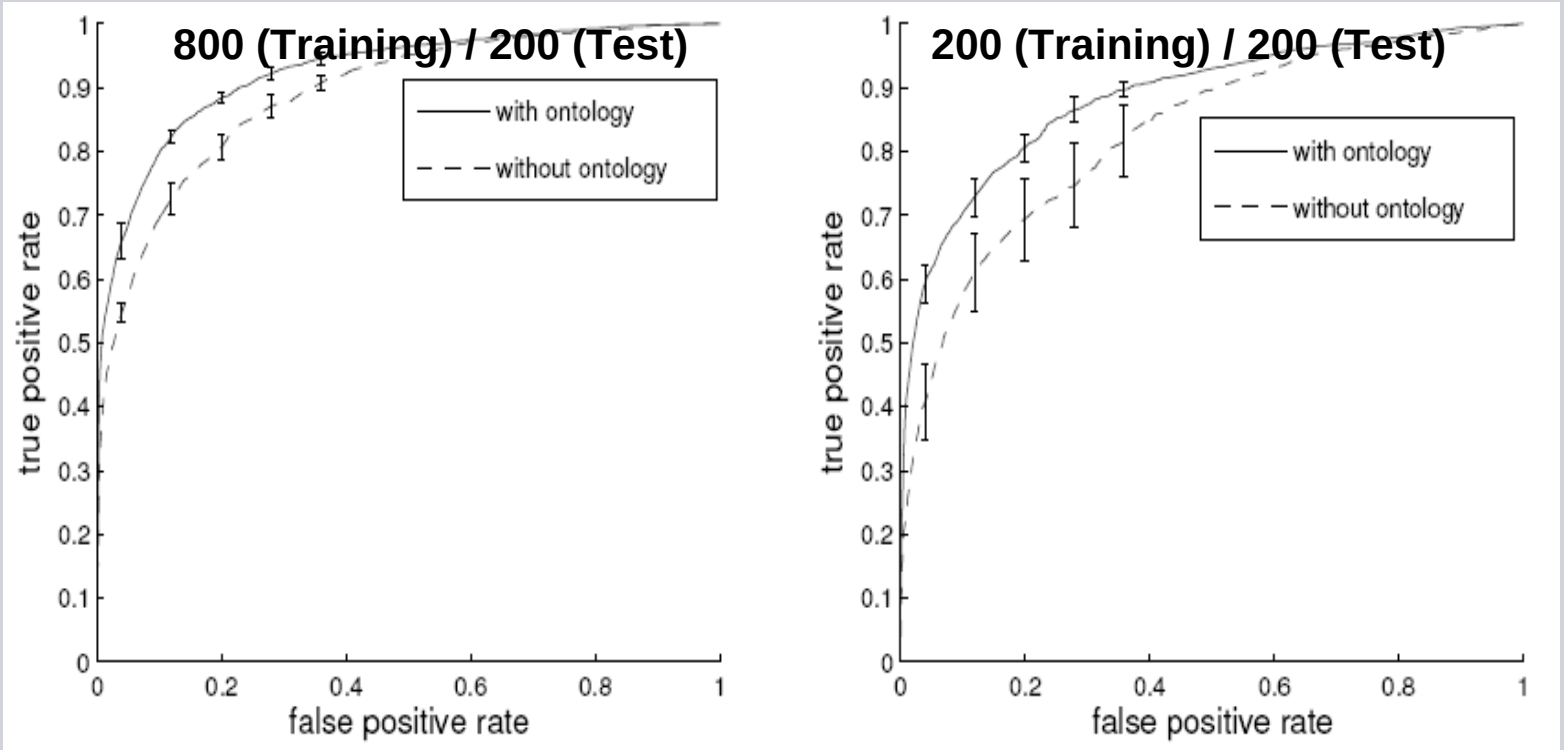
Relationships	Prediction Accuracy (%) (without the relationship)	Importance
Complex	91.13	197
Interaction	92.14	100
Structural Category	92.61	55
Phenotype	92.71	45
Attributes of Gene	93.08	10
Motif	93.12	6

Integration of Ontology into IHRM

Ontologies are a valuable source of prior information



Integration of Ontology into IHRM (2)



AUC	Without Ontology: 0.89	Without Ontology: 0.83
	With Ontology: 0.93	With Ontology: 0.89

Experiment 3:
Context-Dependent Statistical Trust Learning: Who do you trust? When?

- The need for an **evaluation of trustworthiness of agents in future encounters** is getting increasingly important in distributed systems since contemporary developments such as the Semantic Web, Service Oriented Architectures, Pervasive Computing, Ubiquitous Computing and Grid Computing are applied mainly to open and dynamic systems with interacting autonomous agents
- Most existing statistical trust models do not perform well when there is no long history of interactions in a predefined and consistent environment

- We implement and learn **context sensitive trust** from past experience using a probabilistic relational model
  - A seller might be trustworthy if offering a specific product, but not another product.

- Being the most popular online auction and shopping website fraud on eBay is a serious and well-known issue.
- eBay users leave feedback about their experiences

Rettinger, Nickles, Tresp (2008)

Infinite Hidden Relational Trust Model

ATT^a
- % of positive ratings[2]
- eBays feedback score [5]
  - More than x number of positive ratings
- Member since

ATT^s
- Top eBayCategory[47]
- Condition [new/used]

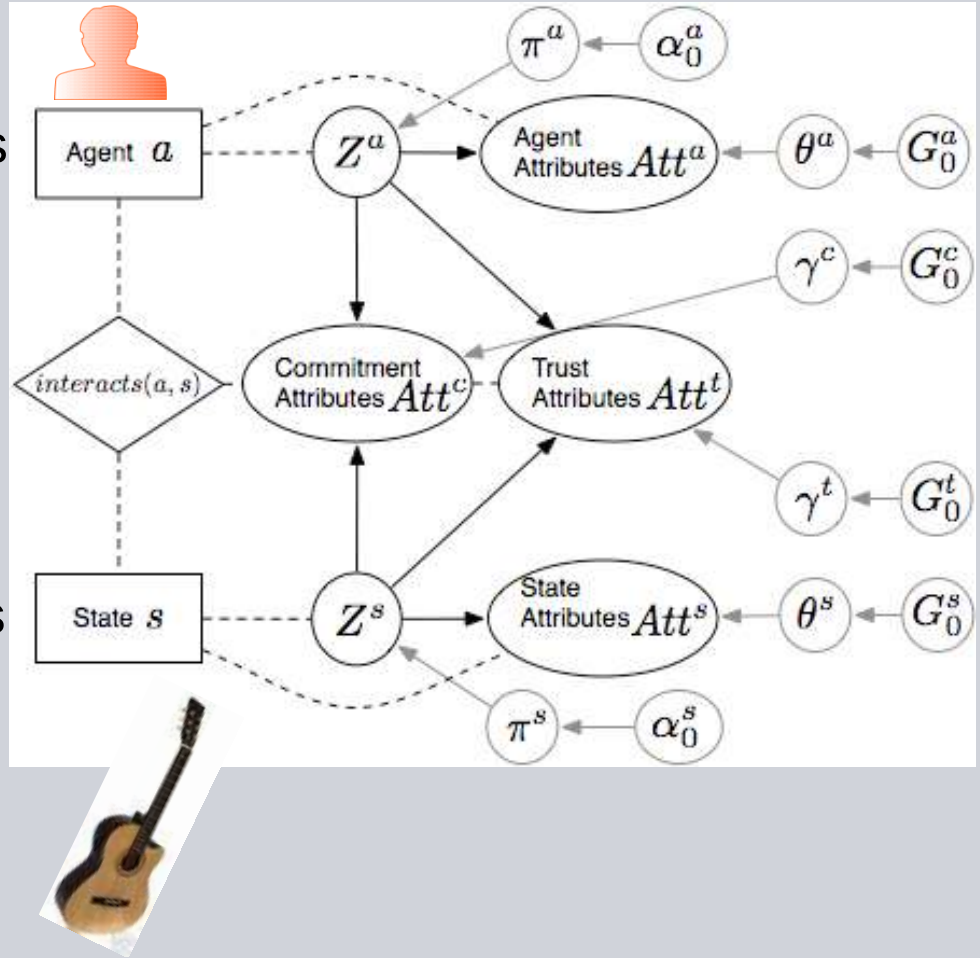
ATT^c
- Final price
- # of bids

ATT^t
- Feedback [2]

- Task:
  - Predict ATT^t for new situation

sellers

items

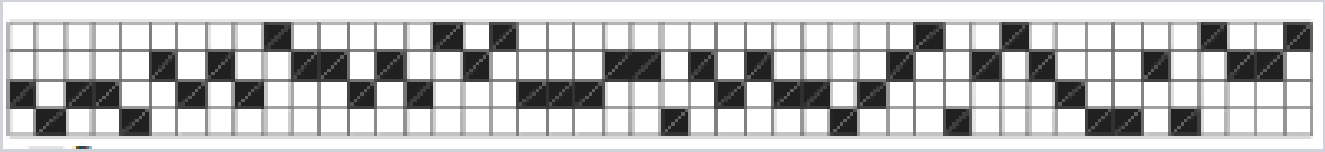


eBay Data

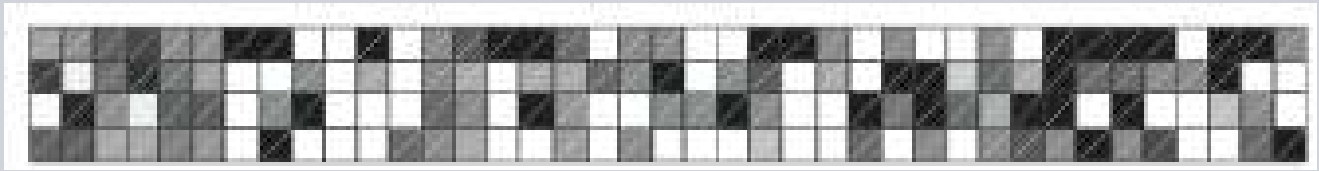
- 47 sellers (agents)
- 631 different items (states)

- 1818 rated sales (47x631 possible sales)

47 agents in 4 agent clusters



4 agent clusters versus 40 item clusters (black: trustworth)



Predictive Performance

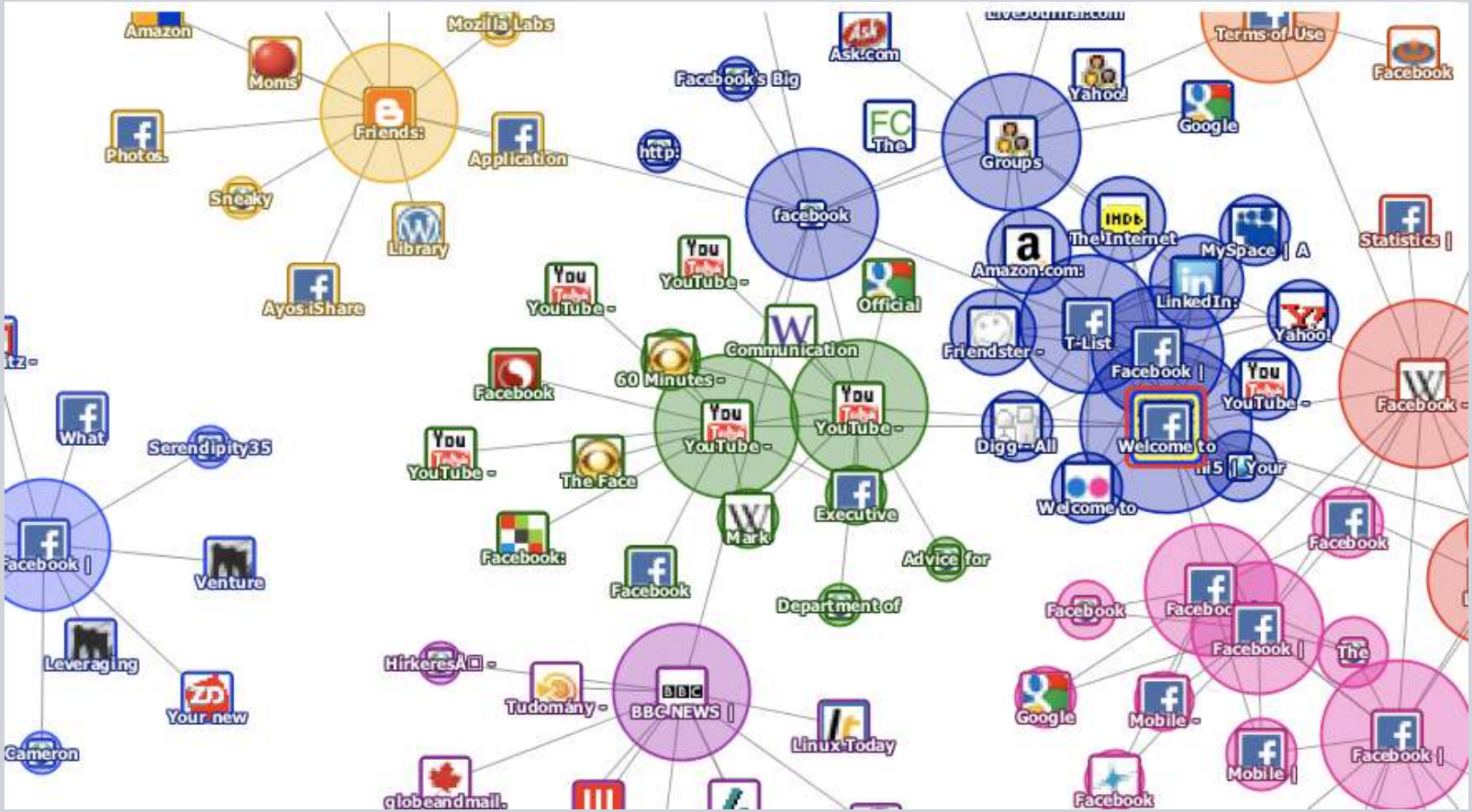
- Predicting Ratings:
  - 95% confidence interval, 5-fold cross-validation
- Ratio: Baseline
- SVM: Support Vector Machine, DecTree: Decision Tree
- +ID: Different way of propositionalizing by adding an ID-number for every entry

	Accuracy	ROC Area
Ratio	48.5334 (± 3.2407)	-
SVM	54.1689 (± 3.5047)	0.512 (± 0.0372)
DecTree	54.6804 (± 5.3826)	0.539 (± 0.0502)
SVM+ID	56.1998 (± 3.5671)	0.5610 (± 0.0362)
DecTree+ID	60.7901 (± 4.9936)	0.6066 (± 0.0473)
IHRM	71.4196 (± 5.5063)	0.7996 (± 0.0526)

Experiment 4:
Integrating DL-Ontologies

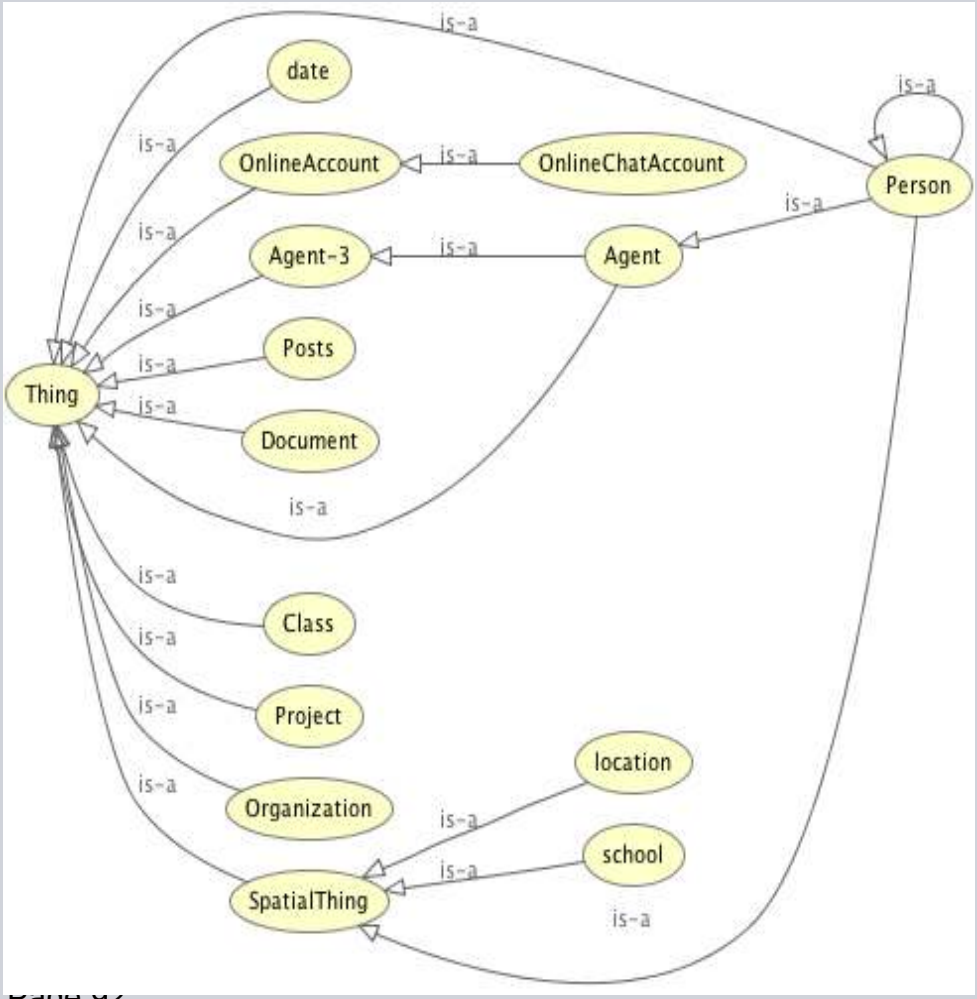
- Integrating ontological background knowledge
- From a formal ontology to a probabilistic relational model i.e., the Infinite Hidden Semantic Model (IHSM) that obeys formal constraints

Experiments - Social Networks

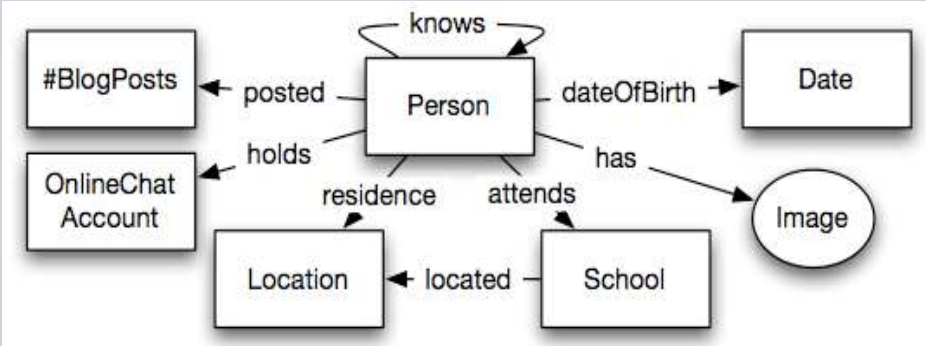


FOAF from livejournal.com

Taxonomie:



Roles:



Statistics:

Concept	#Indivi.	Role	#Inst.
Location	200	*residence*	514
School	747	*attends*	963
OnlineChatAccount	5	*holdsAccount*	427
Person	638	*knows*	8069
		hasImage	574
Date	4	*dateOfBirth*	194
#BlogPosts	5	*posted*	629

SHOIN(D) constraints for FOAF

- Constraints:

$Person \sqsubseteq Agent$	$knows^- \sqsubseteq knows$	$\exists knows.\top \sqsubseteq Person$
$\top \sqsubseteq \forall knows.Person$	$\exists hasBD.\top \sqsubseteq Person$	$\top \sqsubseteq \forall hasBD.DOB$
$\top \sqsubseteq \leq 1\ hasBD$	$\top \sqsubseteq \geq 1\ hasBD$	$\exists yearValue.\top \sqsubseteq DOB$
$\top \sqsubseteq \forall yearValue.gYear$	$\top \sqsubseteq \leq 1\ yearValue$	$\top \sqsubseteq \forall attends.School$

- The age of persons and the schools they are attending is partially known
- The ontology designer specifies that persons under the age of 6 are not allowed to attend a school

$Pupil \sqsubseteq Person$	$Pupil \sqsubseteq \neg UnderSixOld$	$Pupil \sqsubseteq \exists\ attendsSchool$

Software

- Protege: Ontology engineering
- Jena: Triple Store
- Pellet: Deductive Reasoning
- Colt: Linear Algebra, Statistics

Results: Number of Components

Concept	#Indivi.	Role	#Inst.	#C. IHRM	#C. IHSM
Location	200	*residence*	514	18	17
School	747	*attends*	963	**36**	**48**
OnlineChatAccount	5	*holdsAccount*	427	4	4
Person	638	*knows*	8069	**38**	**45**
		hasImage	574		
Date	4	*dateOfBirth*	194	4	**2**
#BlogPosts	5	*posted*	629	4	4

- IHSM (with constraining) needs more components for the concepts affected by constraints compared to IHRM (without constraining)
- For "School" and "Person" additional components were learned for inconsistent individuals
- For "Date" only 2 components were found: A "too young and one "old enough" component

Results: Predictive Performance for Different Roles

Role	attends	dateOfBirth	knows
IHRM	0.577 (±0.013)	0.548 (±0.018)	0.813 (±0.005)
IHSM	**0.608** (±0.017)	**0.561** (±0.011)	**0.824** (±0.002)

- Area under the ROC (AUC) and 95% confidence intervals for predicting relations
  - Those relations where randomly chosen for testing and withhold from training via a 5-fold cross-validation.
- Roles "attends" and "dateOfBirth" are both affected by the constraints, so it is obvious that IHSM achieve a higher performance.
- However, "knows" is not directly affected by the constraints and still IHSM shows improved performance
  - This shows that the constraints chosen by the ontology designer conform with the actual evidence.

Generalization of Gaussian Processes Hierarchical Bayes

Relationship-Centered Prediction with Gaussian Processes

- Gaussian Processes HB can also be generalized to a two-sided solution

$$F = UV^T$$

$$f_{i,j} = \sum u_{i,k} v_{j,k}$$

- Let *Y* be a link matrix and let *F* be its approximations
- With SVD $F = \tilde{U} D \tilde{V}^T$

  one solution is

$$U = \tilde{U}\sqrt{D}$$

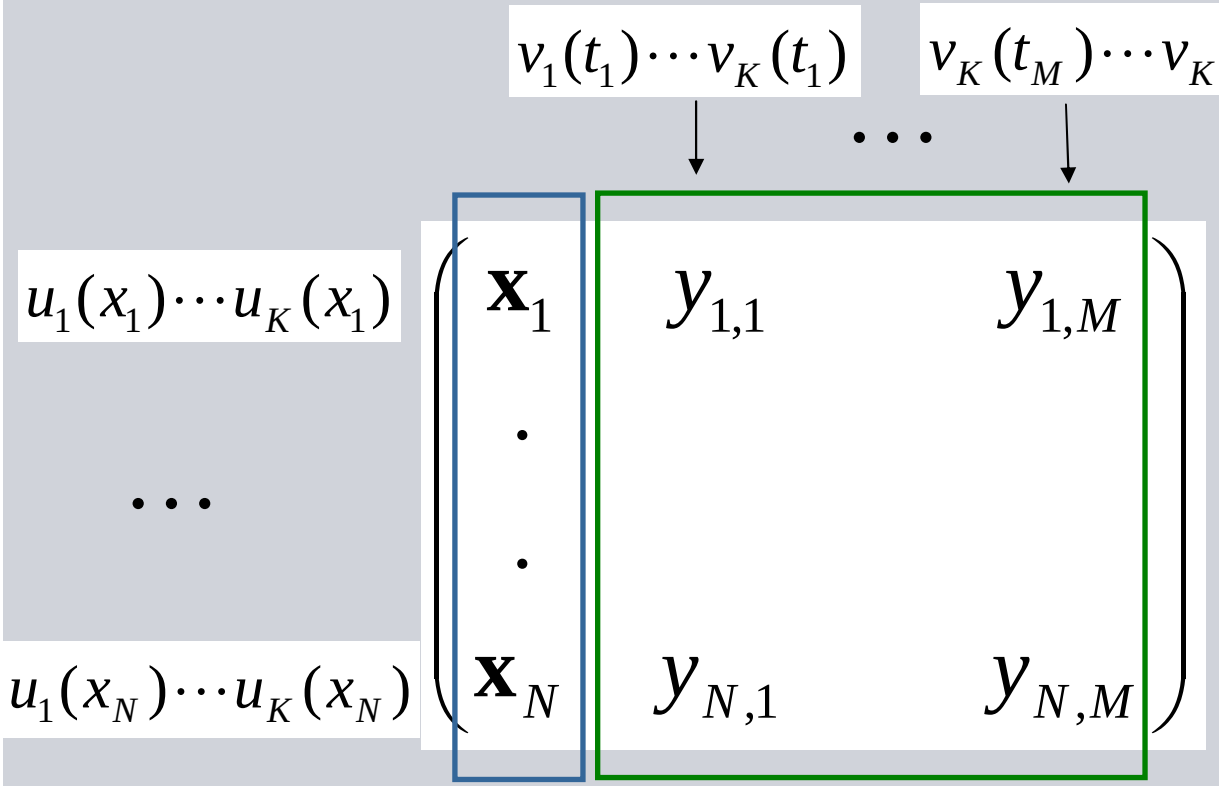
$$V^T = \sqrt{D}\tilde{V}^T$$

- We can integrate attribute information by assuming that $u_{*,k}, v_{*,k}$

  are Gaussian processes (input dependent)

Stochastic Relational Model: Multi-task Learning using Task-specific features

- Similar architecture but the latent components consist of *K* continuous variables generated from Gaussian processes



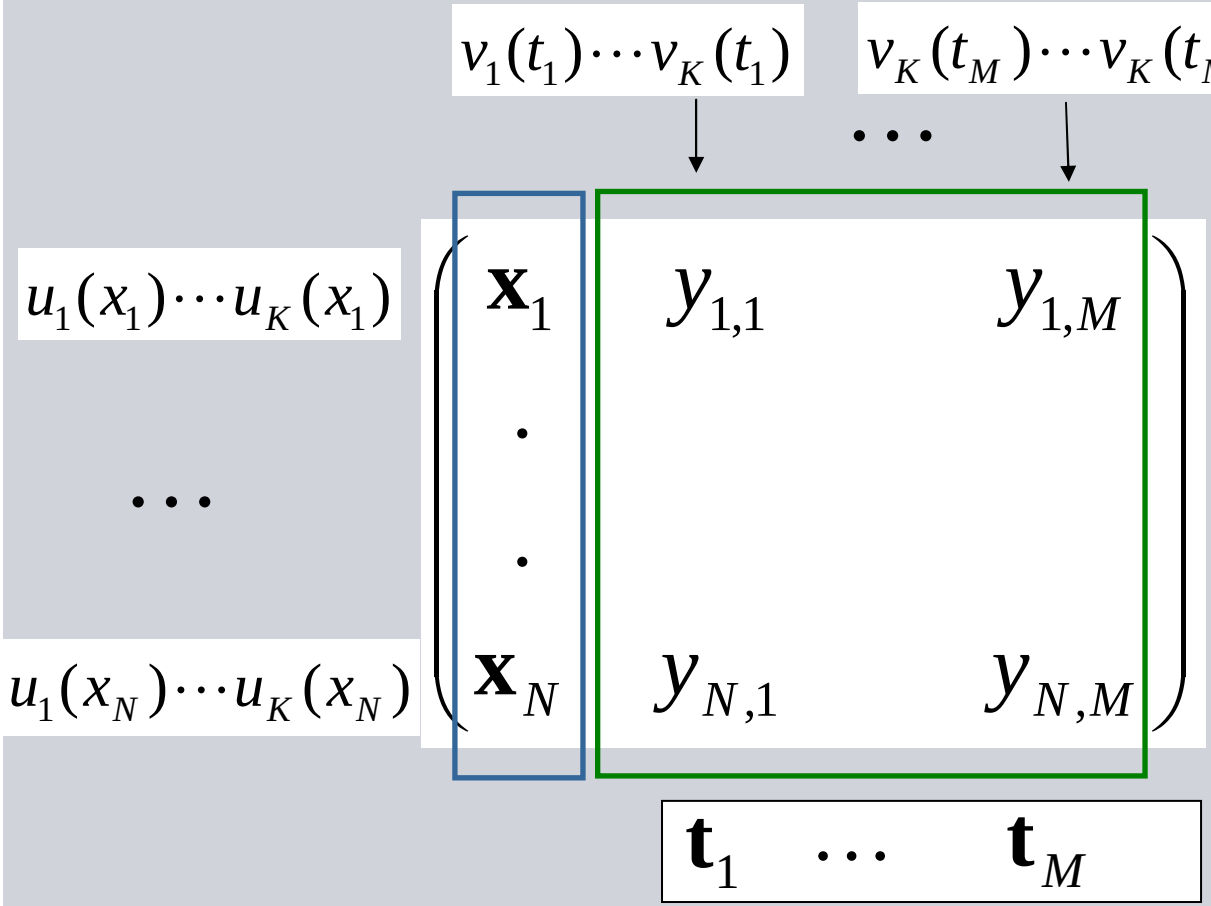
$$f(\mathbf{x}_i, \mathbf{t}_j) = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} u_k(\mathbf{x}_i) v_k(\mathbf{t}_j)$$

$$F = \frac{1}{\sqrt{K}} UV^T$$

- Yu, Chu, Yu, Tresp, Xu (2006)
- Yu and Chu (2007)
- Bonilla, Chai, and Williams (2007)
- Zhu, Yu, Gong (2008)

- Yu, Chu, Yu, Tresp, Xu (2006)
- Zhu, Yu, Gong (2008)

Stochastic Relational Model
Multi-task Learning using Task-specific features (2)



- Given two prior kernel functions based on row & column features:

$$\Omega_0(\mathbf{x}_i, \mathbf{x}_{i'}),\ \Sigma_0(\mathbf{t}_j, \mathbf{t}_{j'})$$

SRM defines a distribution for the rank-k relational function *f(x,t)*

- Generalization of matrix factorization using attributes in a hierarchical Bayesian framework

- Efficient Gibbs sampler is developed to do full Bayesian inference (code is available online)
- Applied to Netflix data (480189x17770), gave excellent performance
- In the limit k->infinity, *f(x,t)* follows a Gaussian process

$$GP(0, \Omega \otimes \Sigma) \qquad Cov(f_{ij}, f_{i',j'}) = \Omega(x_i, x_{i'})\Sigma(t_j, t_{j'})$$

Some Related Work

- C. Lippert, S.-H. Weber, Y. Huang, V. Tresp, M. Schubert, H.-P. Kriegel: *Relation Prediction in Multi-Relational Domains using Matrix Factorization*, NIPS 2008 Workshop on Structured Input Structured Output

- Z. Xu, K. Kersting, V. Tresp. *Multi-relational learning with gaussian processes*. In Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09), July 2009.

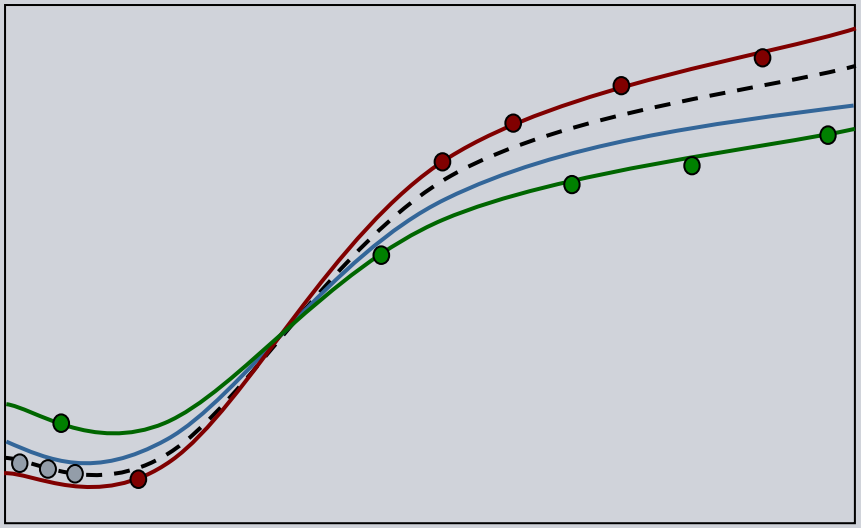
Summary

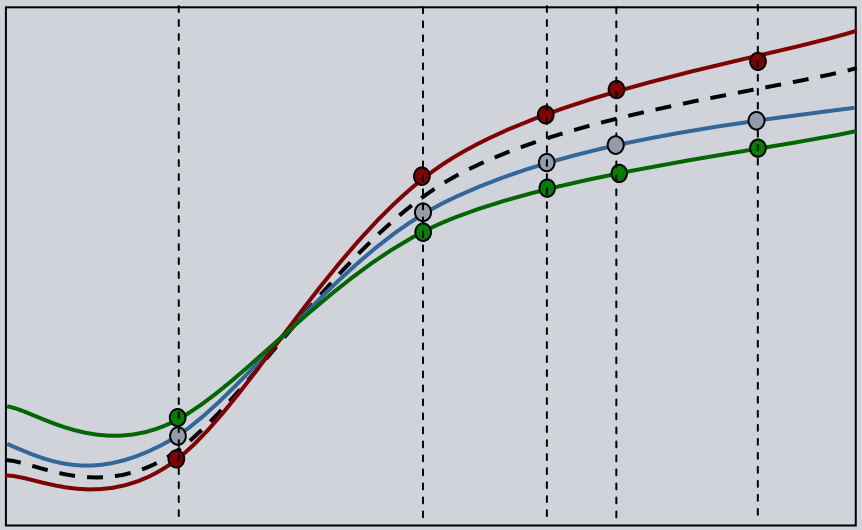
- The IHRM is a natural generalization of mixture models and of nonparametric Bayesian models to relational domains: both attributes and relationships can be predicted
- The SRM is a natural generalization of GP-HB to a relational domain

- Both the IHRM and the SRM can be generalized to Joint Models (as Coupled Multivariate Models)

III. Projection Methods

- For the set of objects all (or many) outputs (labels) are available
  - before
  - now





Projection Methods:

- Recall: Hierarchical Bayes
  defines new derived basis functions

$$\psi_k(\mathbf{x}) = d_{k,k} \sum_l v_{l,k} \phi_l(\mathbf{x})$$

- The projection methods considered here have a similar goal: they define new basis functions as a linear combination of the existing basis functions, such that the (independent) prediction of the outputs is improved

Projection Methods:

- Projection methods are only mentioned in passing
    - They are closely related to HB modeling
    - There is a huge literature
- But
    - Little or no use in relational modeling
    - Require: complete training data

Projection Methods: Principle Component Regression

- Principle component regression (PCR) is based on an optimal approximation of the design matrix

$$\min \| \Phi - W^T V \|_F \quad \text{where} \quad V^T V = I$$

- The derived basis functions are

$$\psi_j(\mathbf{x}) = \sum_k \nu_{k,j} \phi_k(\mathbf{x})$$

- In our context, the disadvantage of PCR is that it only considers input information

Projection Methods: Canonical Correlation

- It is desirable to also take into account output information
- An example is Canonical Correlation Analysis (CCA), which solves

$$\max_{u,v} (\Phi u)^T (Yv) \qquad u^T u = 1, \quad v^T v = 1$$

- The solution is based on a generalized eigenvector problem
- Related: Partial Least Squares (PLS), Linear Discriminant Analysis (LDA)
  - Shawe-Taylor and Christianini (2004)

Summary: Projection Methods

- Suitable when for a given *x*, the target is known at all (or most) situations in training but in testing, no outputs are available
- Close connection to *Hierarchical Bayes* modeling

- Suitable for predicting many labels of objects (text annotaions, image annotations) based on object features!
- Generalization
  - to new objects (inputs) is possible
  - to new situations (output dimensions) is possible

- Output driven dimensionality reduction!
- Limited to models that are linear in the parameters resp. kernel representations
- There is a huge literature on projection methods

  (e.g., papers in Hardoon, Leen, Kaski and Shawe-Taylor (2008)
- *For relational learning not so interesting, since assumes complete data in training*

IV.
Multivariate Modeling: *Unstructured*

Main Difference

- With Hierarchical Bayes and with Projection Methods: after training, there is no coupling between the various outputs

$$P(y_{i,*} \mid \mathbf{x}_i, \mathbf{w}) = \prod_{j=1}^{M} P(y_{i,j} \mid \mathbf{x}_i, \mathbf{w}_i)$$

- Now we consider models, for which -after training- the dependencies between the outputs are part of the model

$$P(y_{i,1}, \cdots, y_{i,M} \mid \mathbf{x}_i, \mathbf{w})$$

- In particular in induction, both approaches behave differently

Predicting a Single Output

- From $P(y_{i,1}, \cdots, y_{i,M} \mid \mathbf{x}_i)$

we can marginalize and obtain

$$P(y_{i,j} \mid \mathbf{x}_i) = \sum_{y_{i,1}, \cdots, y_{i,M} \setminus y_{i,j}} P(y_{i,1}, \cdots, y_{i,M} \mid \mathbf{x}_i)$$

Thus the marginal of a single output variable given the input is, in general, a *complex mixture model*

Mixture Models

Mixture Model

- Joint distribution (complete data)

$$P(y_{i,*}, \mathbf{x}_i, Z_i = l) = P(Z_i = l) P(y_{i,*}, \mathbf{x}_i \mid Z_i = l)$$

- Integration out the latent variable leads to the log-likelihood (EM-training)

$$l = \sum_{i=1}^{N} \log \sum_l P(Z_i = l) P(y_{i,*}, \mathbf{x}_i \mid Z_i = l)$$

  - Prediction of a single output:

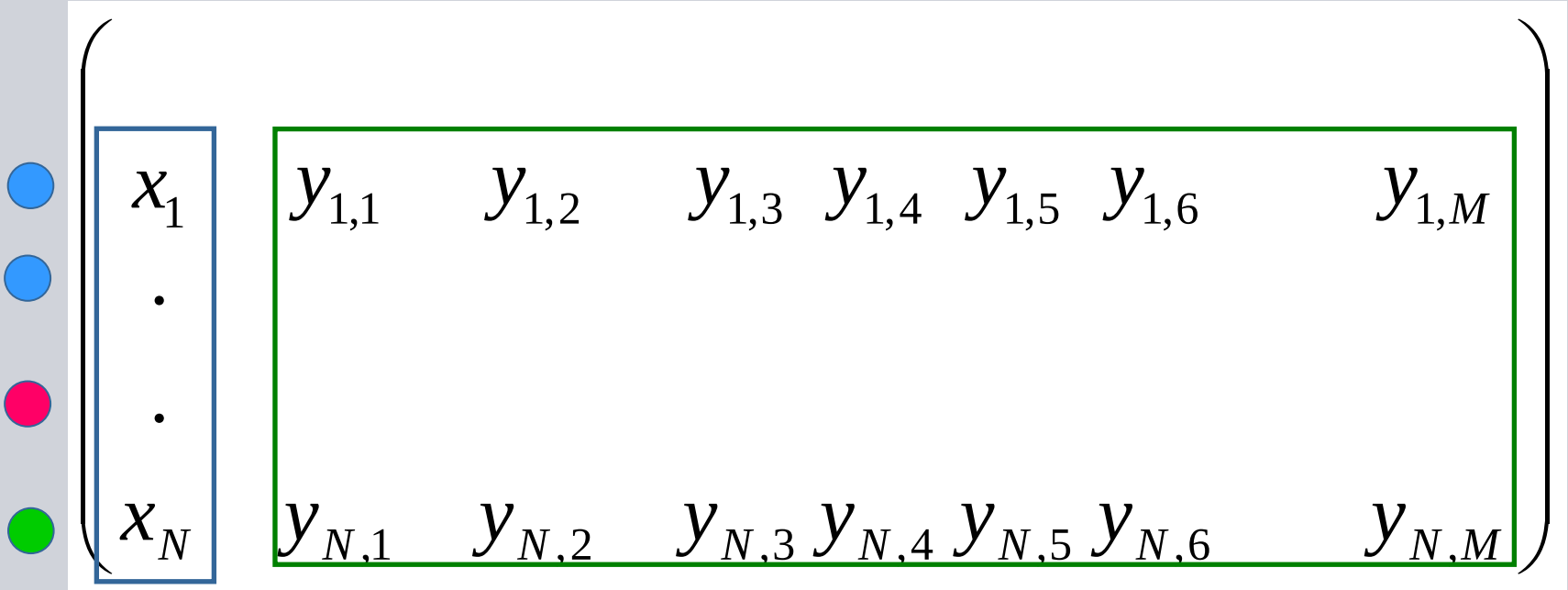
$$P(y_{i,j} \mid \mathbf{x}_i) = \frac{1}{Z(\mathbf{x}_i)} \sum_{l=1}^{l} P(Z_i = l) P(y_{i,j}, \mathbf{x}_i \mid Z = l)$$

  - Sharing strength: component assignments of a data point in training depend on all outputs
  - Infinite number of clusters ->(Another) Dirichlet process mixture model

Conditional from Joint: Mixture Model

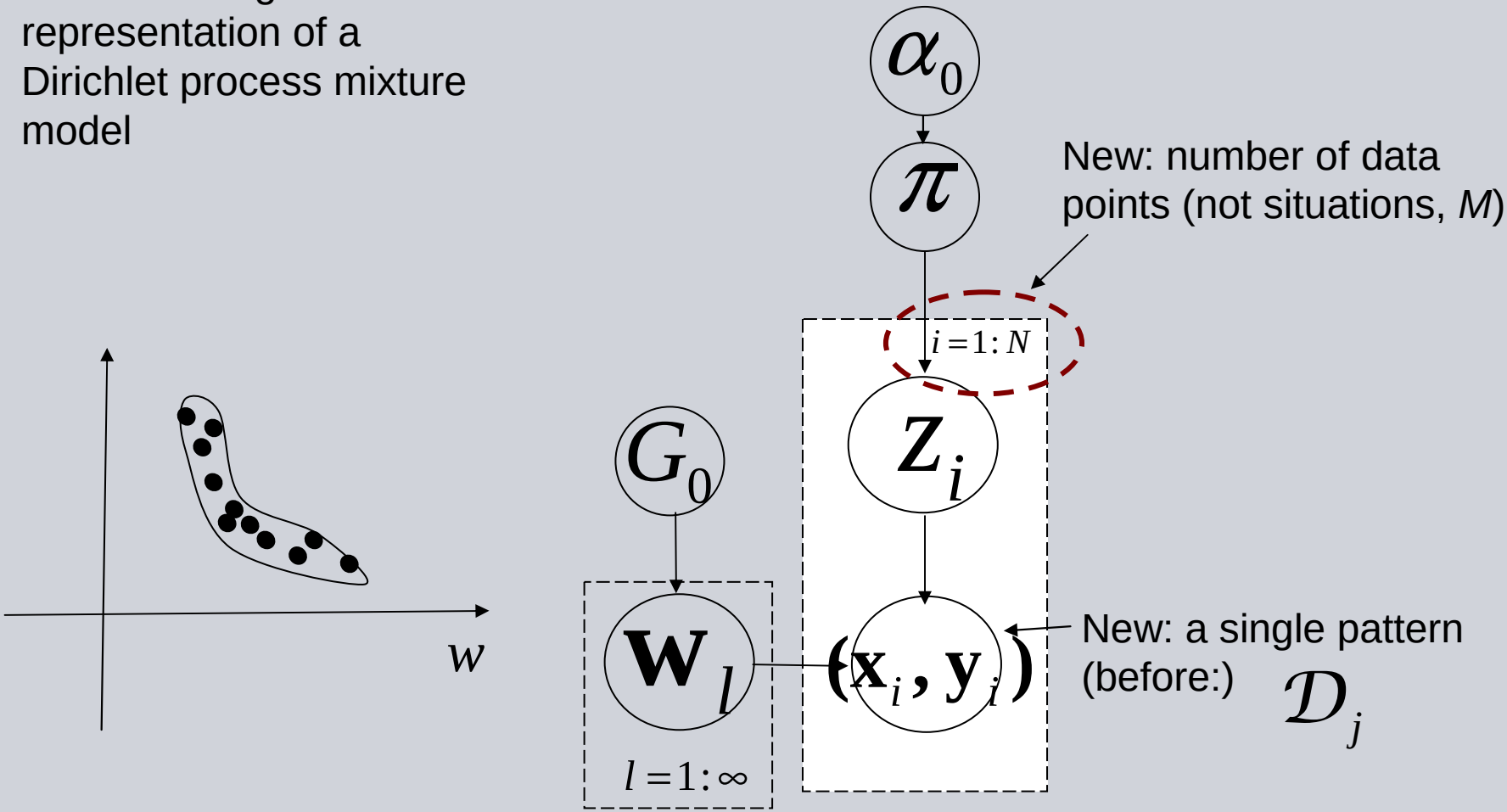
Entity-Centered Prediction

- Colors: cluster assignment (grouping of data points)
- In each cluster (grouping of rows) parameters are shared



Dirichlet Process Mixture Models for Multivariate Learning

Stick breaking representation of a Dirichlet process mixture model



Relational Mixture Models

- J. S. Breese, D. Heckerman, C. M. Kadie. *Empirical Analysis of Predictive Algorithms for Collaborative Filtering*. UAI 1998

- B. Marlin, R. Zemel, S. Roweis, M. Slaney. *Collaborative filtering and the missing at random assumption. UAI-2007*

Reduced Rank Linear Models

Reduced Rank Penalized Model

- Consider the SVD $X = UDV^T$

- The penalized least squares prediction can be written as

$$(\hat{y}_1, \ldots, \hat{y}_M) = (x_1, \ldots, x_K) W$$

- Where

$$W = V_r \operatorname{diag}_r \frac{d_k}{d_k + \lambda} U_r{}^T Y$$

- Reduced Rank: only singular vectors up to rank *r* are considered
- Disadvantage: no sharing of statistical strength

Joint Reduced Rank Penalized Model

Entity-Centered Prediction

- Consider the SVD of the data matrix $[X, Y] = \mathcal{D} = UDV^T$

- The penalized least squares prediction can again be written as

$$(\hat{y}_1, \ldots, \hat{y}_M) = (x_1, \ldots, x_K, y_1, \ldots, y_M) W$$

- Where again $$W = V_r \operatorname{diag}_r \frac{d_k}{d_k + \lambda} U_r^T Y$$

- Reduced Rank: only singular vectors up to rank *r* are considered
- Advantage: Sharing of statistical strength!

- V.Tresp, Y. Huang, M. Bundschus, A. Rettinger. *Materializing and querying learned knowledge*. IRMLeS 2009

Application to the Semantic Web

Semantic Web is based on the Resource Description Framework (RDF)

- **RDF triples** (subject,property,object) present the relationship between things
- The linking structure forms a directed, labeled graph, i.e. **RDF graph**
- An example: *Jack knows Joe*

  subject: <http://www.example.org/Jack>

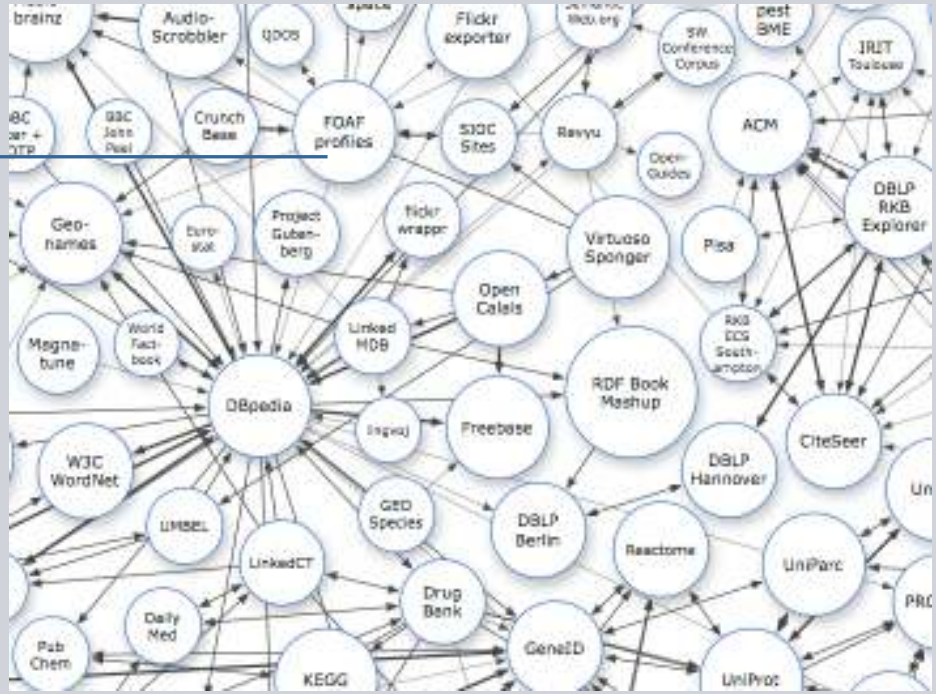
  property: <http://xmlns.com/foaf/0.1/knows>

  object: <http://www.example.org/Joe>

Linked Open Data (LOD) Project

- Each node in this cloud diagram represents a distinct data set published as Linked Data
- The arcs indicate that links exist between items in the two connected data sets.

In our experiments **FOAF** is used, which is a distributed social domain describing persons and their relationships in SW format.



Part of the Linking Open (LOD) Data Project Cloud Diagram

Characteristics and Challenges of Linked Data

Linked Data is	
Dereferenceable	It uses URIs as names for things
Interlinked	It is linked to other external data sets and can in turn be linked to from external data sets
Generic	It uses RDF to make typed statements that link arbitrary things in the world
Structured	It is arranged in an hierarchical ontology

So we can do: data integration, query answering, reasoning, and **learning**, but it is also	
Heterogeneous	Many different entity types and relationships
Extremely sparse	E.g., only a tiny subset of all possible persons are someone's friends
Incomplete	Information is missing, e.g., for privacy reasons
Large scale	The size of the Web

Relationship Prediction on the Semantic Web

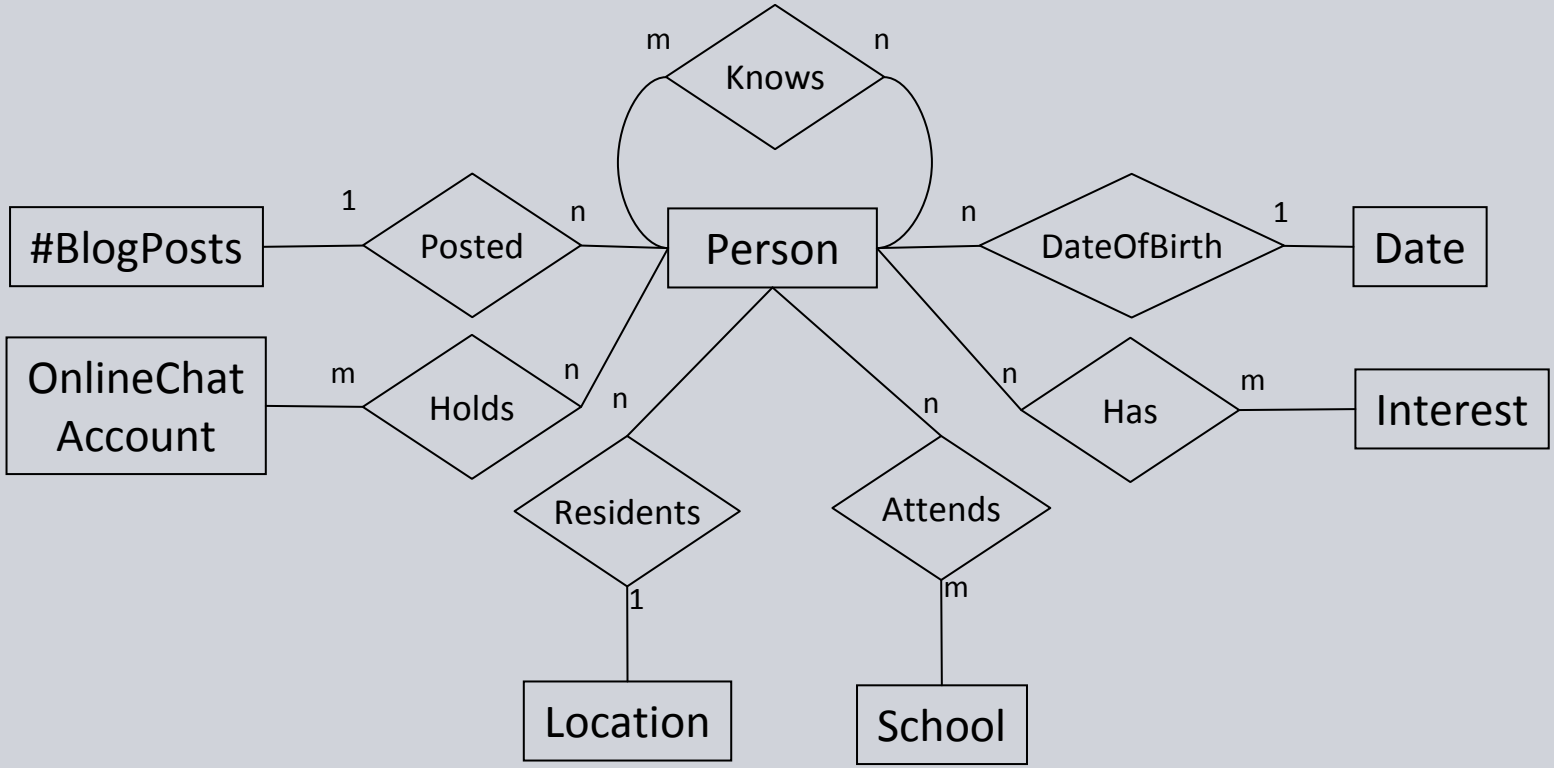
Again, the Linked Data is notoriously **incomplete** and **sparse**. A huge amount of potential inter- and intra-relationships is to uncover.

For doing this, we developed a machine **learning framework** which

- is used to **predict potential relationships** and attributes by exploiting regularities in the data using statistical relational learning algorithms
- is capable to deal with the challenge data situation on the SW, i.e. sparse data and missing information
- scales well with the size of the SW
- is as easy to use as “push-button”
- In addition, the learnt relationships and their probabilities can be easily integrated into standard query language

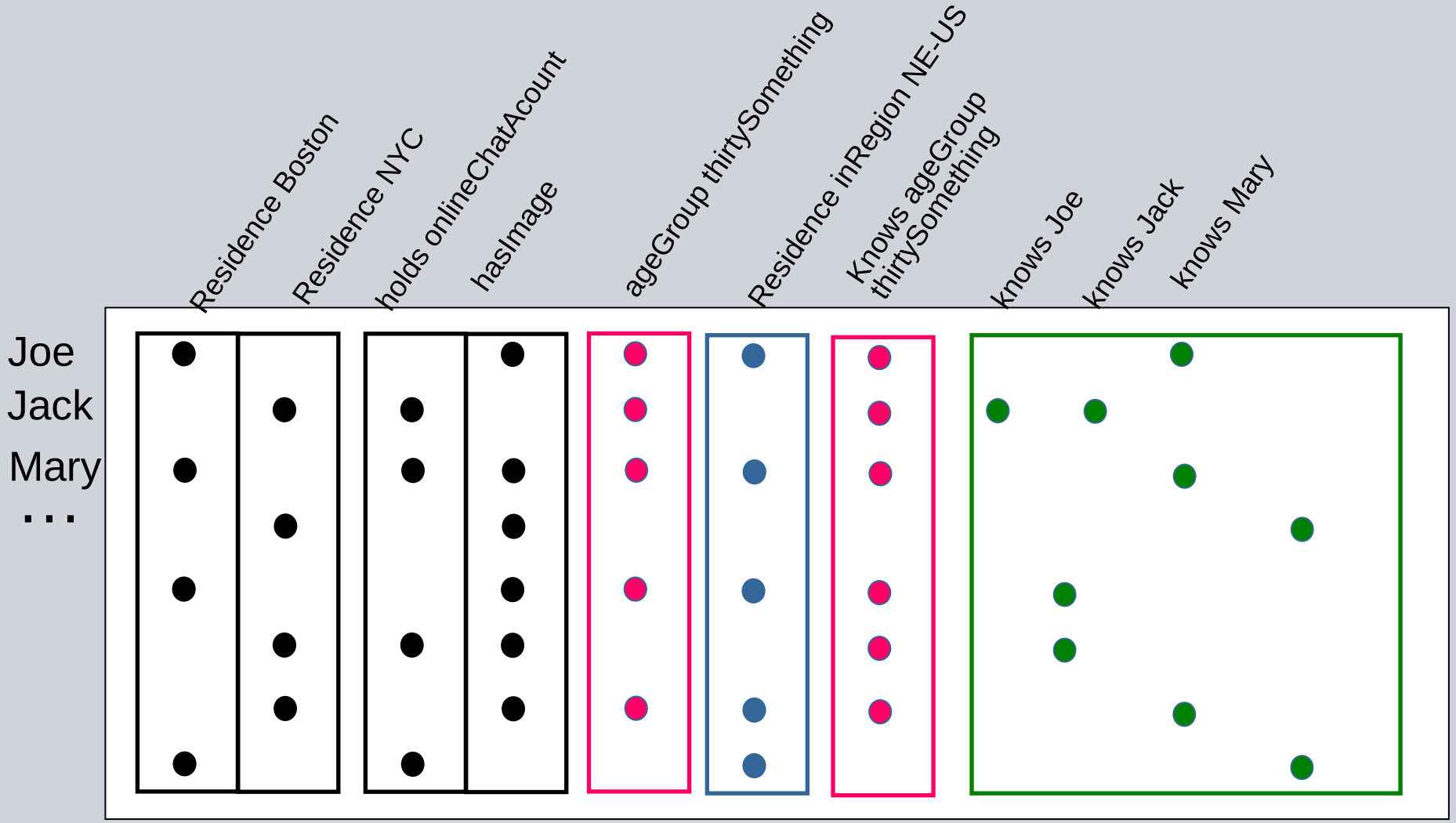
[V. Tresp et al., 2009]

Data Set – ER-Diagram



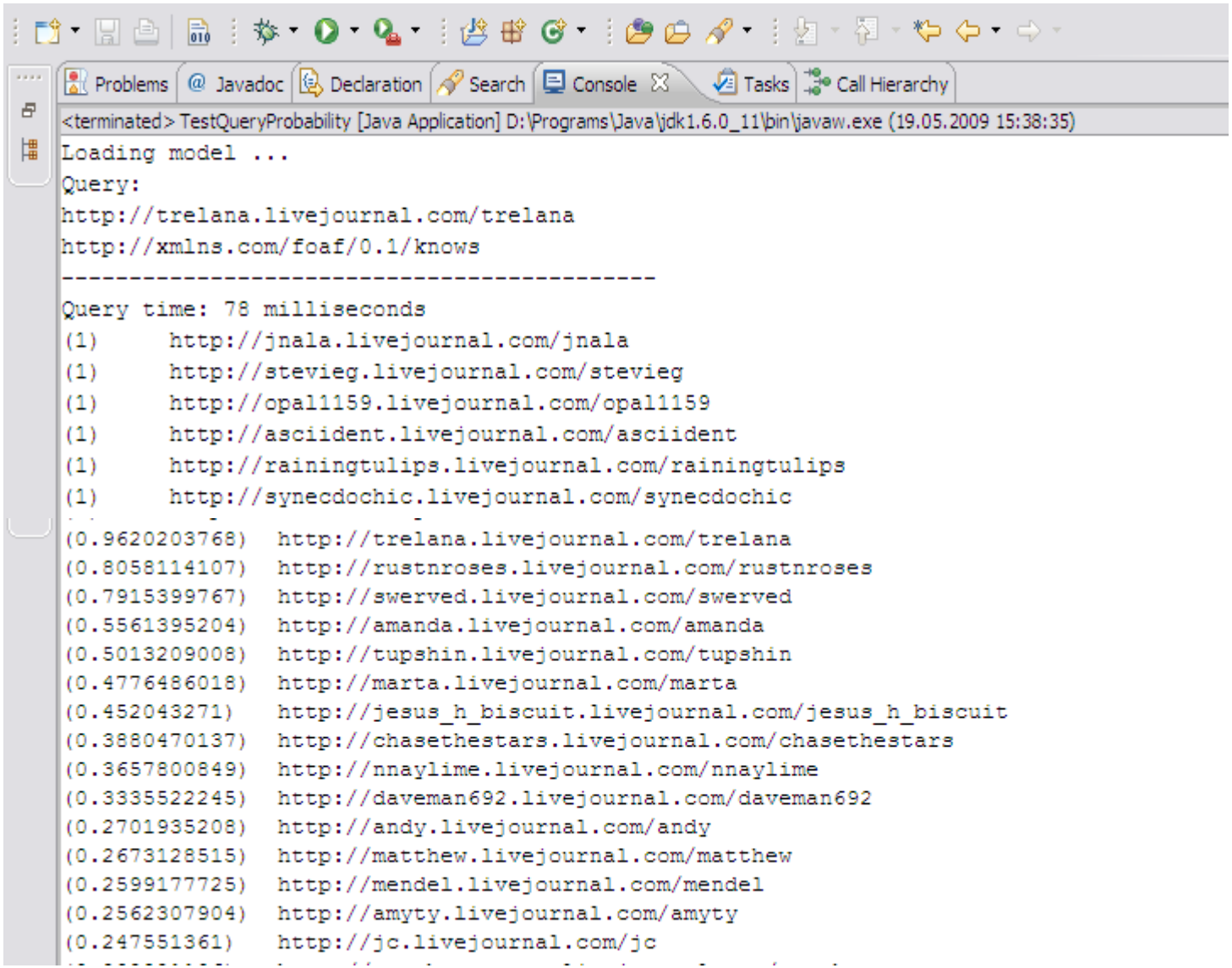
Entity-relationship diagram of LJ-FOAF domain

Data Matrix (FOAF)



Tresp et al. 2009, Huang et al. 2010

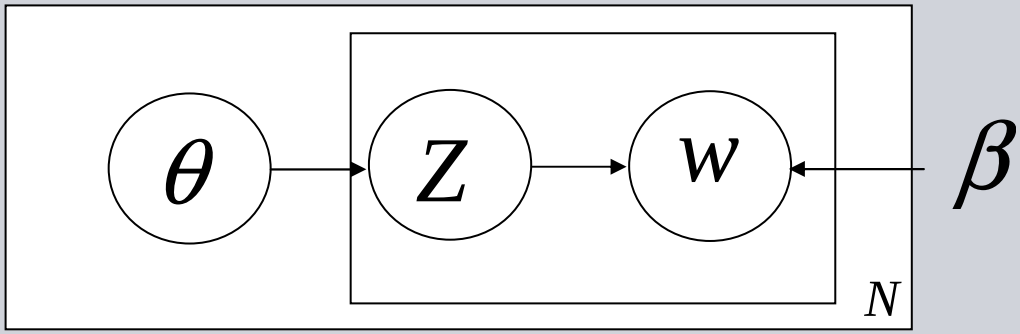
Results: Who wants to be Trelena's Friend



Topic Models

Probabilistic Latent Semantic Indexing (pLSI)

- The pPLSI has been introduced as a probabilistic generative model for document collections (Hofmann, 1999)



$$P(w = j \mid d = i) = \sum_l P(z = l \mid d = i) P(w = j \mid z = l)$$

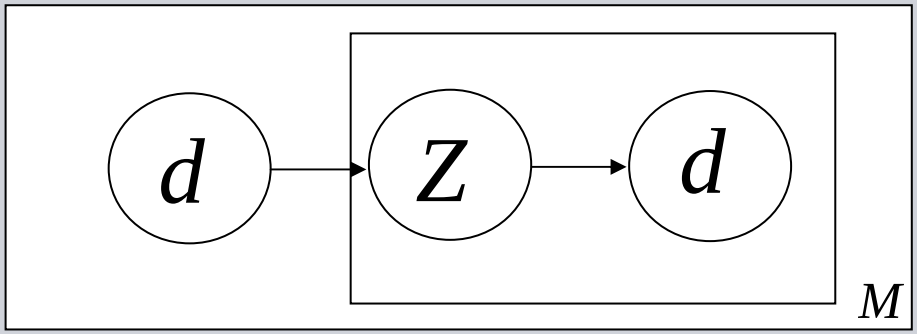
- This is the probability that word *j* is added to document *i*
- *Z* is a latent variable

$$P(Z = l \mid d = i) = \theta_{i,j}$$

$$P(w = j \mid z = l) = \beta_{l,j}$$

pHITS

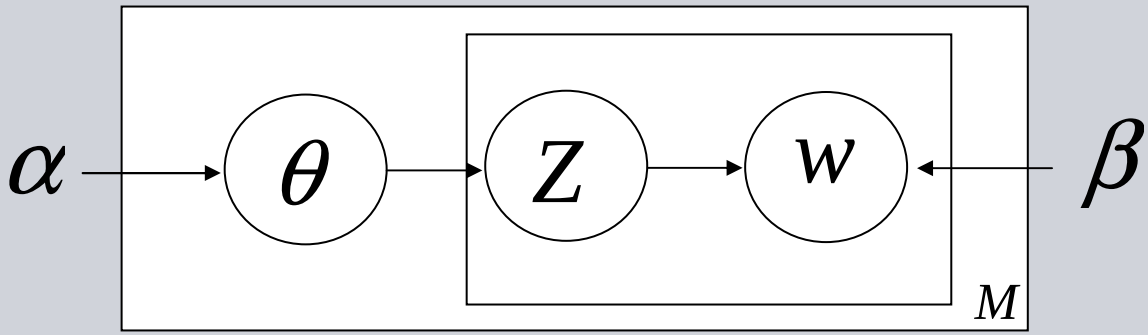
- pHITS uses the same idea to model citation links between documents (Cohn and Chang, 2000)



- Cohn and Hofmann (2001) have combined pLSI and pHITS to model citation networks by including words as document attributes (*Missing Link*)
- Predicting links from word counts and existing links
- A related *infinite* version based on Dirichlet processes has been proposed by Sinkkonen, Parkkinen, Aukia and Kaski (2008)

Latent Dirichlet Allocation (LDA)

- The LDA model is a proper Hierarchical Bayesian version of the pLSI model

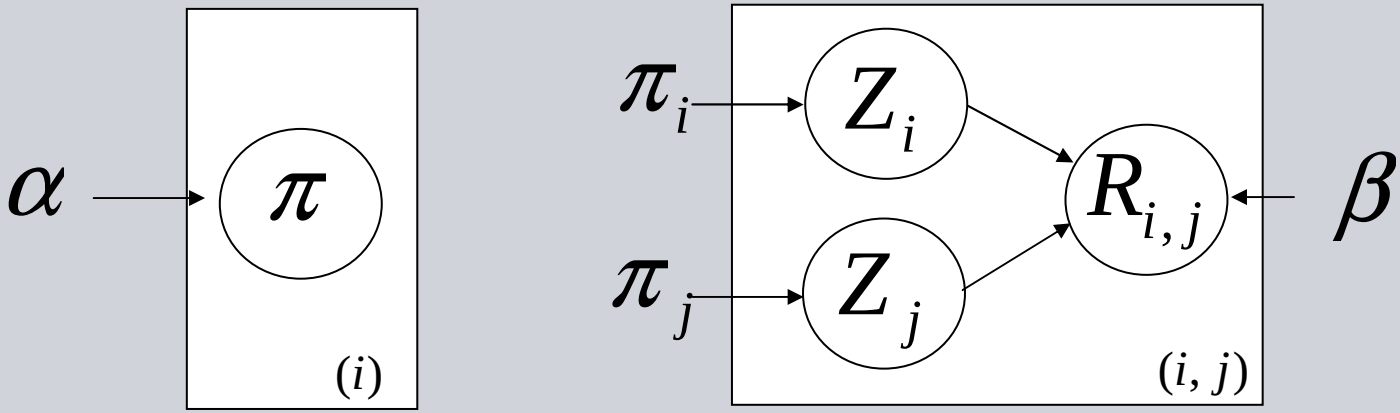


$$P(\theta_d = \theta) = \mathrm{Dir}(\theta \mid \alpha)$$

- The LDA model has been used to model network data in the *infinite* DERL model of Xu and Tresp (2005)

Mixed-membership stochastic block models (MMSB)

- The MMSB is another generalization of the pLSI/LDA models for network modeling (Airoldi et al.)
- Model: link between node *i* and node *j*



$$\pi_i \propto Dir(\alpha)$$

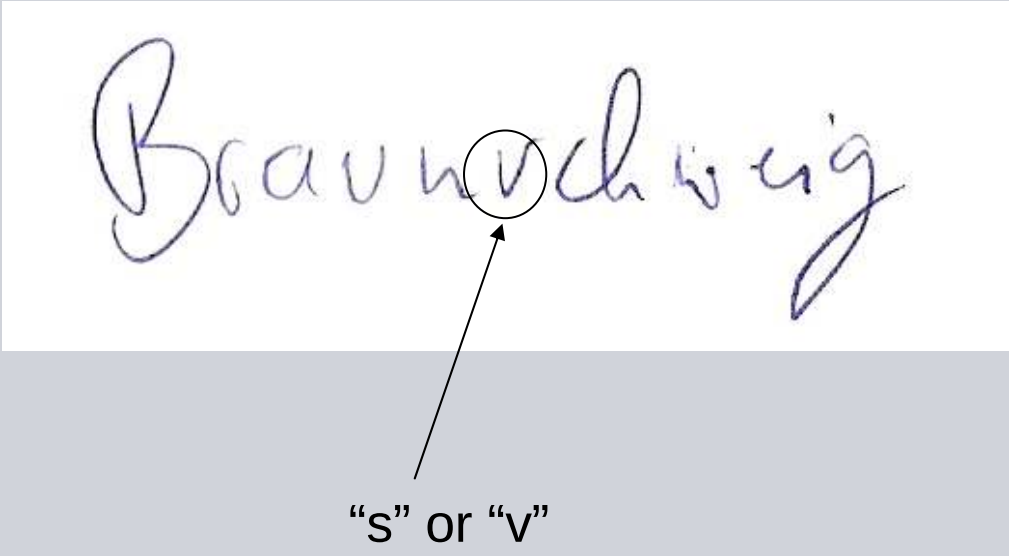
$$Z_i \propto \text{Mult}(\pi_i)$$

$$R_{i,j} \propto \text{Bernoulli}(\beta_{i,j})$$

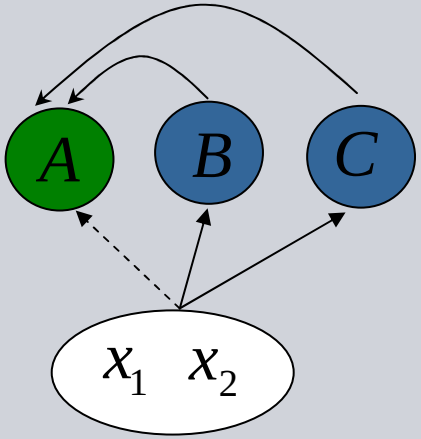
V
Multivariate Modeling: *Structured*

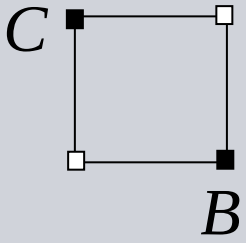
Intuition: Structured Output Prediction Problems

- Exploit correlations and constraints in the outputs
- Based on independent classification, since the “v” had a higher probability than an “s”, an OCR gives “Braunvchweig” as an answer
  - Since “sch” is very common in German, an “s” becomes more likely
  - “Braunschweig” is in the dictionary

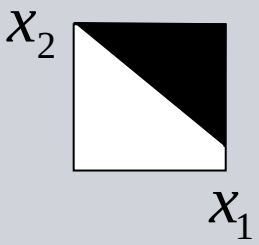


Intuitive Example

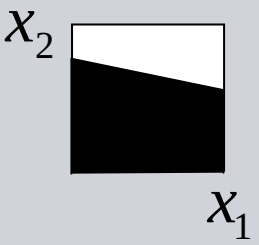




$A \mid B, C$

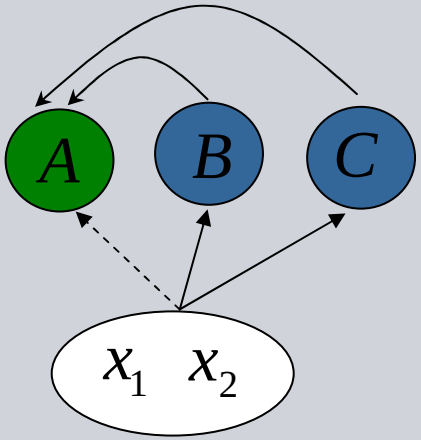


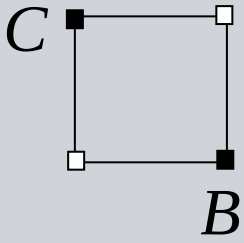
$B \mid x_1, x_2$



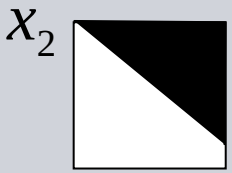
$C \mid x_1, x_2$

Intuitive Example

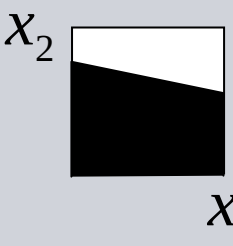




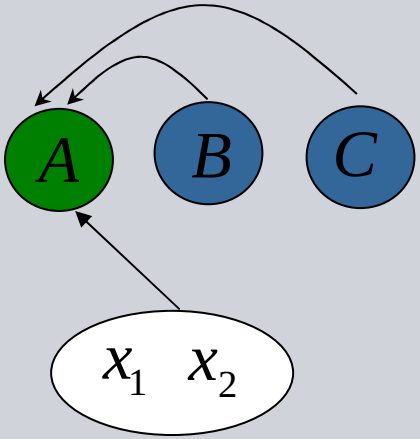
$A \mid B, C$

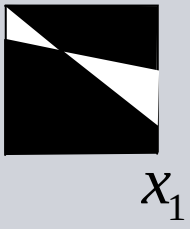


$B \mid x_1, x_2$



$C \mid x_1, x_2$





$A \mid x_1, x_2$

Examples

- Text to text-content (annotation)
- Text to parse trees
- Machine translation: English to French

- Images to image segmentation
- Images to image content
- Images to image annotation
- Images to image 3D pose
- Images to image robot arm coordinates
- From projections to reconstructed de-noised image (CT, MRI)

- DNA to DNA-segmentation
- DNA to protein structure

Important Model Class: Conditional Log-Linear Models

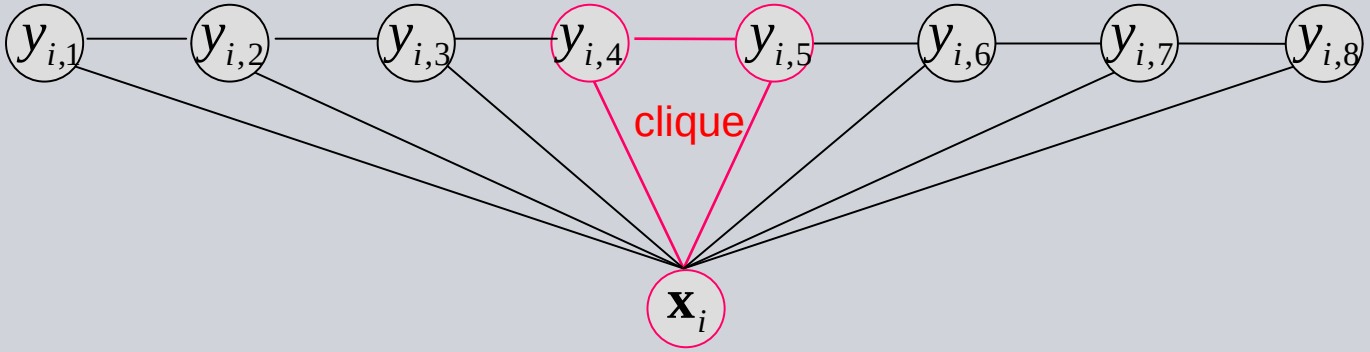
- How does one design interesting multivariate models?
- An interesting class: conditional log-linear models (a.k.a generalized linear models)
- Model design boils down to the design of interesting features

$$P(y_{i,*} \mid \mathbf{x}_i) = \frac{1}{Z(\mathbf{x}_i)} \exp \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}_i, y_{i,*})$$

$$\log P(y_{i,*} \mid \mathbf{x}_i) = -\log Z(\mathbf{x}_i) + \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}_i, y_{i,*})$$

- Feature functions (input, output): $f_k(\mathbf{x}_i, y_{i,*})$ Parameters: λ_k

Conditional Log-Linear Models from Graph Structure

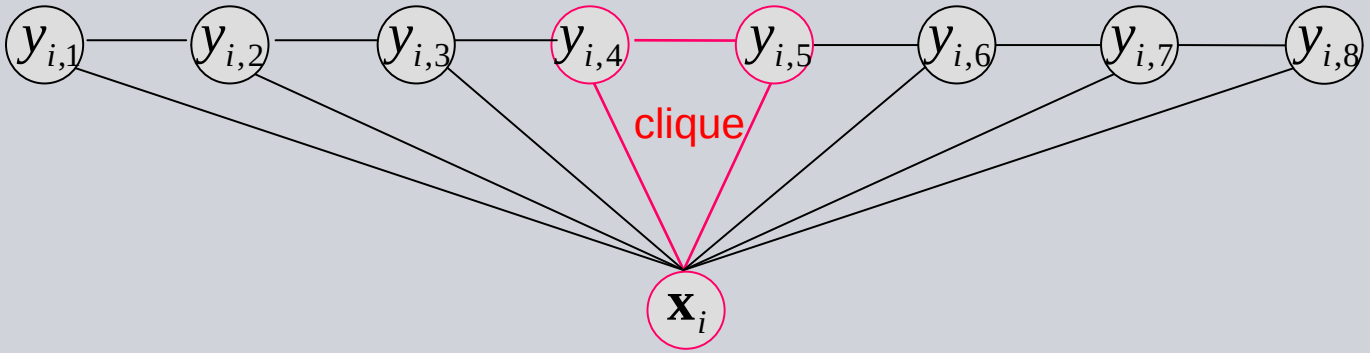


- Given a undirected graphical structure and its independence assumptions, a probability distribution factorizes in clique potentials as

$$P(\mathbf{x}_i, y_{i,*}) = \frac{1}{Z} \prod_c g_c(\mathbf{x}_i^{(c)}, y_{i,*}^{(c)})$$

$$P(y_{i,*} \mid \mathbf{x}_i) = \frac{1}{Z(\mathbf{x}_i)} \prod_c g_c(\mathbf{x}_i^{(c)}, y_{i,*}^{(c)})$$

Conditional Log-Linear Models from Graph Structure



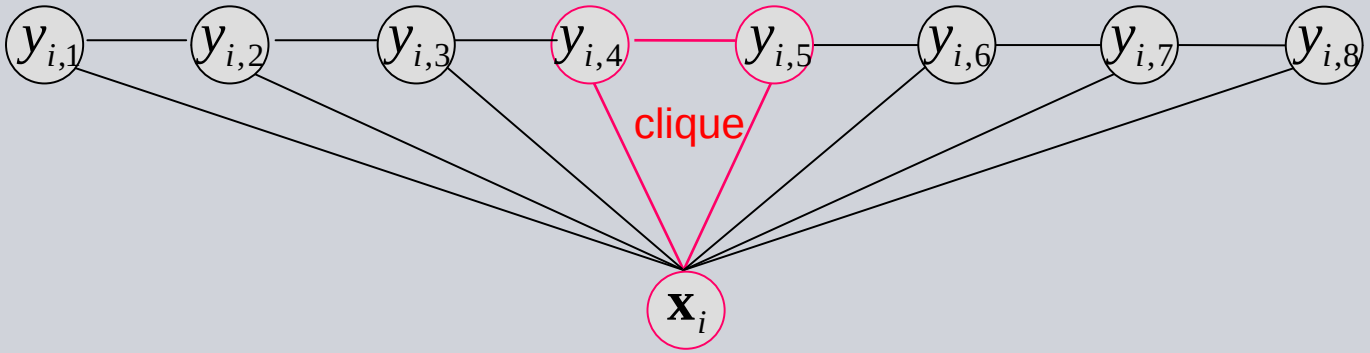
- A particular parameterization

$$g(\mathbf{x}_i^{(c)}, y_{i,*}^{(c)}) = \exp \sum_{k=1}^{K} \lambda_{c,k} f_{c,k}(\mathbf{x}_i^{(c)}, y_{i,*}^{(c)})$$

$$P(y_{i,*} \mid \mathbf{x}_i) = \frac{1}{Z(\mathbf{x}_i)} \exp \sum_c \sum_{k=1}^{K} \lambda_{c,k} f_{c,k}(\mathbf{x}_i^{(c)}, y_{i,*}^{(c)})$$

- If the features imply an independency structure, conditional log-linear models are also known as
  - Conditional Markov networks
  - Conditional (Markov) Random Fields (CRFs)

Parameters Sharing



- Often one assumes some invariance, e.g.,

$$P(y_{i,*} \mid \mathbf{x}_i) = \frac{1}{Z(\mathbf{x}_i)} \exp \sum_c \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}_i^{(c)}, y_{i,*}^{(c)})$$

- Each clique uses the same feature functions
  - Data efficiency
  - Can handle sequences with varying lengths

Training and Recall

Form the conditional version of a joint model or directly formulate a conditional model and *train the conditional model directly*

$$P(y_{i,*} \mid \mathbf{x}_i) = \frac{1}{Z(\mathbf{x}_i)} \exp \sum\nolimits_c \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}_i^{(c)}, y_{i,*}^{(c)})$$

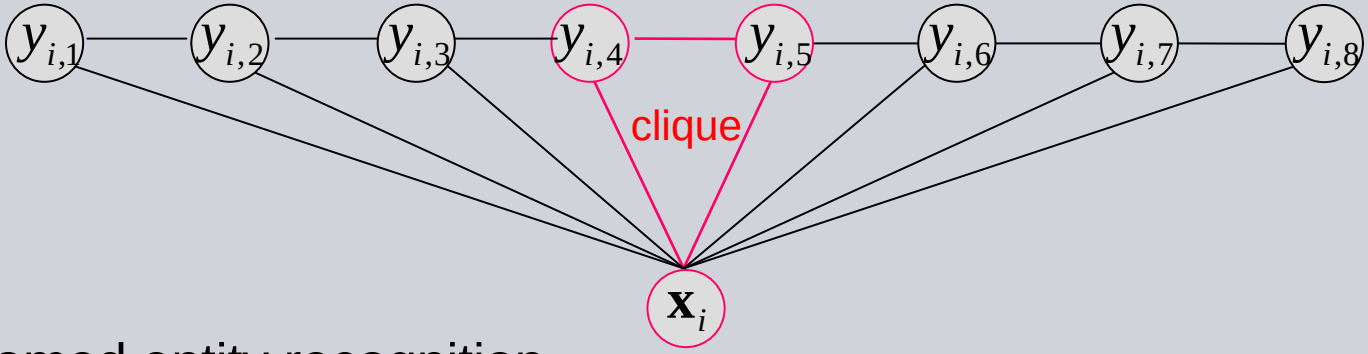
Log-likelihood:

$$l = -\sum_{i=1}^{N} \log Z(\mathbf{x}_i) + \sum\nolimits_{i=1}^{N} \sum\nolimits_c \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}_i^{(c)}, y_{i,*}^{(c)})$$

Prediction: e.g., by finding the most likely configuration:

$$\max\nolimits_{y_{i,*}} \left[-\log Z(\mathbf{x}_i) + \sum\nolimits_c \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}_i^{(c)}, y_{i,*}^{(c)}) \right]$$

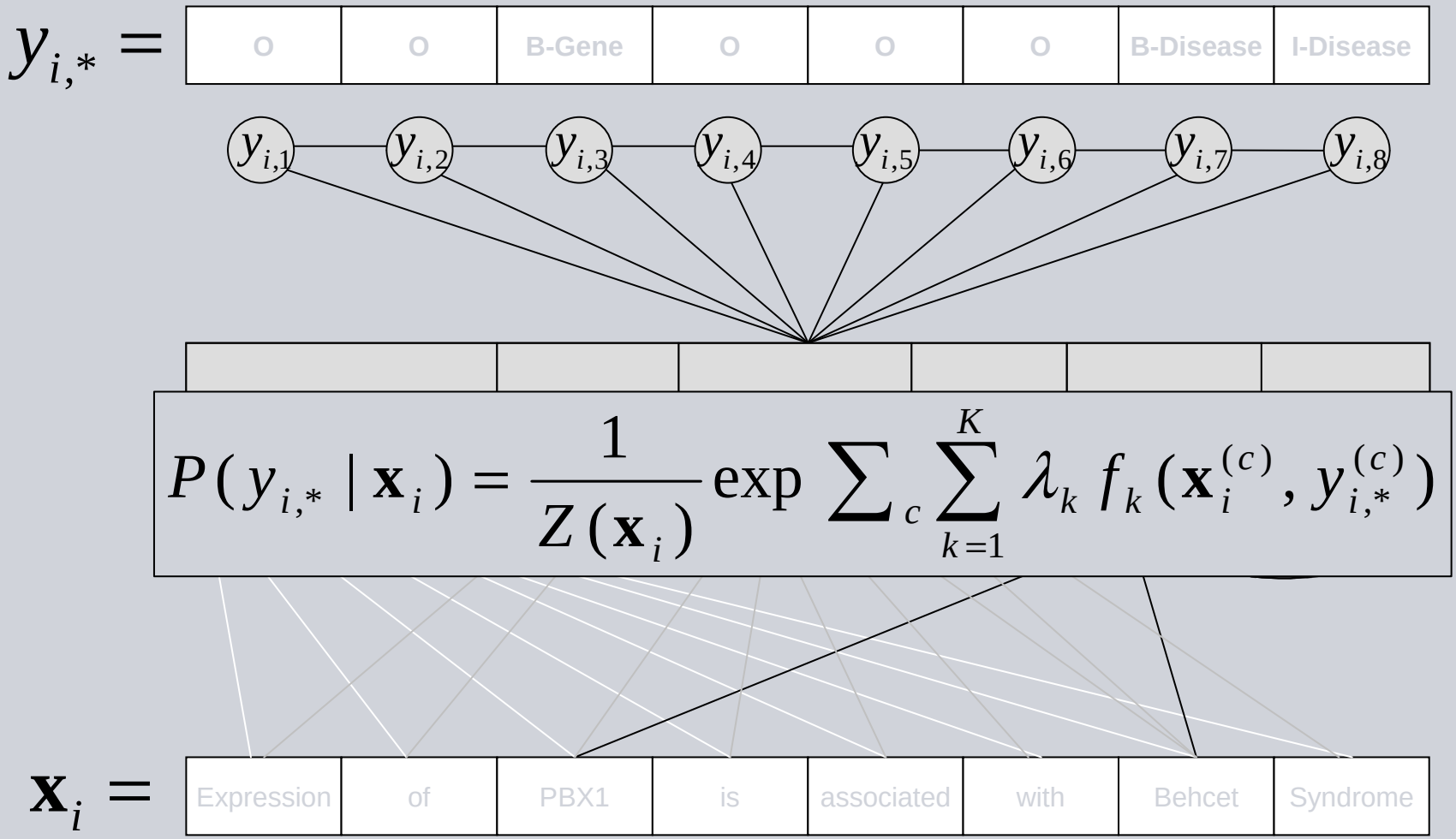
Conditional Random Fields (CRFs)



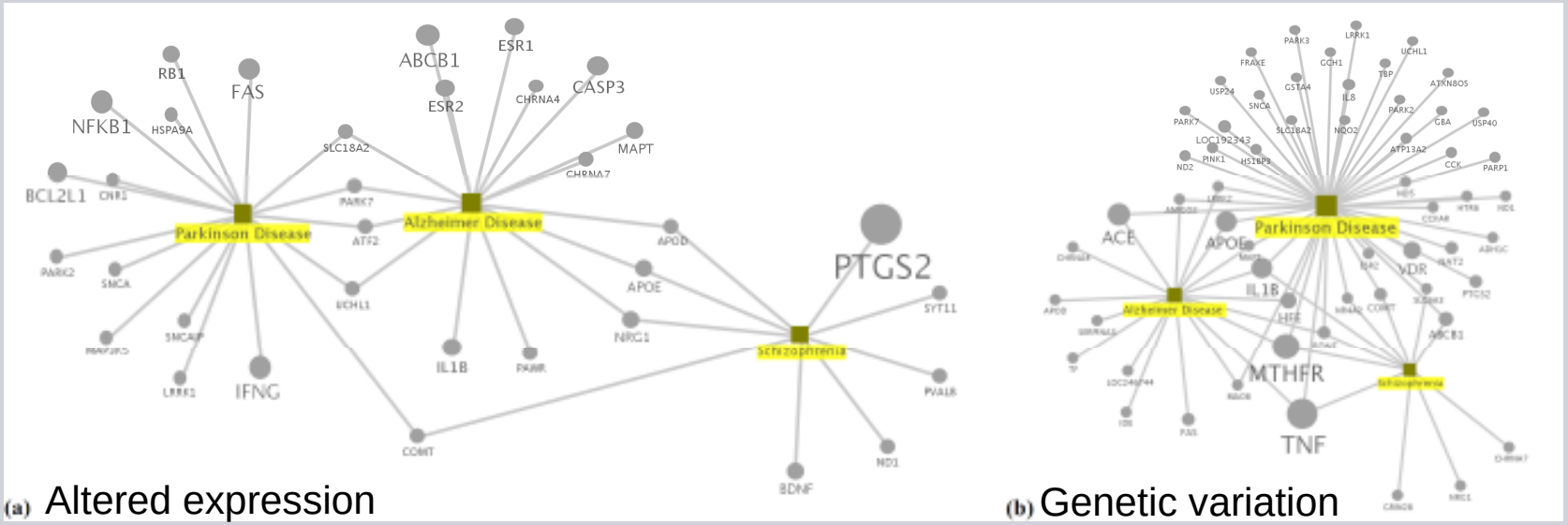
- CRFs for named entity recognition
  - Input: 50 000 and more textual features
  - Output: Sequence of maybe 10 entity classifications (with maybe 5 states for each entity: null, city, organization, person name, occupation) (*Lafferty, McCallum, Pereira, 2001*)

- Increasingly replacing Hidden Markov Models in many applications
- Interactions between outputs are explicitly modeled (since low-dimensional)
- Parameter sharing
- Prediction: iterative process
- Clear performance benefits from training a multivariate model!

Conditional Random Fields

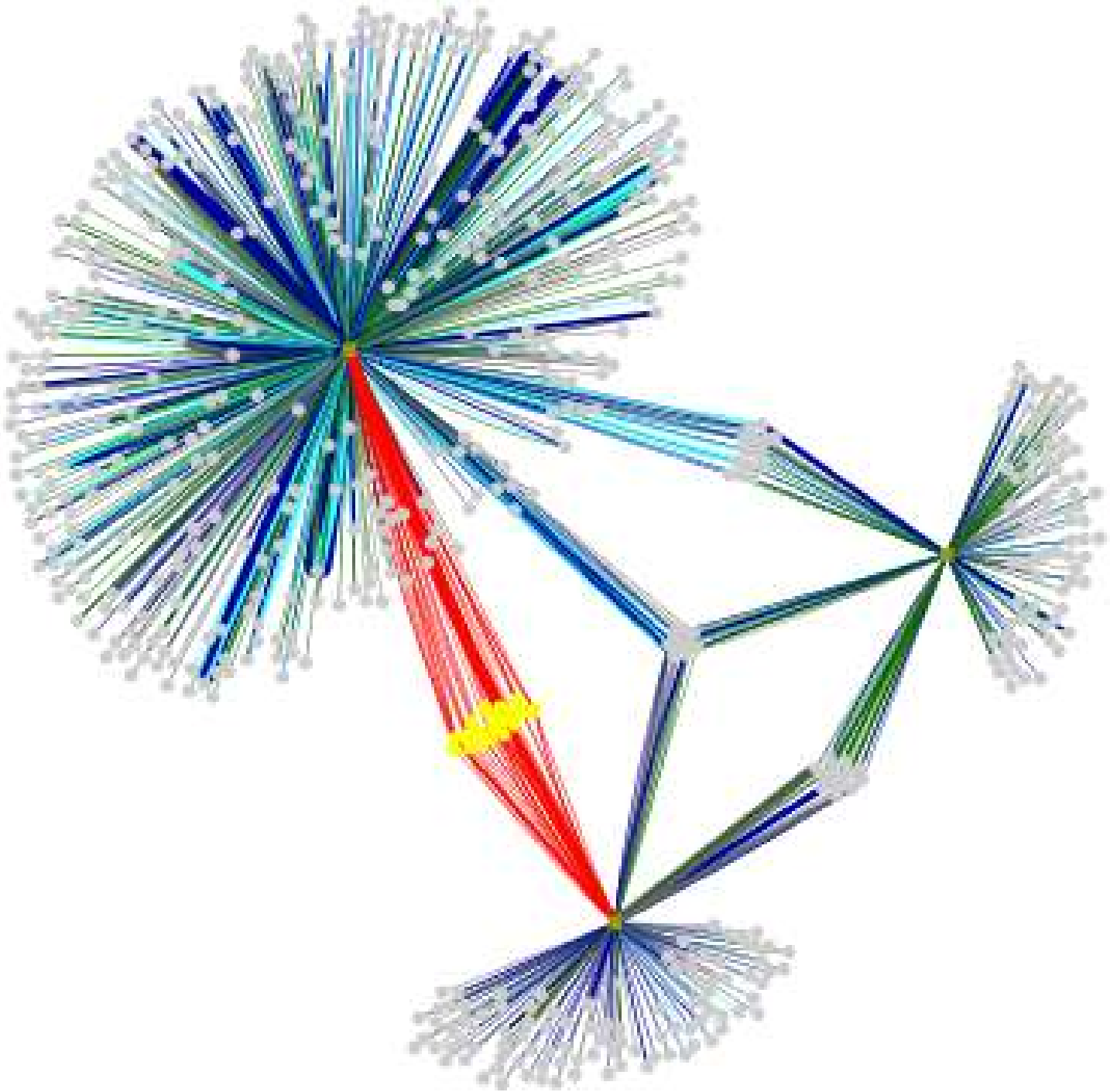


CRFs for Named Entity Recognition and Relation Extraction



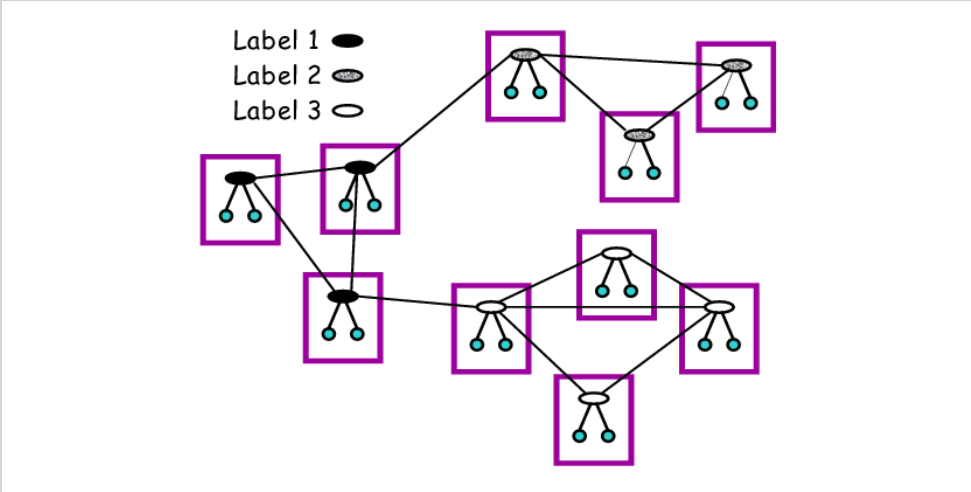
- Mining of the complete *GeneRIF Db* for gene-disease relations
- Known disease gene pairs according to *GeneCards Db* is 3.962 compared to 4.856 in our network (as of May 2009)

- Bundschus, Dejori, Stetter, Tresp and Kriegel (2008)



Social Network Analysis

- Outputs *y* correspond to attributes of entities (wealth, social status)
- Inputs are grouped and describe properties of nodes (e.g., persons)
- Often there is only one network (one data point): learning via parameter sharing
- New challenge since number of neighbors is varying: aggregation



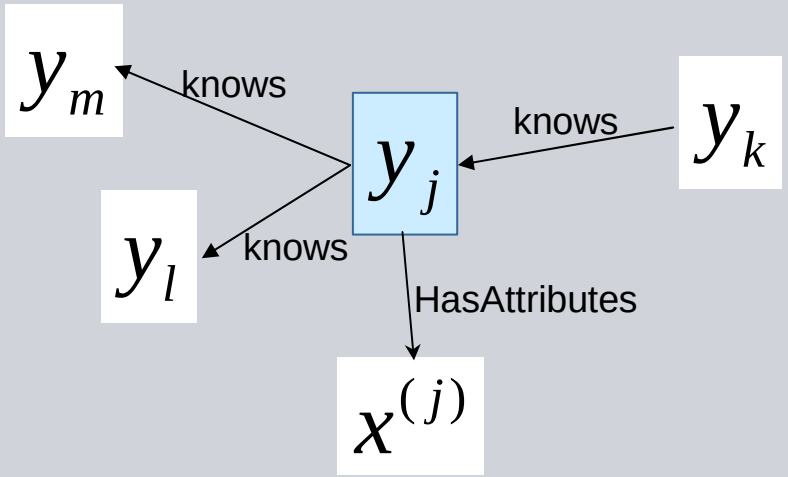
$$\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(M)}, y_1, \cdots y_M$$

- Chakrabarti, Dom and Indyk (1998)
- Neville and Jensen (2000)
- Taskar, Abbeel and Koller (2002)
- Lu and Getoor (2003)
- Neville and Jensen (2004)

Collective Classification in Social Network Analysis

- Collective classification: a class label of an entity depends on the class label of entities to which a relationship exists (“knows”) (homophily)
- Inference in the network via Gibbs sampling, relaxation labeling, iterative classification or loopy belief propagation
- Simple propagation models, e.g., Gaussian random in semi-supervised learning give very competitive results.

$$\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(M)}, y_1,\ldots y_M$$



Examples

- The wealth of person *j* depends on features of the person *j*, and on the wealth of the persons that person *j* knows (person *m* and person *l*) and the wealth of persons which know person *j* (person *k*)
- The classification of document *j* depends on the classes of cited and citing documents and on document attributes (hypertext classification)

Summary: Structured Output Prediction

- In structured output prediction constraints between outputs implied by a graphical model are exploited, which leads to a reduction in model complexity (exploitation of independencies)
- Parameter sharing leads to data efficient models
- At the same time, the dependency between input and a single output variable can be highly complex (highly complex mixture model)
- Highly active area of research (e.g., Gökhan, Hofmann, Schölkopf, Smola, Taskar, Vishwanathan, 2007, Borgwardt, Tsuda, Vishwanathan, Yan, 2008)

What We Did Not Cover

What we Did Not Cover: Max Margin Approaches

These approaches are related to CRFs but optimize a margin-based cost function

$$\forall i,\ \forall \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}_i : \left\langle w, \delta \mathbf{f}_i(\mathbf{y}) \right\rangle > 0,$$

$$\delta \mathbf{f}_i(\mathbf{y}) \equiv \mathbf{f}(\mathbf{x}_i, \mathbf{y}_i) - \mathbf{f}(\mathbf{x}_i, \mathbf{y})$$

- No normalization function
- Potentially: advantages in terms of accuracy and tunability to specific loss functions

- Taskar, Guestrin and Koller (2004)
- Tsochantaridis, Hofmann, Joachims and Altun (2004)
- Tsochantaridis, Joachims, Hofmann, and Altun (2006)
- Rousu, Saunders, Szedmak and Shawe-Taylor (2006)
- Rousu, Saunders, Szedmak and Shawe-Taylor (2007)
- Altun, Hofmann and Tsochantaridis (2007)
- Weston, Bakir, Bousquet, Mann, Noble and Schölkopf (2007)

What we Did Not Cover: Neural Networks

- The very first Neural Networks had multiple outputs (e.g., Nettalk)
- There are Neural Networks for multi-task learning and for structured prediction
  - E.g., papers by *Yann LeCun, Yoshua Bengio, Leon Bottou, Patrick Haffner*, …

- Also *ICML 2009 Workshop on Learning Feature Hierarchies. Organizers: Kai Yu, Ruslan Salakhutdinov, Yann LeCun, Geoff Hinton, Yoshua Bengio*

Conclusions

- In many situations it makes sense to predict *M* outputs than to only predict one
- This is also true in SRL where many correlated variables can be predicted

- We discussed Hierarchical Bayes
- Nonparametric Hierarchical Bayes (Gaussian processes, Dirichlet process mixture models) provide flexible model classes
- We discussed applications to SRL

- Multivariate modeling exploits dependencies between inputs and outputs but also dependencies in between outputs
- Often all outputs are sensitive to a parameter and learning is data efficient
- They are most effective for the prediction of relationships

- Structures Output Prediction exploits both prior knowledge about the structural independencies between outputs and parameter sharing
- An important model class concerns conditional random fields (CRFs)
- Structures Output Prediction has been shown to be effective in Social Network modeling

Acknowledgements

Collaborators:

- Anton Schwaighofer, Microsoft Research
- Kai Yu, NEC Research
- Shipeng Yu, Siemens Healthcare
- Zhao Xu, FhG IAIS
- Markus Bundschus, Roche
- Achim Rettinger, KIT
- Christoph Lippert, Max Planck Institute for Biological Cybernetics
- Yi Huang, Siemens CT

References: General

- L. Fahrmeier, G. Tutz. Multivariate Statistical Modeling Based on Generalized Linear Models. Springer, 1994.
- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin. Bayesian Data Analysis, 2nd edition. Chapman, 2003
- T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning. Springer 2001

References: Hierarchical Bayes / Multitask (1)

- A. Argyriou, T. Evgeniou, M. Pontil. Multi-task Feature Learning. NIPS 2006
- Z. Barutcuoglu, R. Schapire, O. Troyanskaya. Hierarchical multi-label prediction of gene function. Bioinformatics 22, 2006
- D. M. Blei, A. Y. Ng, M. I. Jordan. Latent Dirichlet Allocation. JMLR 2003
- E.V. Bonilla, K.M.V. Chai, C.K.I. Williams. Multi-task Gaussian process prediction. NIPS-07
- R. Caruana. Learning many related tasks at the same time with backpropagation. NIPS 1995
- T. Evgeniou, C.A. Micchelli, M. Pontil. Learning multiple tasks with kernel methods. JMLR 2006
- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin. Bayesian Data Analysis, 2nd edition. Chapman 2003
- N. D. Lawrence, J. C. Platt. Learning to learn with the informative vector machine. ICML 2004
- B. Krishnapuram, S. Yu, O. Yakhnenko, R. B. Rao, L. Carin. NIPS Workshop: Cost Sensitive Learning. 2008
- C. Lippert, S. Weber, Y. Huang, V. Tresp, M. Schubert, H.-P. Kriegel. Relation-Prediction in Multi-Relational Domains using Matrix-Factorization. NIPS Workshop: Structured Input - Structured Output 2008
- R. Raina, A. Y. Ng, D. Koller. Constructing informative priors using transfer learning. ICML 2006
- J. Rousu, C. Saunders, S. Szedmak, J. Shawe-Taylor. Kernel-based Learning of Hierarchical Multilabel Classification Models. JLMR 2006
- A. Schwaighofer, V. Tresp, K. Yu. Learning gaussian process kernels via hierarchical bayes. NIPS 2004
- S. Thrun. Is Learning the n-th Thing Any Easier Than Learning the First? NIPS 1996

References: Hierarchical Bayes / Multitask (2)

- Z. Xu, K. Kersting, and V. Tresp. Multi-relational learning with gaussian processes. In Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09), July 2009. Collaborative Filtering. SIGIR 2009
- K. Yu, V. Tresp, A. Schwaighofer. Learning gaussian processes from multiple tasks. ICML 2005
- K. Yu, W. Chu, S. Yu, V. Tresp, Z. Xu. Stochastic relational models for discriminative link prediction. NIPS 2006.
- S. Yu, K. Yu, V. Tresp. Collaborative ordinal regression. ICML 2006
- K. Yu, S. Zhu, J. Lafferty, Y. Gong. Large-scale Collaborative Prediction
- Using a Nonparametric Random Effects ModelJ. Zhang. Sparsity Models for Multi-task Learning. NIPS Workshop on Inductive Transfer 2005
- J. Zhang, Z. Ghahramani, Y. Yang. Learning Multiple Related Tasks using Latent Independent Component Analysis. NIPS 2005
- S. Zhu, K. Yu, Y. Gong. Stochastic Relational Models for Large-scale Dyadic Data using MCMC. NIPS 2008

References: Dirichlet Process Mixture Models

- M. I. Jordan. Dirichlet Processes, Chinese Restaurant Processes and All. Tutorial at NIPS 2005
- C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, N. Ueda. Learning Systems of Concepts with an Infinite Relational Model. National Conference on Artificial Intelligence 2006
- S. Reckow, V. Tresp. Integrating ontological prior knowledge into relational learning. NIPS Workshop: Structured Input - Structured Output 2008
- Achim Rettinger, Matthias Nickles, and Volker Tresp. Statistical relational learning with formal ontologies. ECML PKDD, 2009.
- V. Tresp, K. Yu. An introduction to nonparametric hierarchical Bayesian modeling with a focus on multi-agent learning. Hamilton Summer School on Switching and Learning in Feedback Systems. 2004
- V. Tresp. Dirichlet processes and nonparametric bayesian modelling. Tutorial at the Machine Learning Summer School 2006
- Y. Xue, X. Liao, L. Carin, B. Krishnapuram. Multi-task learning for classification with Dirichlet process priors. JMLR 2007
- Z. Xu, V. Tresp, K. Yu, H.-P. Kriegel. Infinite hidden relational models. UAI 2006
- Z. Xu, V. Tresp, S. Yu, K. Yu. Nonparametric relational learning for social network analysis. SNA-KDD 2008
- Z. Xu, V., A. Rettinger, K. Kersting. Social network mining with nonparametric relational models. In Advances in Social Network Mining and Analysis, Springer, 2009.
- K. Yu, W.-Y. Ma, V. Tresp, Z. Xu, X. He, H. J. Zhang, H.-P. Kriegel. Knowing a tree from the forest: Art image retrieval using a society of profiles. ACM Multimedia 2003
- K. Yu, V. Tresp, S. Yu. A nonparametric hierarchical bayesian framework for information filtering. SIGIR 2004
- K. Yu, S. Yu, V. Tresp. Dirichlet enhanced latent semantic analysis. AISTAT 2005.
- K. Yu, A. Schwaighofer, V. Tresp, W.-Y. Ma, H. J.Zhang. Collaborative ensemble learning: Combining collaborative and content-based information filtering via hierarchical bayes. UAI 2003

References: Projection Methods

- R. K. Ando, T. Zhang. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. JMLR 2005
- D. Hardoon, G. Leen, S. Kaski, J. Shawe-Taylor. NIPS Workshop: Learning from Multiple Sources. 2008
- J. Shawe-Taylor, N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press 2004
- K. Yu, S. Yu, V. Tresp. Multi-output regularized projection. IEEE CVPR 2005
- K. Yu, S. Yu, V. Tresp. Multi-label informed latent semantic indexing. SIGIR 2005
- S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, M. Wu. Supervised probabilistic principal component analysis. KDD 2006
- S. Yu, K. Yu, V. Tresp, H.-P. Kriegel. Multi-output regularized feature projection. IEEE TKDE 2006

References: Multivariate Models and Structured Outputs (1)

- E.M- Airoldi, D.M. Blei, S.E. Fienberg, E.P. Xing. Mixed-membership stochastic blockmodels. Journal of Machine Learning Research, 2008.
- Y. Altun, T. Hofmann, I. Tsochantaridis: Support Vector Machine Learning for Interdependent and Structured Output Spaces. Chapter 5 in Predicting Structured Data, MIT Press 2007
- K. Borgwardt, K. Tsuda, S. V. N. Vishwanathan, X. Yan. NIPS Workshop: Structured Input - Structured Output 2008
- J. S. Breese, D. Heckerman, C. M. Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. UAI 1998
- M. Bundschus, M. Dejori, M. Stetter, V. Tresp, H.-P. Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics 2008
- M. Bundschus, M. Dejori, S. Yu, V. Tresp, H.-P. Kriegel. Statistical modeling of medical indexing processes for biomedical knowledge information discovery from text. BIOKDD 2008
- S. Chakrabarti, S. Dom, P. Indyk. Enhanced hypertext categorization using hyperlinks. SIGMOD 1998
- D. A. Cohn, T. Hofmann . The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity.NIPS 2000.
- D. A. Cohn, H. Chang. Learning to probabilistically identify authoritative documents. ICML 2000.
- L. Getoor, N. Friedman, D. Koller, B. Taskar. Learning Probabilistic Models of Link Structure. JMLR 2002
- H. B. Gökhan, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar S. V. N. Vishwanathan (editors). Predicting Structured Data. MIT press 2007
- T. Hofmann, Probabilistic Latent Semantic Indexing, SIGIR, 1999

References: Multivariate Models and Structured Outputs (2)

- J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ICML 2001
- Q. Lu, L. Getoor. Link-based classification. ICML 2003
- B. Marlin, R. Zemel, S- Roweis, M. Slaney. Collaborative filtering and the missing at random assumption. UAI-2007.
- J. Neville, D. Jensen. Dependency networks for relational data. ICDM 2004
- J. Neville, D. Jensen. Iterative classification in relational data. AAAI 2000
- J. Rousu, C. Saunders, S. Szedmak, J. Shawe-Taylor. Efficient algorithms for Max-Margin Structured Classification. Chapter 6 in Predicting Structured Data, MIT Press 2007
- J. Sinkkonen, J. Parkkinen, J. Aukia, S. Kaski. A simple infinite topic mixture for rich graphs and relational data. NIPS Workshop: Applications for Topic Models: Text and Beyond, 2009
- C. Sutton, A. McCallum: An Introduction to Conditional Random Fields for Relational Learning. In: Lise Getoor, Ben Taskar. Introduction to Statistical Relational Learning. MIT Press, 2006.
- B. Taskar , M.-F. Wong , P. Abbeel, D. Koller. Link Prediction in Relational Data. NIPS 2003
- B. Taskar, P. Abbeel, D. Koller. Discriminative probabilistic models for relational data. UAI 2002
- B. Taskar, C. Guestrin, D. Koller. Max-Margin Markov Networks. NIPS 2004
- B. Taskar, V. Chatalbashev, D. Koller, C. Guestrin. Learning Structured Prediction Models: A Large Margin Approach. Tutorial at ICML 2005

References: Multivariate Models and Structured Outputs (3)

- V. Tresp, Y. Huang, M. Bundschus, A. Rettinger. Materializing and querying learned knowledge. In Proceedings of the First ESWC Workshop on Inductive Reasoning and Machine Learning on the Semantic Web (IRMLeS 2009), 2009
- V. Tresp. Mixtures of gaussian processes. NIPS 2001
- I. Tsochantaridis, T. Hofmann, T. Joachims, Y. Altun. Support vector machine learning for interdependent and structured output spaces. ICML 2004
- I. Tsochantaridis, T. Joachims, T. Hofmann, T. Y. Altun. Large margin methods for structured and interdependent output variables. JMLR 2006
- J. Weston, G. Bakir, O. Bousquet, T. Mann, W.S. Noble, B. Schölkopf. Joint Kernel Maps. Chapter 4 in Predicting Structured Data, MIT Press 2007
- Z. Xu, V. Tresp, K. Yu, S. Yu, H.-P. Kriegel. Dirichlet enhanced relational learning. ICML 2005
- X. Zhu. Semi-supervised learning literature survey. TR University of Wisconsin 2005