Convex Point Estimation using Undirected Bayesian Transfer Hierarchies

Gal Elidan Ben Packer

Geremy Heitz Daphne Koller

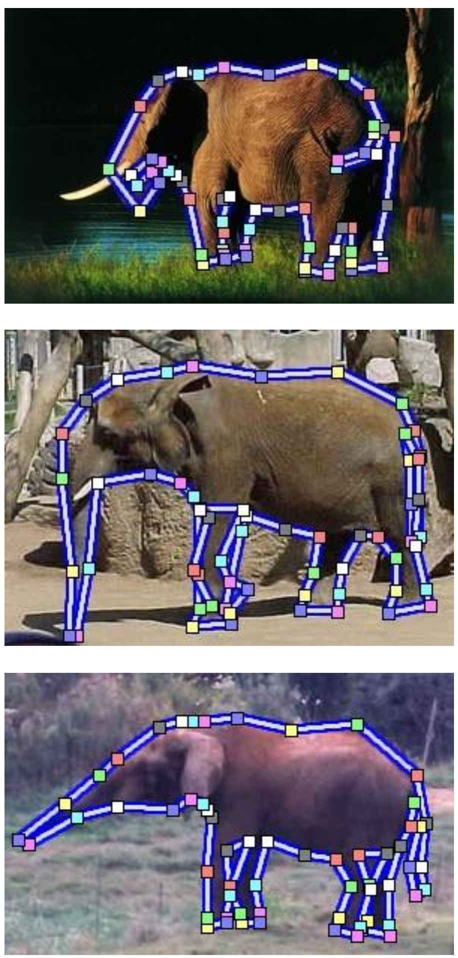
STANFORD COMPUTER SCIENCE

Stanford AI Lab

Motivation

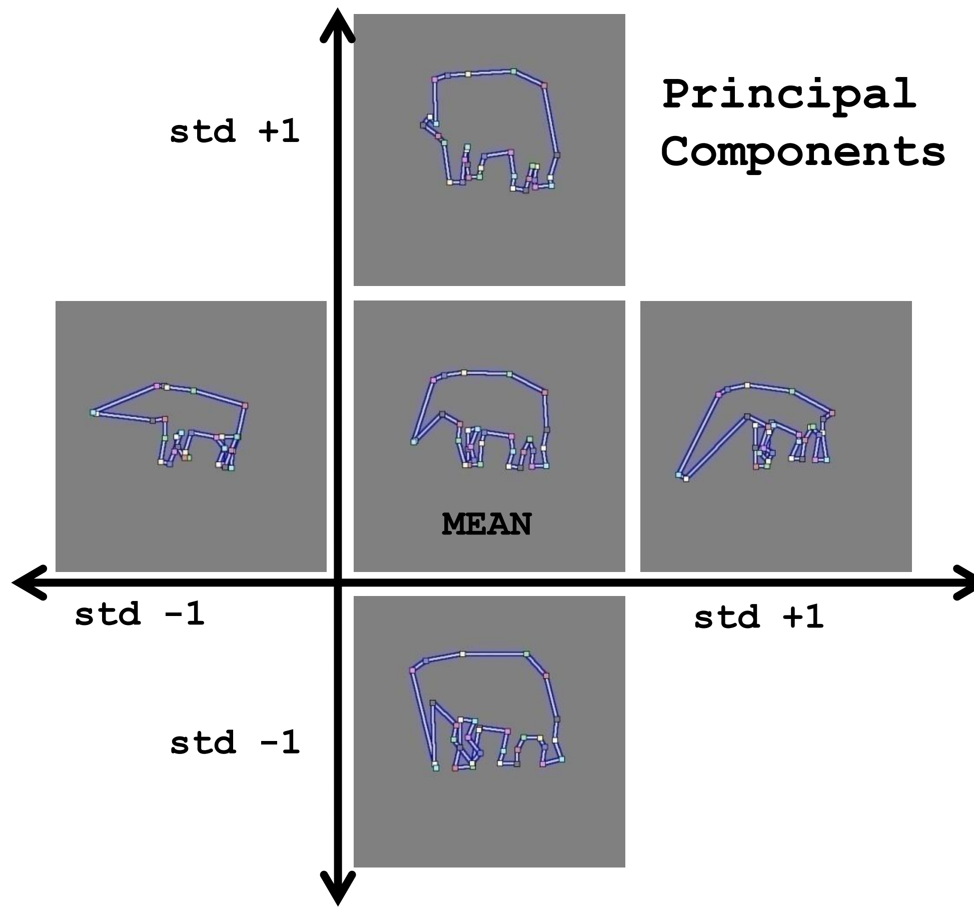
Task:

Shape modeling



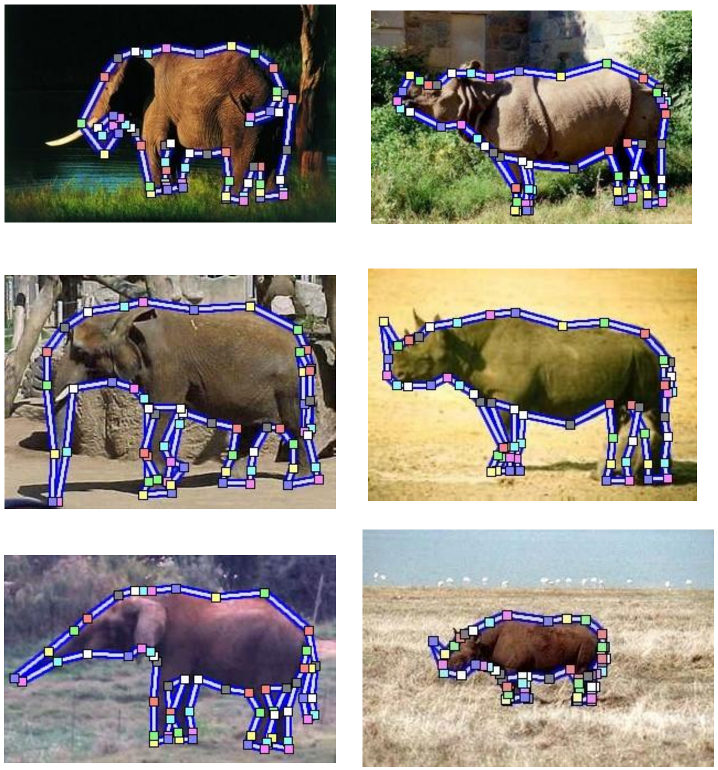
Problem:

With few instances, learned models aren't robust

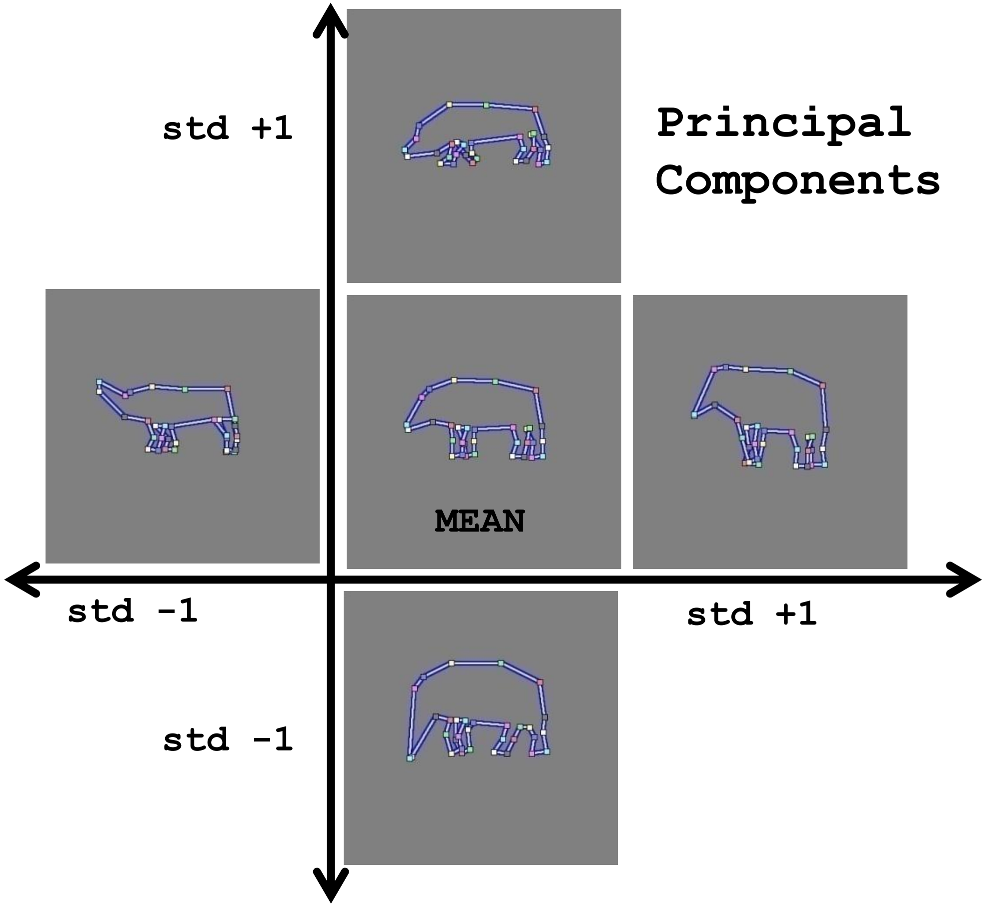


Transfer Learning

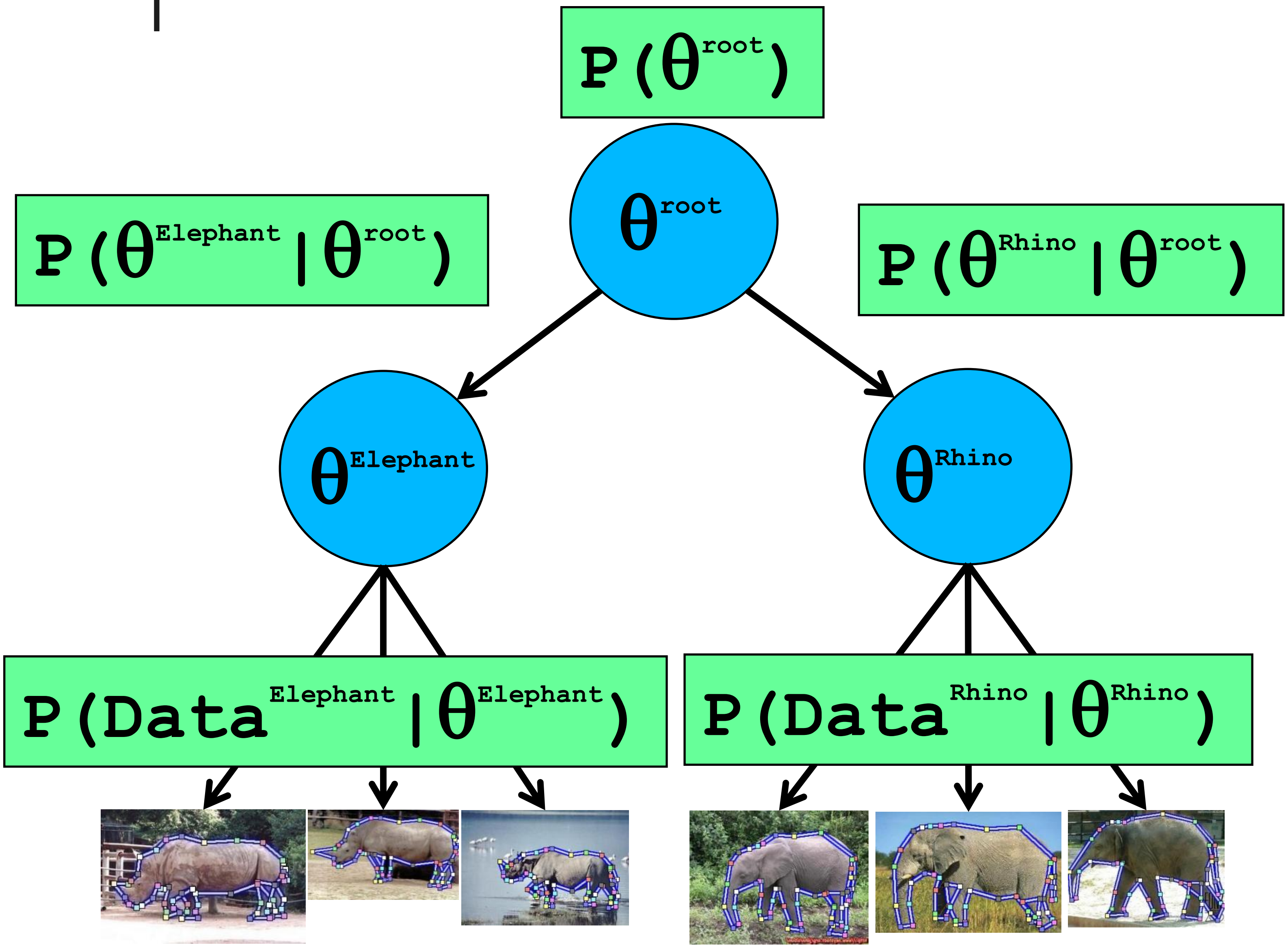
Can we use rhinos to help elephants?



Shape is stabilized, but doesn't look like an elephant



Hierarchical Bayes

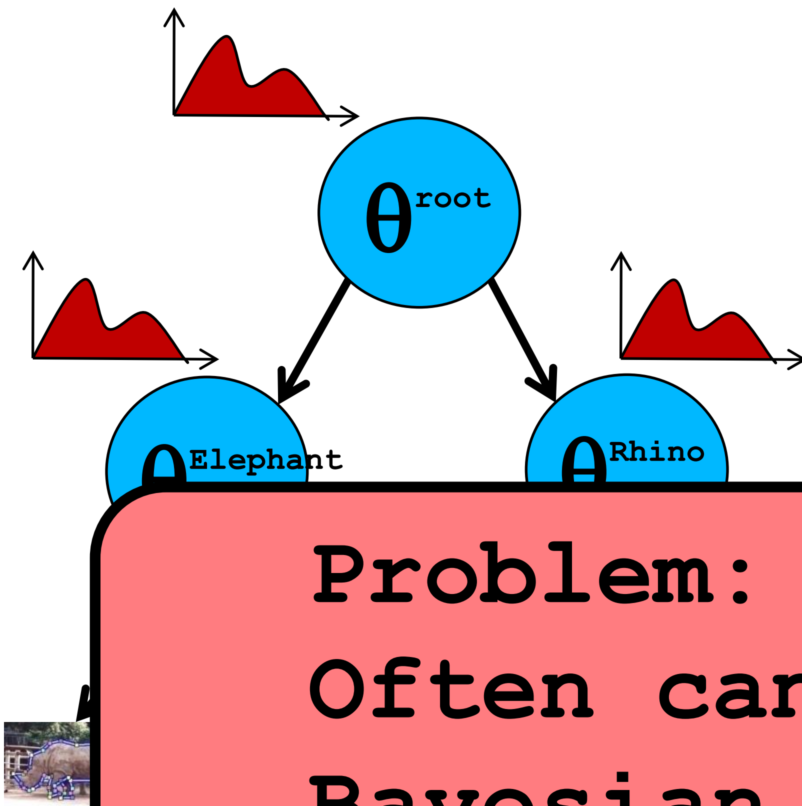


Goals

- Transfer between related classes ✓
- Range of settings, tasks ✓
- Probabilistic motivation ✓
- Multilevel, complex hierarchies
- Simple, efficient computation
- Automatically learn what to transfer

Hierarchical Bayes

$$P(\mathcal{D},\Theta) = \prod_{c\in\mathcal{L}} P(\mathcal{D}^c \mid \Theta^c) \times \prod_{c\in\mathcal{C}} P(\Theta^c \mid \Theta^{par(c)})$$

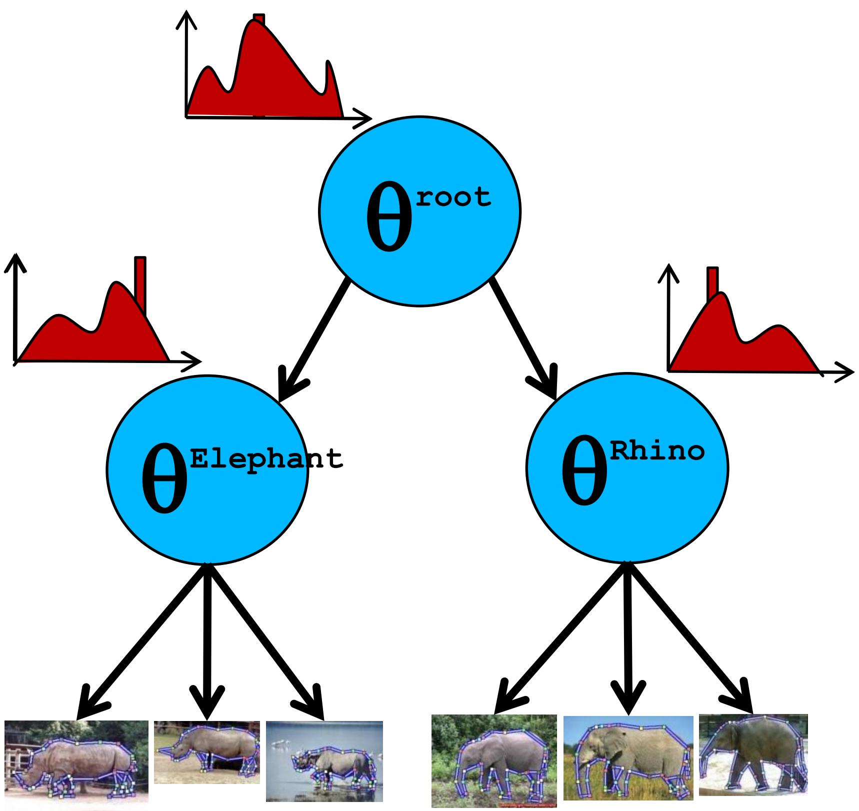


- Compute full posterior $P(\Theta|\mathcal{D})$
- $P(\Theta^c|\Theta^{root})$ must be conjugate to $P(\mathcal{D}|\Theta^c)$

Problem:
Often can't perform full Bayesian computations

Approx.: Point estimation

Best parameters are good enough; don't need full distribution



- Empirical Bayes
- Point estimation

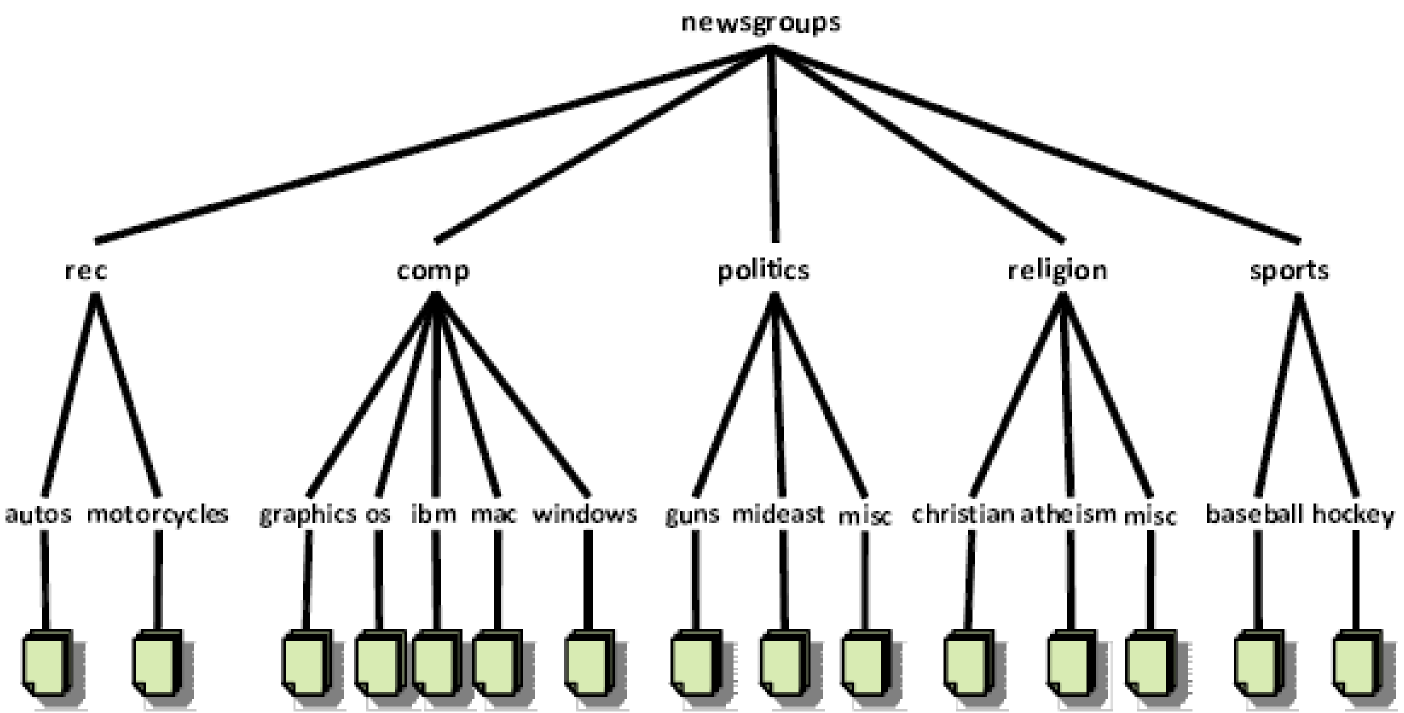
Other approximations:

Posterior as normal, sampling, etc.

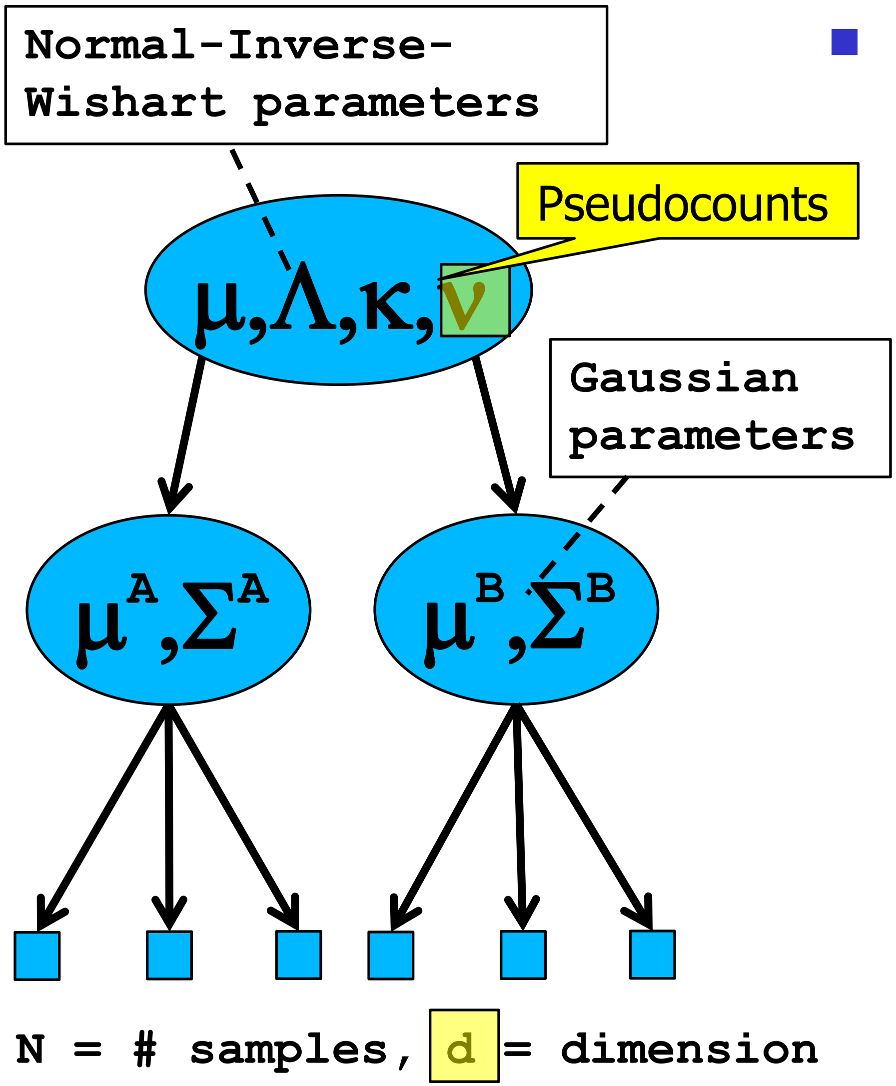
More Issues: Multiple Levels

Conjugate priors usually can’t be extended to multiple levels (e.g., Dirichlet, inverse-Wishart)

Exception: Thibeaux and Jordan (‘05)



More Issues: Restrictive Priors

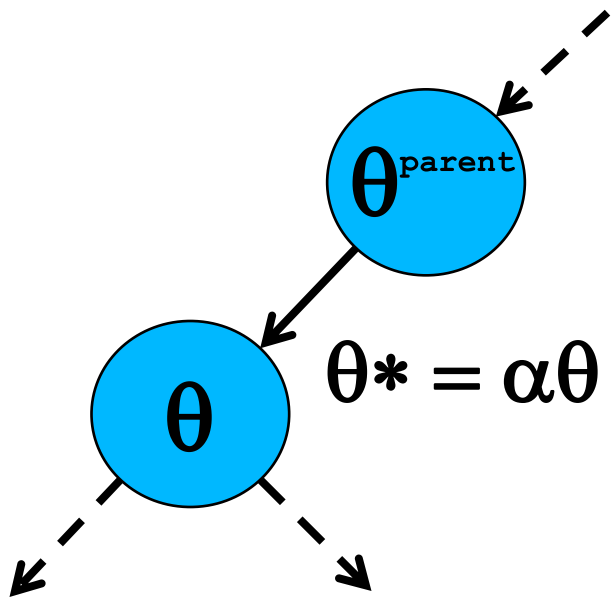


- Example: inverse-Wishart
  - Pseudocount restriction
    - $\nu >= d$
    - If d is large, N is small, signal from prior overwhelms data
    - We show experiments with N=3, d=20

Alternative: Shrinkage

McCallum et al. ('98)

1. Compute maximum likelihood at each node
2. "Shrink" each node toward its parent
   - Linear combination of θ and θ^{parent}
   - Uses cross-validatio



$\theta* = \alpha\theta + (1-$

Pros:

- Simple to compute
- Handles multiple levels

Cons:

- Naive heuristic for transfer
- Averaging not always appropriate

Undirected HB Reformulation

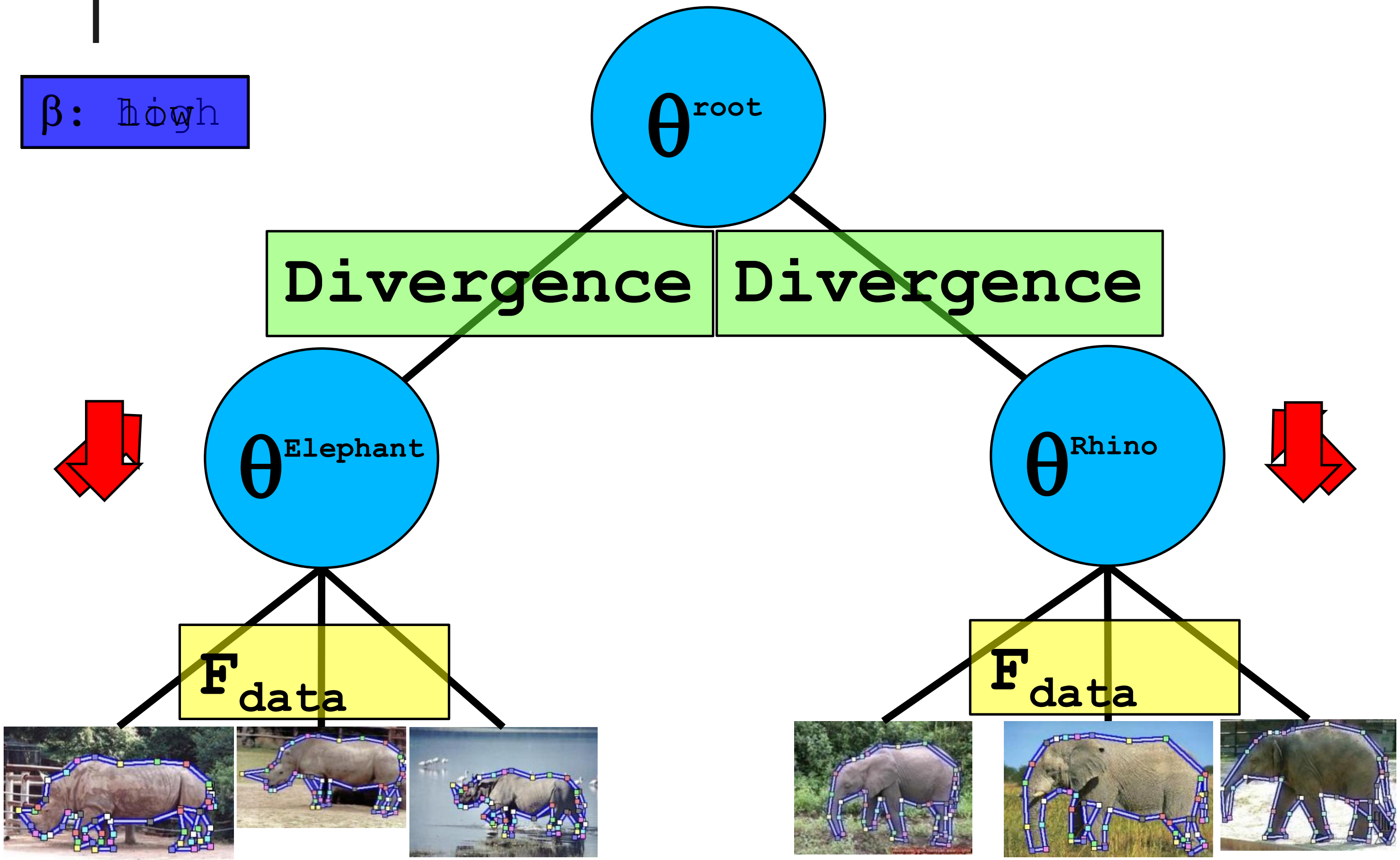
$$\log P(\mathcal{D}, \Theta) = \sum_{c \in \mathcal{L}} \log P(\mathcal{D}^c \mid \Theta^c) + \sum_{c \in \mathcal{C}} \log P(\Theta^c \mid \Theta^{par(c)})$$

$$F_{\text{joint}}(\Theta; \mathcal{D}) = -\sum_{c \in \mathcal{L}} \mathcal{F}_{\text{data}}(\mathcal{D}^c, \Theta^c) + \beta \sum_{c \in \mathcal{C}} \text{Div}(\Theta^c, \Theta^{par(c)})$$

Probabilistic Abstraction Hierarchies (Segal et al. '01)

Defines an undirected Markov random field model over Θ, D

Undirected Probabilistic Model



F_{data}:

Encourage parameters to explain data

Divergence:

Encourage parameters to be similar to parents

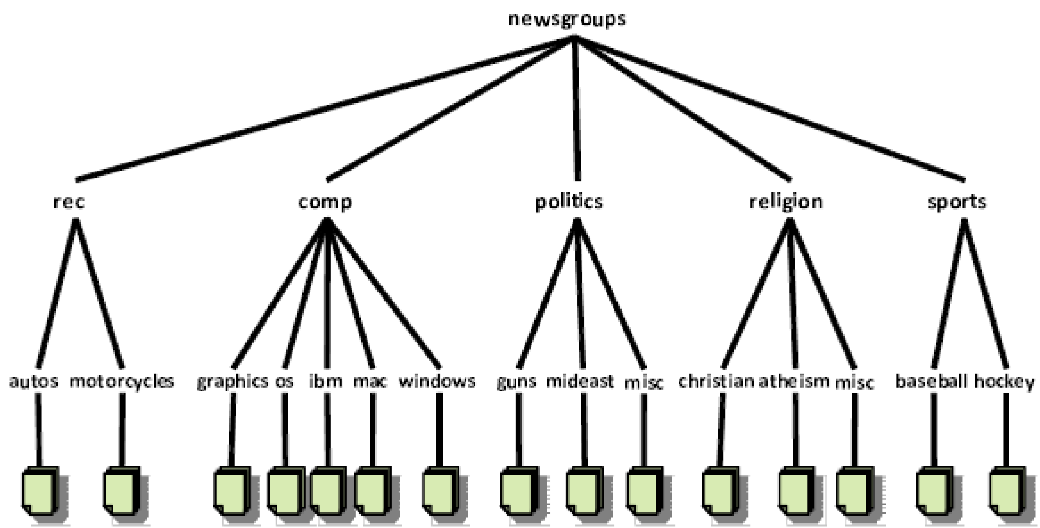
Purpose of Reformulation

$$F_{\text{joint}}(\Theta; \mathcal{D}) = -\sum_{c \in \mathcal{L}} \mathcal{F}_{\text{data}}(\mathcal{D}^c, \Theta^c) + \beta \sum_{c \in \mathcal{C}} \text{Div}(\Theta^c, \Theta^{par(c)})$$

- Easy to specify
  - F_{data} can be likelihood, classification, or other objective
  - Divergence can be L1-distance, L2-distance, ε-insensitive loss, KL divergence, etc.
  - No conjugacy or proper prior restrictions
- Easy to optimize
  - Convex over Θ if F_{data} is convex and Divergence is concave

Application: Text categorization

Task: Categorize Documents



Newsgroup20 Dataset

Bag-of-words model

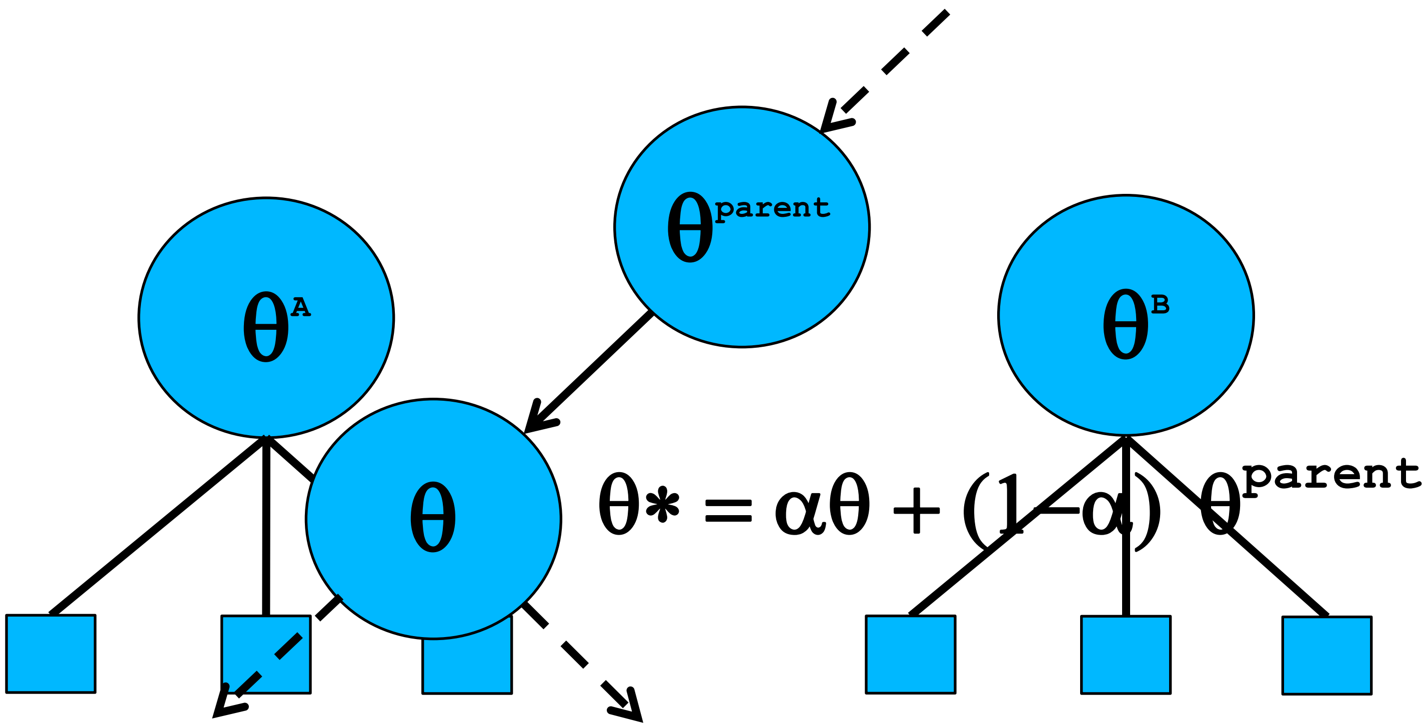
F_{data} : Multinomial log likelihood (regularized)

θ_i represents frequency of word i

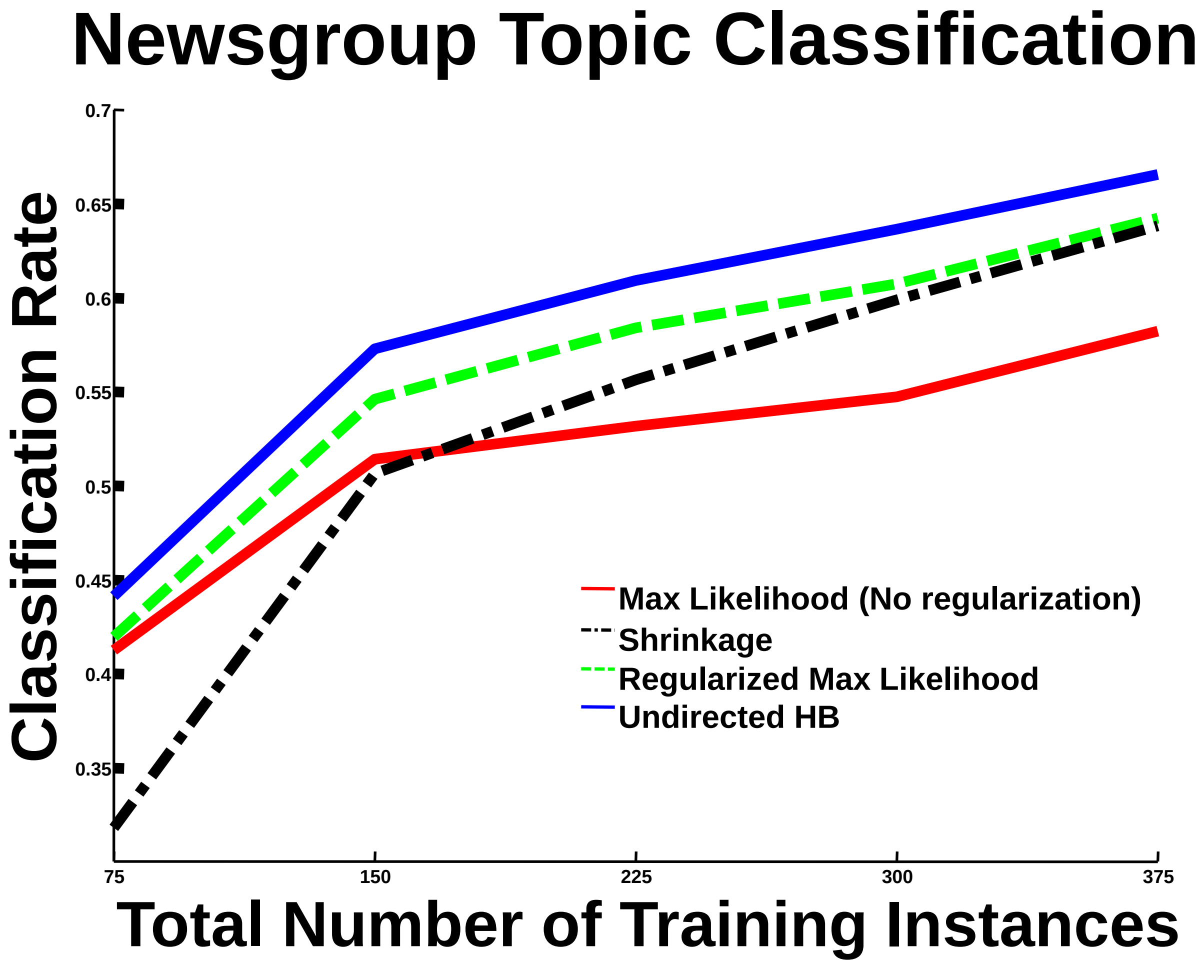
Divergence: L2 norm

Baselines

1. Maximum likelihood at each node (no hierarchy)
2. Cross-validate regularization (no hierarchy)
3. Shrinkage (McCallum et al. ’98, with hierarchy)



Can It Handle Multiple Levels?



Application: Shape Modeling

Task: Learn shape

(Density estimation – test likelihood)

Instances represented by 60 x-y coordinates of landmarks on outline

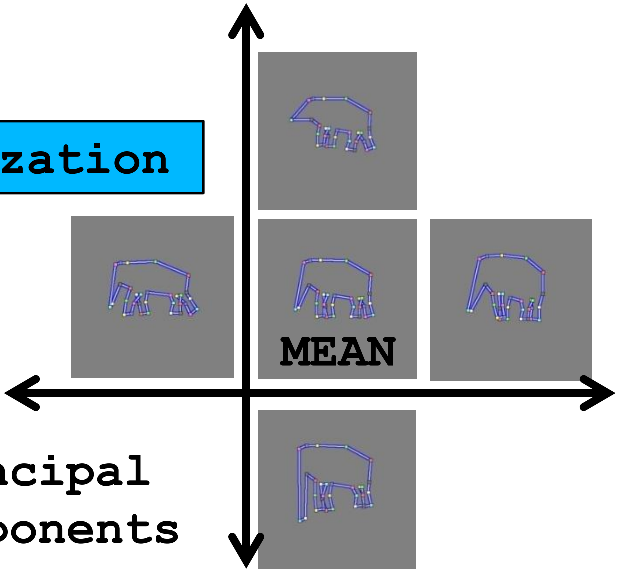
Mammals Dataset (Fink, '05)

$$\mathcal{F}_{\text{data}}(\mathcal{D}^c, \Theta^c) = \sum_{m}^{M_c} \log \mathcal{N}(\mathbf{x}[m] \mid \mu^c, \Sigma^c + \alpha \mathcal{I})$$

Mean landmark location

Covariance over landmarks

Regularization

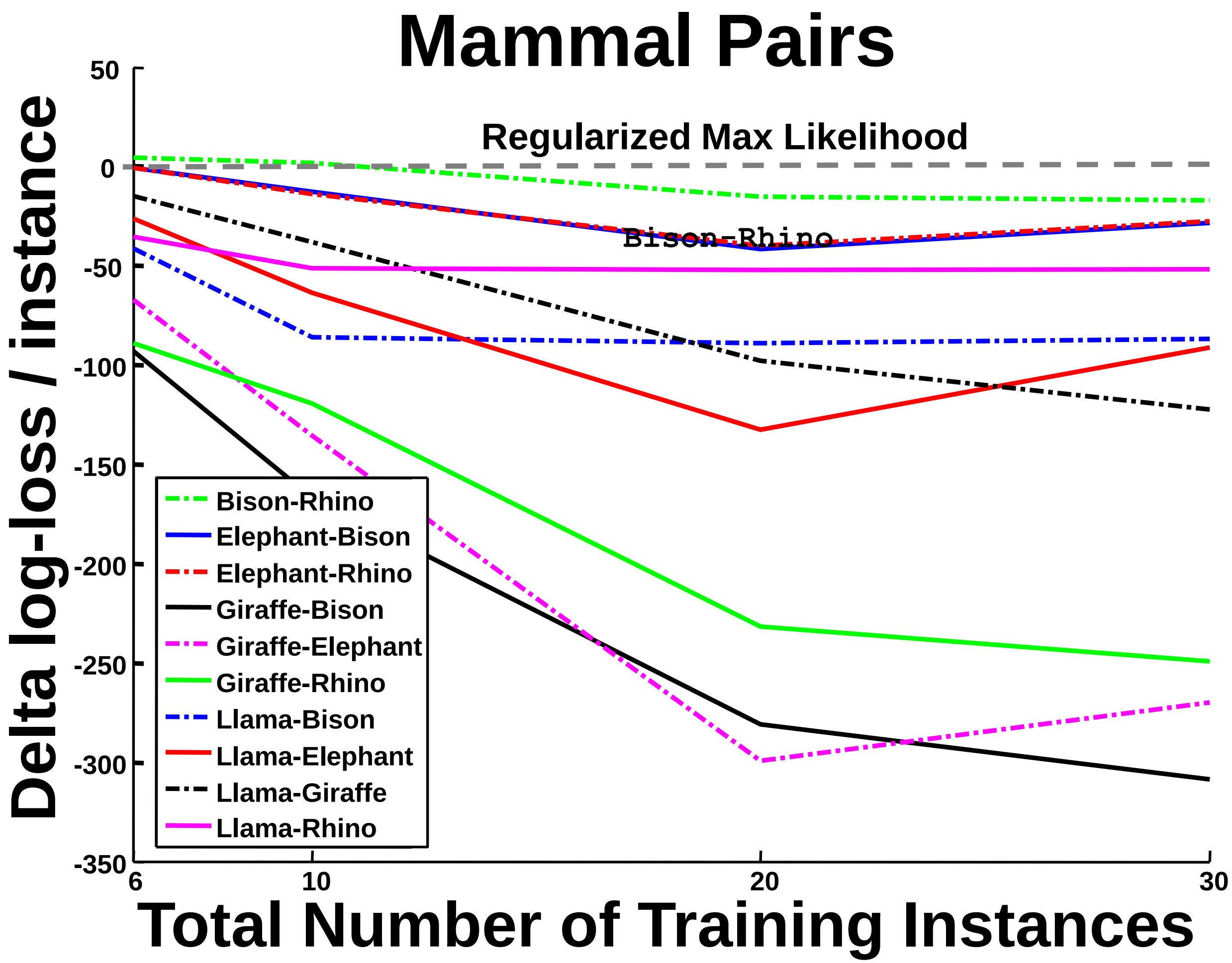


Principal Components

Divergence:

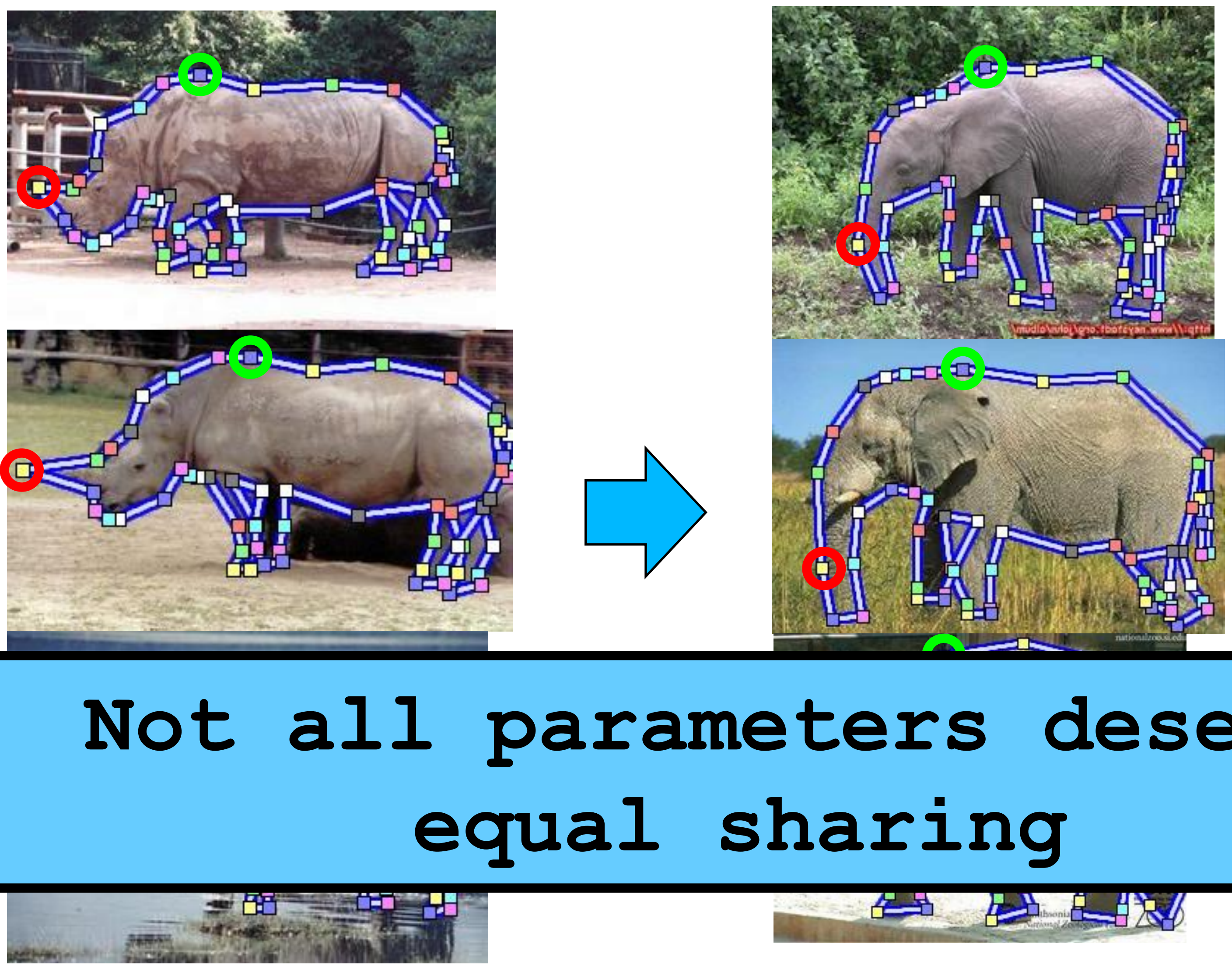
L2 norm over mean and variance

Does Hierarchy Help?



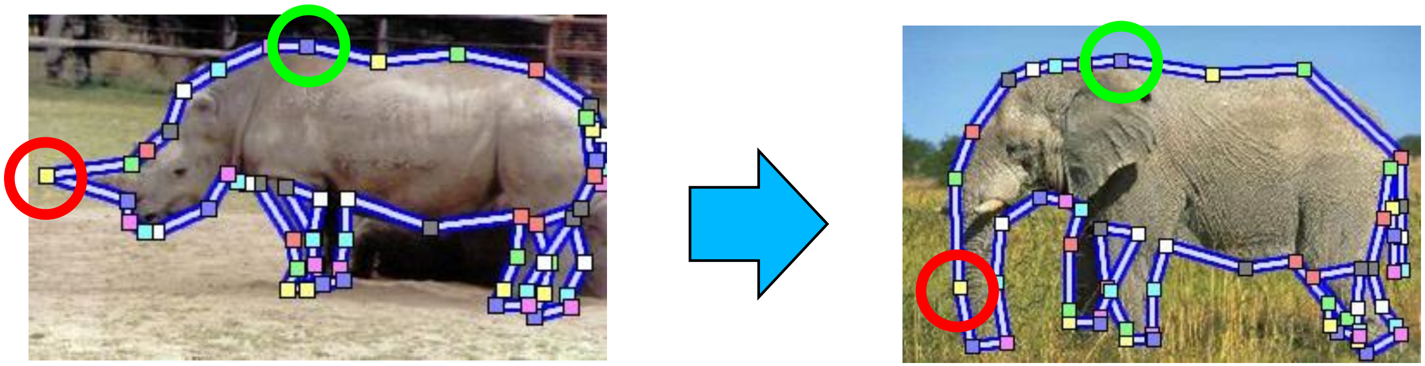
Unregularized max likelihood, shrinkage: Much worse, not shown

Transfer



Not all parameters deserve equal sharing

Degrees of Transfer



$$F_{\text{joint}}(\Theta, \Lambda; \mathcal{D}) = -\sum_{c\in\mathcal{L}} \mathcal{F}_{\text{data}}(\mathcal{D}^c, \Theta^c) + \beta \sum_{c\in\mathcal{C}} \sum_i \frac{1}{\lambda_i^{c,par(c)}} \text{Div}(\theta_i^c, \theta_i^{par(c)})$$

How do we estimate all these parameters?

Sp

Al

subcomponents, child-parent pairs

Learning Degrees of Transfer

- Bootstrap approach

  If θ_i^c and $\theta_i^{par(c)}$ have a consistent relationship, want to encourage them to be similar

- Hyper-prior approach

  Bayesian idea:

  Put prior on λ

  Add λ as parameter to optimization along with Θ

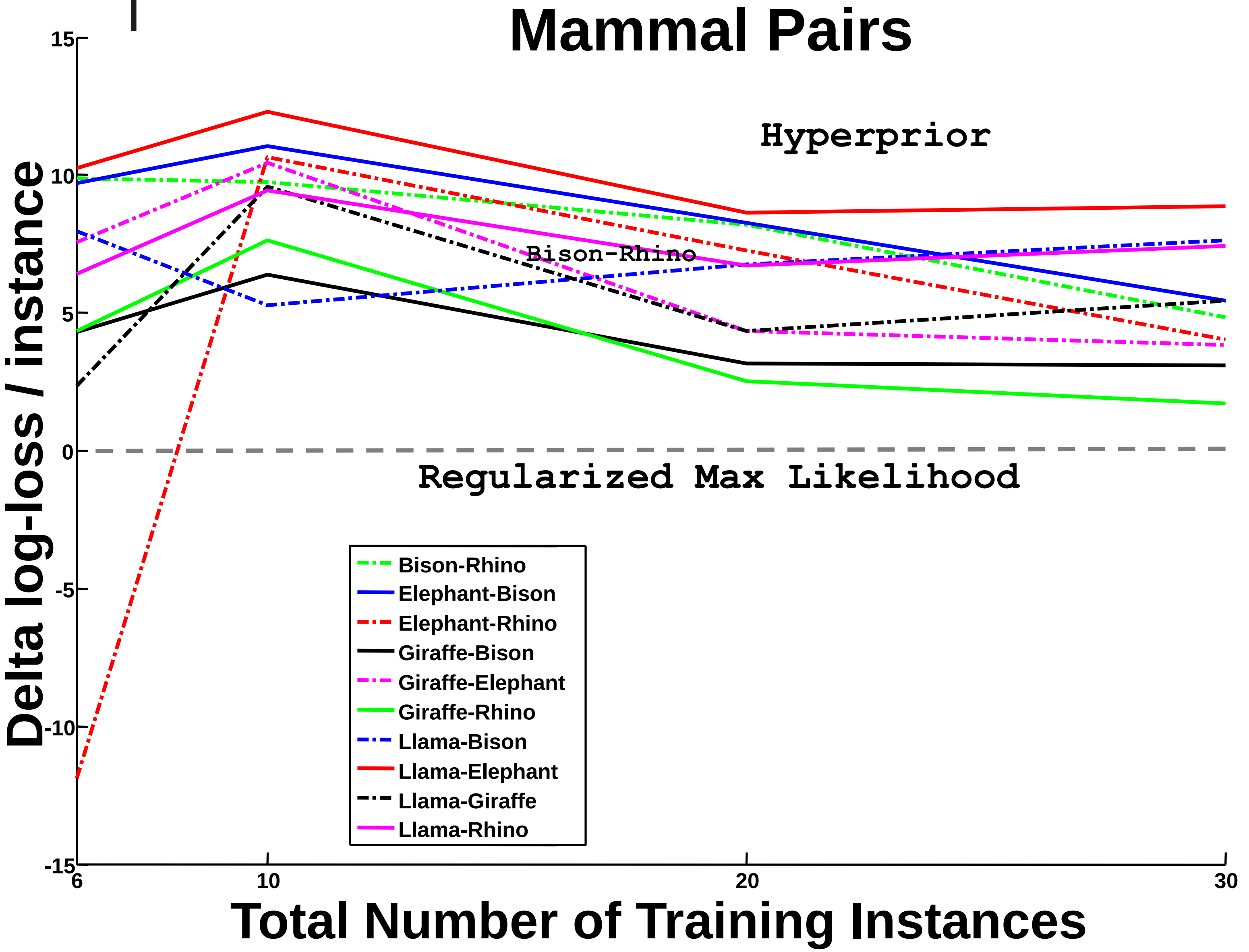
  Concretely: inverse-Gamma prior (forced to be positive)

$$F_{\text{joint}}(\Theta, \Lambda; \mathcal{D}) = \sum_c -\ell\left(\mathcal{D}^c; \Theta^c\right) + \beta \sum_{c \in \mathcal{L}} \sum_i \frac{(\theta_i^c - \theta_i^{par(c)})^2}{\lambda_i^{c,par(c)}} - \sum_{c \in \mathcal{C}} \sum_i \log G^{-1}(\lambda_i^{c,par(c)})$$

Prior on Degree of Transfer

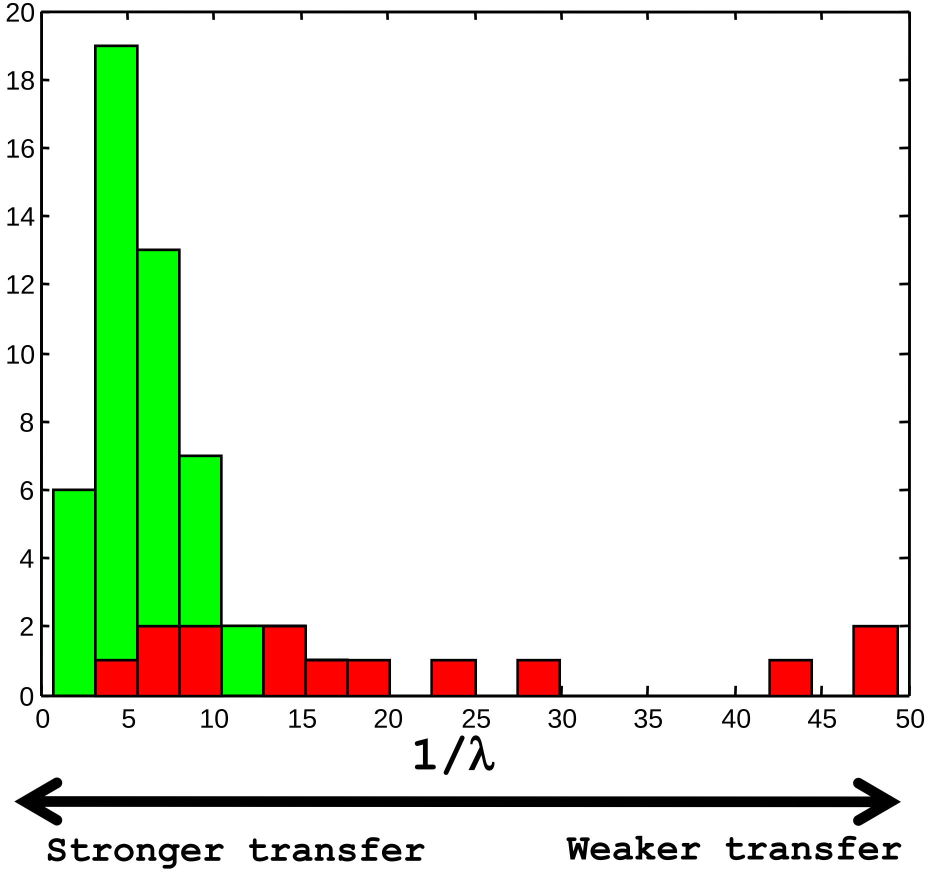
If likelihood is concave, entire objective is convex!

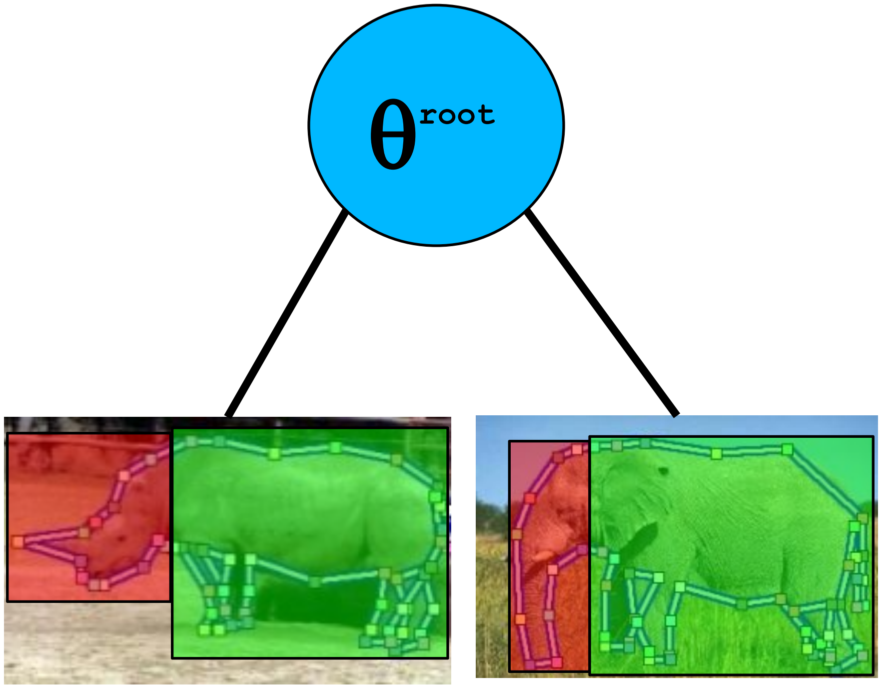
Do Degrees of Transfer Help?



Degrees of Transfer

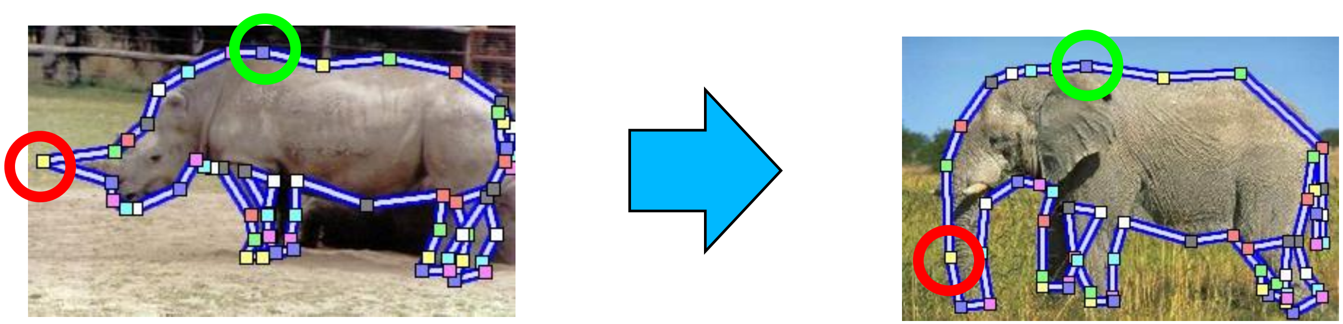
Distribution of DOT coefficients using Hyperprior





Summary

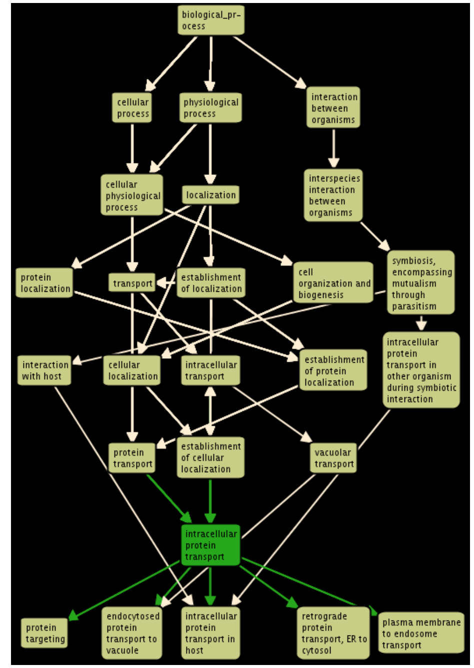
- Transfer between related classes
- Range of settings, tasks
- Probabilistic motivation
- Multilevel, complex hierarchies
- Simple, efficient computation
- Refined transfer of components



Future Work

- Non-tree hierarchies (multiple inheritance)

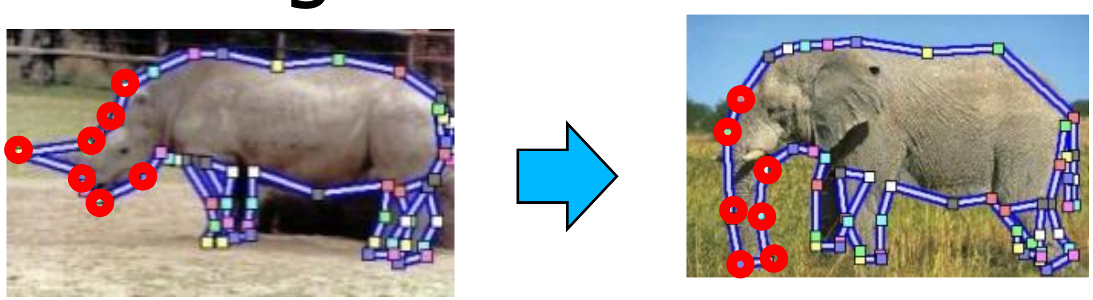
General undirected model doesn't require tree structure



Gene Ontology (GO) network

WordNet Hierarchy

- Block degrees of transfer



Part discovery

- Structure learning