Advanced Hierarchical Models

Russ Salakhutdinov

Department of Statistics and Computer Science
University of Toronto

Motivation

- Learning abstract representations that support transfer to novel tasks, lies at the core of many problems in computer vision, speech perception, natural language processing, and machine learning.

- In many machine learning applications performance is measured using hundreds or thousands of training examples.

- For human learners, a single example of a novel category is often sufficient to make meaningful generalizations to novel instances.

Goal: Transfer higher-order knowledge abstracted from previously learned concept to infer parameters of a novel concept from few examples.

One-shot Learning

How can we learn a novel concept – a high dimensional statistical object – from few examples.

(Lake, Salakhutdinov, Gross, Tenenbaum, CogSci 2011)

Traditional Supervised Learning

Segway

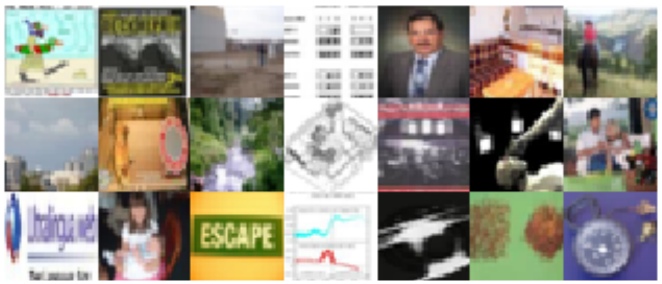
Motorcycle

Test:
What is this?

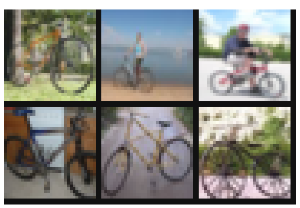
Learning to Transfer

Background Knowledge

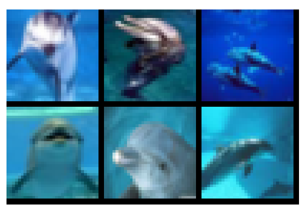
Millions of unlabeled images



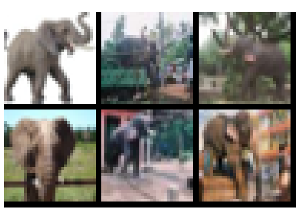
Some labeled images



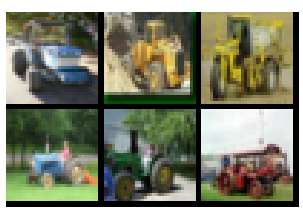
Bicycle



Dolphin



Elephant



Tractor

Learn to Transfer Knowledge

Learn novel concept from one example

Test:
What is this?

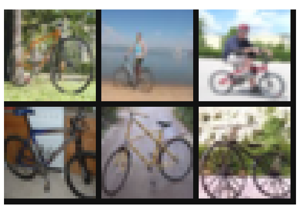
Learning to Transfer

Background Knowledge

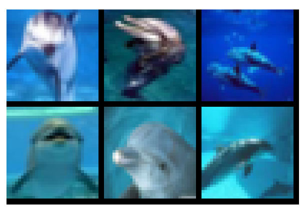
Millions of unlabeled images

Key problem in computer vision, speech perception, natural language processing, and many other domains.

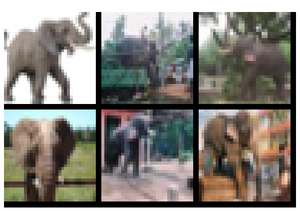
Some labeled images



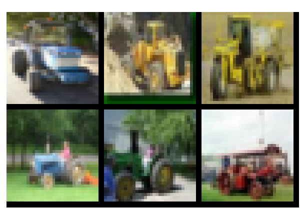
Bicycle



Dolphin



Elephant



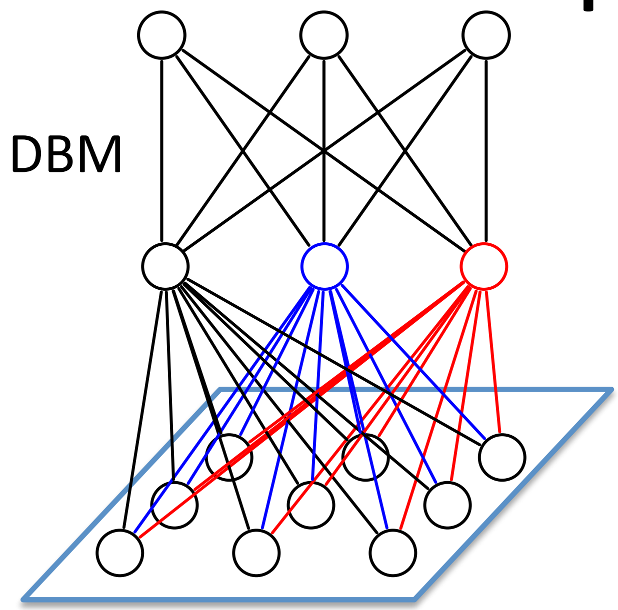
Tractor

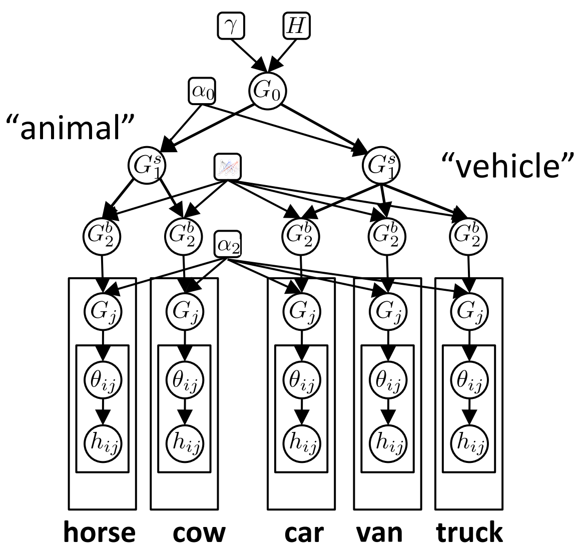
Learn to Transfer Knowledge

Learn novel concept from one example

Test:
What is this?

Talk Roadmap

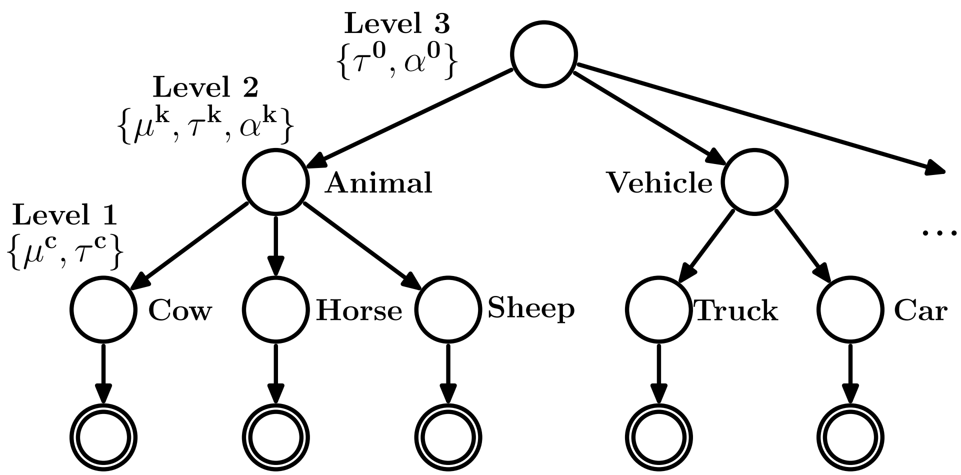




Part 2: Advanced Hierarchical Models

- Introduction: Transfer Learning/ One-Shot Learning.
- Compound Hierarchical Deep Models:
  - Deep Boltzmann Machines.
  - Hierarchical Latent Dirichlet Allocation Model.
- Applications.
- Conclusions

Hierarchical Bayes



Hierarchical Bayesian Models

Hierarchical Prior.

Probability of observed data given parameters

Prior probability of weight vector W

Posterior probability of parameters given the training data D.

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})P(\mathbf{w})}{P(\mathcal{D})}$$

- Fei-Fei, Fergus, and Perona, TPAMI 2006
- E. Bart, I. Porteous, P. Perona, and M. Welling, CVPR 2007
- Miller, Matsakis, and Viola, CVPR 2000
- Sivic, Russell, Zisserman, Freeman, and Efros, CVPR 2008

Hierarchical-Deep Models

HD Models: Compose hierarchical Bayesian models with deep networks, two influential approaches from unsupervised learning

Deep Networks:

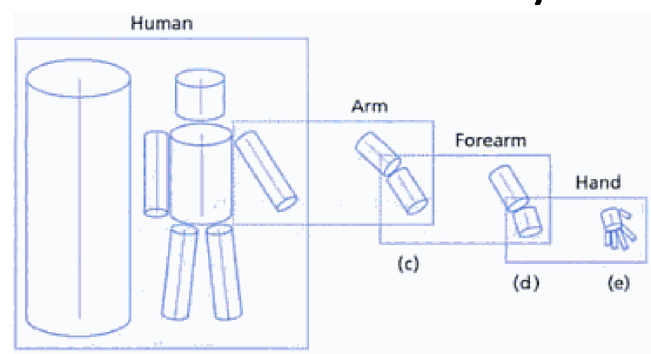
- learn multiple **layers of nonlinearities.**
- trained in unsupervised fashion -- **unsupervised feature learning** – no need to rely on human-crafted input representations.
- **labeled data** is used to slightly adjust the model for a specific task.

Hierarchical Bayes:

- **explicitly represent category hierarchies** for sharing abstract knowledge.
- explicitly identify only a **small number of parameters** that are relevant to the new concept being learned.

Deep Nets

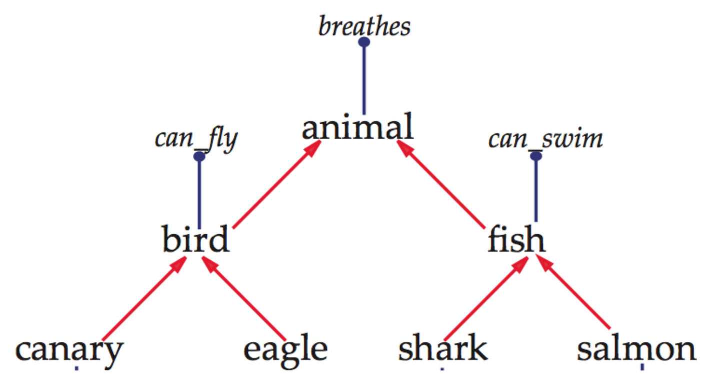
Part-based Hierarchy



Marr and Nishihara (1978)

Hierarchical Bayes

Category-based Hierarchy



Collins & Quillian (1969)

(Salakhutdinov, Tenenbaum, Torralba, NIPS 2011)

Motivation for Our Approach

Learning to transfer knowledge:

Hierarchical

- Super-category: “A segway looks like a funny kind of vehicle”.

- Higher-level features, or parts, shared with other classes:
  - ➢ wheel, handle, post

- Lower-level features:
  - ➢ edges, composition of edges

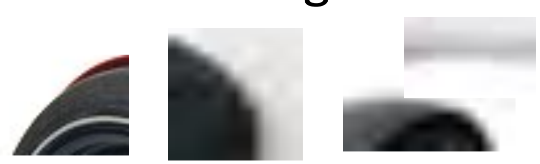
Deep

Segway

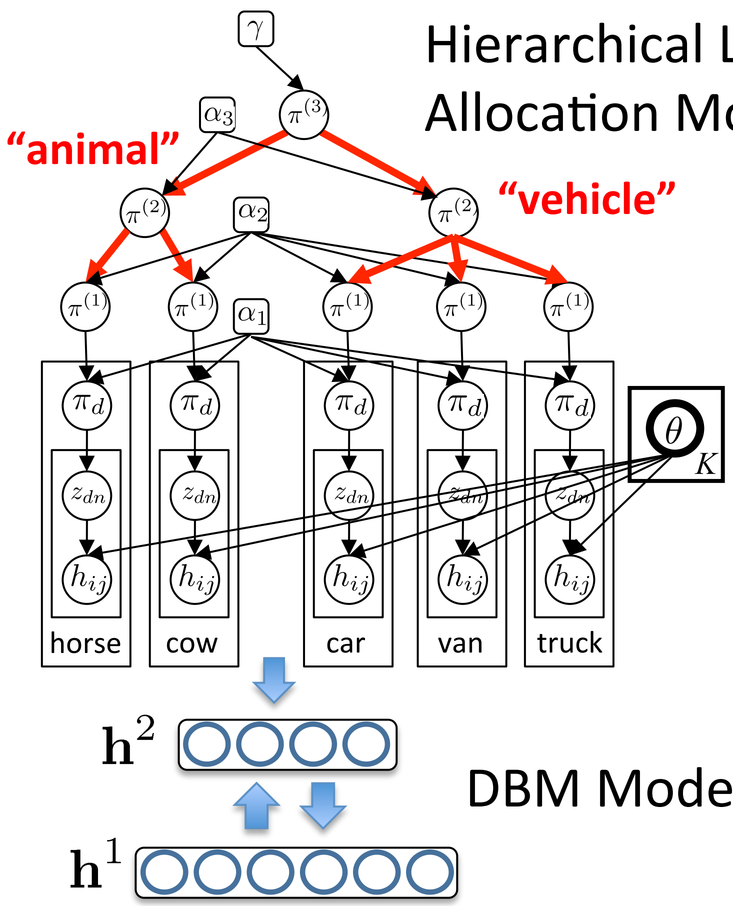
Super-class

Parts

Edges



Hierarchical Generative Model



Hierarchical Latent Dirichlet Allocation Model

DBM Model

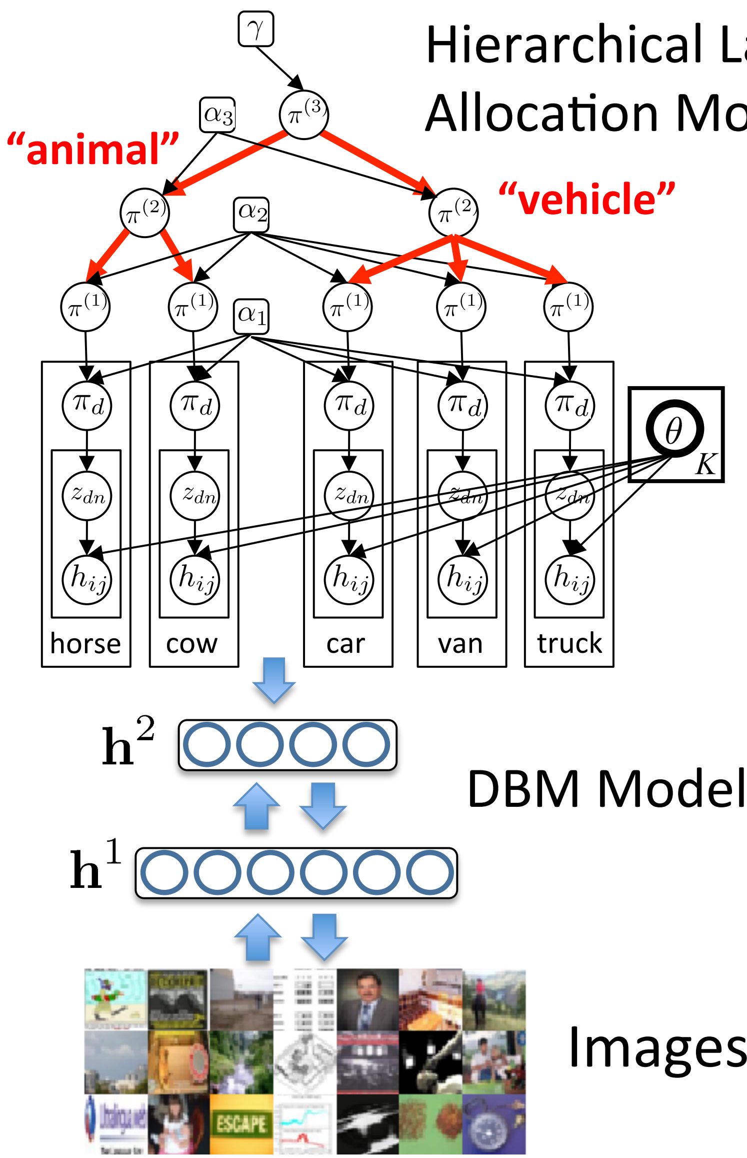
Images

Lower-level generic features:

- edges, combination of edges

Hierarchical Generative Model

Hierarchical Latent Dirichlet Allocation Model



Hierarchical Organization of Categories:

- express priors on the features that are typical of different kinds of concepts
- modular data-parameter relations

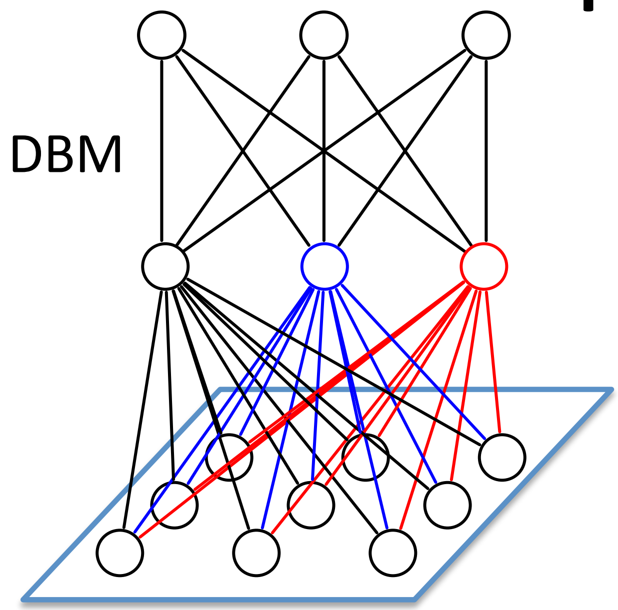
Higher-level class-sensitive features:

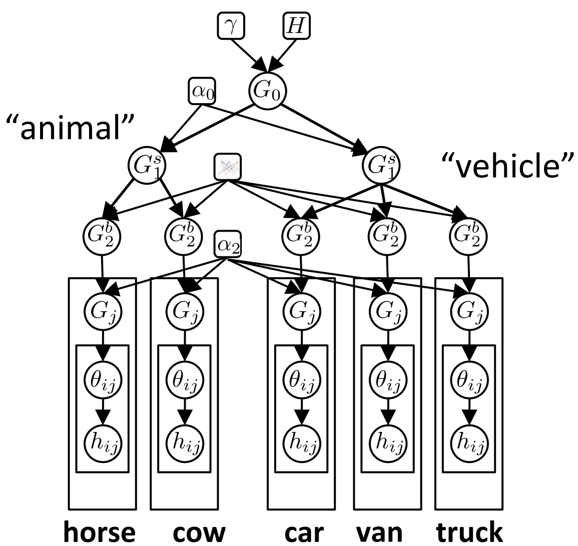
- capture distinctive perceptual structure of a specific concept

Lower-level generic features:

- edges, combination of edges

Talk Roadmap



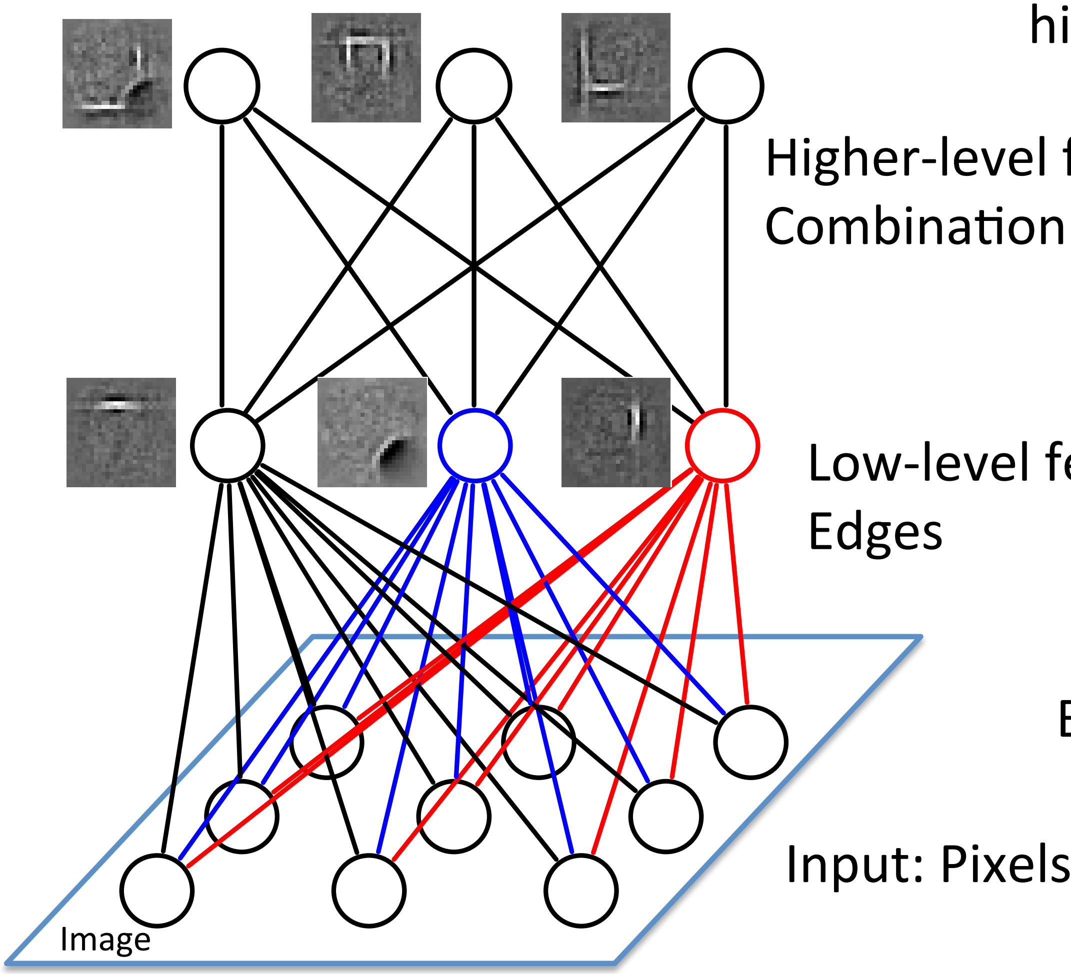


Part 2: Advanced Hierarchical Models

- Introduction: Transfer Learning/ One-Shot Learning.
- Hierarchical Deep Models:
  - Deep Boltzmann Machines.
  - Hierarchical Latent Dirichlet Allocation Model.
- Applications.
- Conclusions

Deep Boltzmann Machines

Internal representations capture higher-order statistical structure

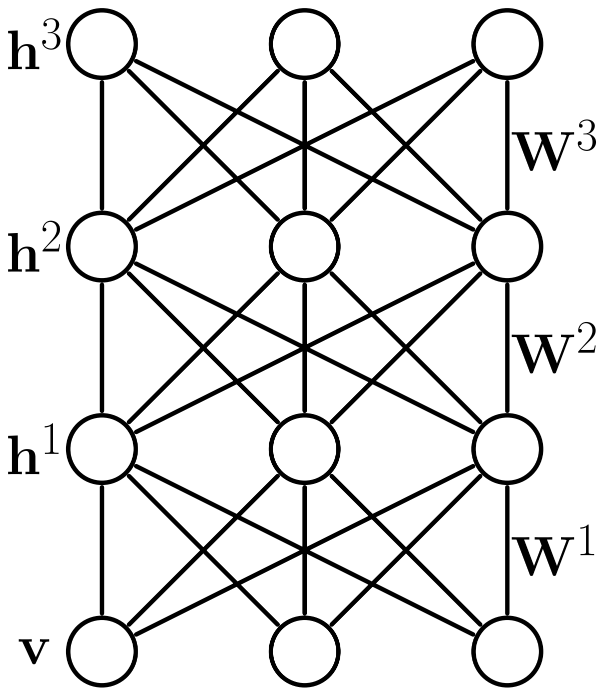


Built from **unlabeled** inputs.

A Brief Review

$$P_\theta(\mathbf{v}) = \frac{P^*(\mathbf{v})}{\mathcal{Z}(\theta)} = \frac{1}{\mathcal{Z}(\theta)} \sum_{\mathbf{h}^1,\mathbf{h}^2,\mathbf{h}^3} \exp\left[\mathbf{v}^\top W^1 \mathbf{h}^1 + \underline{\mathbf{h}^{1\top} W^2 \mathbf{h}^2} + \underline{\mathbf{h}^{2\top} W^3 \mathbf{h}^3}\right]$$

Deep Boltzmann Machine



Input

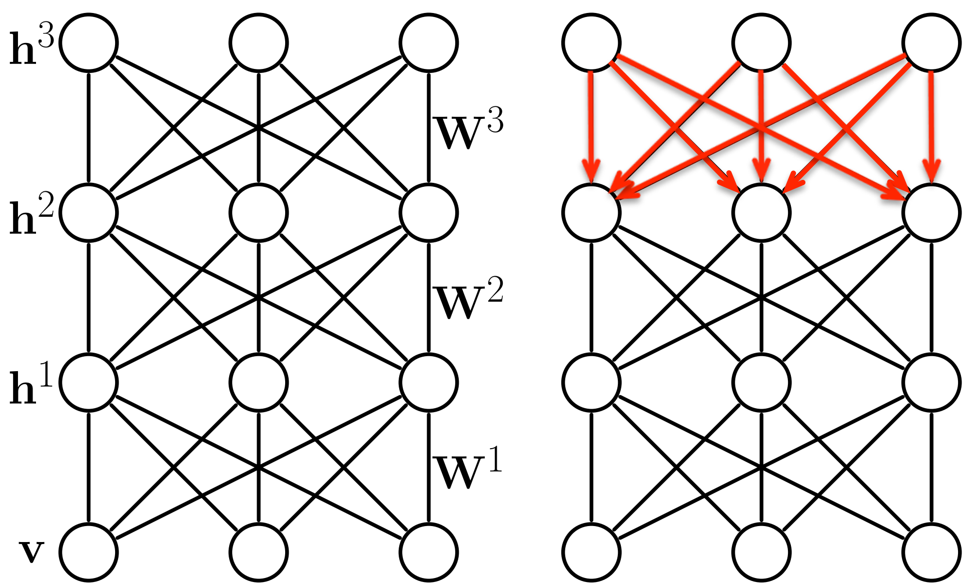
$\theta = \{W^1, W^2, W^3\}$ model parameters

- Dependencies between hidden variables.
- All connections are undirected.
- Bottom-up and Top-down:

Decomposition

The joint probability can be decomposed:

$$P_\theta(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3) = \underbrace{P_\theta(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2 | \mathbf{h}^3)}_{\text{Conditional DBM}} \underbrace{P_\theta(\mathbf{h}^3)}_{\text{Prior term}}$$



DBM Conditional DBM

Replace the last term with more structured hierarchical prior.

$$P_\theta(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2 | \mathbf{h}^3) = \frac{1}{\mathcal{Z}(\theta, \mathbf{h}^3)} \exp\left[\mathbf{v}^\top W^1 \mathbf{h}^1 + \mathbf{h}^{1\top} W^2 \mathbf{h}^2 + \mathbf{h}^{2\top} W^3 \mathbf{h}^3\right]$$

Stage-wise Learning

The joint probability can be decomposed:

$$P_\theta(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3) = \underbrace{P_\theta(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2 | \mathbf{h}^3)}_{\text{Conditional DBM}} \underbrace{P_\theta(\mathbf{h}^3)}_{\text{Prior term}}$$

DBMs approximate intractable posterior $P_\theta(\mathbf{h}|\mathbf{v})$ with fully factorized tractable distribution $Q_\mu(\mathbf{h}|\mathbf{v})$. The variational lower-bound takes form:

$$\log P_\theta(\mathbf{v}) \geq \underbrace{\sum_{\mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3} Q_\mu(\mathbf{h}^1, \mathbf{h}^2 | \mathbf{v}) \Big[\log P_\theta(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2 | \mathbf{h}^3) \Big]}_{\text{Likelihood term}} + \underbrace{\mathcal{H}(Q_\mu(\mathbf{h}|\mathbf{v}))}_{\text{Entropy functional}}$$

$$+ \underbrace{\sum_{\mathbf{h}^3} Q_\mu(\mathbf{h}^3 | \mathbf{v}) \log P_\theta(\mathbf{h}^3)}_{\text{Fit Hierarchical LDA prior}}$$

$$\mathcal{H}(Q_\mu(\mathbf{h}|\mathbf{v})) = \sum_{\mathbf{h}} Q_\mu(\mathbf{h}|\mathbf{v}) \log \frac{1}{Q_\mu(\mathbf{h}|\mathbf{v})}$$

Stage-wise Learning

The joint probability can be decomposed:

$$P_\theta(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3) = \underbrace{P_\theta(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2 | \mathbf{h}^3)}_{\text{Conditional DBM}} \underbrace{P_\theta(\mathbf{h}^3)}_{\text{Prior term}}$$

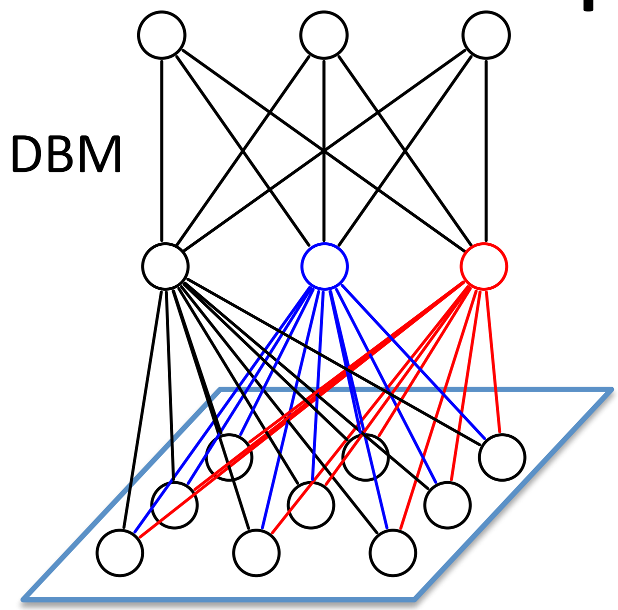
DBMs approximate intractable posterior $P_\theta(\mathbf{h}|\mathbf{v})$ with fully factorized tractable distribution $Q_\mu(\mathbf{h}|\mathbf{v})$. The variational lower-bound takes form:

$$\log P_\theta(\mathbf{v}) \geq \underbrace{\sum_{\mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3} Q_\mu(\mathbf{h}^1, \mathbf{h}^2 | \mathbf{v}) \Big[\log P_\theta(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2 | \mathbf{h}^3) \Big]}_{\text{Likelihood term}} + \underbrace{\mathcal{H}(Q_\mu(\mathbf{h}|\mathbf{v}))}_{\text{Entropy functional}}$$

- Learn DBM.
- Using variational inference, infer the states of the top-level variables and fit an LDA prior.

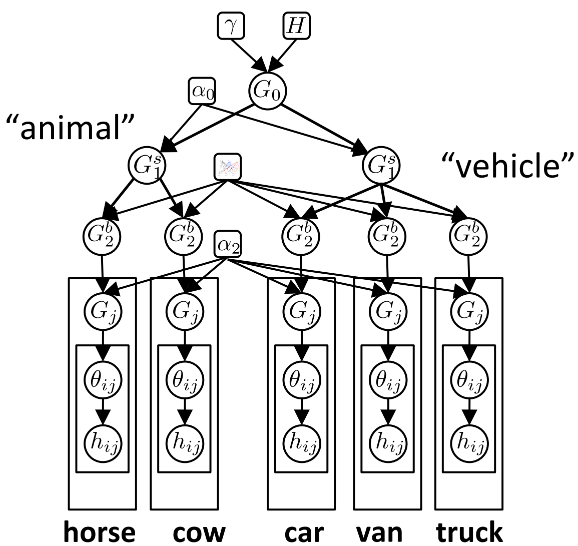
$$\underbrace{Q_\mu(\mathbf{h}^3|\mathbf{v}) \log P_\theta(\mathbf{h}^3)}_{\text{Fit Hierarchical LDA prior}}$$

Talk Roadmap



Part 2: Advanced Hierarchical Models

- Introduction: Transfer Learning/ One-Shot Learning.
- Compound Hierarchical Deep Models:
  - Deep Boltzmann Machines.
  - Hierarchical Latent Dirichlet Allocation Model.
- Applications.
- Conclusions.

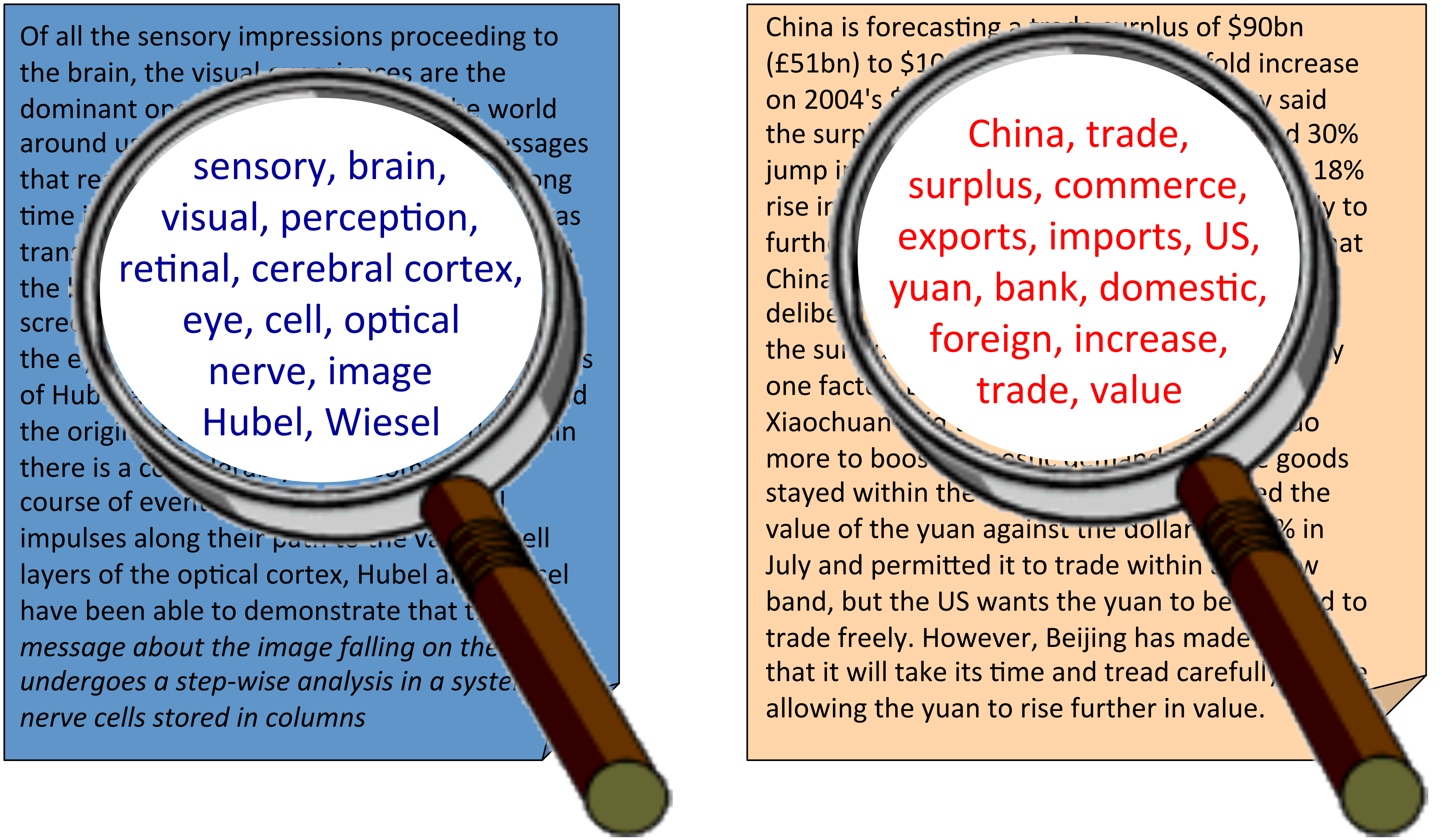


Bag of Words Representation

Object → Bag of 'words'

Slide credit: Fei fei

Analogy to Documents



Intuition: Documents contain multiple topics.

Slide credit: Fei fei

Latent Dirichlet Allocation

Text document

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Discovered topics

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Blei, et al. 2003

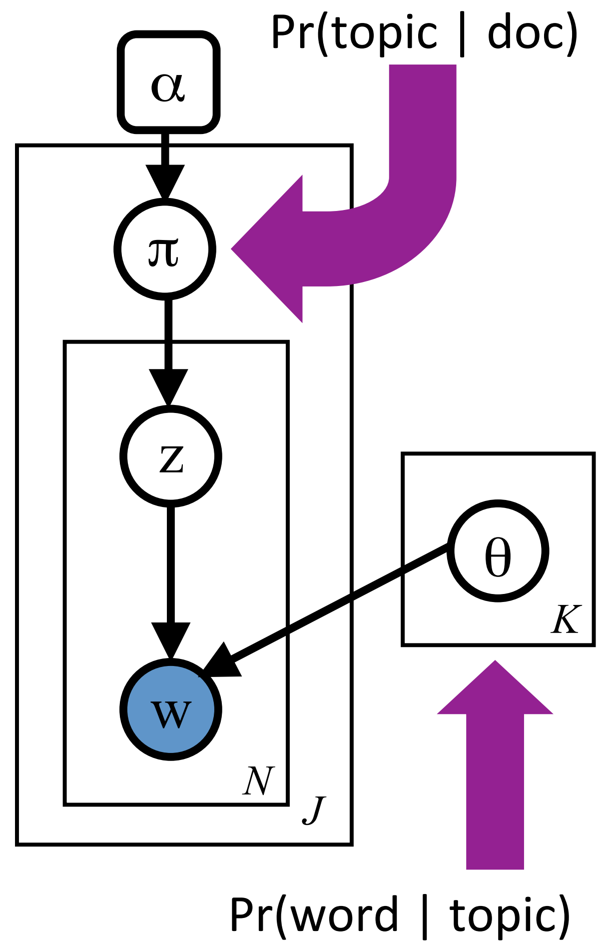
Latent Dirichlet Allocation

Generative Process: $\mathbf{w} \sim \mathrm{LDA}$

Draw each topic $\theta_k \sim \mathrm{Dir}(\eta)$ for $k = 1..., K$

For each document d:

- Draw topic proportions $\pi_d \sim \mathrm{Dir}(\alpha)$
- For each word:
  - Draw topic indicator $z_{d,n} \sim \mathrm{Mult}(\pi_d)$
  - Draw word $w_{d,n} \sim \mathrm{Mult}(\theta_{z_{d,n}})$



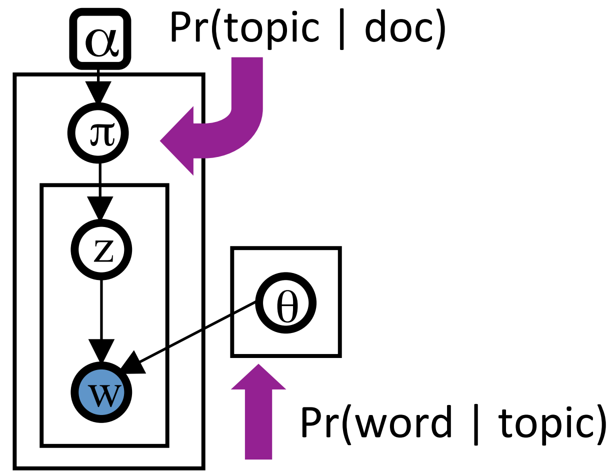
Latent Dirichlet Allocation

Generative Process: $\mathbf{w} \sim \text{LDA}$

Draw each topic $\theta_k \sim \text{Dir}(\eta)$ for $k = 1..., K$

For each document:

- Draw topic proportions $\pi_d \sim \text{Dir}(\alpha)$
- For each word:
  - Draw topic indicator $z_{d,n} \sim \text{Mult}(\pi_d)$
  - Draw word $w_{d,n} \sim \text{Mult}(\theta_{z_{d,n}})$



The William Randolph Hearst Foundation will give $1.25
tan Opera Co., New York Philharmonic and Juilliard Sc
real opportunity to make a mark on the future of the per
every bit as important as our traditional areas of support i
and the social services," Hearst Foundation President I
announcing the grants. Lincoln Center's share will be $
will house young artists and provide new public faciliti
New York Philharmonic will receive $400,000 each. Th
the performing arts are taught, will get $250,000. The He
of the Lincoln Center Consolidated Corporate Fund, v
donation, too.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

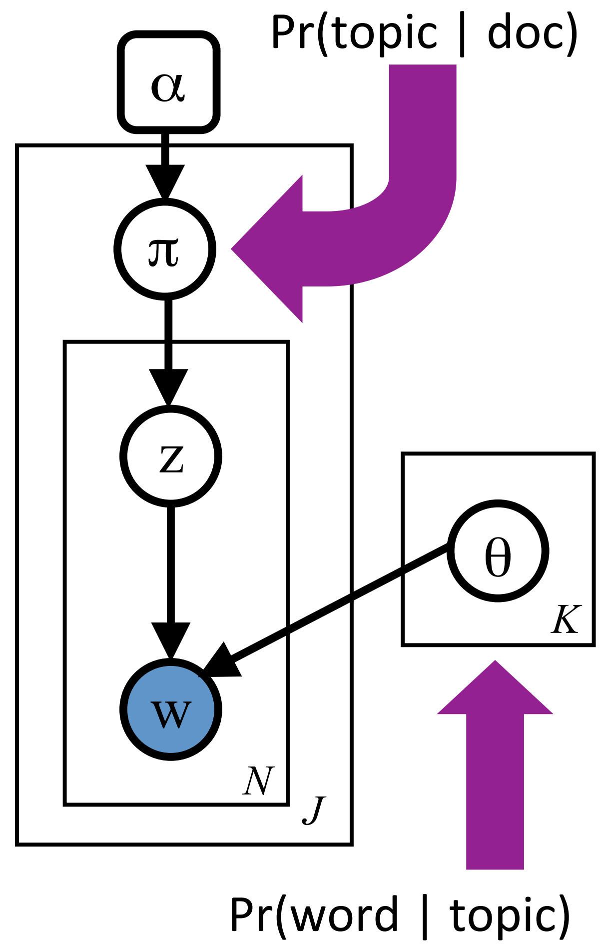
Latent Dirichlet Allocation

Generative Process: $\mathbf{w} \sim \mathrm{LDA}$

Draw each topic $\theta_k \sim \mathrm{Dir}(\eta)$ for $k = 1..., K$

For each document:

- Draw topic proportions $\pi_d \sim \mathrm{Dir}(\alpha)$
- For each word:
  - Draw topic indicator $z_{d,n} \sim \mathrm{Mult}(\pi_d)$
  - Draw word $w_{d,n} \sim \mathrm{Mult}(\theta_{z_{d,n}})$

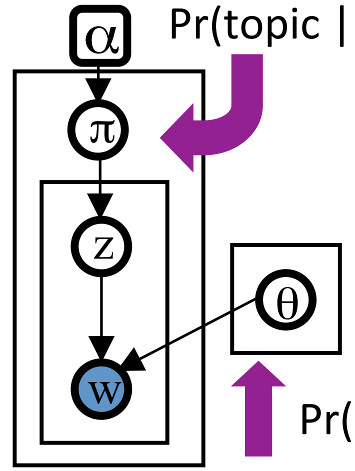


Remember: compound HD model:

$\mathbf{h}^3 \sim \mathrm{LDA\ prior}$

Words ⇔ activations of DBM's top-level units.
Topics ⇔ distributions over top-level units, or higher-level parts.

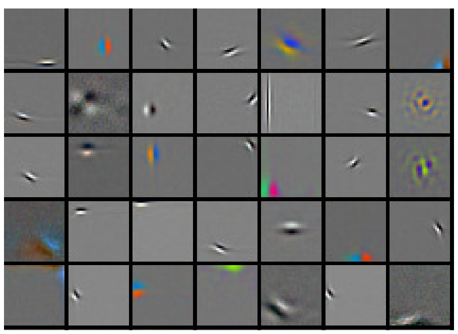
Intuition



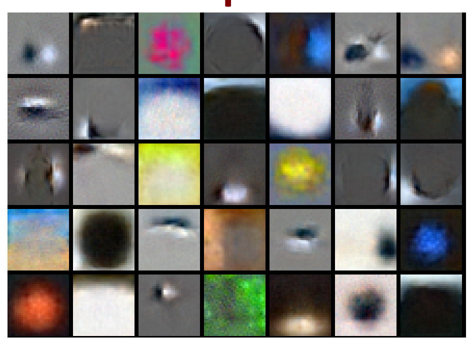
$\mathbf{h}^3 \sim \text{LDA prior}$

Words ⇔ activations of DBM's top-level units.
Topics ⇔ distributions over top-level units, or higher-level parts.

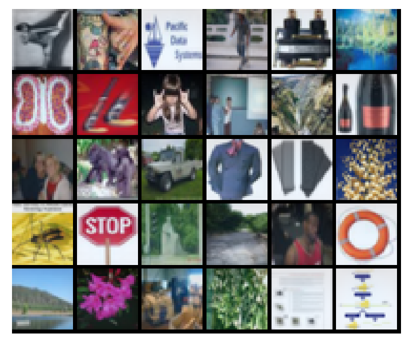
DBM generic features:
Words



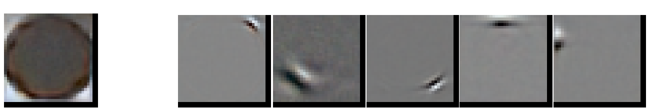
LDA high-level features:
Topics



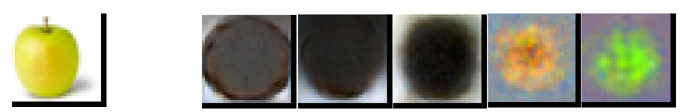
Images
Documents



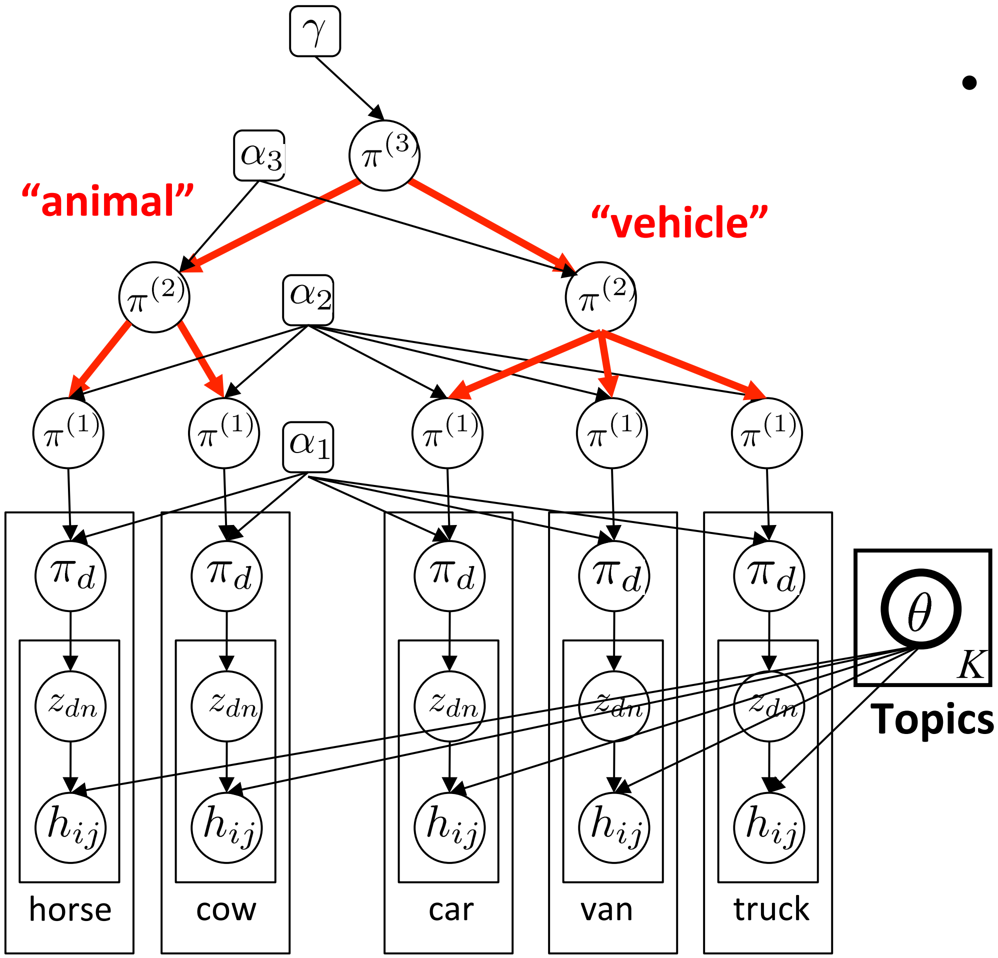
Each topic is made up of words.



Each document is made up of topics.



Hierarchical LDA
Modeling Super-Category Structure



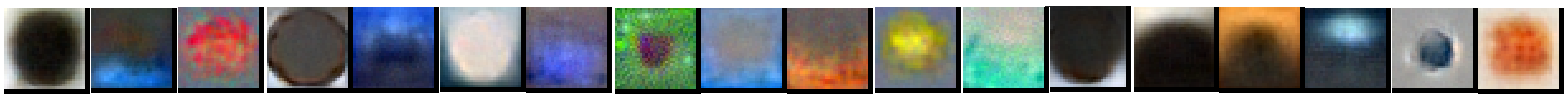
- Draw **global** topic proportions: $\pi^{(3)} \sim \mathrm{Dir}(\gamma)$
  - Draw **super-class specific** topic proportions:
  $$\pi^{(2)}|\pi^{(3)} \sim \mathrm{Dir}(\alpha^{(3)}\pi^{(3)})$$
    - Draw **class-class specific** topic proportions:
    $$\pi^{(1)}|\pi^{(2)} \sim \mathrm{Dir}(\alpha^{(2)}\pi^{(2)})$$
      - Draw **document specific** topic proportions:
      $$\pi_d|\pi^{(1)} \sim \mathrm{Dir}(\alpha^{(1)}\pi^{(1)})$$

Nonparametric extension:

Hierarchical Dirichlet Process (HDP).

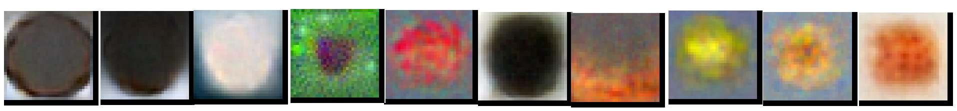
Hierarchical LDA: Example

Global topic proportions:

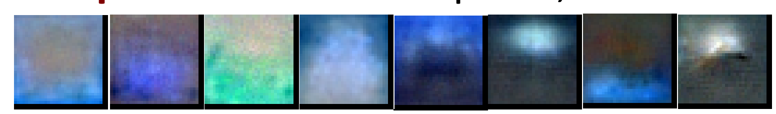


Super-class specific topic proportions:

Fruits: apples, oranges, pears

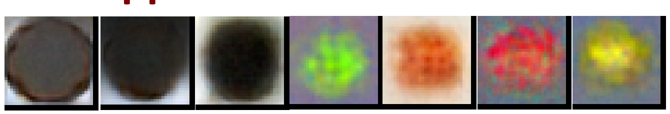


Aquatic animals: dolphins, sharks.



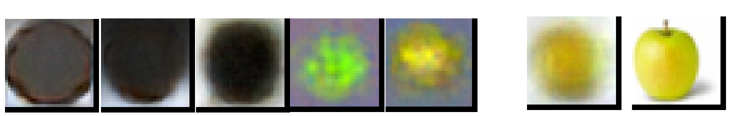
Class specific topic proportions:

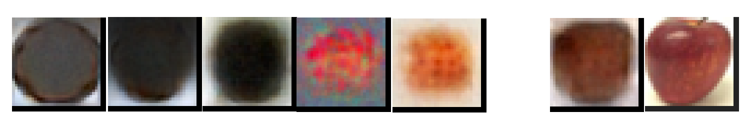
Apples:



Oranges:

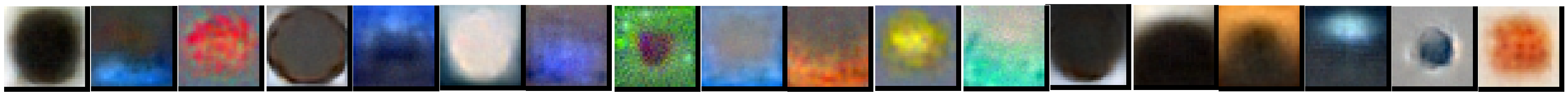
Image specific topic proportions:





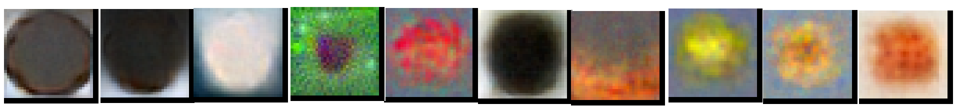
Hierarchical LDA: Example

Global topic proportions:

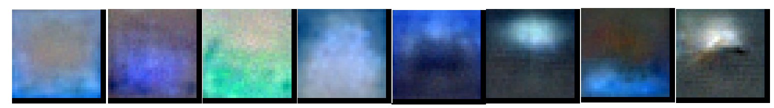


Super-class specific topic proportions:

Fruits: apples, oranges, pears

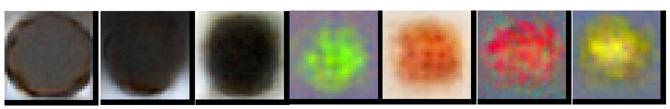


Aquatic animals: dolphins, sharks.

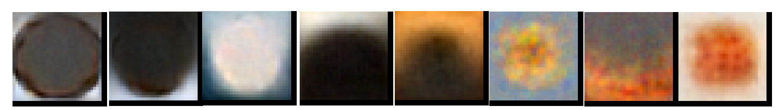


Class specific topic proportions:

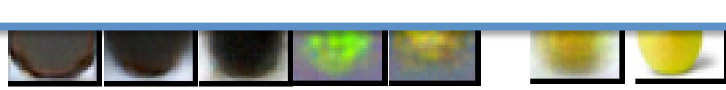
Apples:

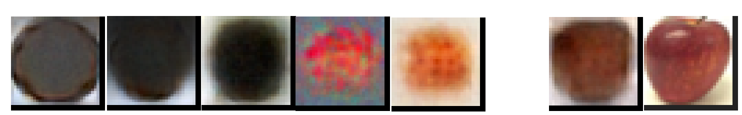


Oranges:



So far we have assumed a **fixed** hierarchy





Modeling the Number of Super-Categories

Place Chinese Restaurant Process (CRP) Prior over the number of super-classes.

CRP defines a distribution on partition of integers.

Generating from $\mathrm{CRP}(\alpha)$:

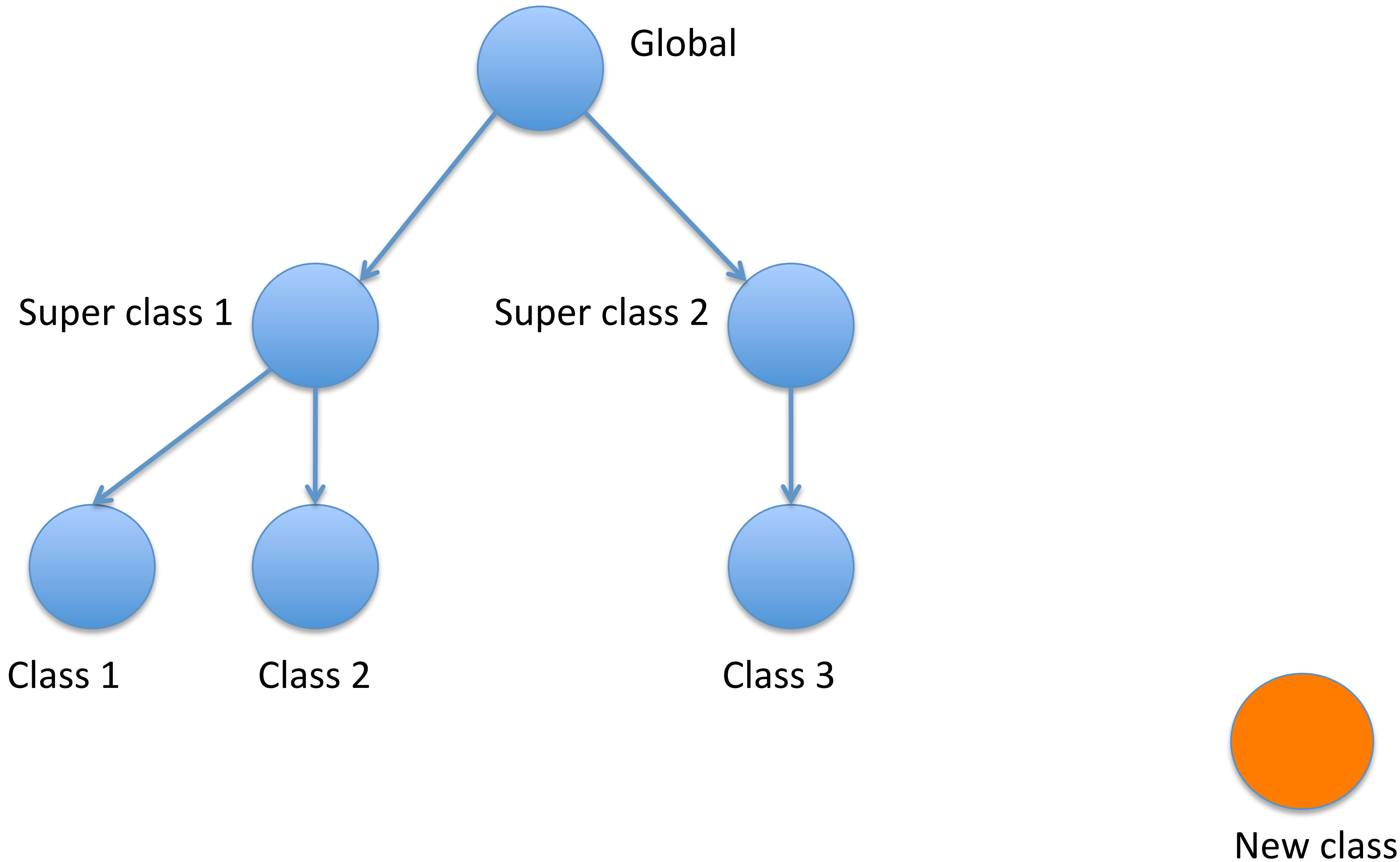
Customers enter a restaurant with an unbounded number of tables, where the n^{th} customer occupies a table k drawn from:

$$P(z_n = k | z_1, ..., z_{n-1}) = \begin{cases} \frac{n^k}{n-1+\alpha} & n^k > 0 \\ \frac{\alpha}{n-1+\alpha} & k \text{ is new} \end{cases}$$

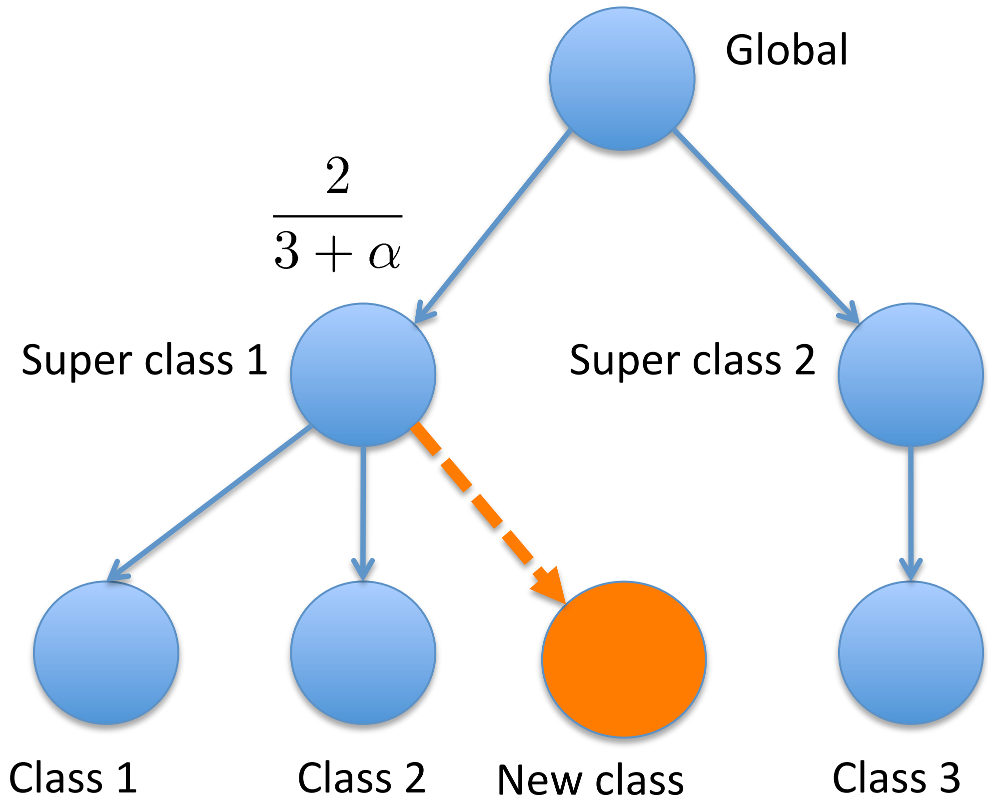
where n^k is the number of previous customers at table k and α is the concentration parameter.

Customers ⇔ integers, tables ⇔ clusters.

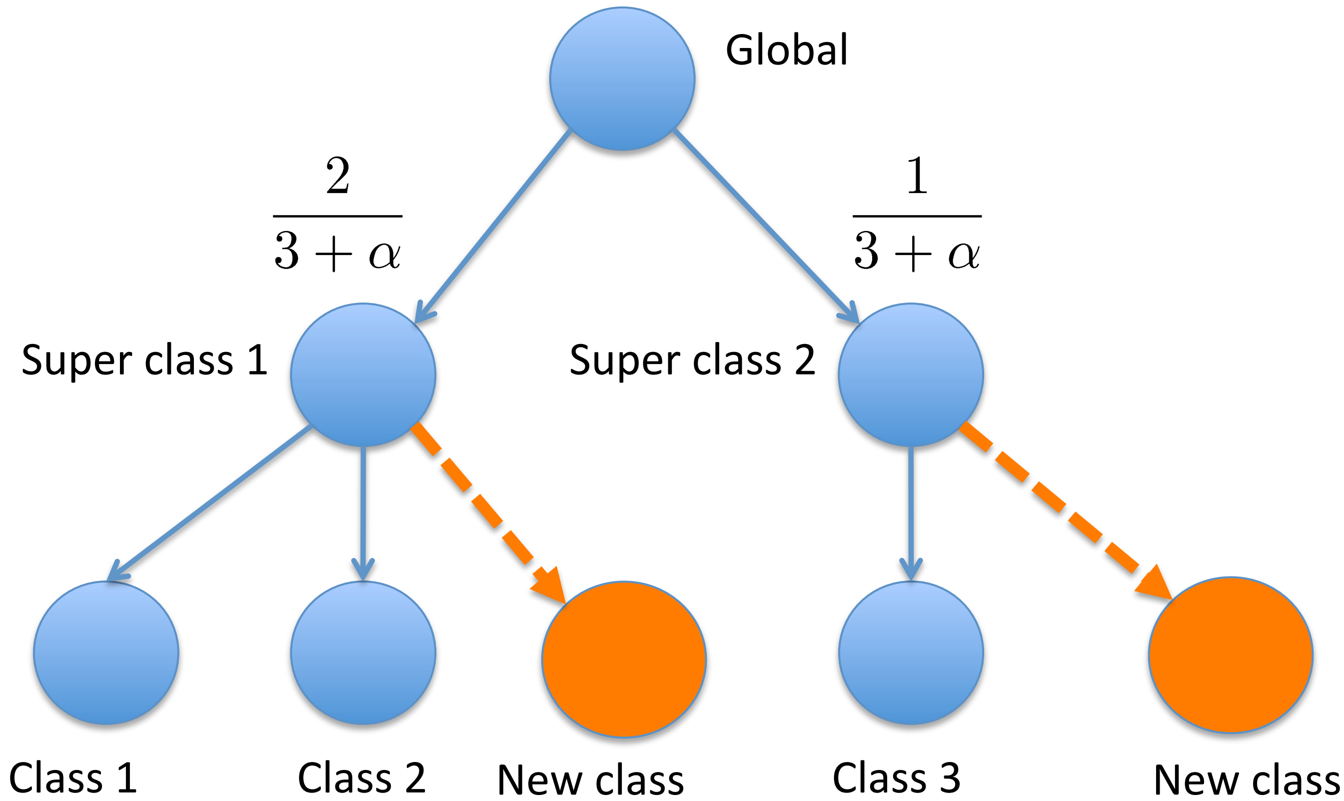
Modeling the Hierarchy



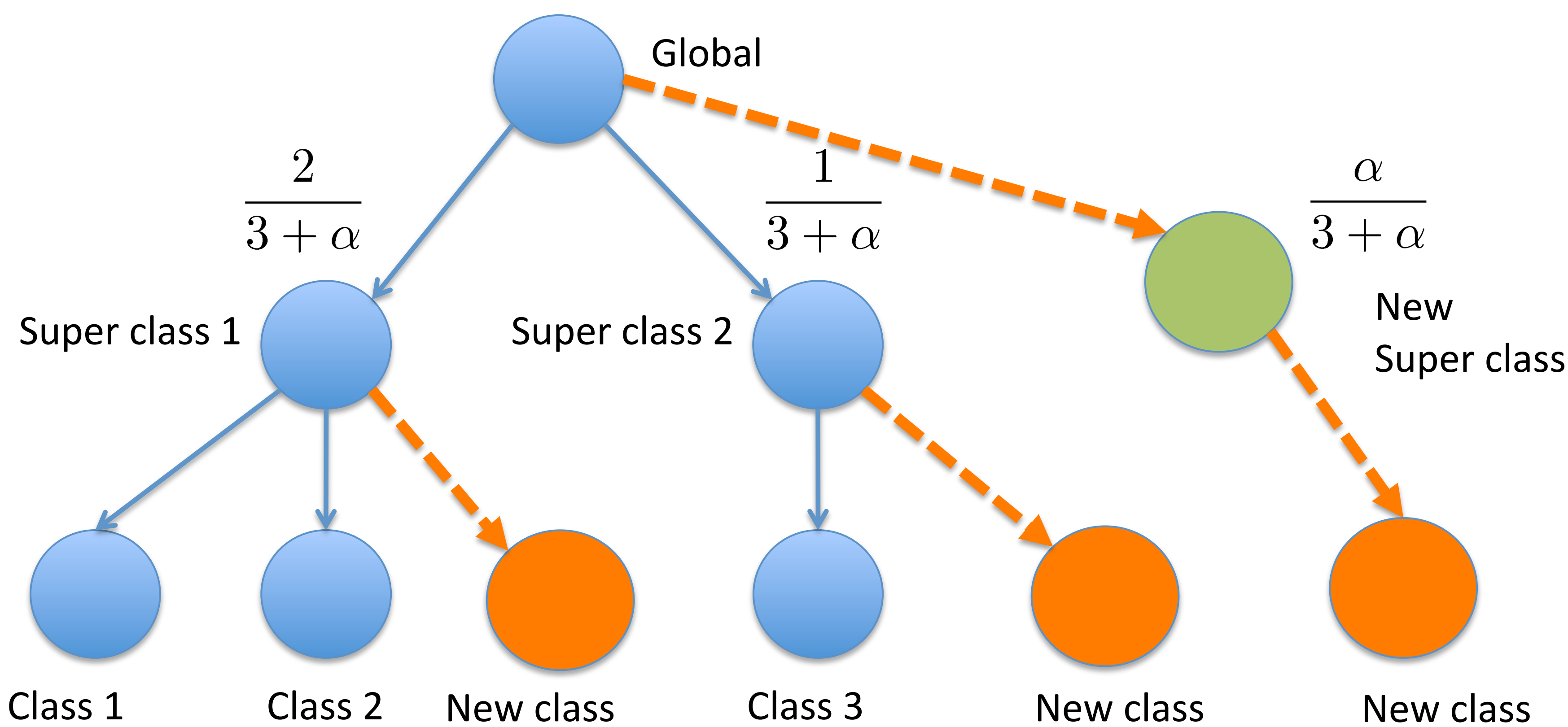
Modeling the Hierarchy



Modeling the Hierarchy



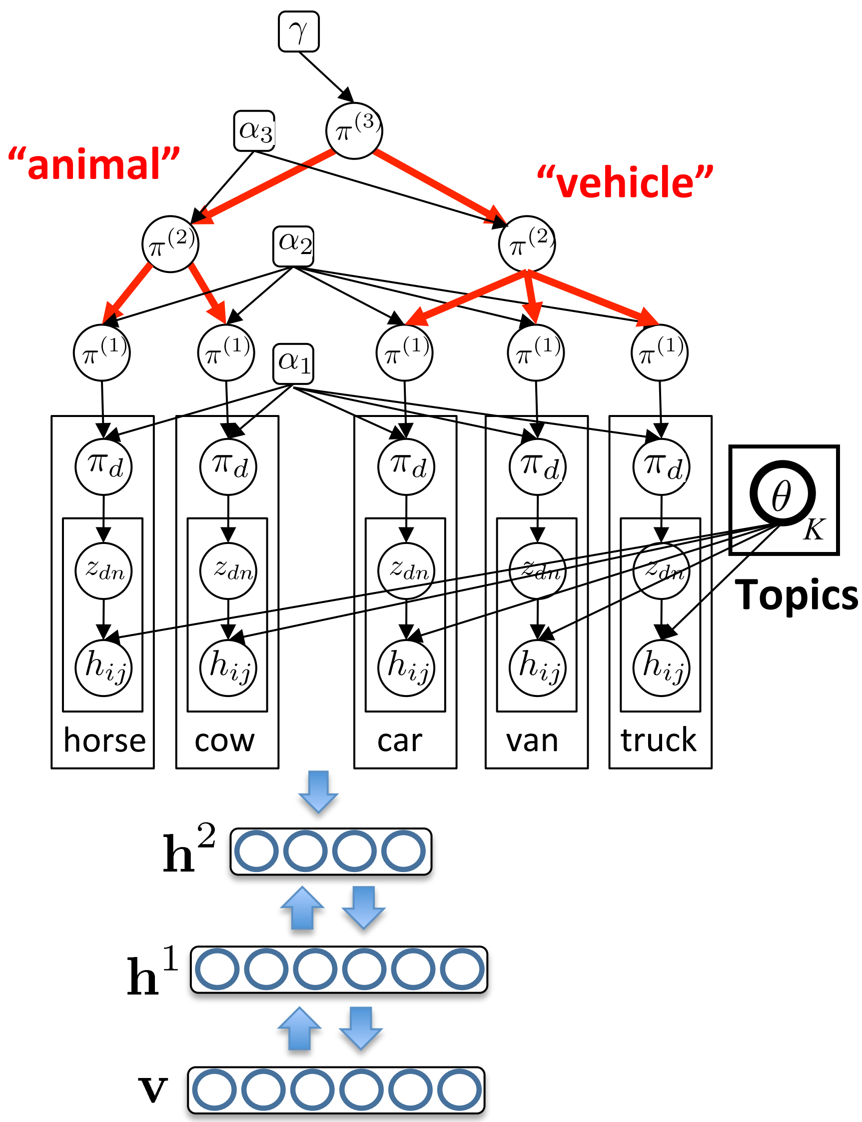
Modeling the Hierarchy



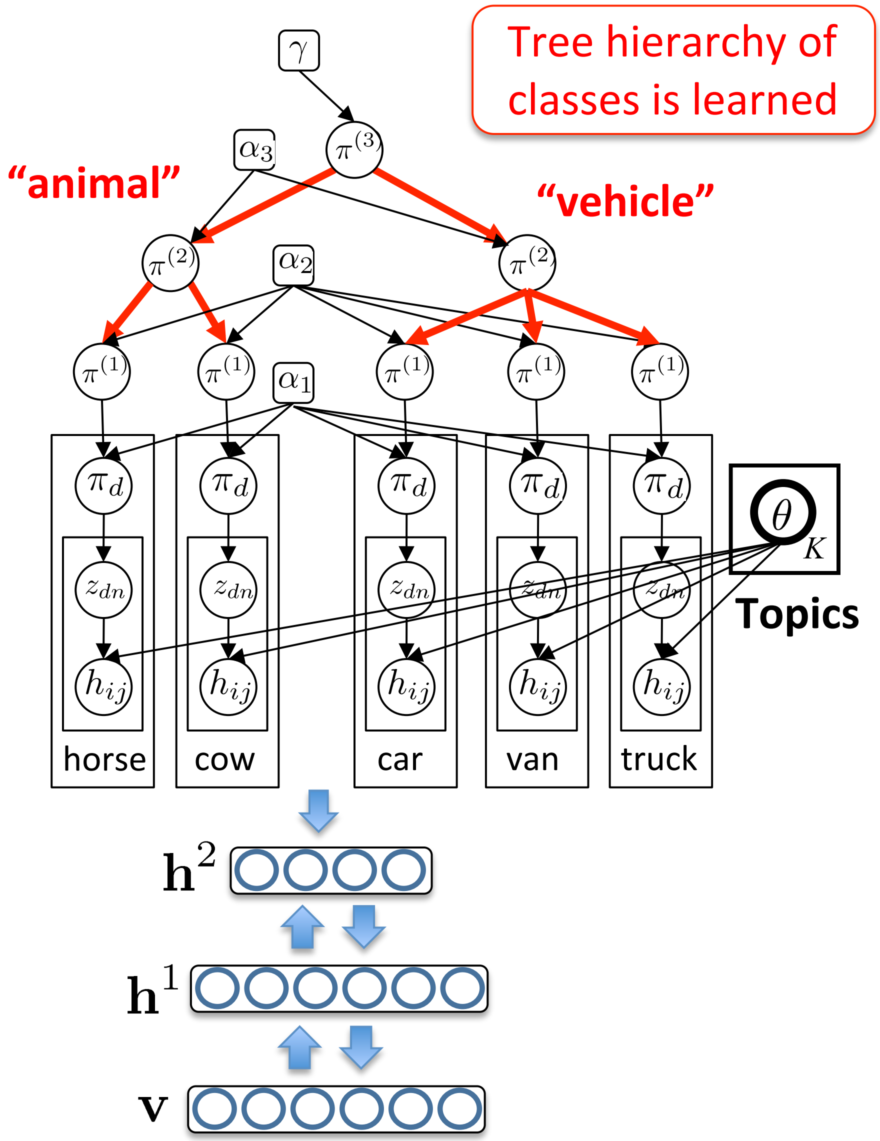
Expected number of clusters: $O(\alpha \log n)$

The nested CRP, nCRP, extends CRP to nested sequence of partitions, one for each level of the tree (Blei et.al. NIPS 2003).

Hierarchical Deep Model

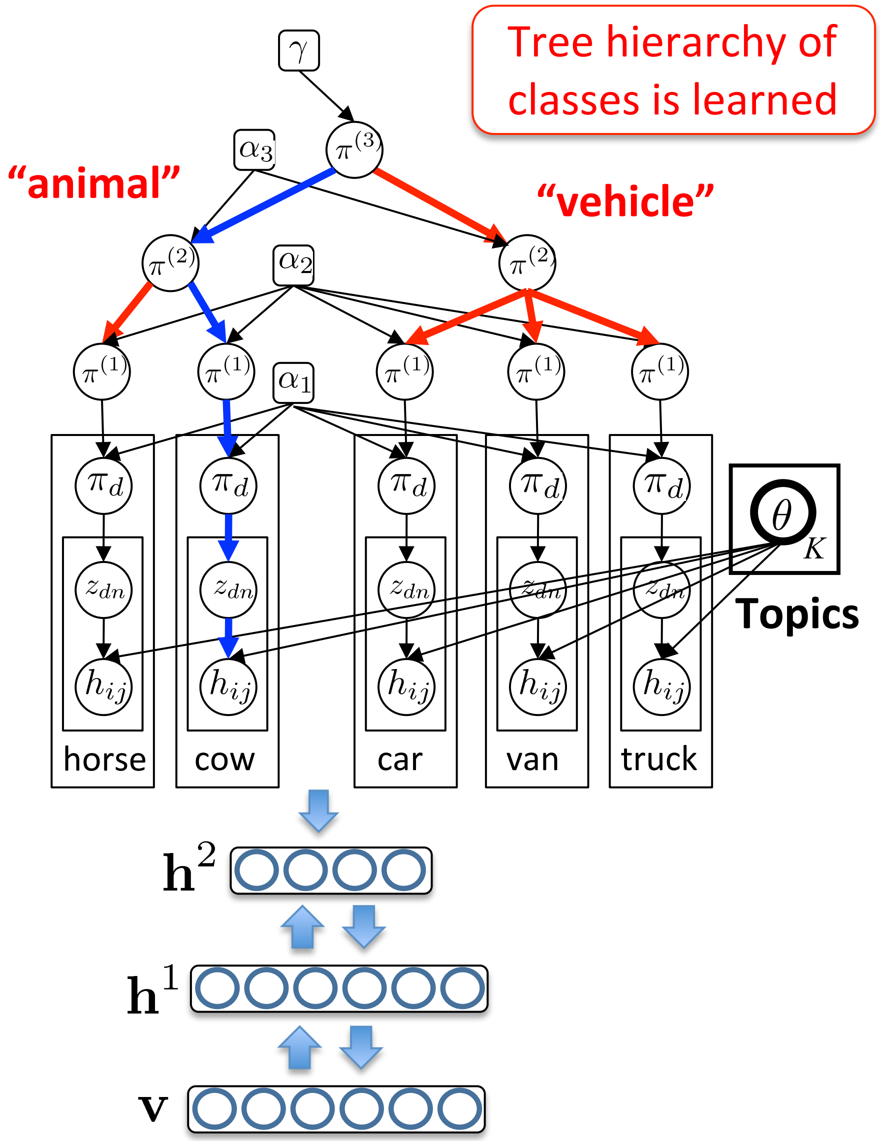


Hierarchical Deep Model



$\mathbf{z} \sim \mathrm{nCRP}$ (**Nested Chinese Restaurant Process**)
prior: a nonparametric prior over tree structures.

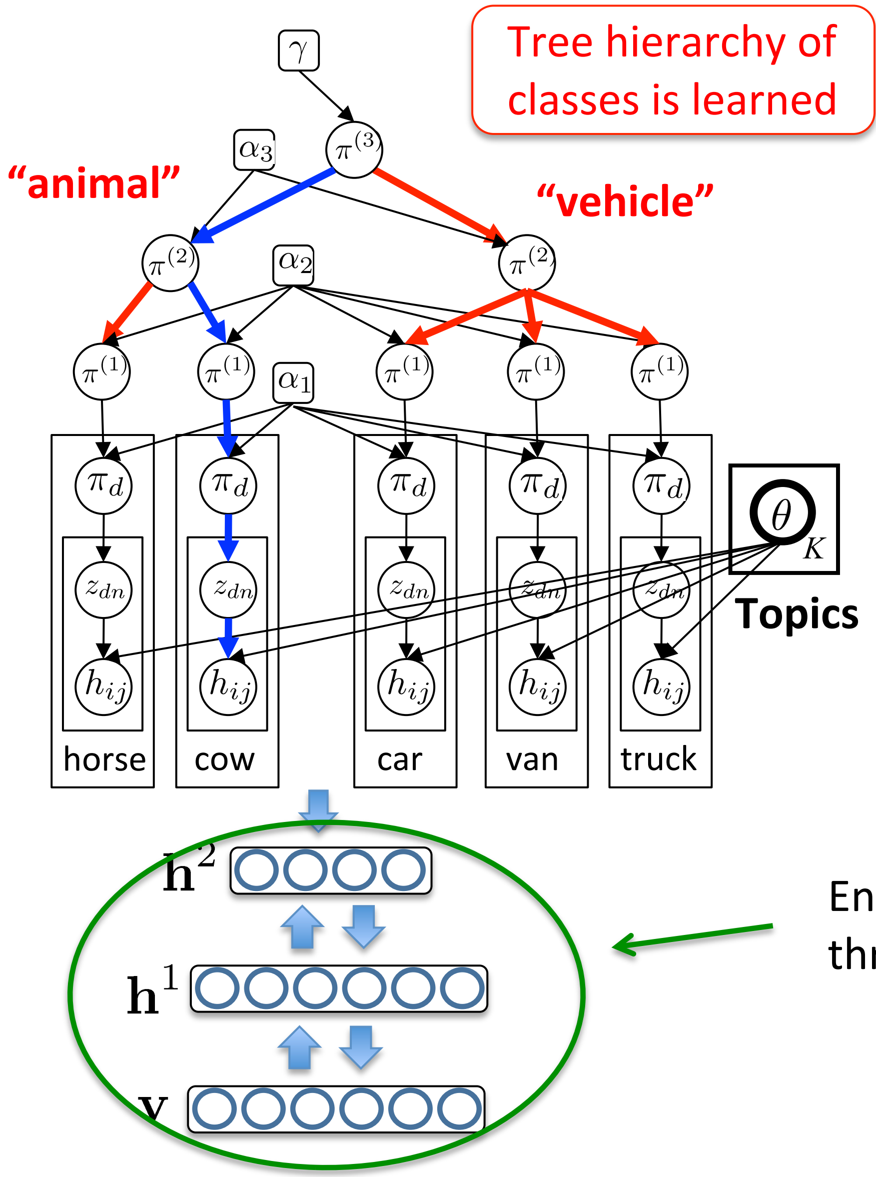
Hierarchical Deep Model



$\mathbf{z} \sim \text{nCRP}$ (**Nested Chinese Restaurant Process**)
prior: a nonparametric prior over tree structures.

$\mathbf{h}^3 | \mathbf{z} \sim \text{HDP}$(**Hierarchical Dirichlet Process**) prior:
a nonparametric prior allowing categories to share higher-level features, or parts.

Hierarchical Deep Model



$\mathbf{z} \sim \mathrm{nCRP}$ (**Nested Chinese Restaurant Process**)
prior: a nonparametric prior over tree structures

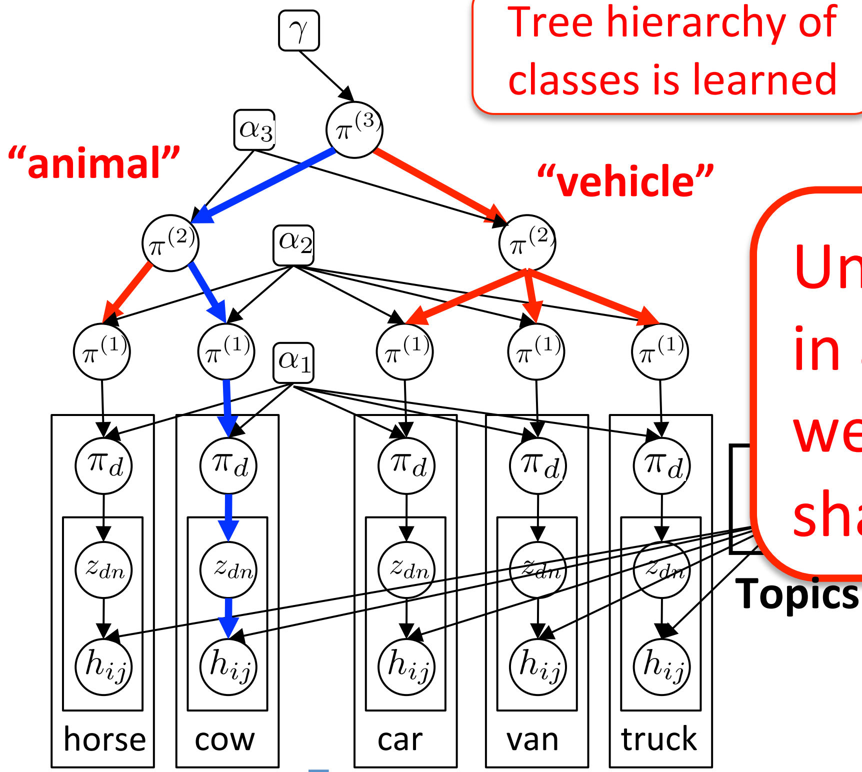
$\mathbf{h}^3|\mathbf{z} \sim \mathrm{HDP}$(**Hierarchical Dirichlet Process**) prior:
a nonparametric prior allowing categories to share higher-level features, or parts.

$\mathbf{v}|\mathbf{h}^3 \sim \mathrm{DBM}$ **Conditional Deep Boltzmann Machine.**

Enforce (approximate) global consistency through many local constraints.

Hierarchical Deep Model

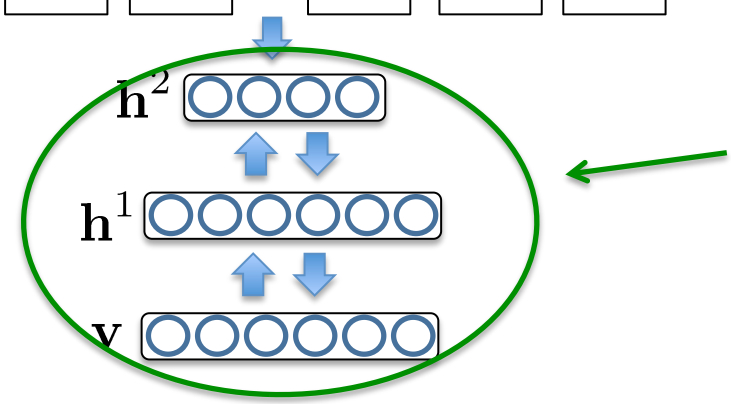
Tree hierarchy of classes is learned



Unlike standard statistical models, in addition to inferring parameters, we also infer the hierarchy for sharing those parameters.

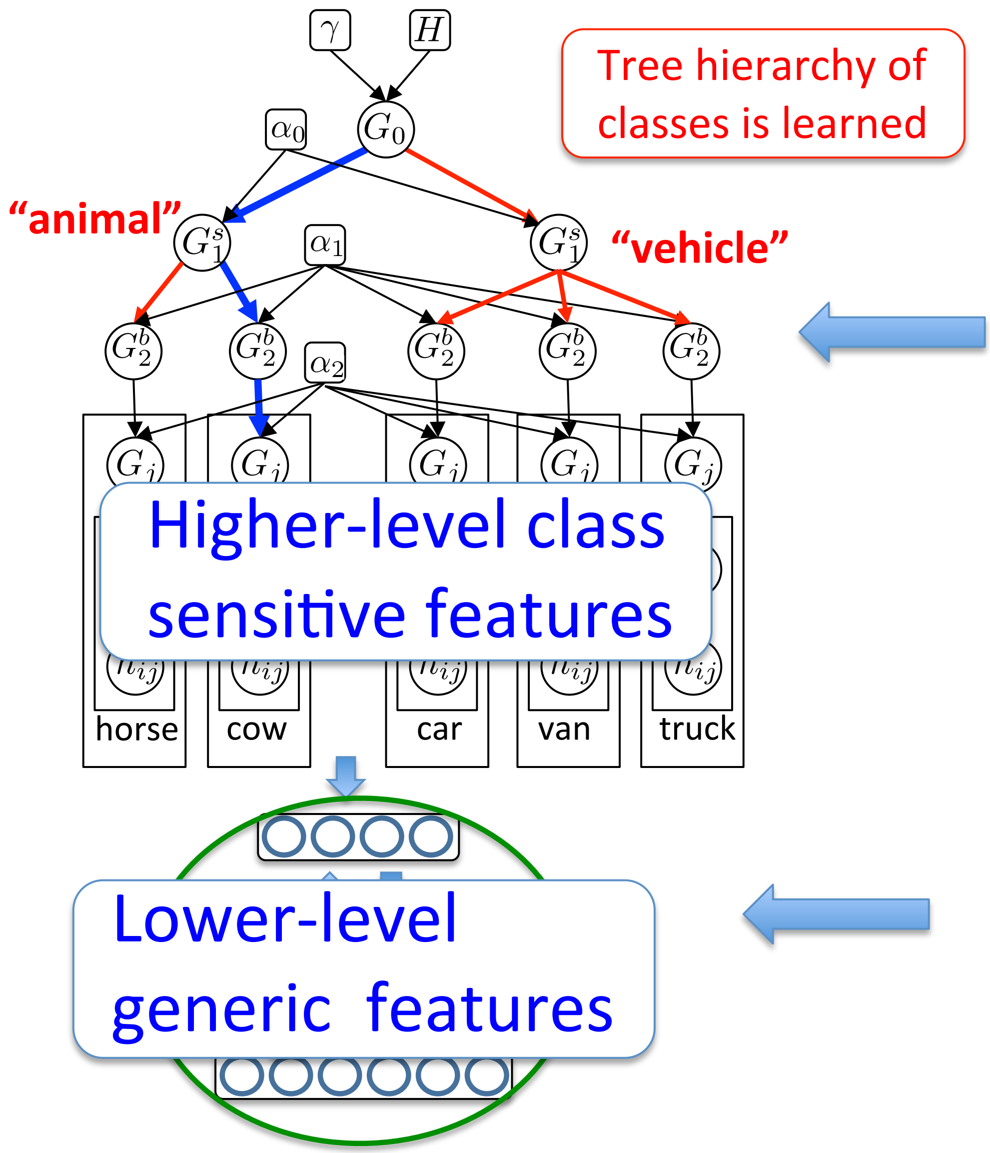
share higher-level features, or parts.

$\mathbf{v}|\mathbf{h}^3 \sim \mathrm{DBM}$ **Conditional Deep Boltzmann Machine.**

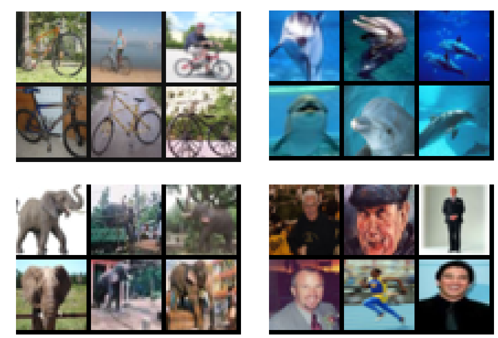


Enforce (approximate) global consistency through many local constraints.

CIFAR Object Recognition

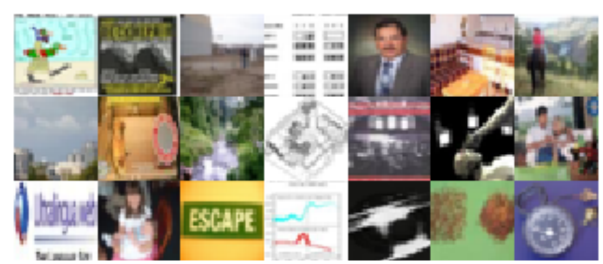


50,000 images of 100 classes



Inference: Markov chain Monte Carlo – Later!

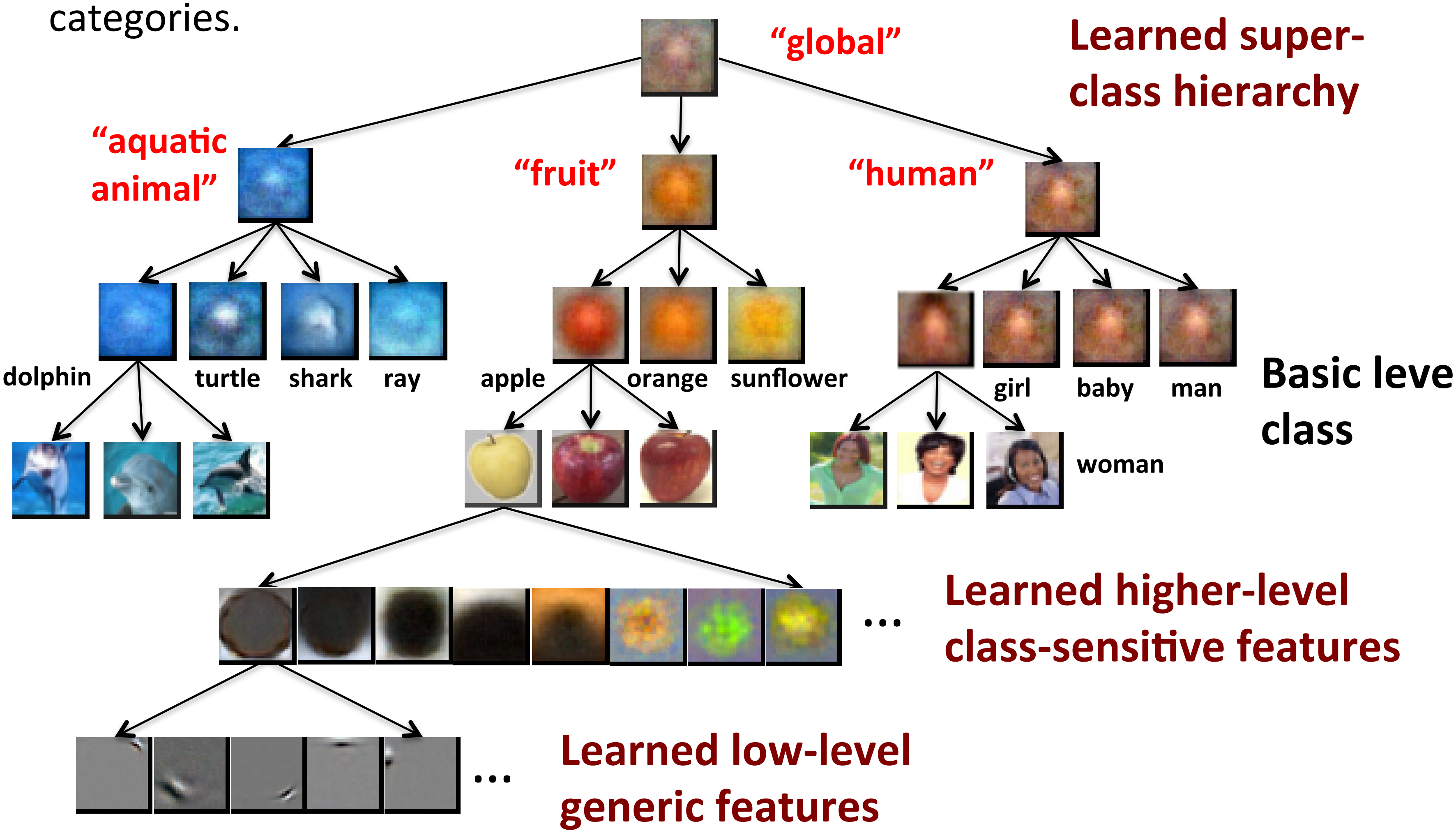
4 million unlabeled images



32 x 32 pixels x 3 RGB

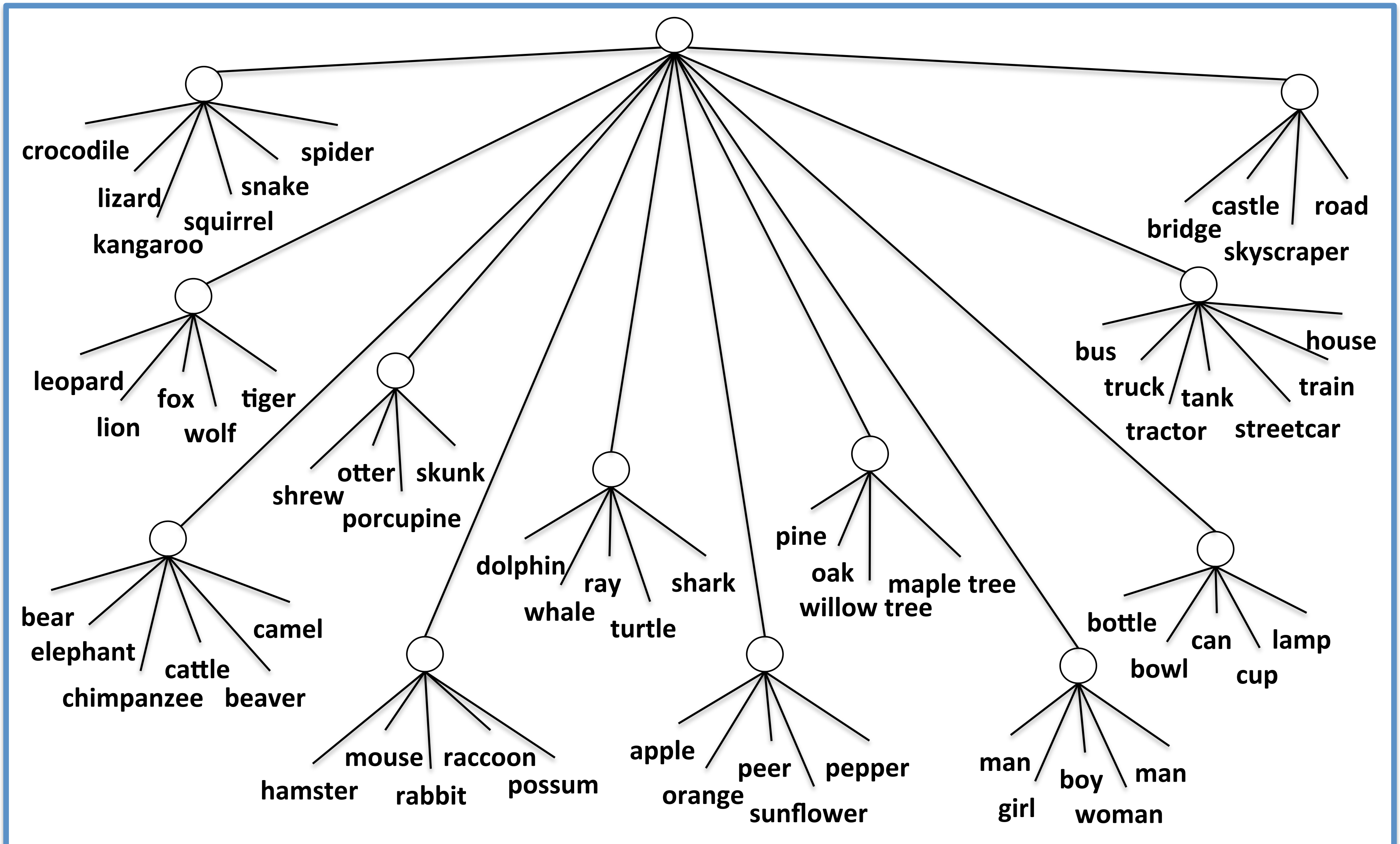
Learning to Learn

The model learns how to share the knowledge across many visual categories.

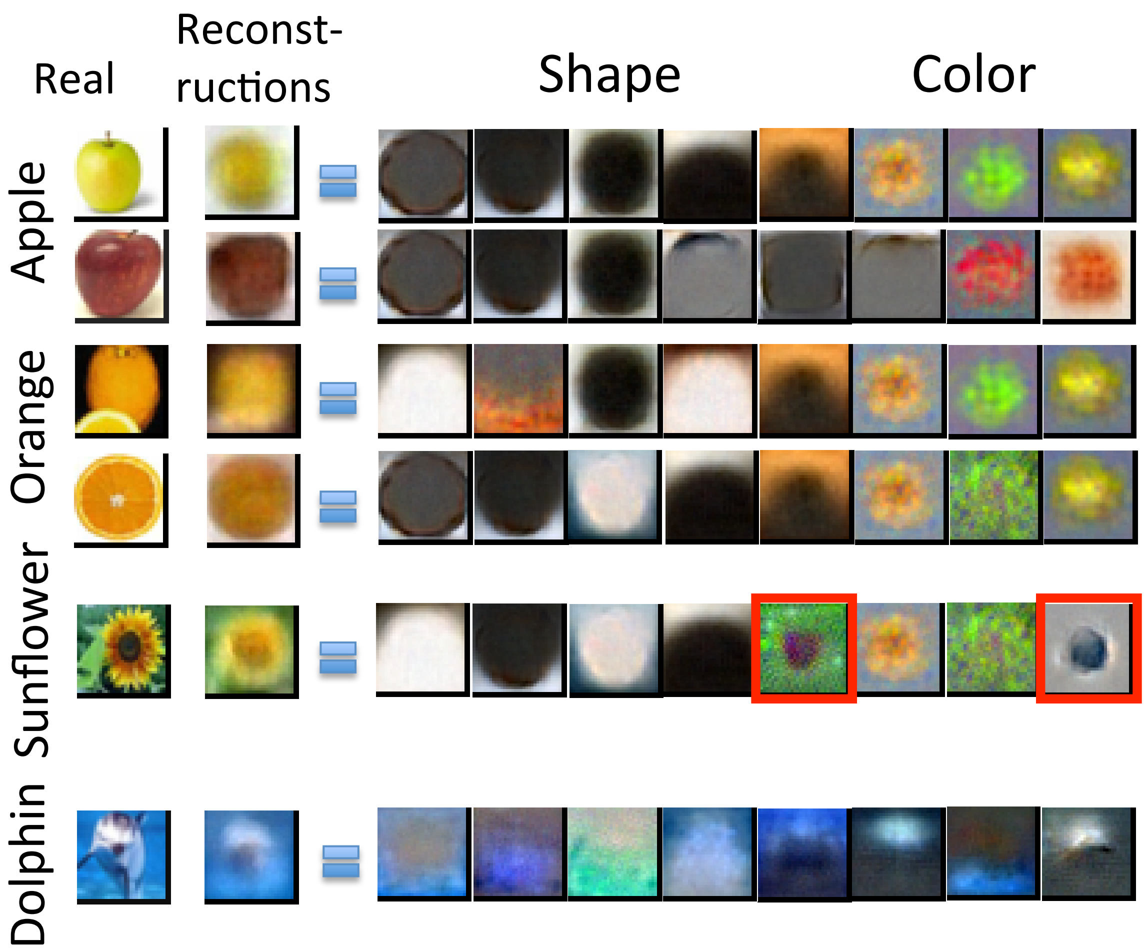


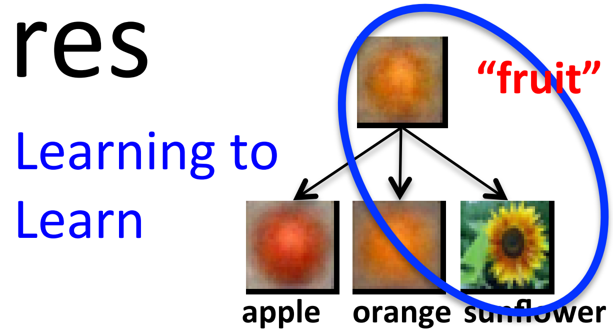
Learning to Learn

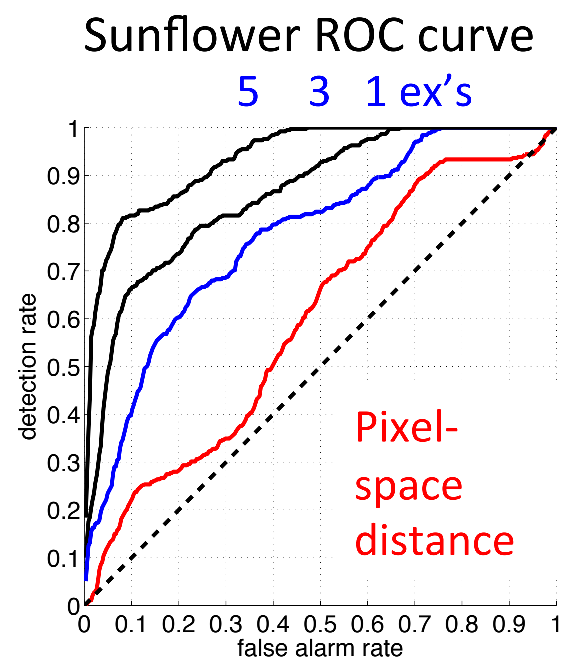
The model learns how to share the knowledge across many visual



Sharing Features



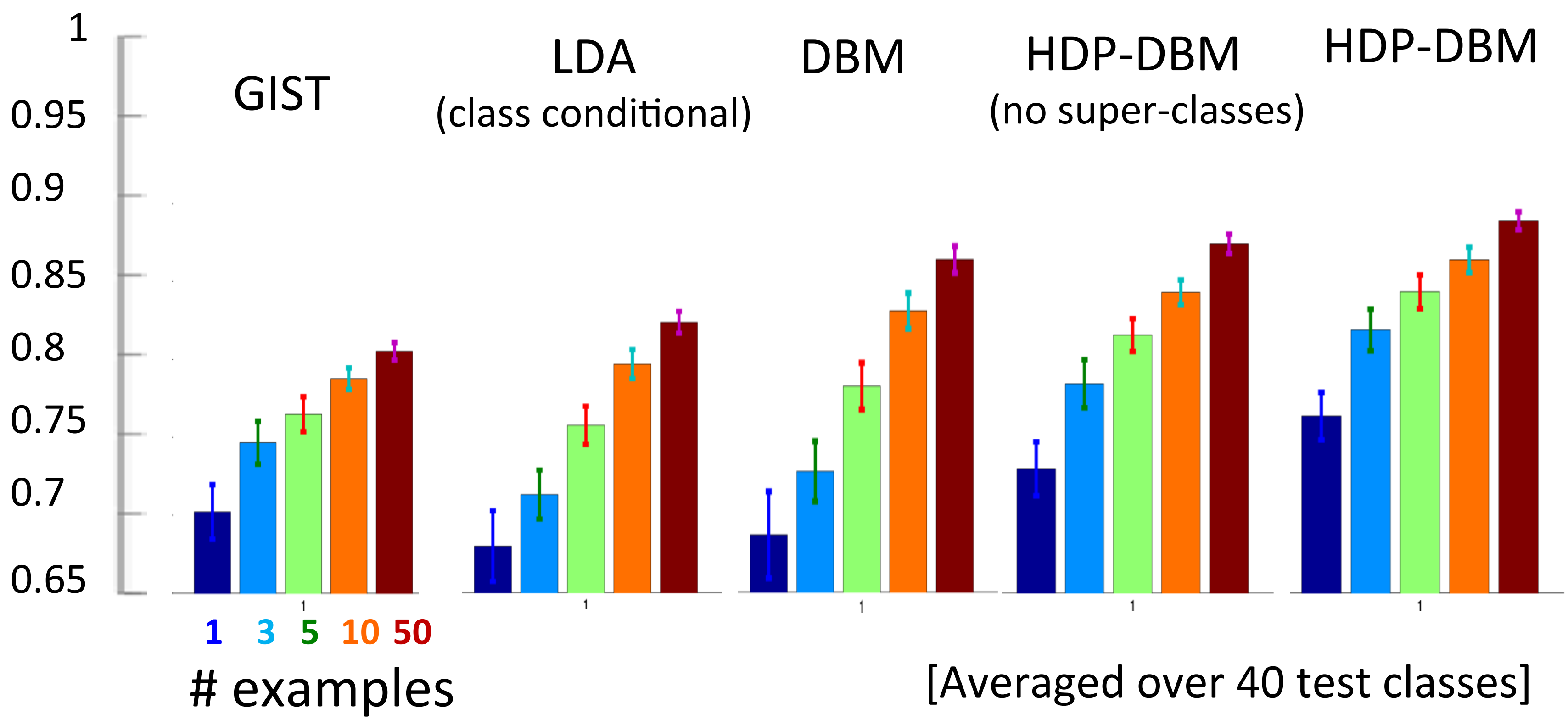




Learning to Learn: Learning a hierarchy for sharing parameters – rapid learning of a novel concept.

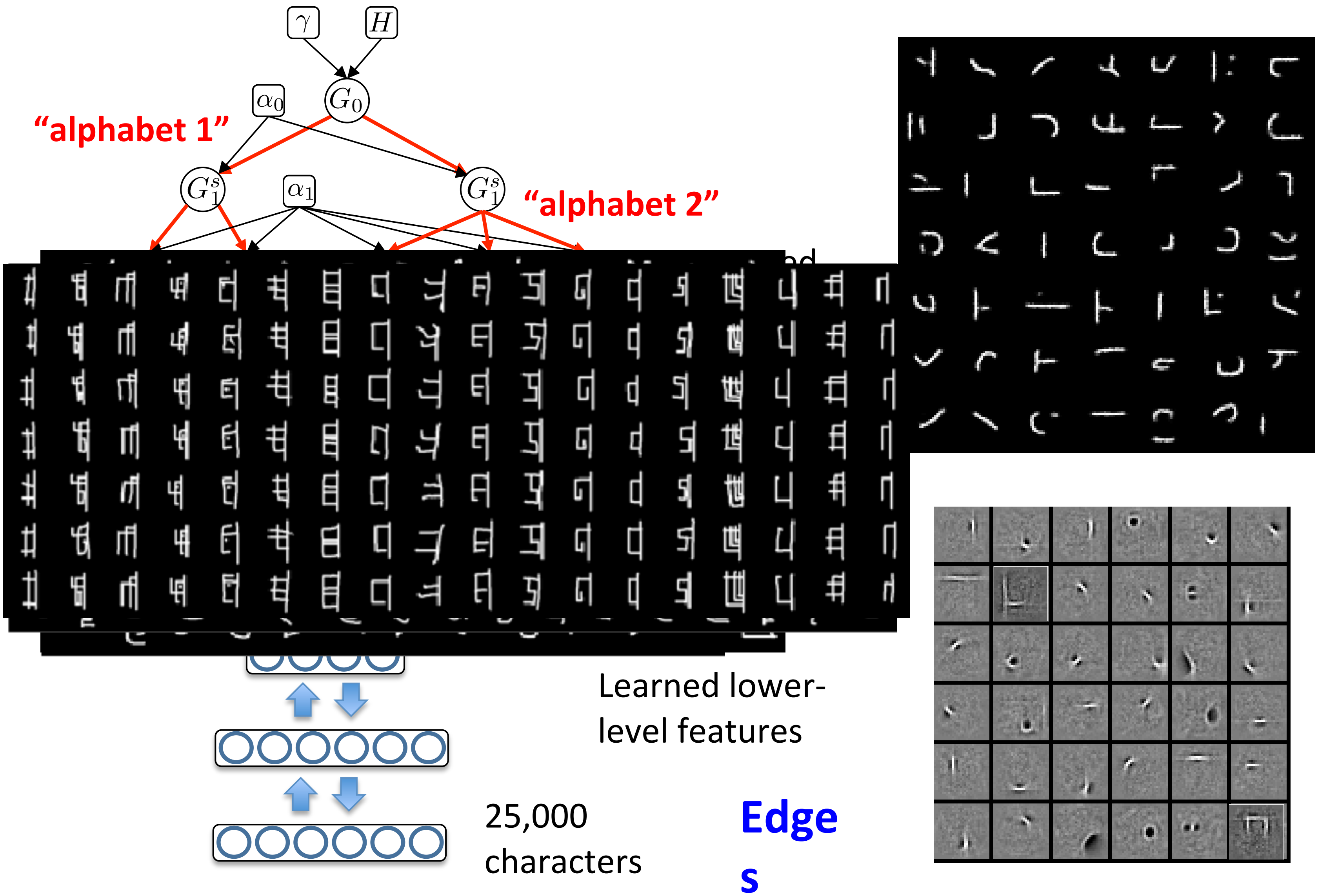
Object Recognition

Area under ROC curve for same/different
(1 new class vs. 99 distractor classes)

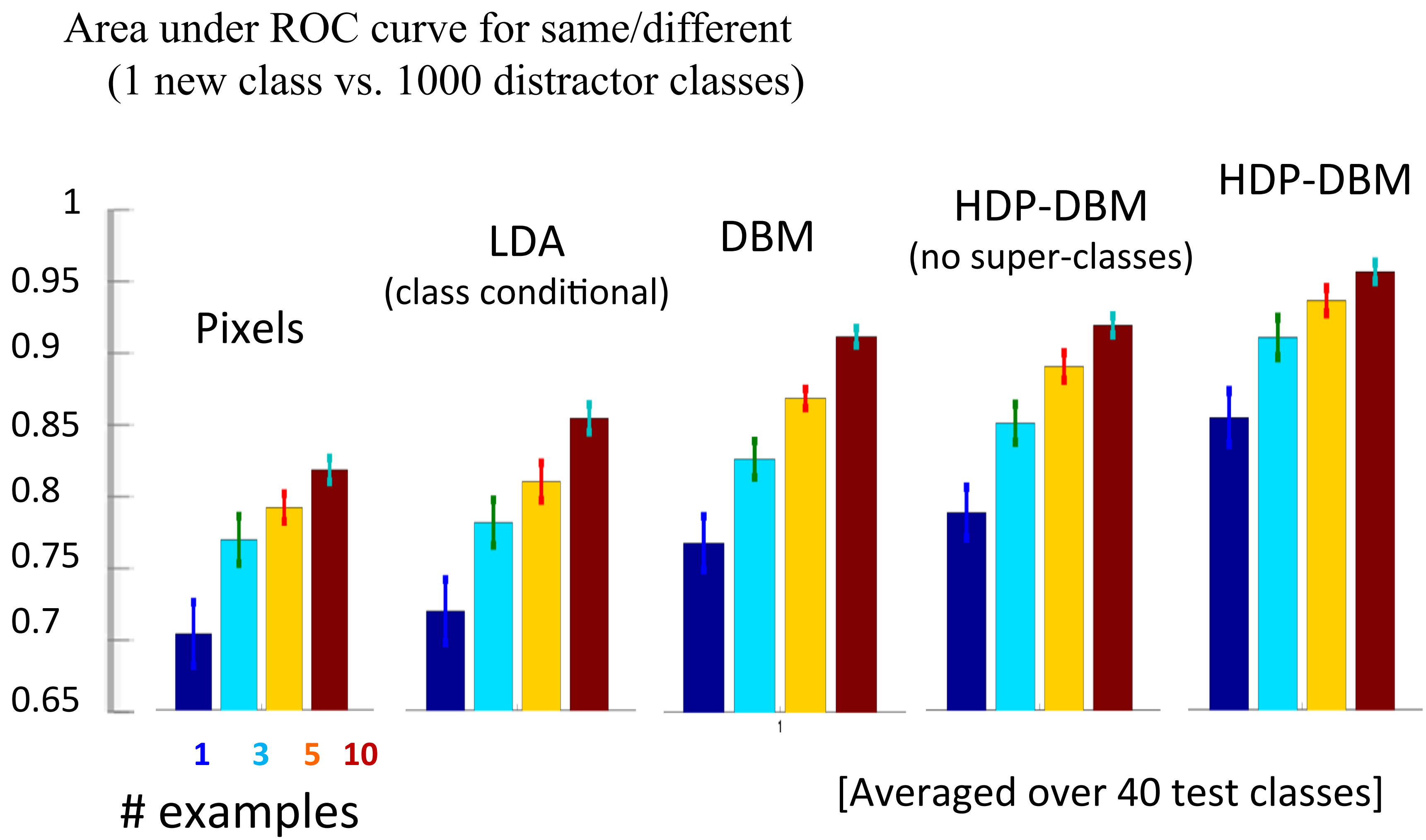


Our model outperforms standard computer vision features (e.g. GIST).

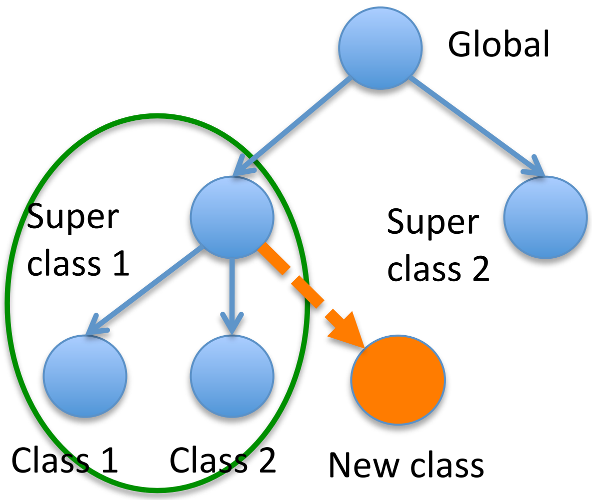
Handwritten Character Recognition



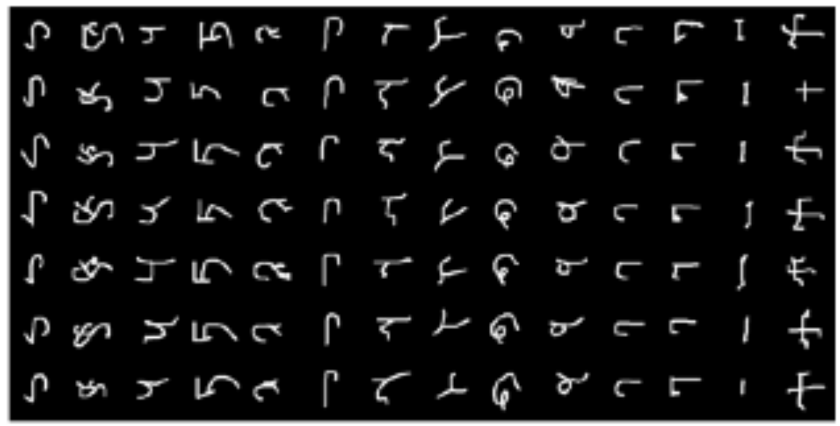
Handwritten Character Recognition



Simulating New Characters

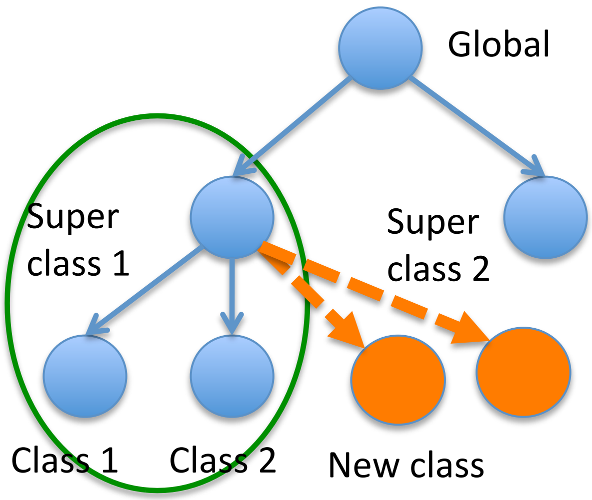


Real data within super class

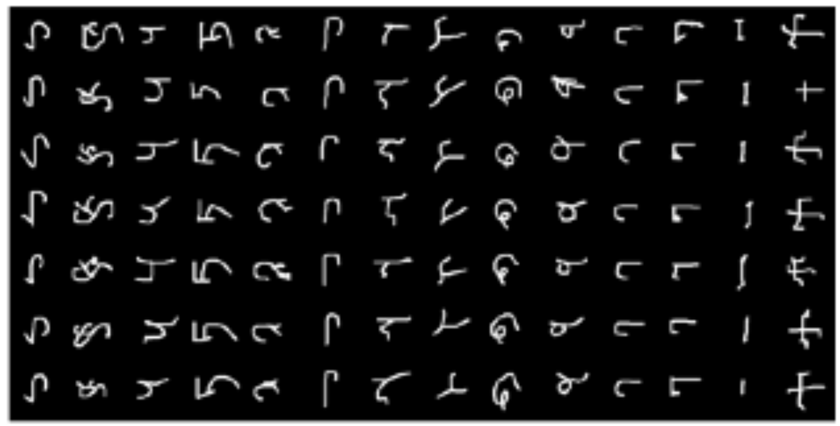


Simulated new characters

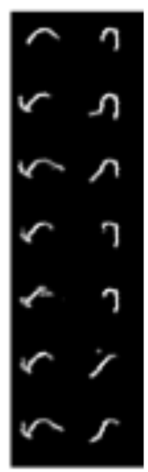
Simulating New Characters



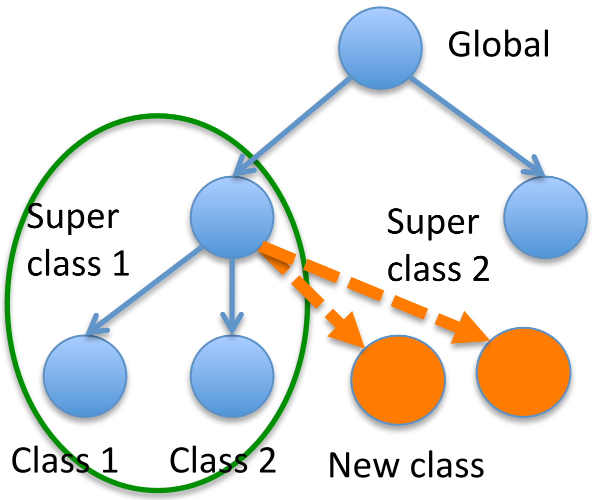
Real data within super class



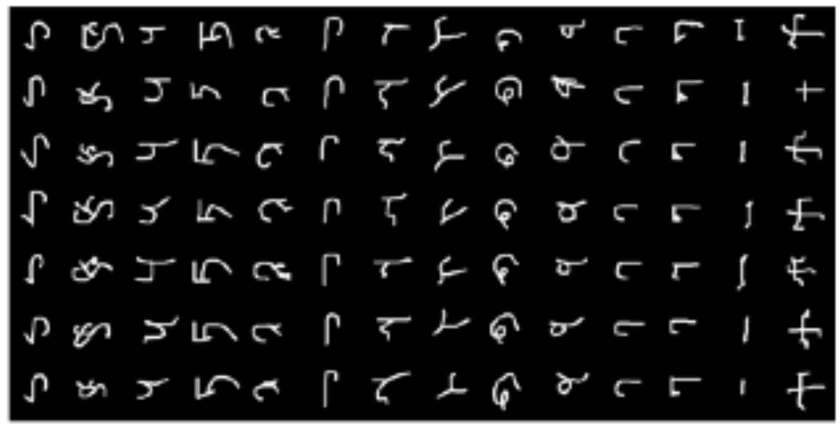
Simulated new characters



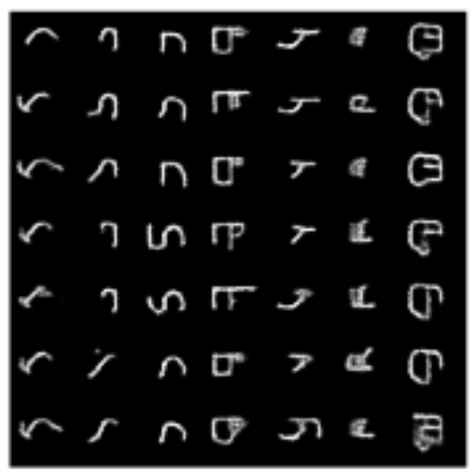
Simulating New Characters



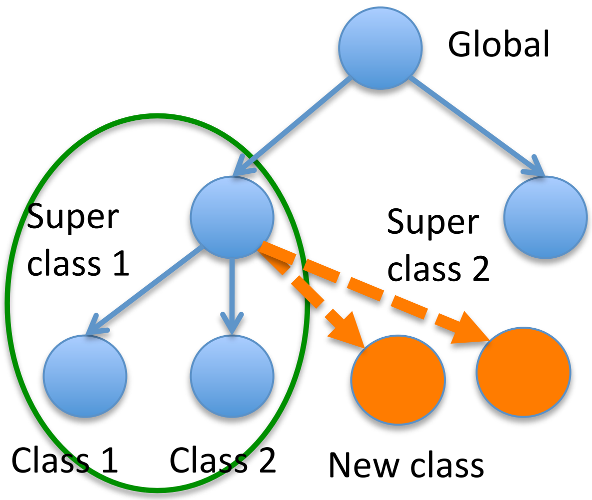
Real data within super class



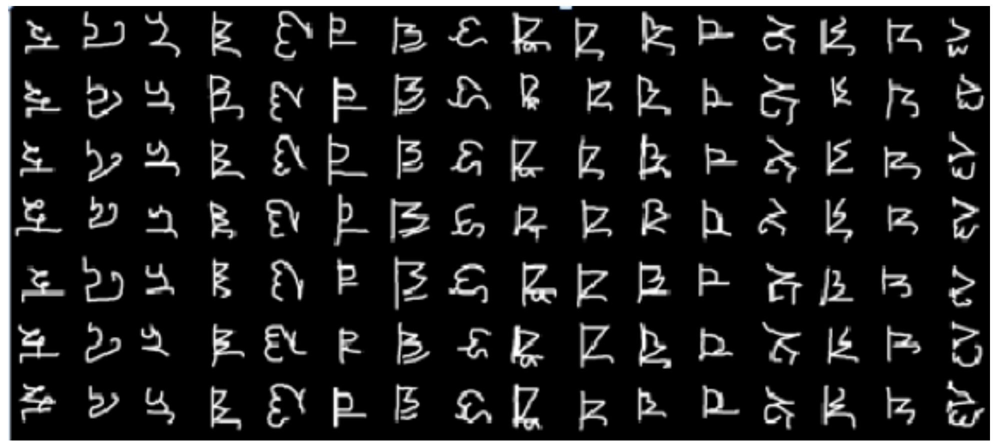
Simulated new characters



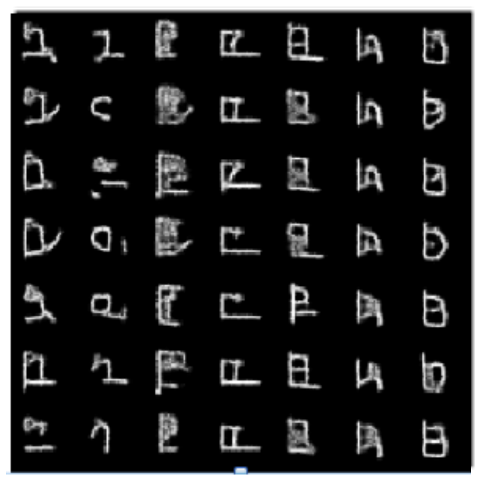
Simulating New Characters



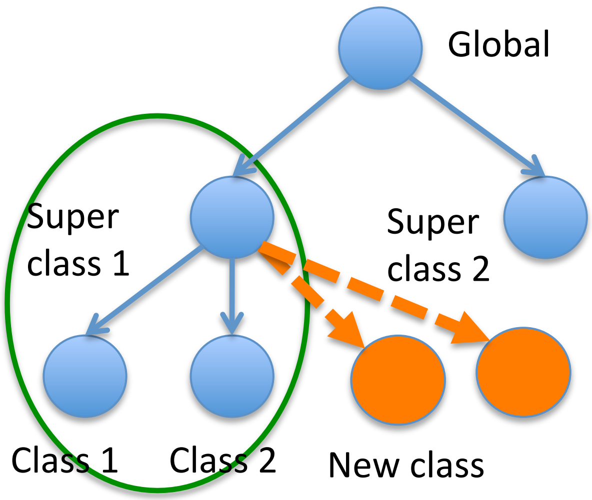
Real data within super class



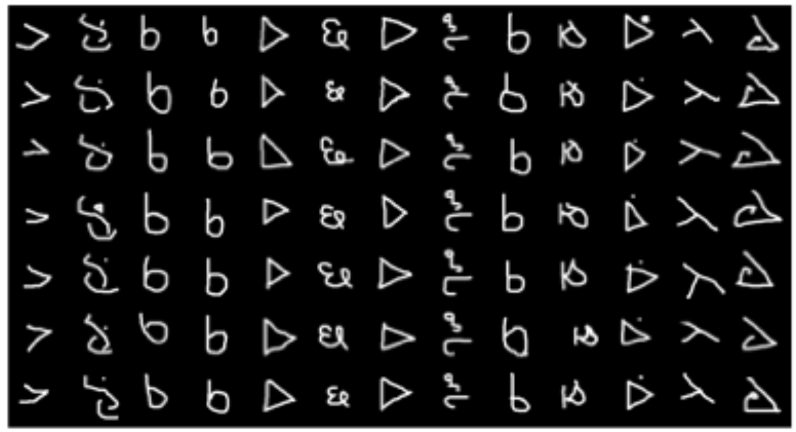
Simulated new characters



Simulating New Characters

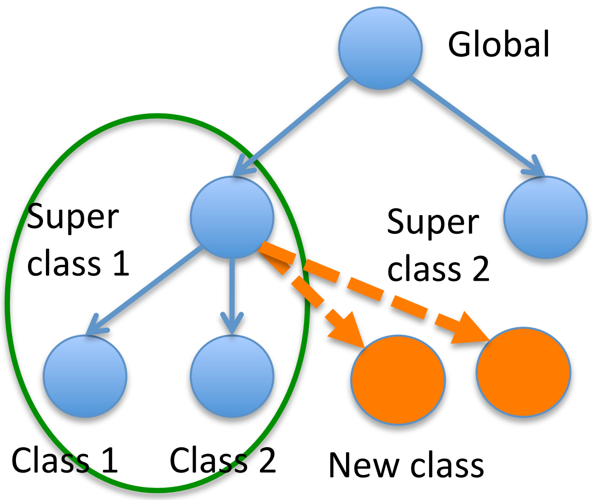


Real data within super class

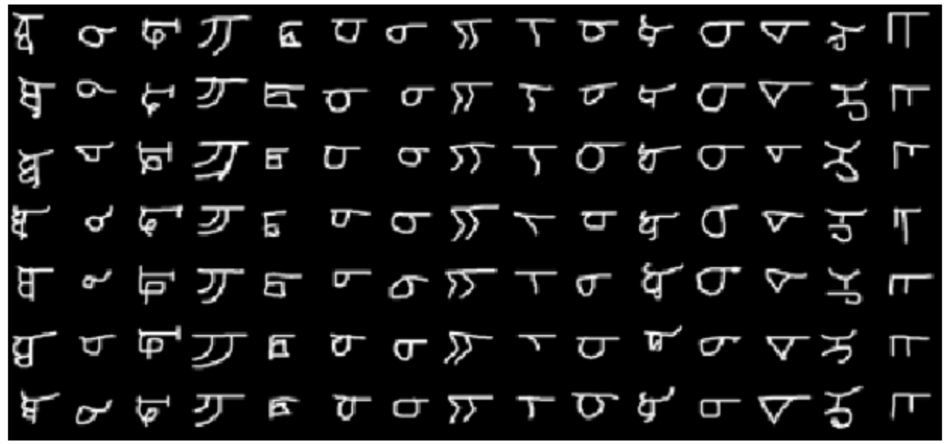


Simulated new characters

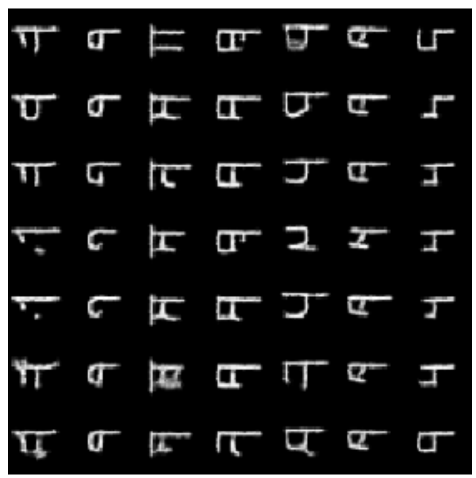
Simulating New Characters



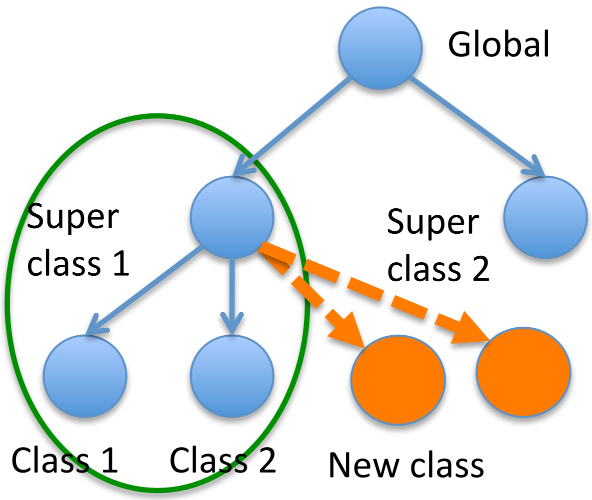
Real data within super class



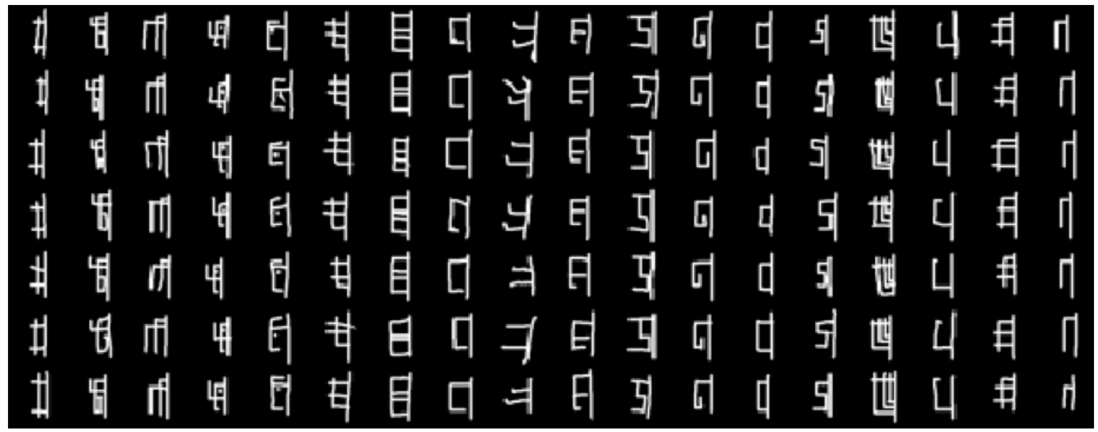
Simulated new characters



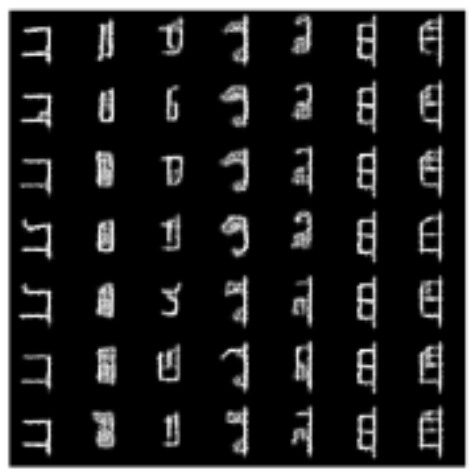
Simulating New Characters



Real data within super class



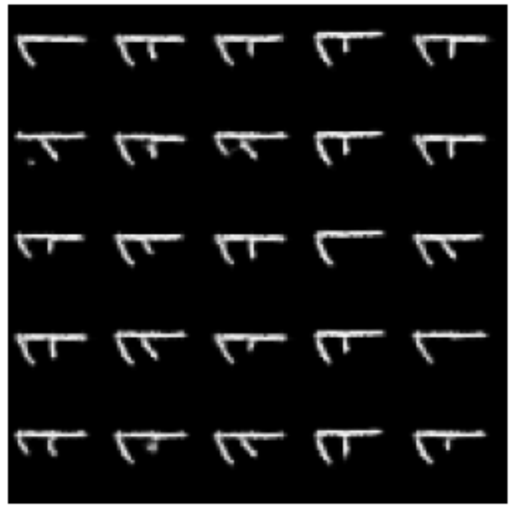
Simulated new characters



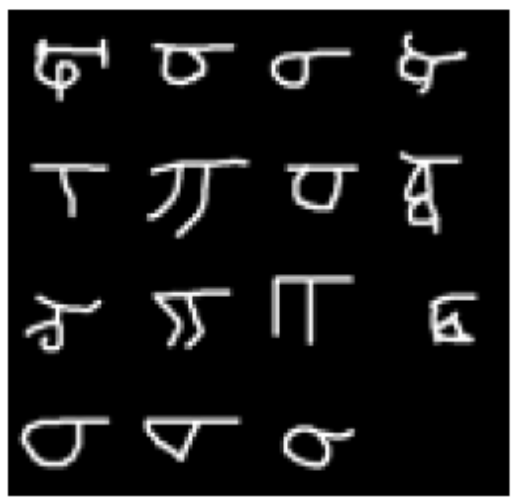
Learning from very few examples

3 examples of
a new class

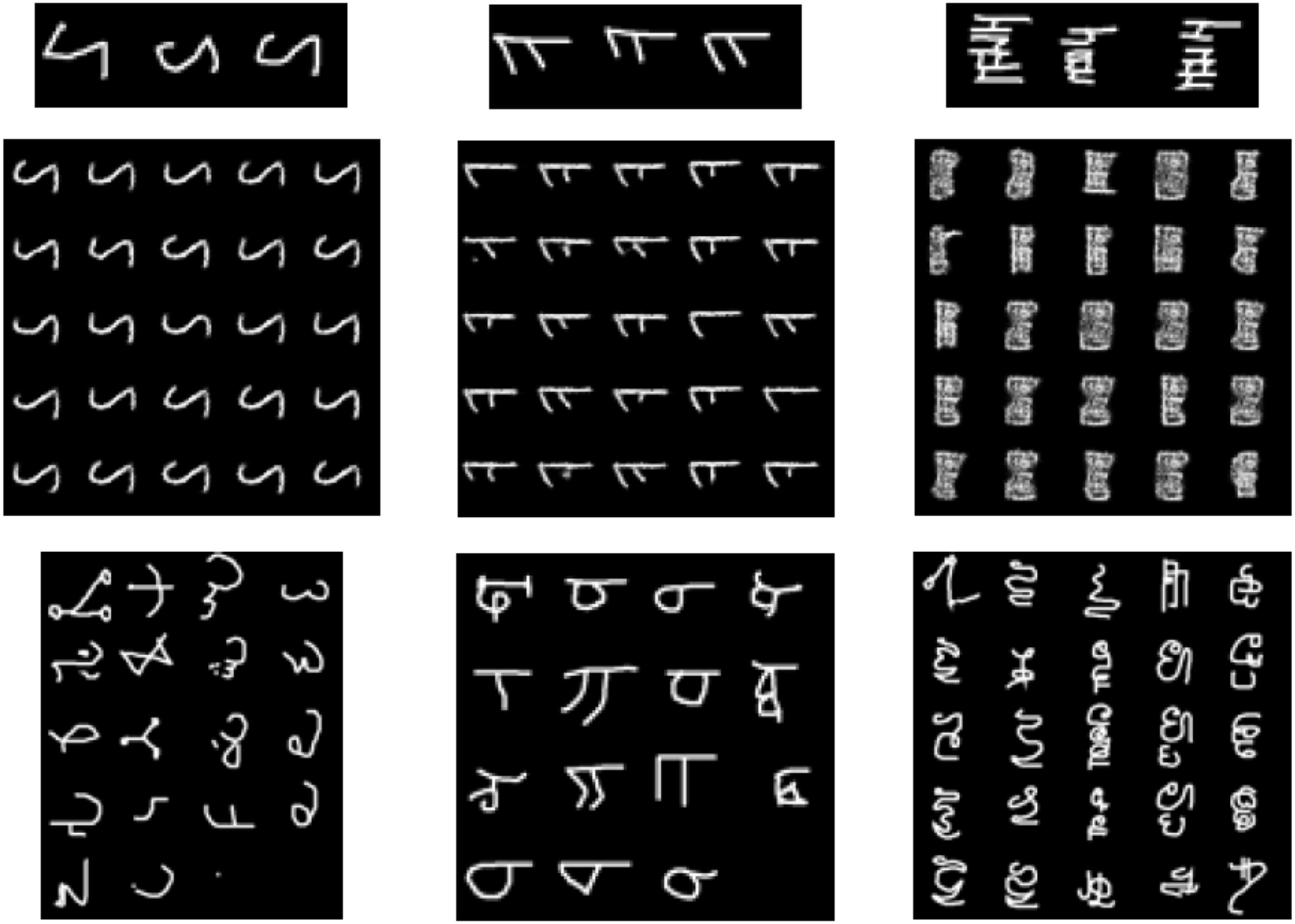
Conditional samples
in the same class



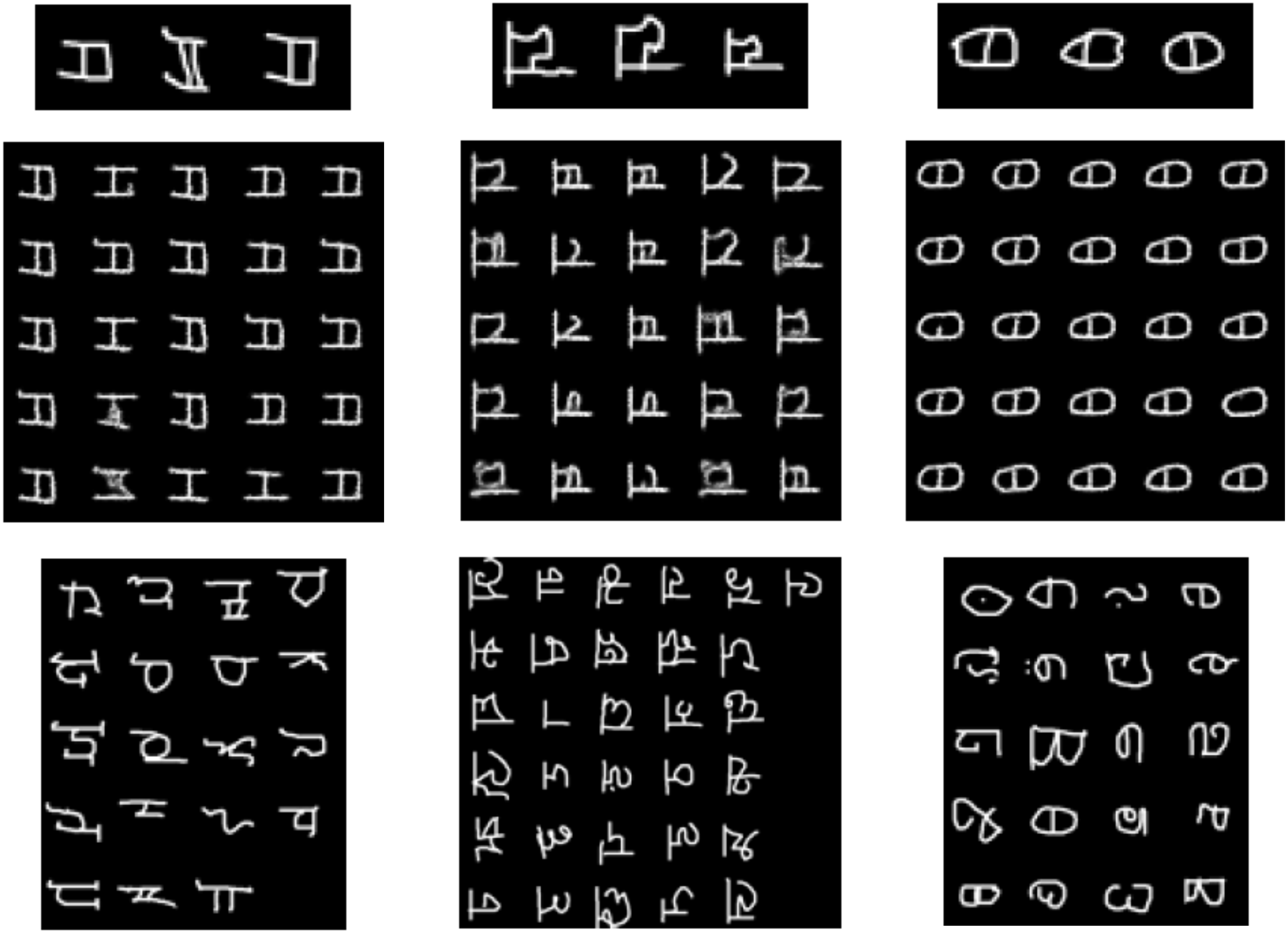
Inferred super-class



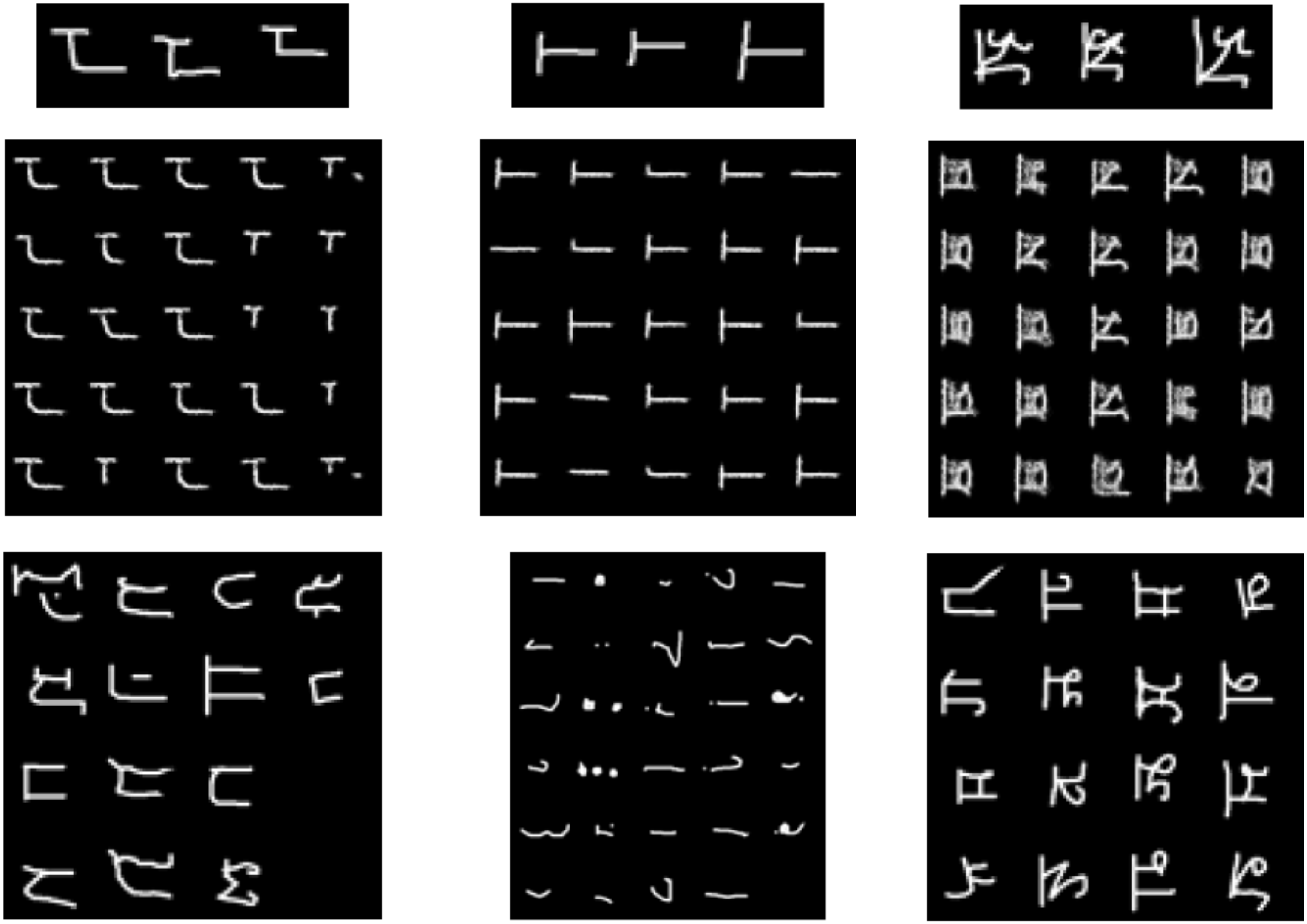
Learning from very few examples



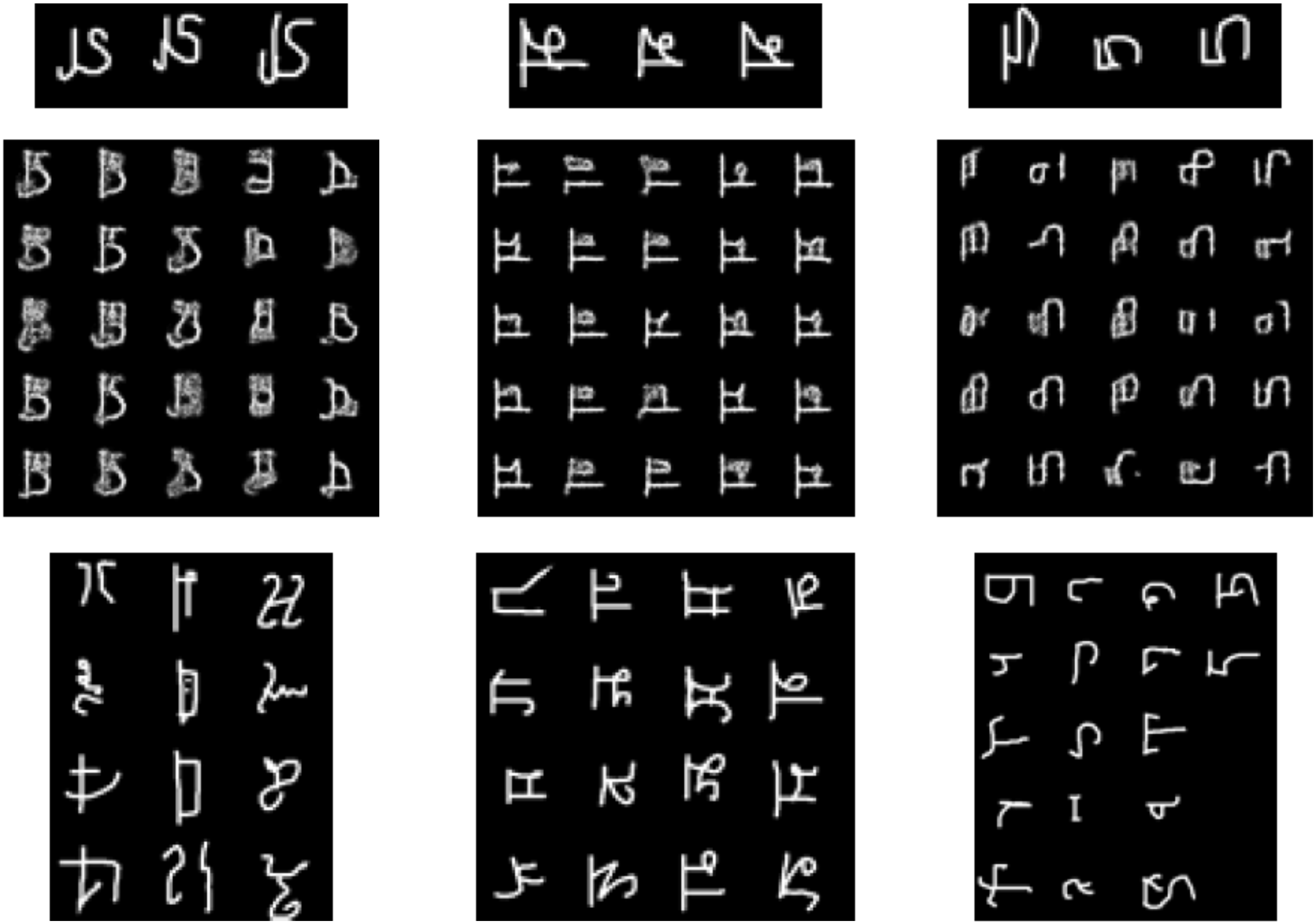
Learning from very few examples



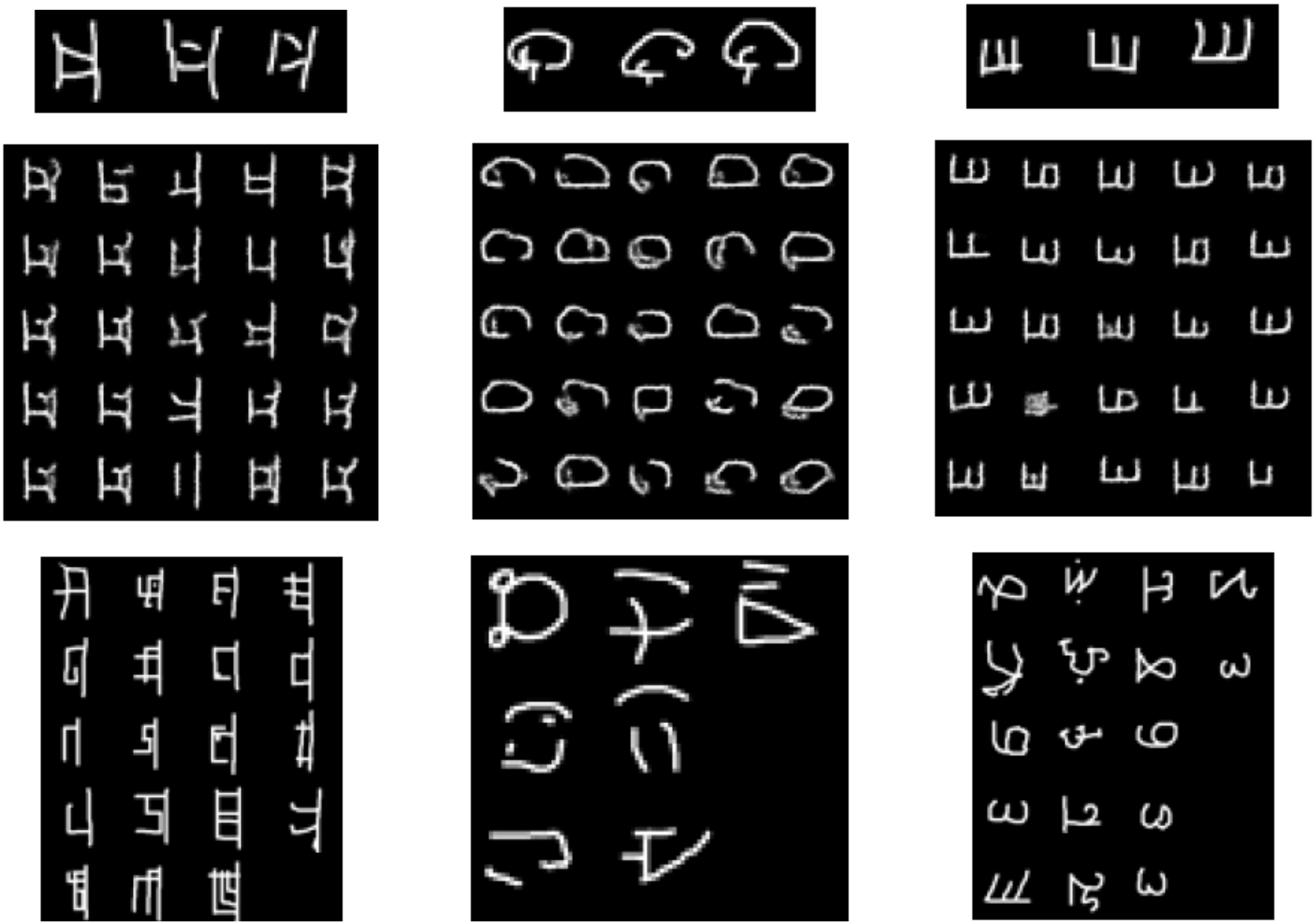
Learning from very few examples



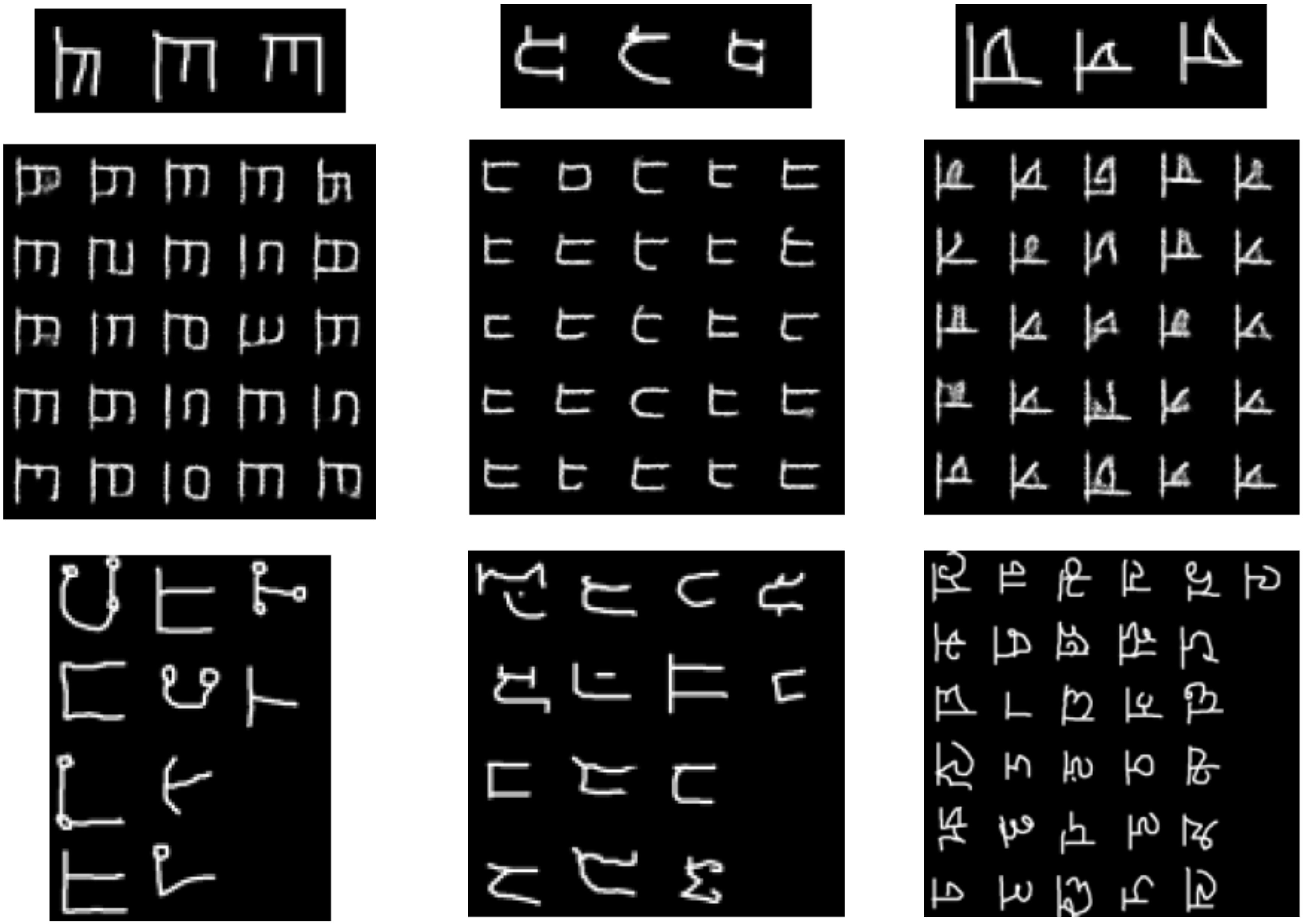
Learning from very few examples



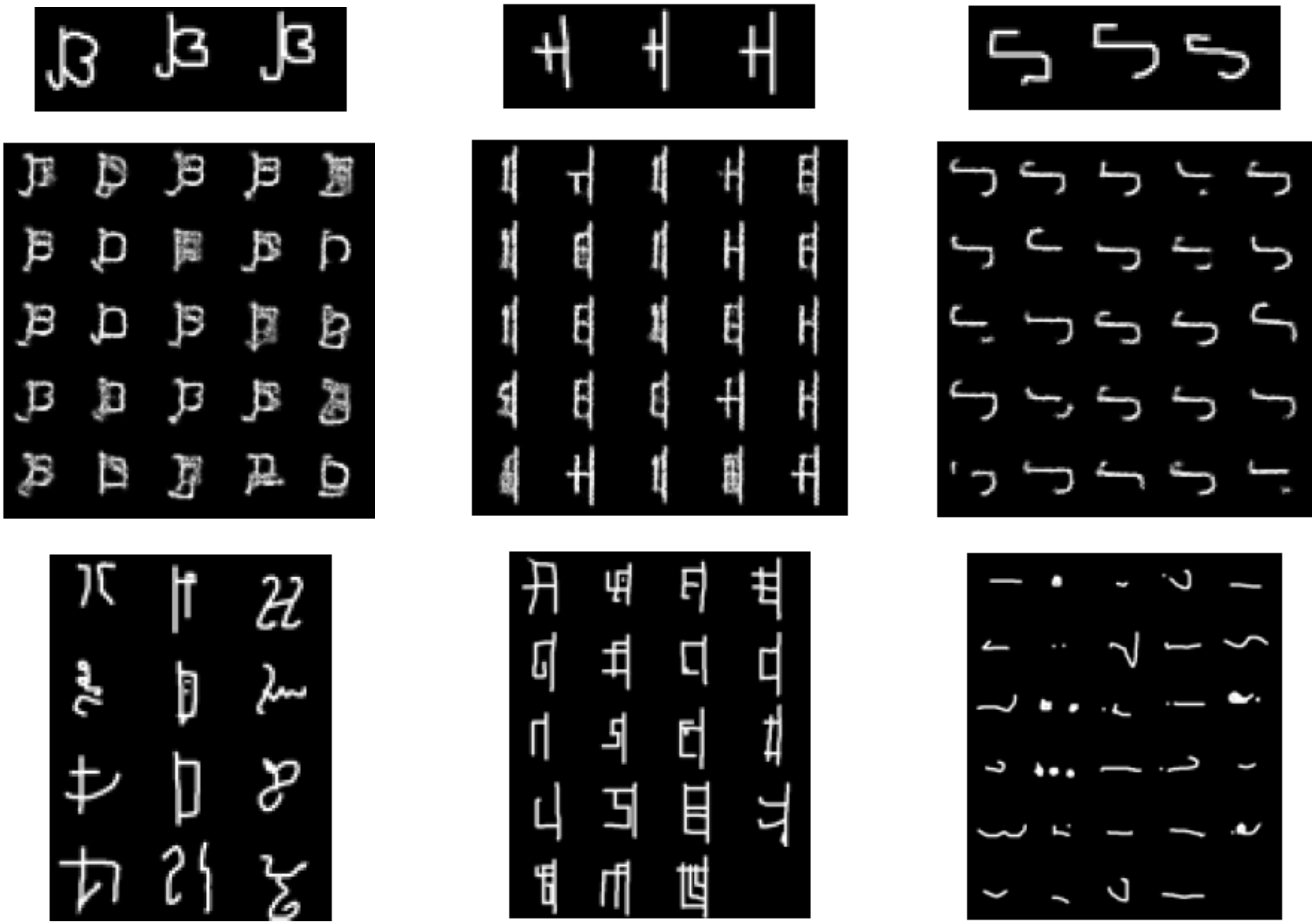
Learning from very few examples



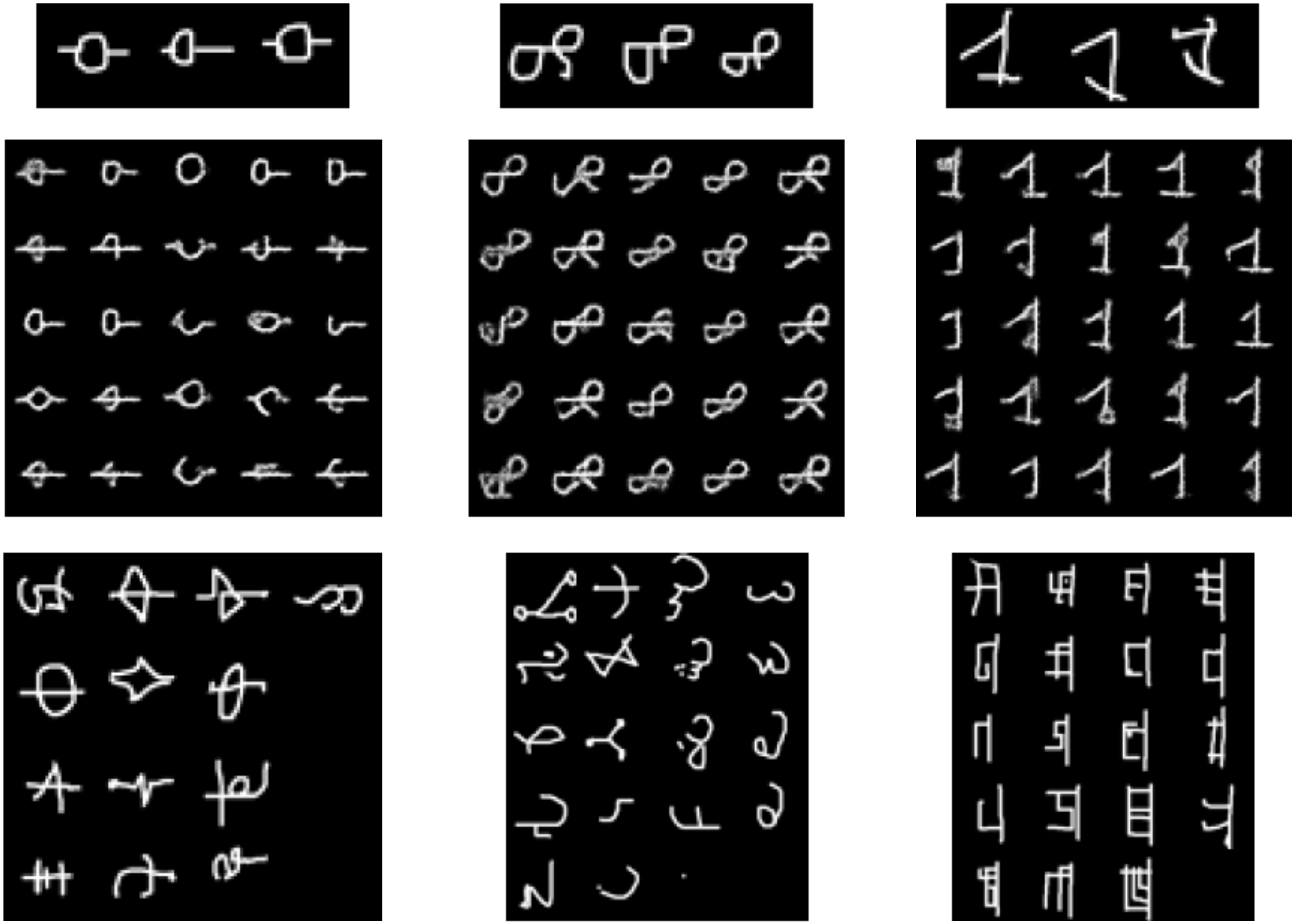
Learning from very few examples



Learning from very few examples

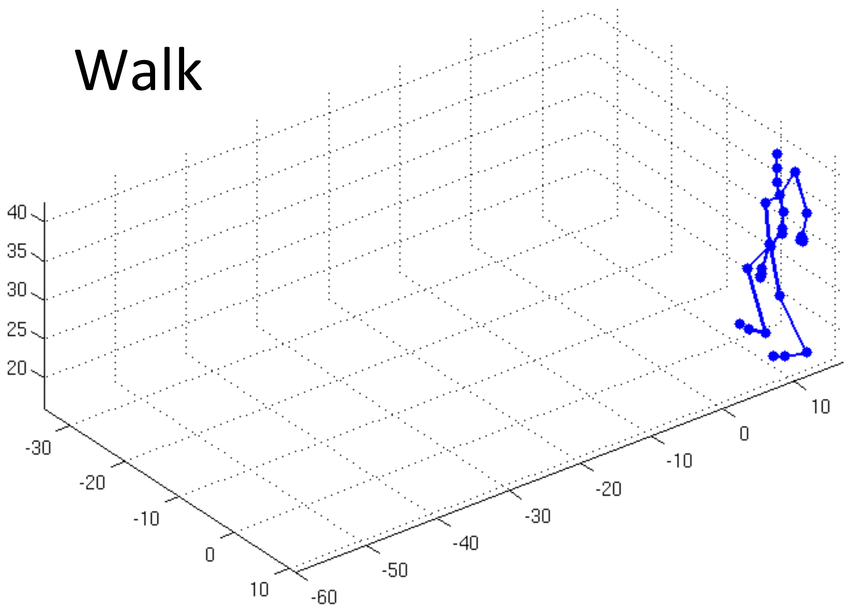


Learning from very few examples

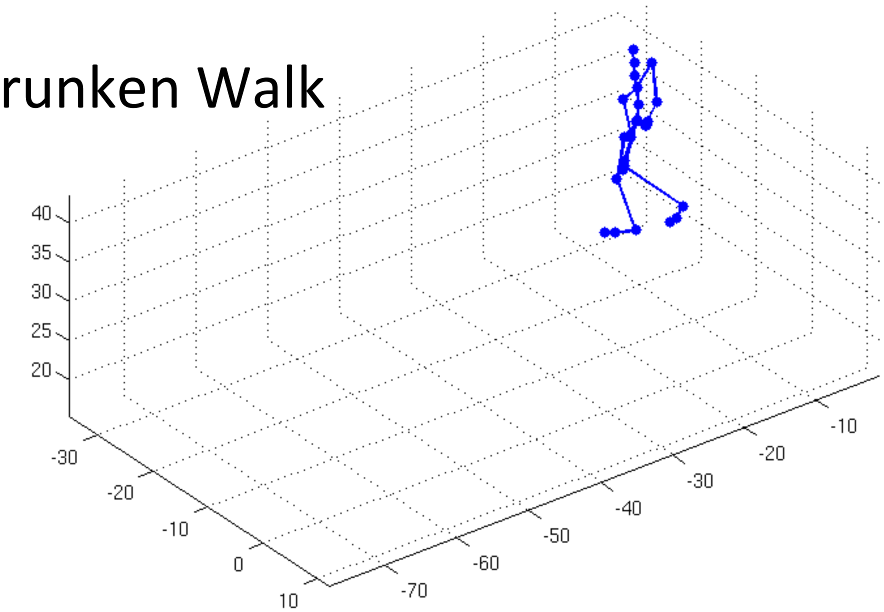


Motion Capture

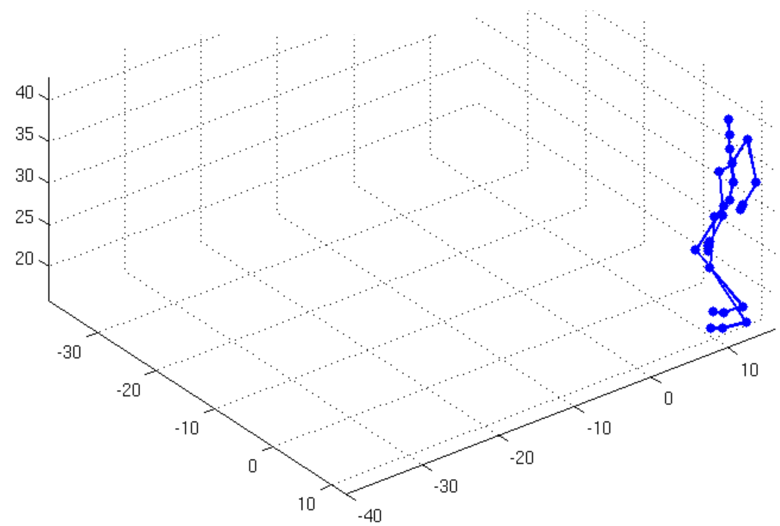
Walk



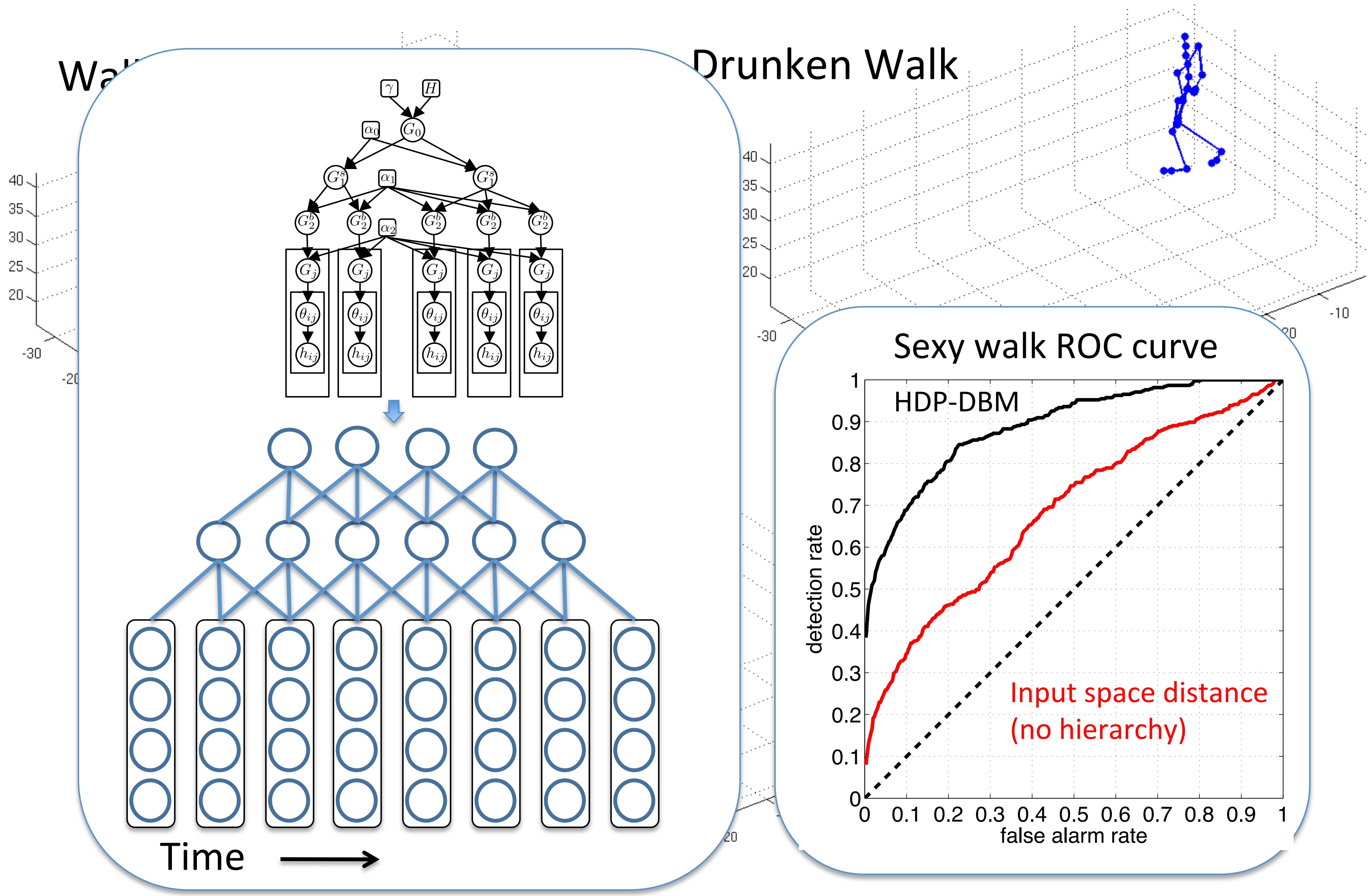
Drunken Walk



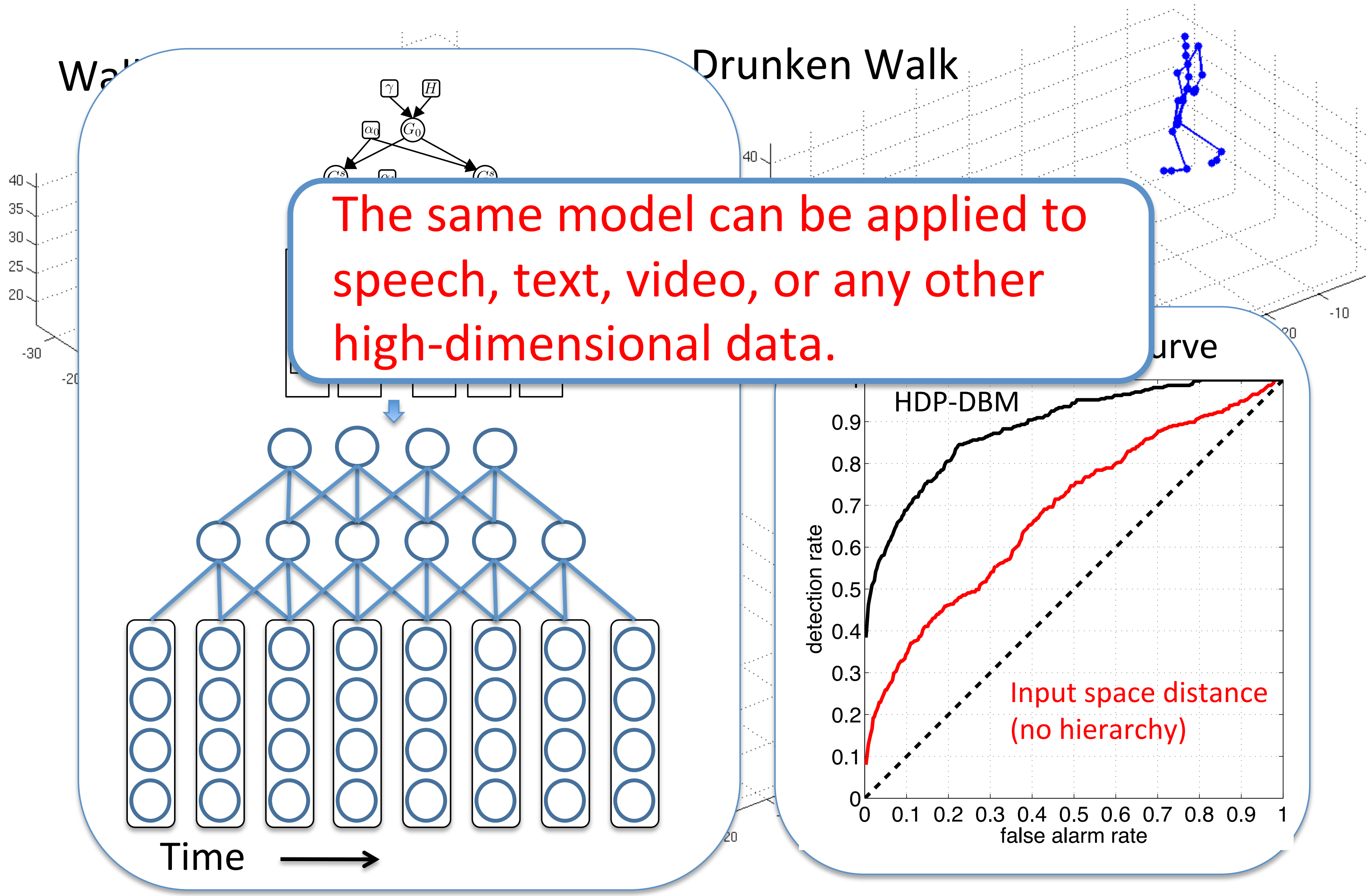
Sexy Walk



Motion Capture

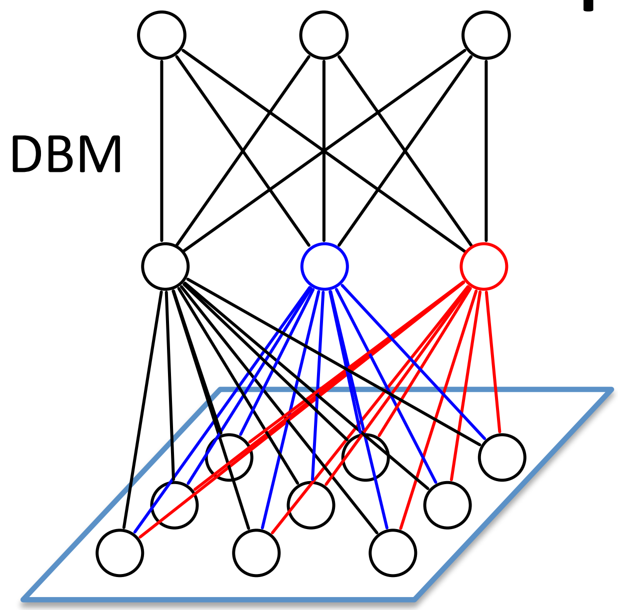


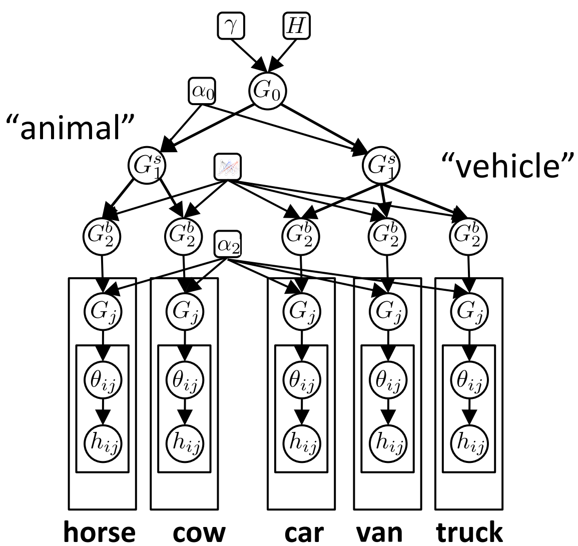
Motion Capture



The same model can be applied to speech, text, video, or any other high-dimensional data.

Talk Roadmap





Part 2: Advanced Hierarchical Models

- Introduction: Transfer Learning/ One-Shot Learning.
- Compound Hierarchical Deep Models:
  - Deep Boltzmann Machines.
  - Hierarchical Latent Dirichlet Allocation Model.
- Applications.
- Conclusions

Other Hierarchical Models

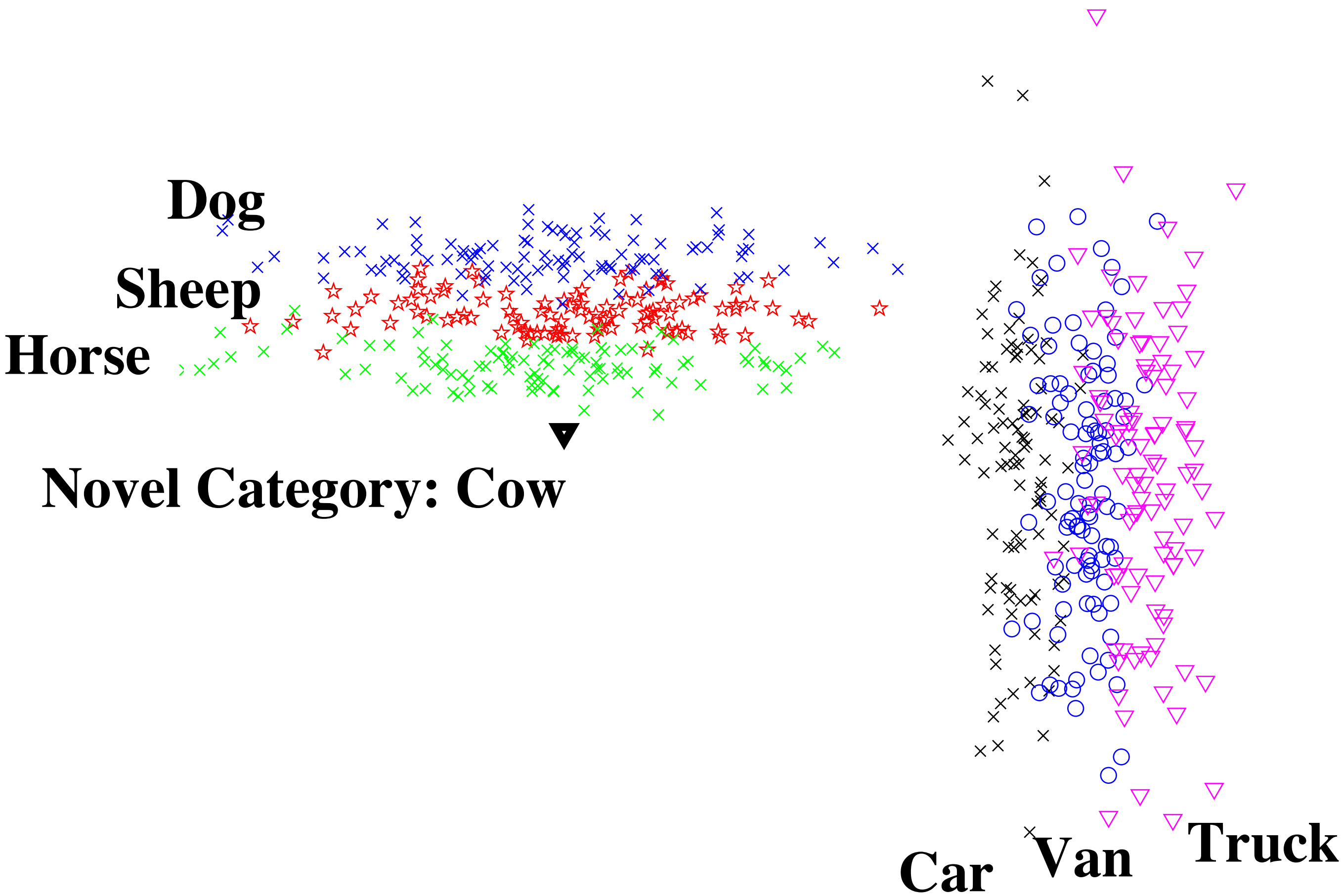
At a minimum, object categorization requires information about
- category mean (prototype)
- variances along each dimension (similarity metric)



Color features vary strongly, whereas shape features vary weakly.

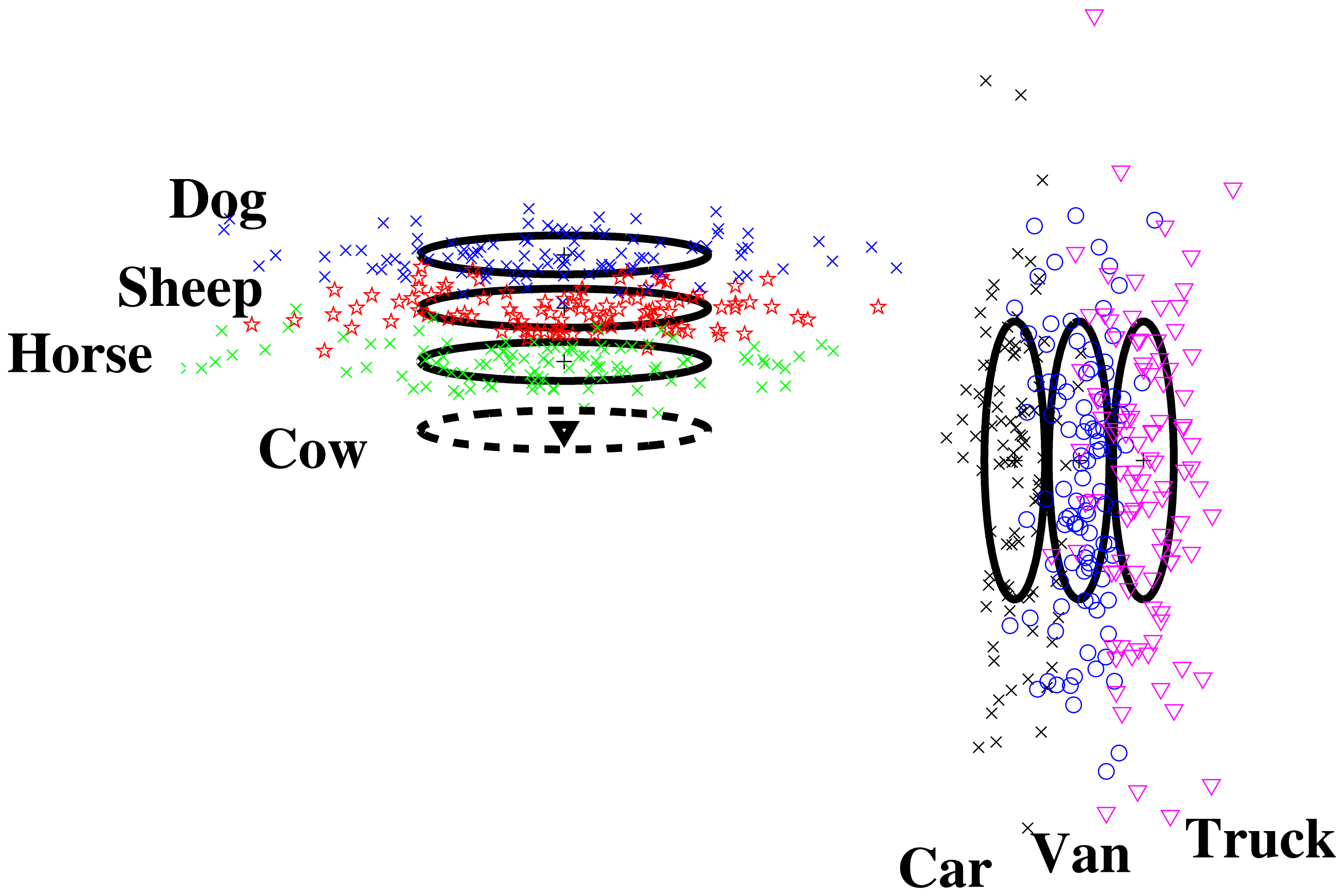
A single example provides some information about the prototype, but not about the variances.

Learning Class-Specific Similarity Metrics



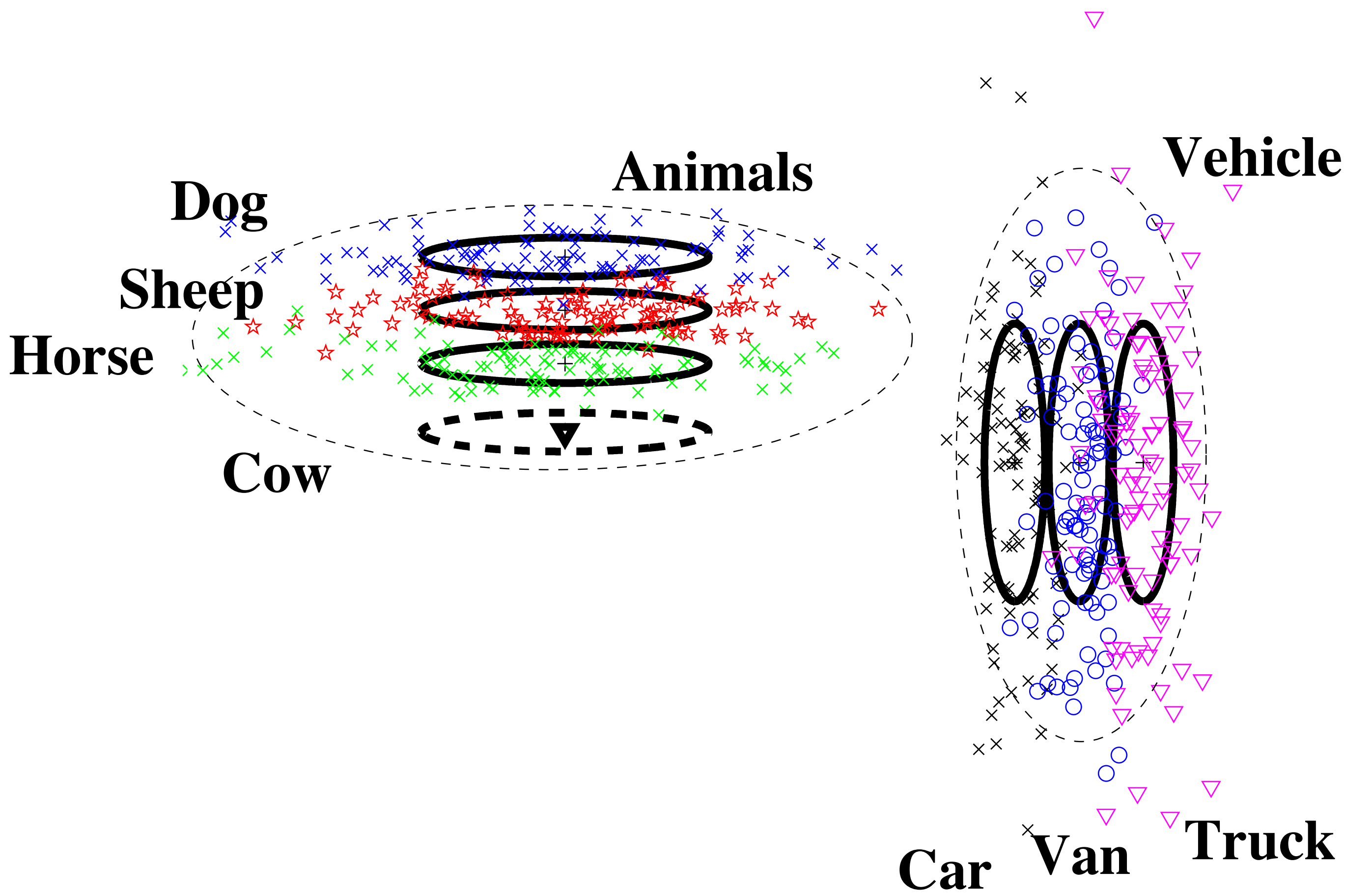
(Salakhutdinov, Tenenbaum, & Torralba, JMLR WC&P 2012)

Learning Class-Specific Similarity Metrics



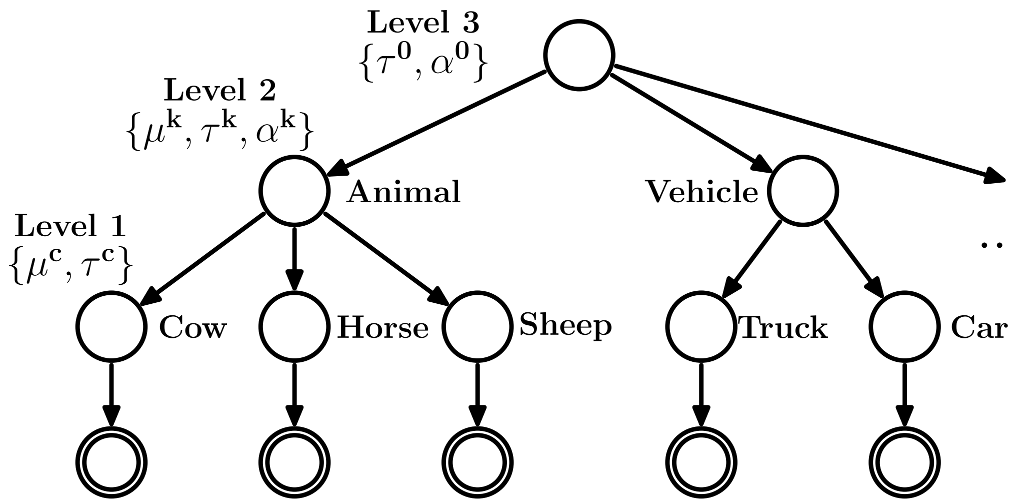
(Salakhutdinov, Tenenbaum, & Torralba, JMLR WC&P 2012)

Learning Class-Specific Similarity Metrics



In order to transfer appropriate similarity metric, the model needs to discover how to group related categories into super-categories.

Hierarchical Bayes



- Probabilistic linear model with Gaussian observation noise:

$$P(x|z = c) = N(\mu^c, 1/\tau^c)$$

- Place a conjugate Normal-Gamma prior over the means and precision parameters:

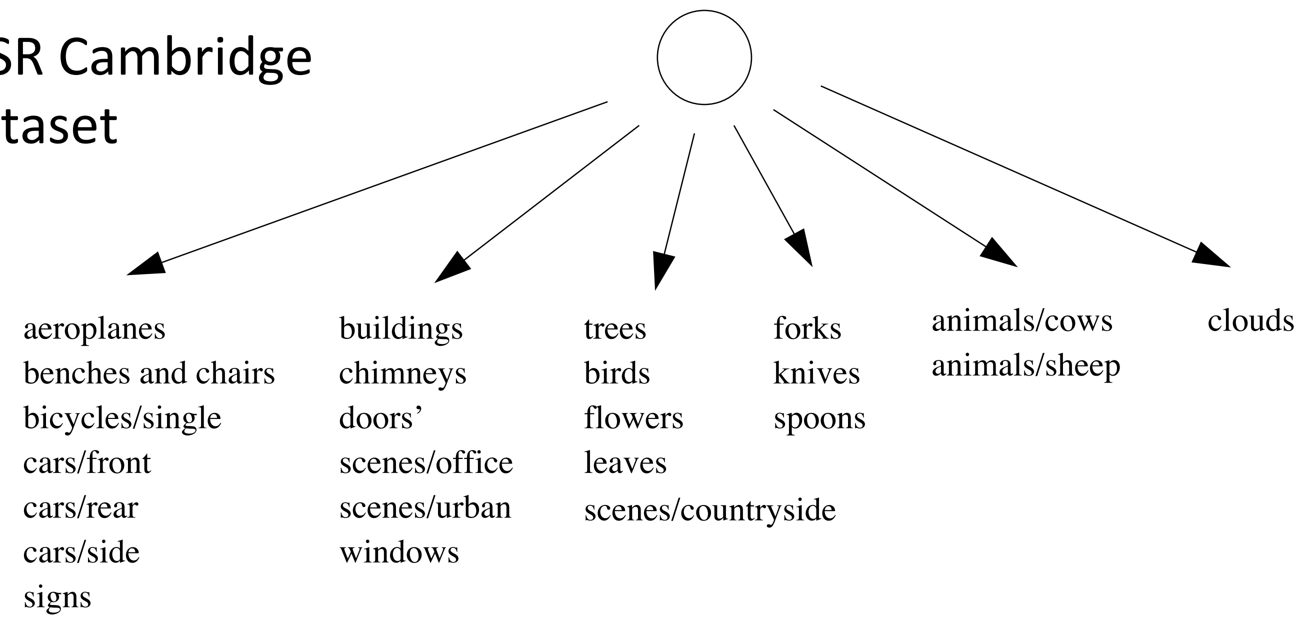
$$P(\mu^c, \tau^c) = \mathcal{N}(\mu^k 1/(\nu\tau^c))\Gamma(\alpha^k, \tau^k)$$

Hierarchical Prior.

As before, infer the hierarchy.

Image Retrieval

MSR Cambridge Dataset

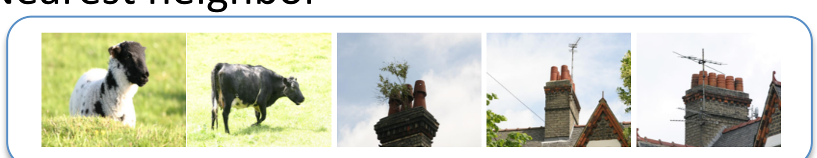


Query image

Given only one examples of a cpw

Retrieved images with our model

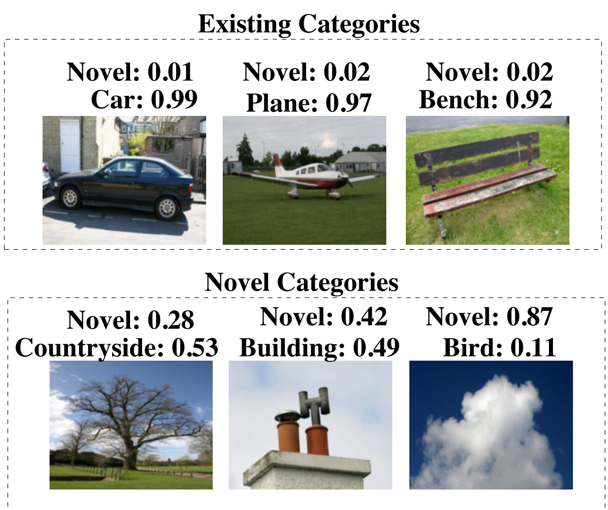
Nearest neighbor



Unsupervised Category Discovery

Can we discover when the model has encountered novel categories, and how can we break up new instances into novel categories?

The test set consists of **many unlabeled examples from an unknown number of basic-level classes.**



With 18 unlabeled test images the model correctly places nine familiar images in nine different basic-level categories, while also correctly forming three novel categories with 3 examples each.

Object Detection Challenge

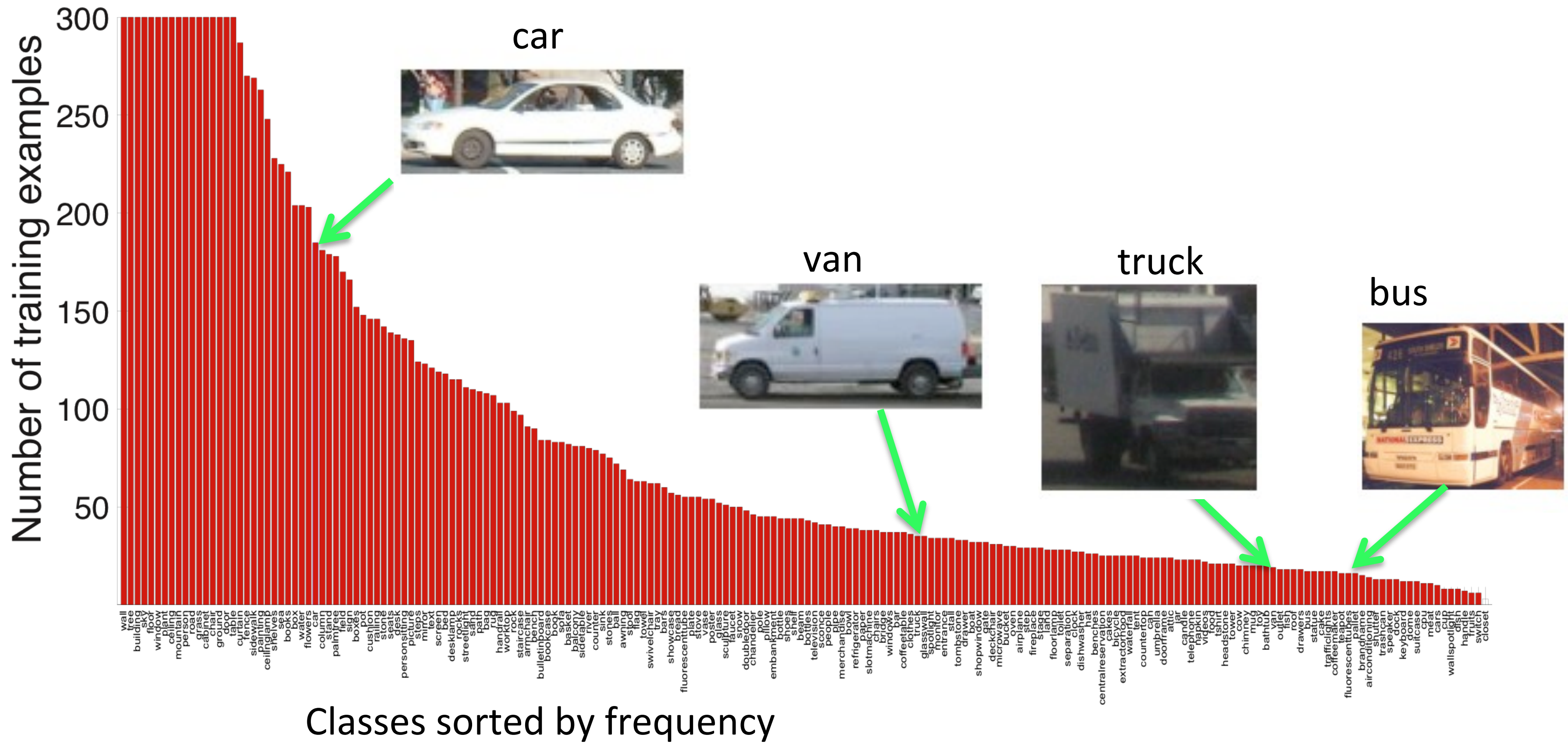
Consider challenging object detection task.

By looking at the output of a detector, can you guess which object is it trying to detect?

Slide credit: Antonio Torralba

Learning from Few Examples

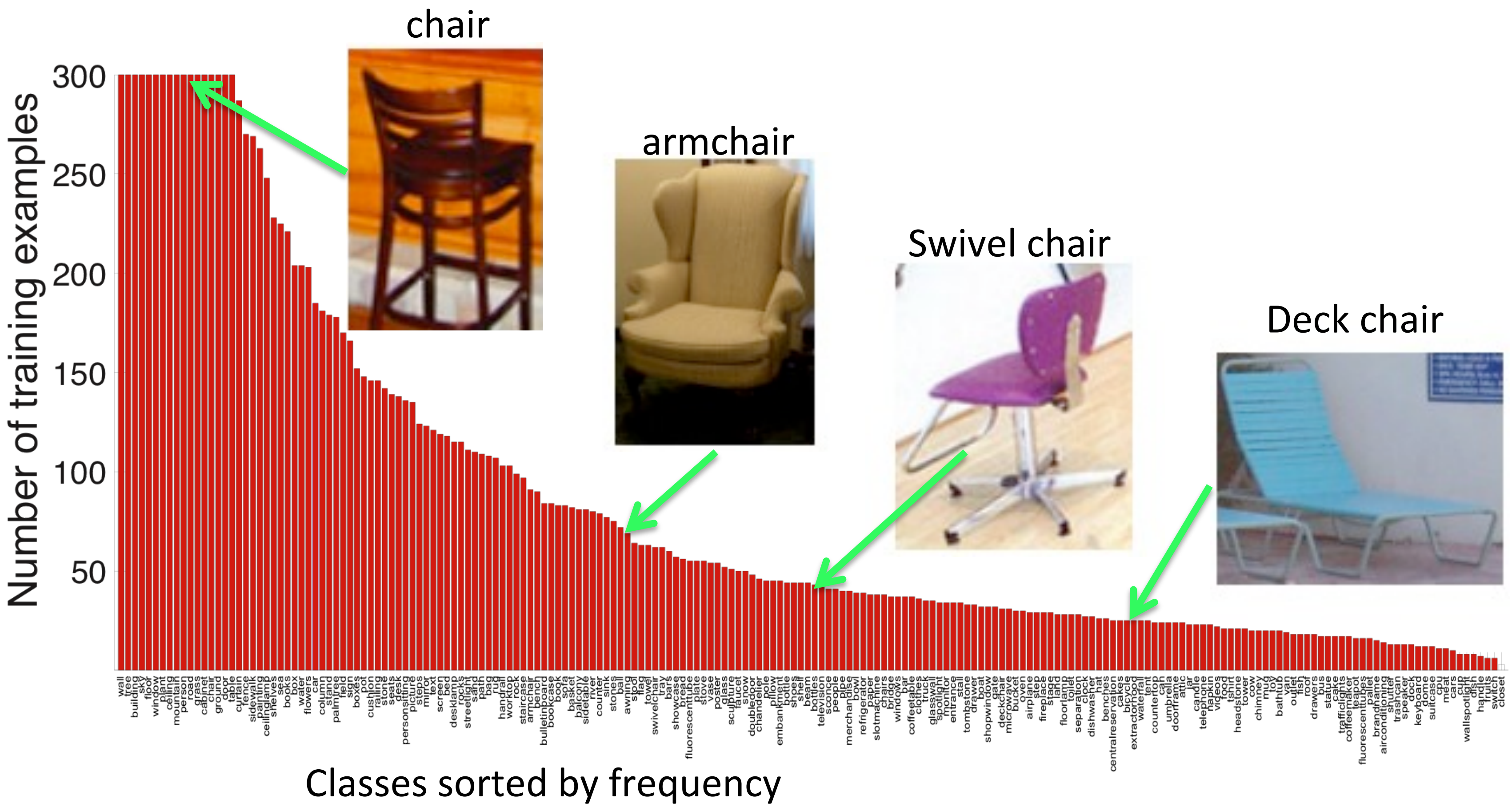
SUN database



Rare objects are similar to frequent objects

(Salakhutdinov, Torralba, & Tenenbaum, CVPR 2011)

Learning from Few Examples



Generative Model of Classifier Parameters

Many state-of-the-art object detection systems use sophisticated models, based on multiple parts with separate appearance and shape components.

$$y = \beta^\top \Phi(\mathbf{x})$$

Detect objects by testing sub-windows and scoring corresponding test patches with a linear function.

We can define hierarchical prior over parameters of discriminative model and learn the hierarchy.

Image Specific: concatenation of the HOG feature pyramid at multiple scales.

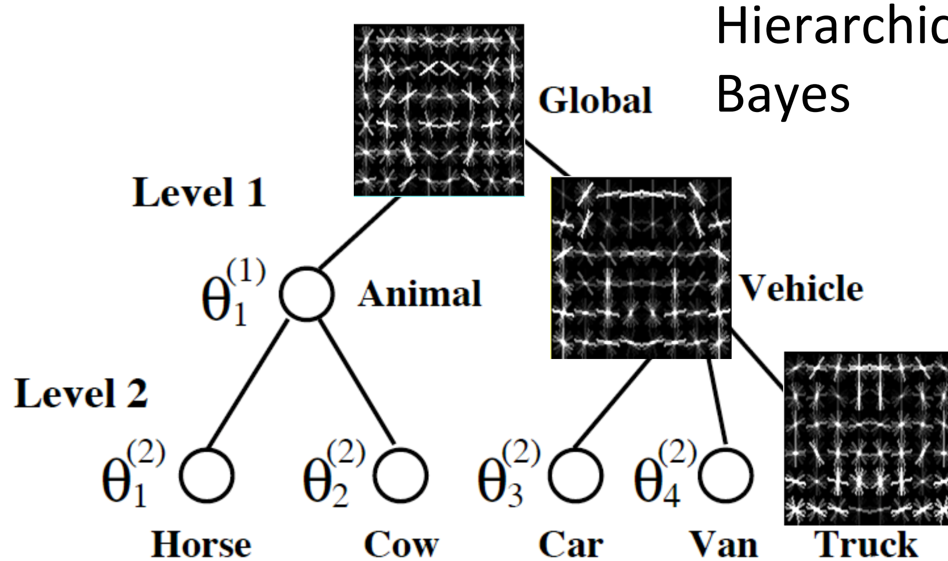
Felzenszwalb, McAllester & Ramanan, 2008

Generative Model of Classifier Parameters

By learning hierarchical structure, we can improve the current state-of-the-art.

Sun Dataset: 32,855 examples of 200 categories

Hierarchical Bayes



185 ex   27 ex   12 ex

Hierarchical Model

Single Class

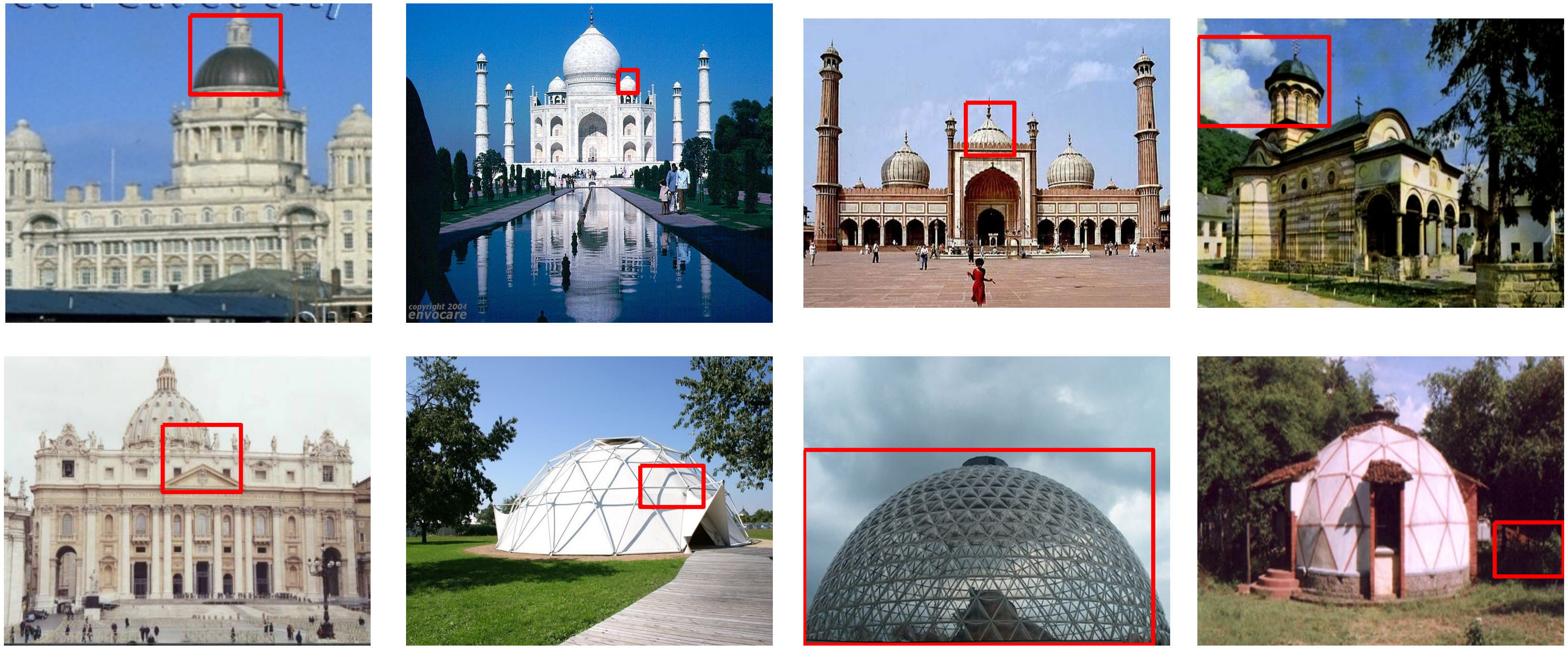
Truck

Single classifier

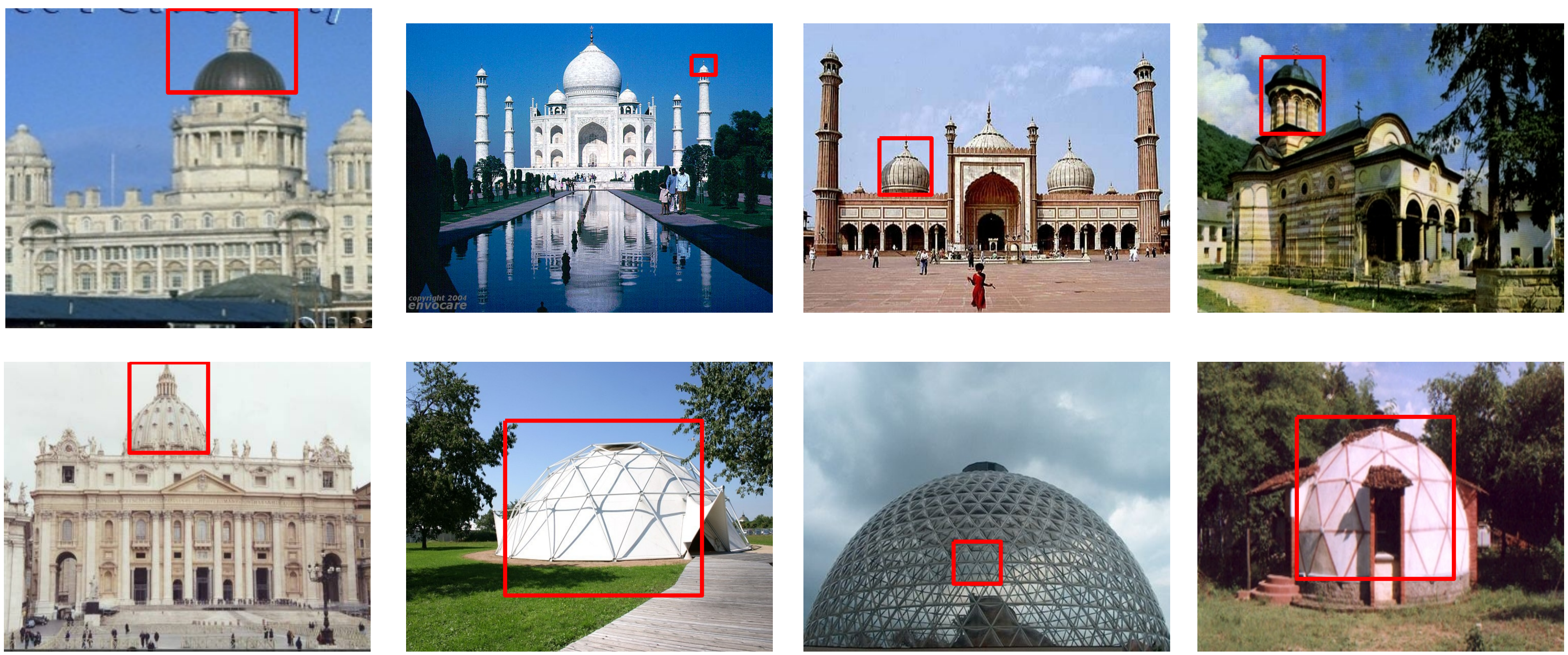
Hierarchical Model

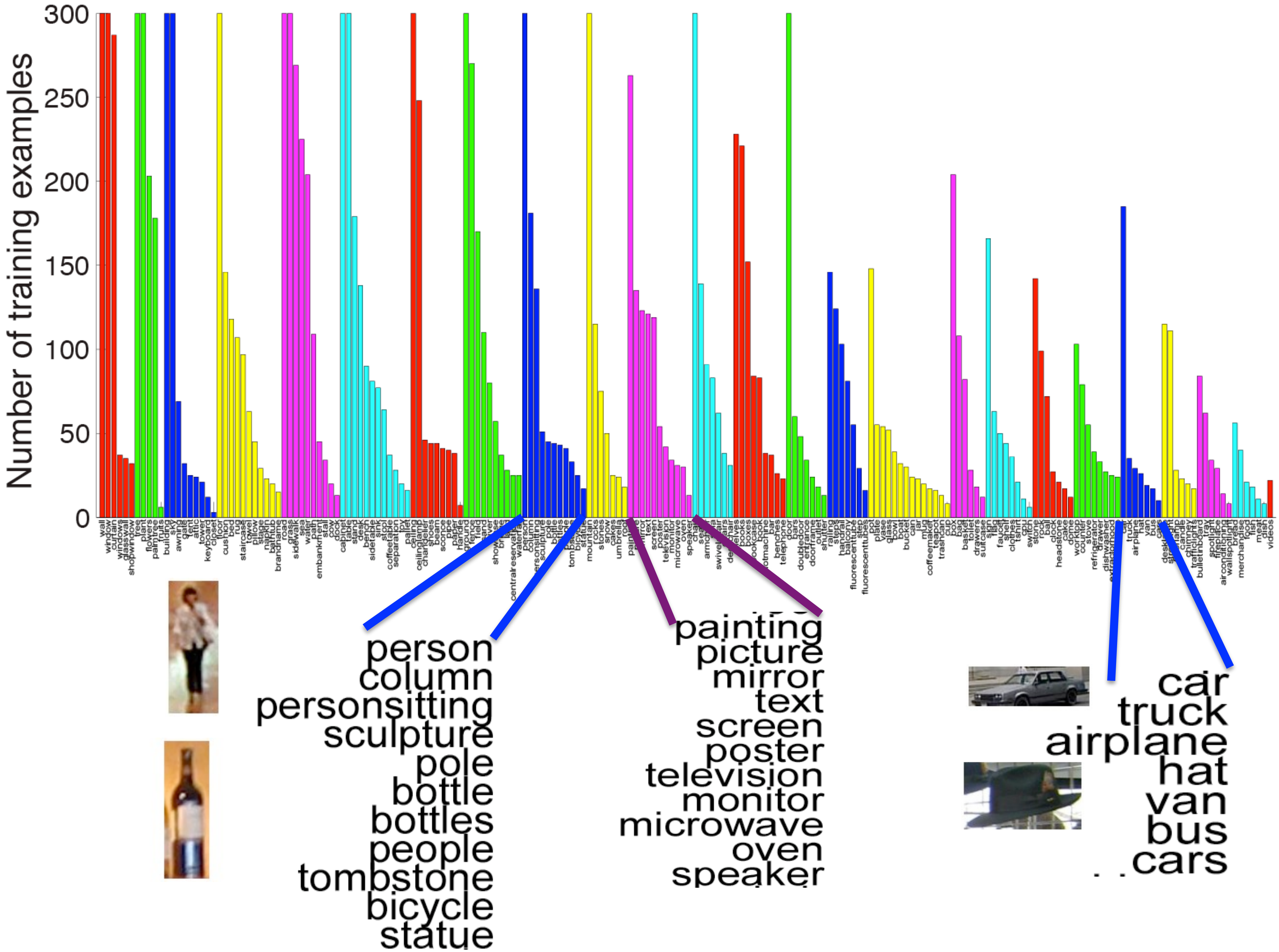
Dome

Single classifier



Hierarchical Model

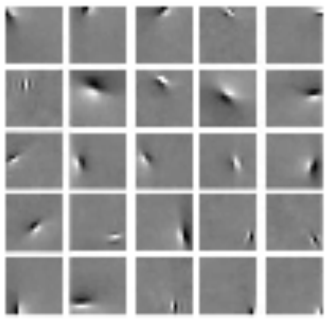




Number of training examples
300
250
200
150
100
50
0
person
column
personsitting
sculpture
pole
bottle
bottles
people
tombstone
bicycle
statue
painting
picture
mirror
text
screen
poster
television
monitor
microwave
oven
speaker
car
truck
airplane
hat
van
bus
cars

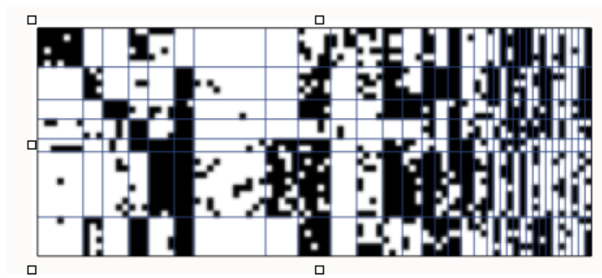
Generative Model of Matrix Factorizations

Image bases



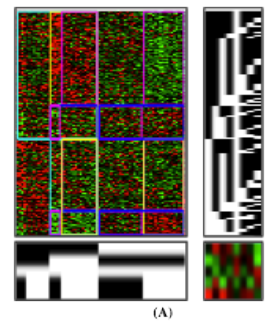
Karklin and Lewicki (2009)

Relational data



Kemp et.al. (2006)

Gene expression data



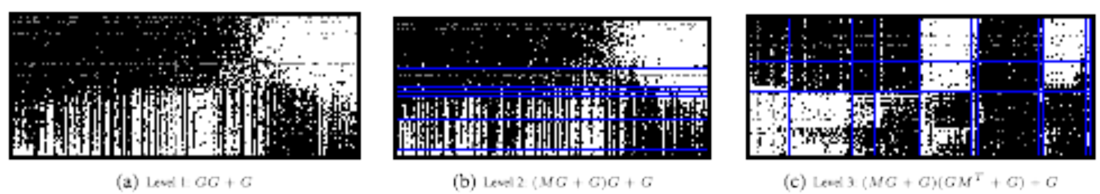
Meeds et al. (2007)

How can we automatically choose the right structure from raw data?

Context free grammar:

low-rank	$G \rightarrow GG + G$	(1)
clustering	$G \rightarrow MG + G \mid GM^T + G$	(2)
	$M \rightarrow MG + G$	(3)
linear dynamics	$G \rightarrow CG + G \mid GC^T + G$	(4)
	$C \rightarrow CG + G$	(5)
sparsity	$G \rightarrow \exp(G) \circ G$	(6)
binary factors	$G \rightarrow BG + G \mid GB^T + G$	(7)
	$M \rightarrow B$	(8)

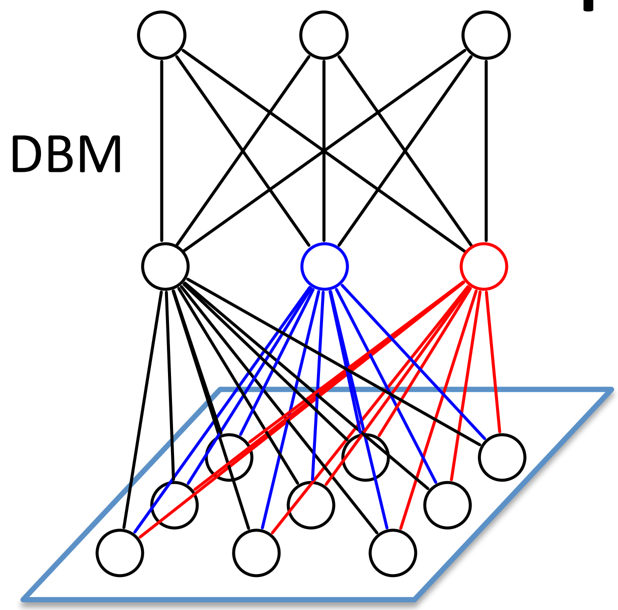
US Senate votes:

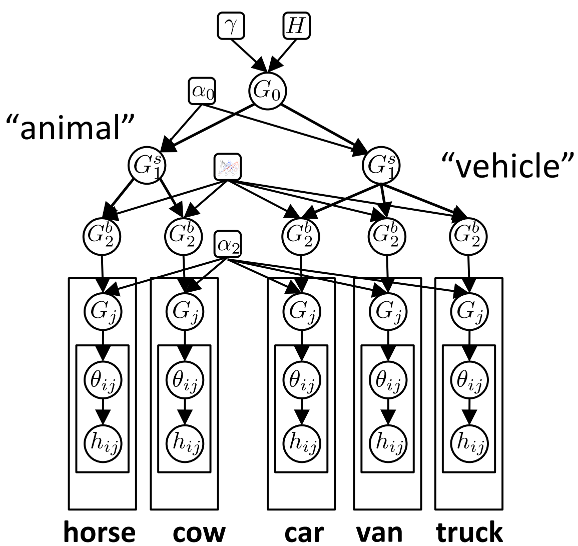


Evolution of structure discovery

Grosse, Salakhutdinov, Freeman, and Tenenbaum, UAI 2012

Talk Roadmap

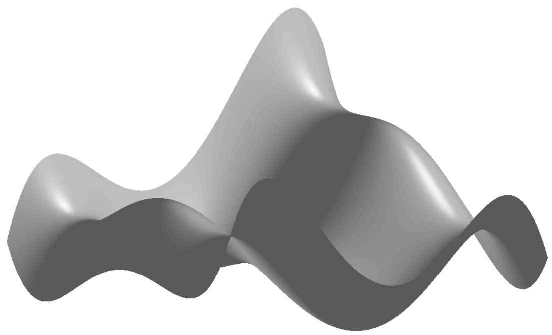




Part 2: Advanced Hierarchical Models

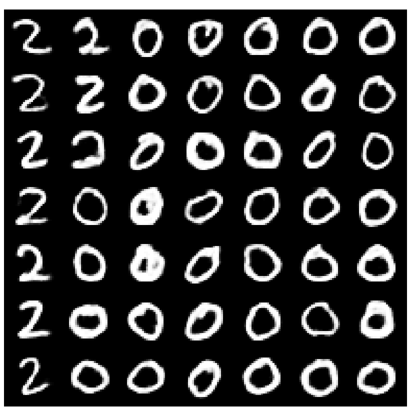
- Introduction: Transfer Learning/ One-Shot Learning.
- Compound Hierarchical Deep Models:
  - Deep Boltzmann Machines.
  - Hierarchical Latent Dirichlet Allocation Model.
- Applications.
- MCMC techniques.

Inference



Problem: When dealing with complex high-dimensional data: the probability landscape is highly multimodal.

Gibbs Sampler



Inability to efficiently explore a distribution with many isolated modes.

Problem for both directed and undirected graphical models.

- Posterior distribution: $P(\theta|\mathcal{D}) = \frac{1}{P(\mathcal{D})} P(\mathcal{D}|\theta) P(\theta)$
- Boltzmann machine: $P(z) = \frac{1}{Z} \exp(-E(z))$

Tempered Transitions

(Radford Neal, 1994)

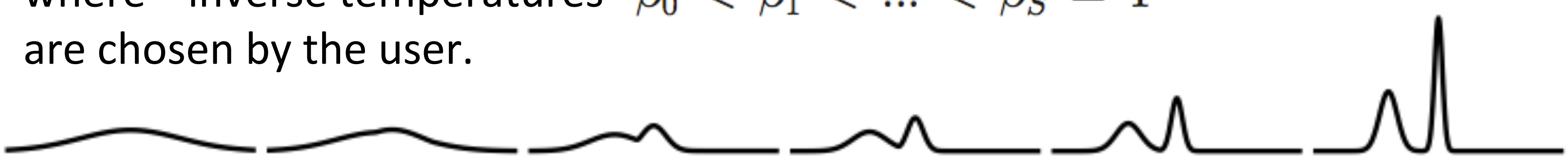
Define a sequence of intermediate probability distributions $p_0, \ldots, p_S$ where:

- $p_S = p(\mathbf{x}; \theta)$ is the original complicated distribution.
- p_0 is more spread out and easier to sample from.

One way is to define:

$$p_s(\mathbf{x}) \;\propto\; p^*(\mathbf{x}; \theta)^{\beta_s},$$

where "inverse temperatures" $\beta_0 < \beta_1 < \ldots < \beta_S = 1$ are chosen by the user.

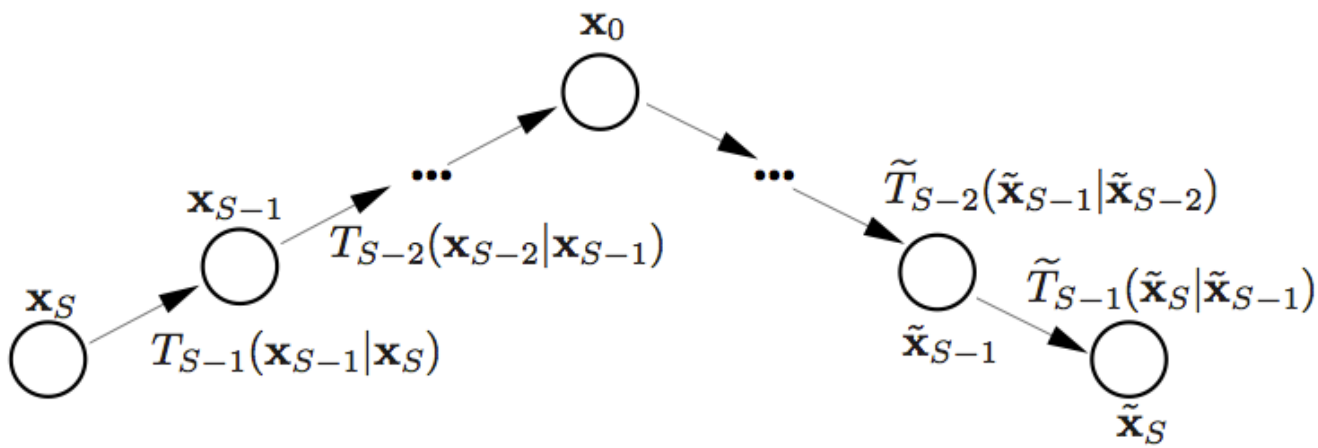


$\beta = 0$ $\beta = 0.01$ $\beta = 0.1$ $\beta = 0.25$ $\beta = 0.5$ $\beta = 1$

For each $s = 1, .., S-1$ we define a transition operator $T_s(\mathbf{x}' \leftarrow \mathbf{x})$ that leaves p_s invariant.

Tempered Transitions

Define reverse transition operator $p_s(\mathbf{x})T_s(\mathbf{x}' \leftarrow \mathbf{x}) = \widetilde{T}_s(\mathbf{x} \leftarrow \mathbf{x}')p_s(\mathbf{x}')$.

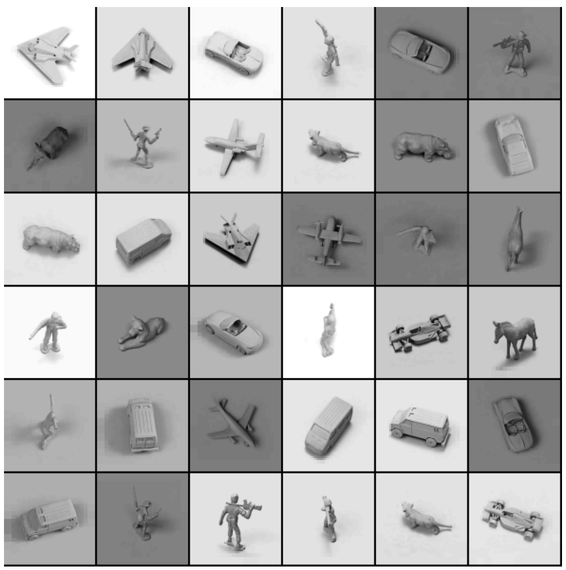


- Given a current state, apply a sequence of transition operators: $T_{S-1} \ldots T_0 \widetilde{T}_0 \ldots \widetilde{T}_{S-1}$.
- Systematically "move" the sample from the complicated distribution to the easily sampled distribution and back.
- Accept a new state $\tilde{\mathbf{x}}^S$ with probability:

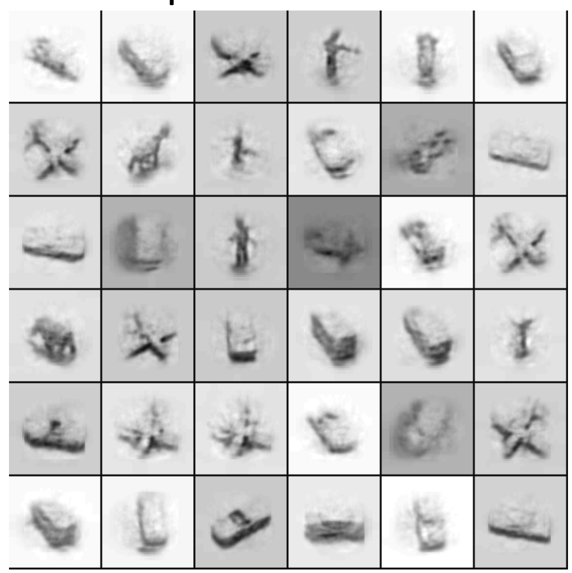
$$\min\left[1, \prod_{s=1}^{S} p^*(\mathbf{x}_s)^{\beta_{s-1}-\beta_s} p^*(\tilde{\mathbf{x}}_s)^{\beta_s-\beta_{s-1}}\right].$$

Learning MRFs using Tempered Transitions

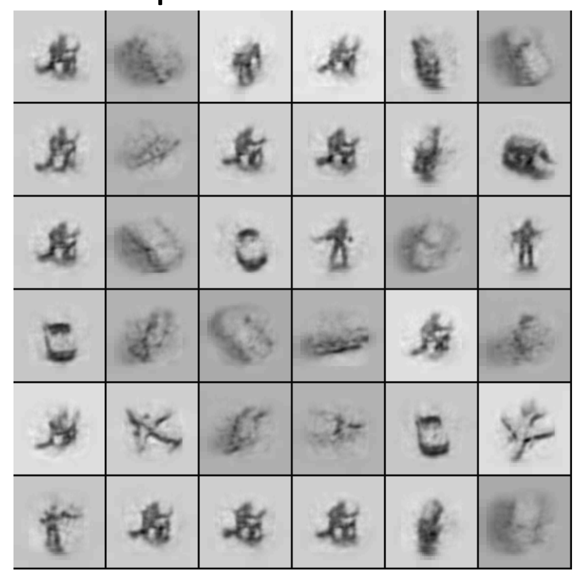
Training data



Samples with Tempered Transitions



Samples without Tempered Transitions



Plain stochastic approximation using simple Gibbs works badly.

A large fraction of the model's probability mass is placed on images of humans.

(Salakhutdinov, NIPS 2010)

Simulated Tempering (ST)

- Simulated tempering: Sample from the joint distribution:

$$p(\mathbf{x}, k) \propto w_k \exp(-\beta_k E(\mathbf{x})),$$

where w_k are pre-specified constants, and $0 < \beta_K < \beta_{K-1} < ... < \beta_1 = 1$ represent the K "inverse temperatures".

Simulated Tempering (ST)

- Simulated tempering: Sample from the joint distribution:

$$p(\mathbf{x}, k) \propto w_k \exp(-\beta_k E(\mathbf{x})),$$

where w_k are pre-specified constants, and $0 < \beta_K < \beta_{K-1} < ... < \beta_1 = 1$ represent the K "inverse temperatures".

- The main problem of ST:

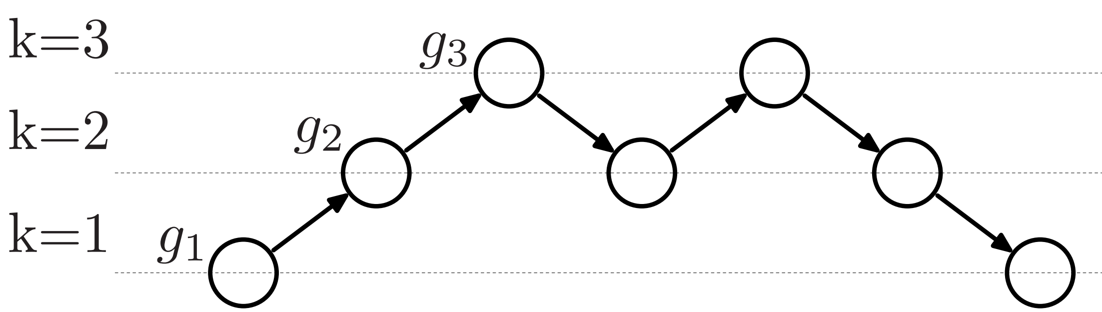
$$p(k) \propto \sum_{\mathbf{x}} w_k \exp(-\beta_k E(\mathbf{x})) = w_k \mathcal{Z}_k$$

- To be efficient, it is important for the Markov chain to spend roughly equal amount of time at each temperature level.

- Hence w_k needs to be proportional to $1/\mathcal{Z}_k$.

Adaptive Simulated Tempering (AST)

- Partitioning the state space into K sets $\{k\} \cup \mathcal{X}$, each corresponding to a different temperature value.
- If the move into a different partition (temperature) is rejected:
  - The adaptive weight g_k for the current partition k will increase.
  - This will (exponentially) increase the probability of accepting the next move into a different temperature level.



Atchade and Liu, 2004,
Famong Liang, 2005

Adaptive Simulated Tempering (AST)

- Given k^t, sample k^{t+1} from proposal distribution: $q(k^{t+1} \leftarrow k^t)$

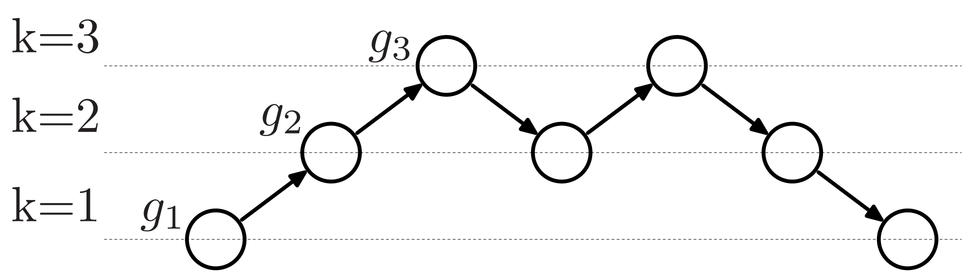
  Accept with probability:

$$\min\left(1, \underbrace{\frac{p(\mathbf{x}^t, k^{t+1}) q(k^t \leftarrow k^{t+1})}{p(\mathbf{x}^t, k^t) q(k^{t+1} \leftarrow k^t)}}_{\text{Standard M-H update}} \times \underbrace{\frac{g_{k^t}}{g_{k^{t+1}}}}_{\text{Adaptive factor}}\right)$$

- Update adaptive weights:

$$g_i^{t+1} = g_i^t(1 + \gamma_t \mathbb{I}(k^{t+1} \in \{i\})),\ i = 1, ..., K.$$

- It can be verified: $g_i^t / g_j^t \longrightarrow \mathcal{Z}_i / \mathcal{Z}_j$ as $\gamma_t \longrightarrow 0$.



Atchade and Liu, 2004,
Famong Liang, 2005

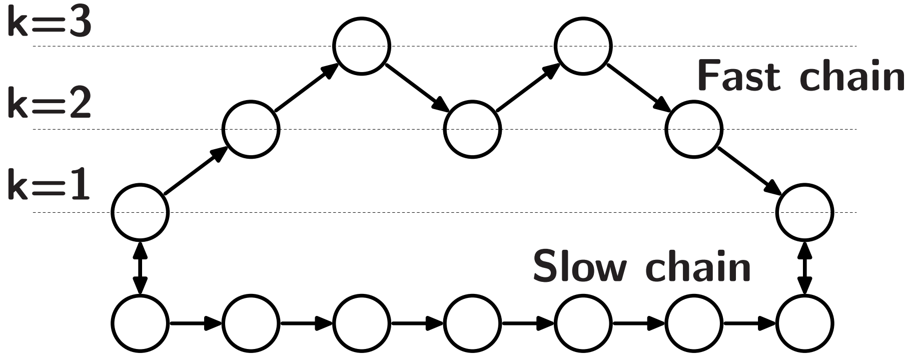
Fast-Slow AST

- When using AST for learning, it is hard to balance between:
  - Exploration: waiting until adaptive ST escapes from the local mode.
  - Exploitation: learning model parameters.

Fast-Slow AST

- When using AST for learning, it is hard to balance between:
  - Exploration: waiting until adaptive ST escapes from the local mode.
  - Exploitation: learning model parameters.

- Consider two chains, sampling from the same target distribution.

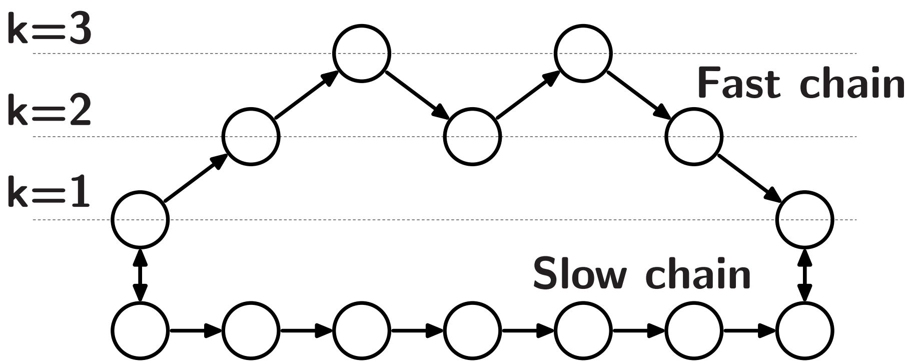


Slow chain evolves according to the standard Gibbs updates.

Fast chain uses adaptive ST.

- Parameters are updated based on the slow chain. The role of the fast chain is to explore different modes.

Fast-Slow AST



Slow chain evolves according to the standard Gibbs updates.

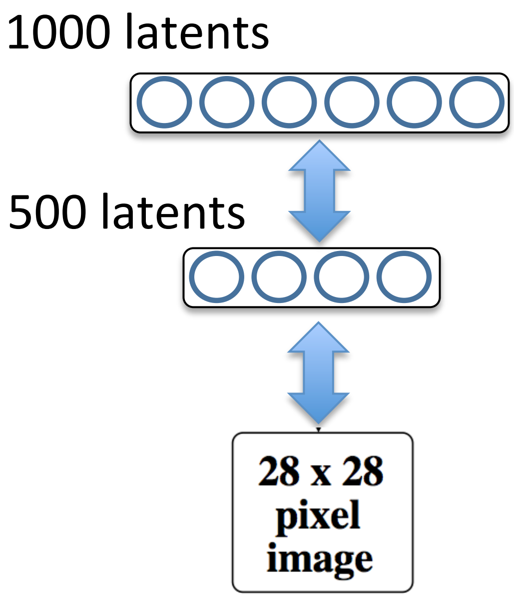
Fast chain uses adaptive ST.

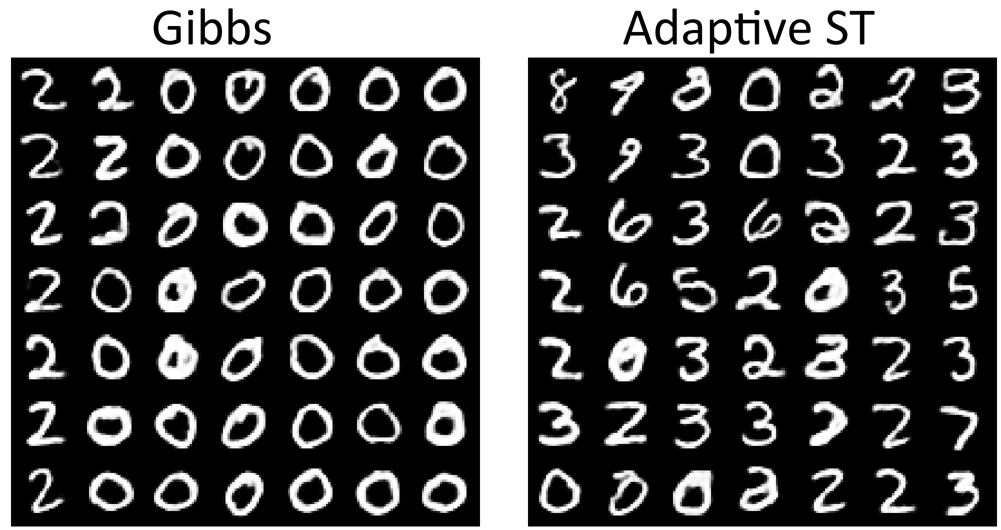
- The algorithm is only twice as expensive compared to the standard stochastic approximation algorithm.

- Parameters are updated after every Gibbs update, while the fast chain runs in parallel, adaptively mixing between different modes of the energy landscape.

- Unlike fast Persistent Contrastive Divergence (PCD), the fast chain is likely to visit spurious modes that may reside far away from the data.

MNIST Dataset

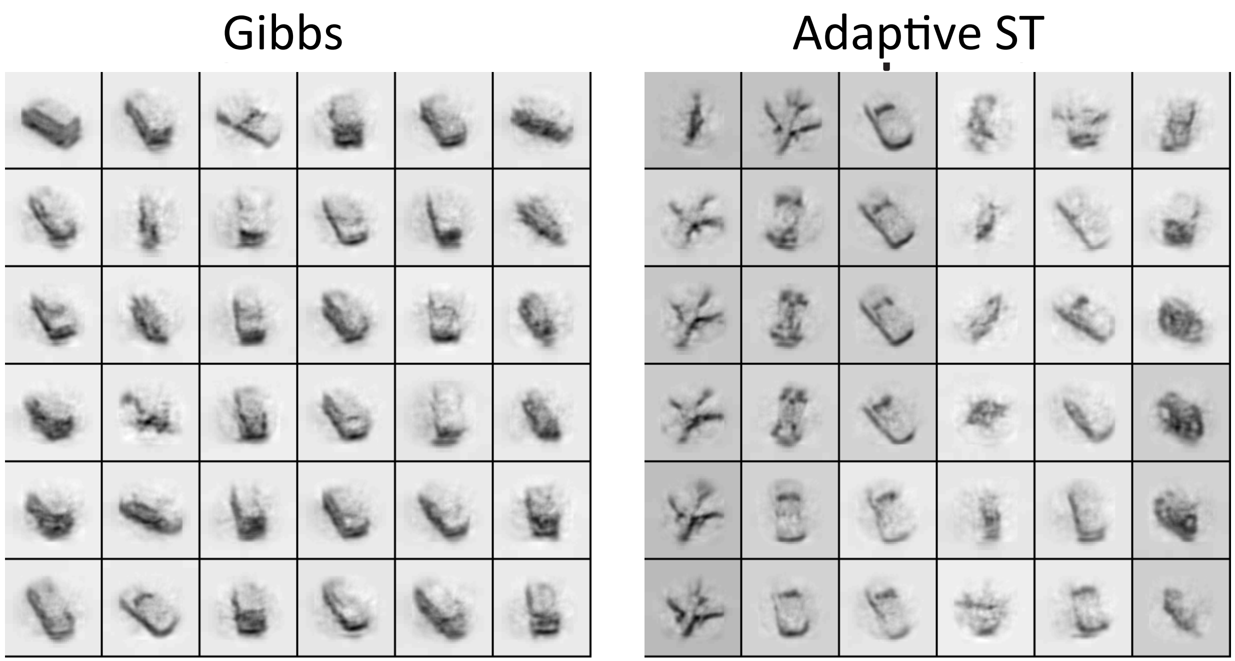




About 890,000
parameters

- Samples from the two-hidden-layer DBM (1000-500-784) produced by the Gibbs and adaptive ST with 300 Gibbs steps between consecutive images (by column).

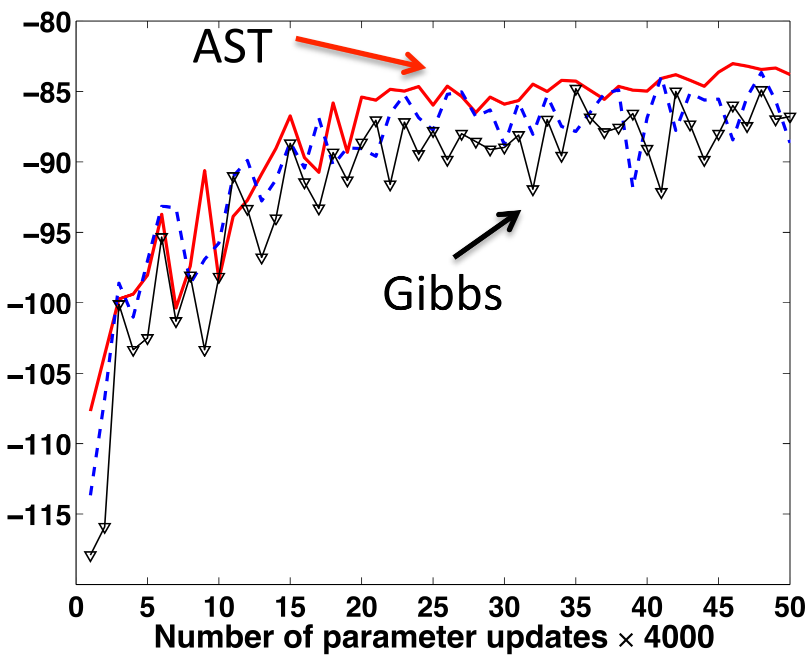
NORB Dataset



• Samples from two-hidden-layer DBM: 4000-4000-(96x96), produced by the Gibbs and fast-slow adaptive ST with 500 Gibbs steps between consecutive images (by column). About 3 million parameters.

Learning DBMs

The estimates of the average test log-probabilities per image (in nats) for different learning algorithms.

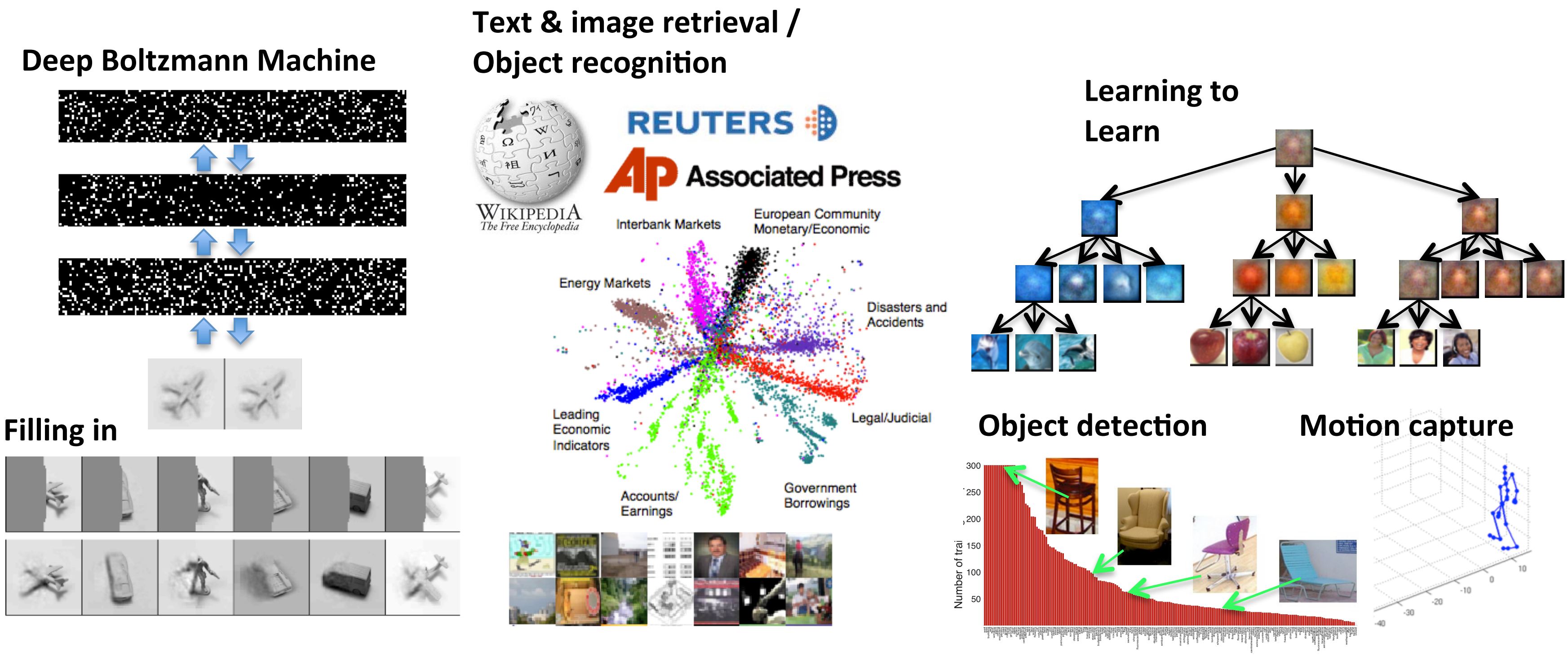


Algorithm	MNIST (+/- 0.5)	NORB (+/- 1.1)
Gibbs	-87.23	-596.92
Fast PCD	-86.72	-597.12
Tempered Transitions	-85.41	-595.54
Fast-Slow AST	-84.12	-591.18

- Fast-Slow AST tends to exhibit a more stable behavior during learning.

Recap

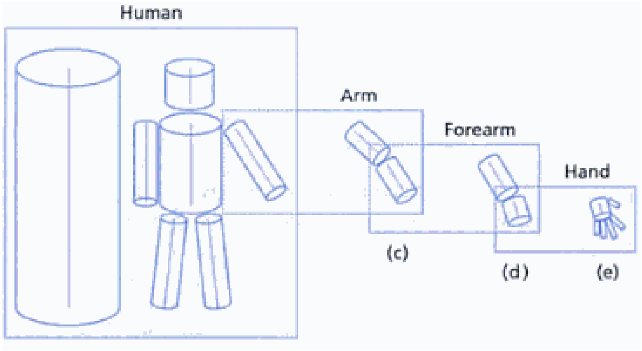
- Efficient learning algorithms for Hierarchical Generative Models.



- Deep generative models can improve current state-of-the art in many application domains:
  - Object recognition and detection, text and image retrieval, handwritten character recognition, motion capture, and others.

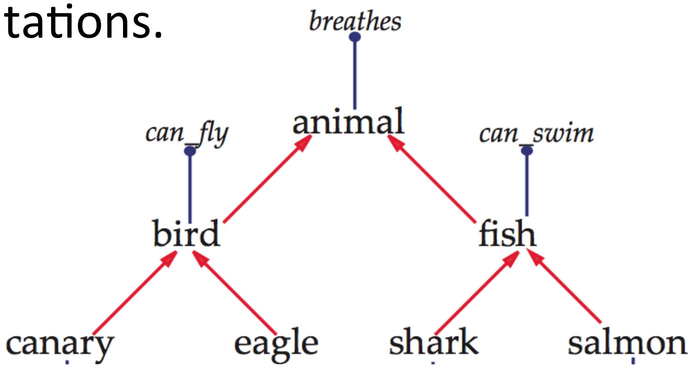
Summary

Compose hierarchical Bayesian models with deep networks for transfer learning / one-shot learning.



Deep Networks: Learning Part-based Hierarchy:

- multiple **layers of nonlinearities.**
- **distributed representations.**
- **unsupervised learning of generic features** -- no need to rely on human-crafted input representations.



Hierarchical Bayes: Learning Category Hierarchy:

- **explicitly learn category hierarchies** for sharing abstract knowledge.
- **modular data-parameter relations**.
- higher-level **class sensitive features.**

Thank you