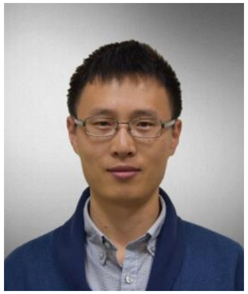
Poster ID 4

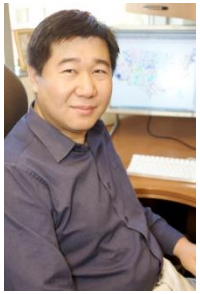
Synthesized Classifiers for Zero-shot Learning

Soravit (Beer) Changpinyo[*1]

Wei-Lun (Harry) Chao[*1]



Boqing Gong[2]



Fei Sha[3]

1 USC

2 UCF

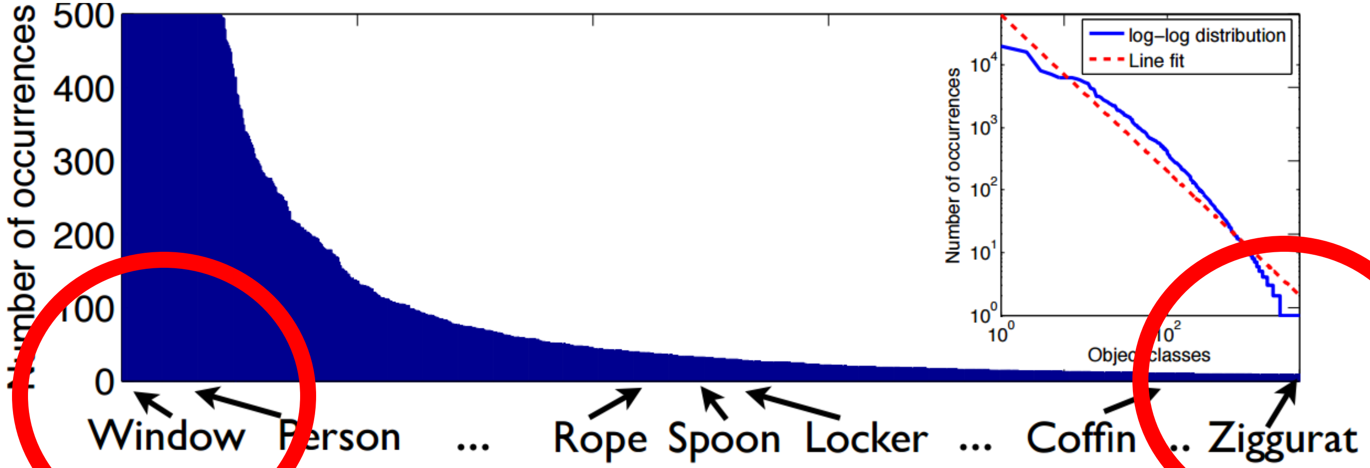
3 UCLA

Challenge for Recognition in the Wild

Figures from Wikipedia

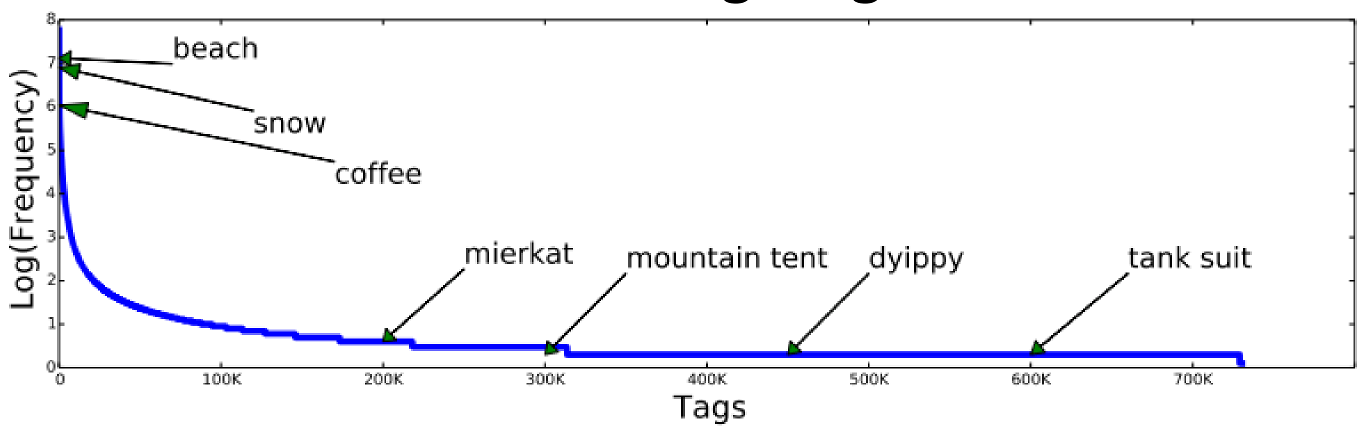
The Long Tail Phenomena

Objects in SUN dataset



Zhu et al.
CVPR 2014

Flickr image tags



Kordumova et al.
MM 2015

The Long Tail Phenomena

Problem for the tail

How to train a good classifier when **few labeled examples** are available?

Extreme case

How to train a good classifier when **no labeled examples** are available?

Zero-shot Learning

Zero-shot Learning

- **Two** types of classes
  - Seen: **with** labeled examples
  - Unseen: **without** examples

Cat

Horse

Dog

Zebra

?

Seen

Unseen

Figures from Derek Hoiem's slides

Zero-shot Learning: Challenges

- How to relate seen and unseen classes?
- How to attain discriminative performance on the unseen classes?

Zero-shot Learning: Challenges

- How to relate seen and unseen classes?

  Semantic information that describes each object, including unseen ones.

- How to attain discriminative performance on the unseen classes?

Semantic Embeddings

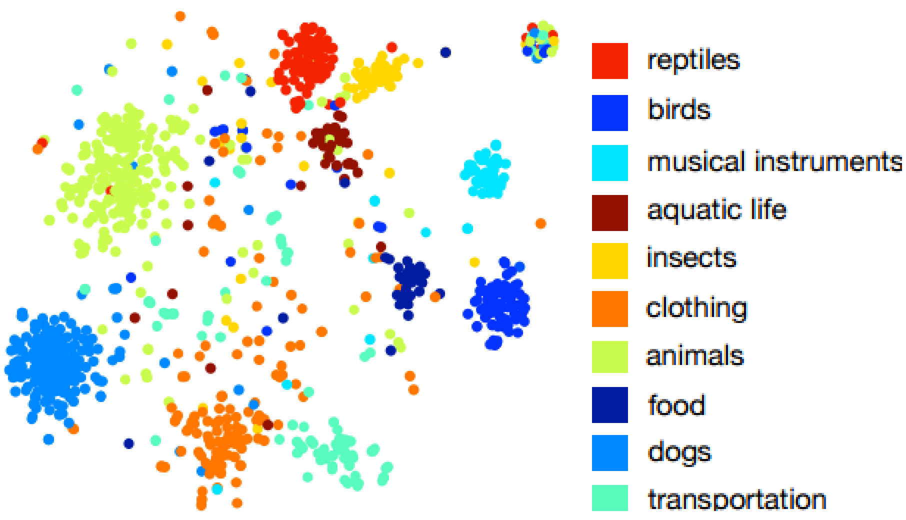
- **Attributes** (*Farhadi et al. 09, Lampert et al. 09, Parikh & Grauman 11, …*)

Bird
"has beak"
"has wing"
"feather"
"has head"
"has leg"

Cow
"has ear"
"has snout"
"furry"
"has head"
"has leg"

- **Word vectors** (*Mikolov et al. 13, Socher et al. 13, Frome et al. 13, …*)

WIKIPEDIA
The Free Encyclopedia



Zero-shot Learning: Challenges

- How to relate seen and unseen classes?

  Semantic embeddings (attributes, word vectors, etc.)

- How to attain discriminative performance on the unseen classes?

Zero-shot Learning: Challenges

- How to relate seen and unseen classes?

  Semantic embeddings (attributes, word vectors, etc.)

- How to attain discriminative performance on the unseen classes?

  Zero-shot learning algorithms

Zero-shot Learning

Seen Objects

Has Stripes
Has Ears
Has Eyes

Has Four Legs
Has Mane
Has Tail

Brown
Muscular
Has Snout

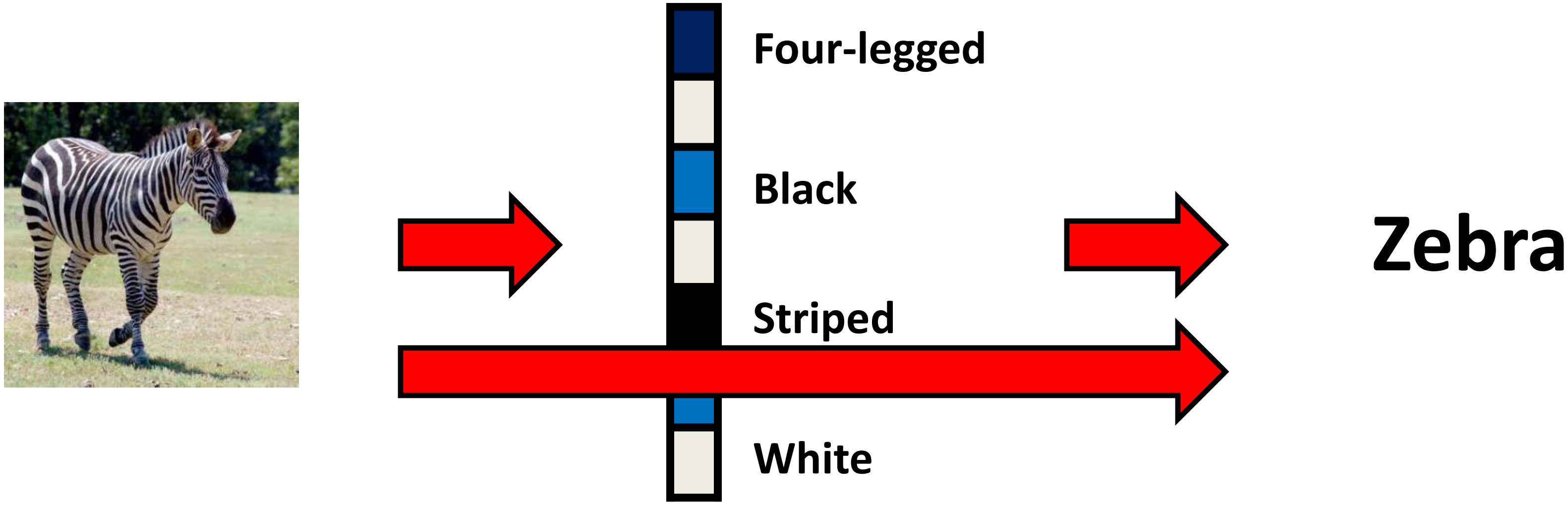
Unseen Object

Has Stripes (like cat)
Has Mane (like horse)
Has Snout (like dog)

How to effectively construct a model for zebra?

Figures from Derek Hoiem's slides

Given A Novel Image…



Separate (*Lampert et al. 09, Frome et al. 13, Norouzi et al. 14, …*)

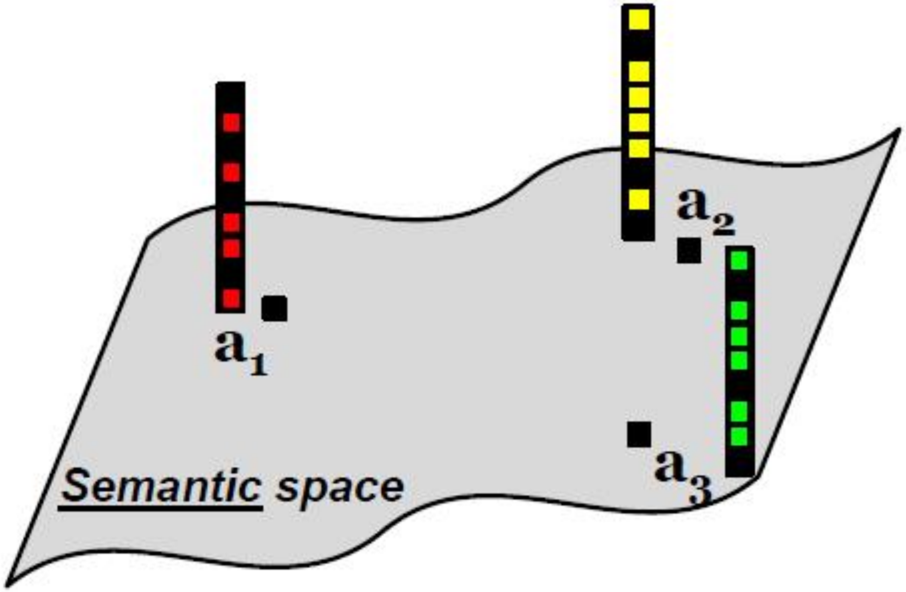
Unified (*Akata et al. 13 and 15, Mensink et al. 14, Romera-Paredes et al. 15, …*)

Our unified model uses *highly flexible bases* for *synthesizing* classifiers

Our Approach: Manifold Learning

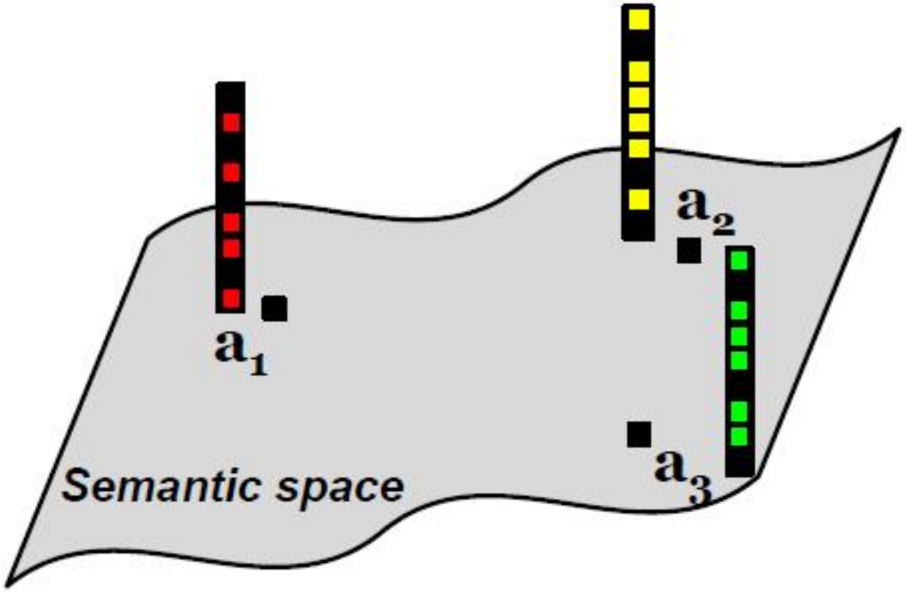
Our Approach: Manifold Learning

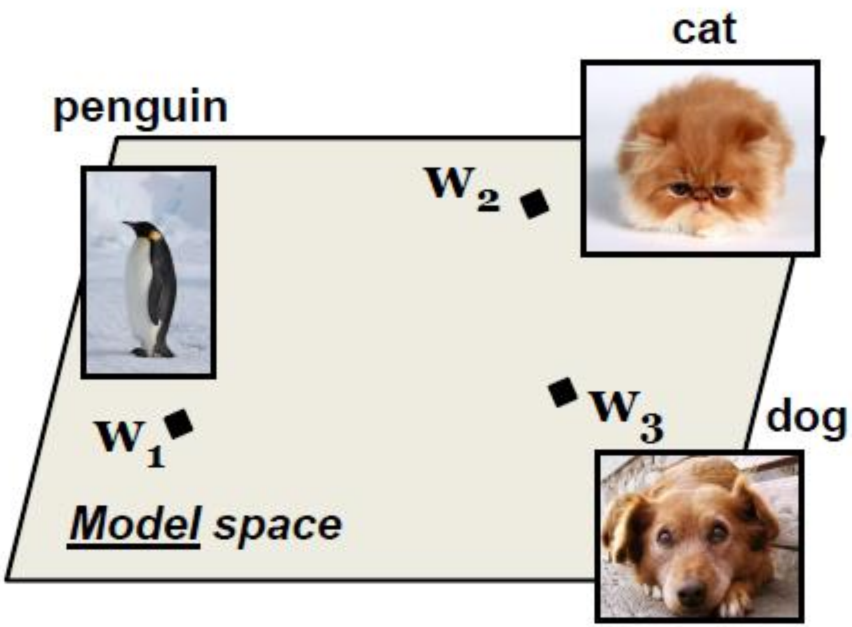
Semantic



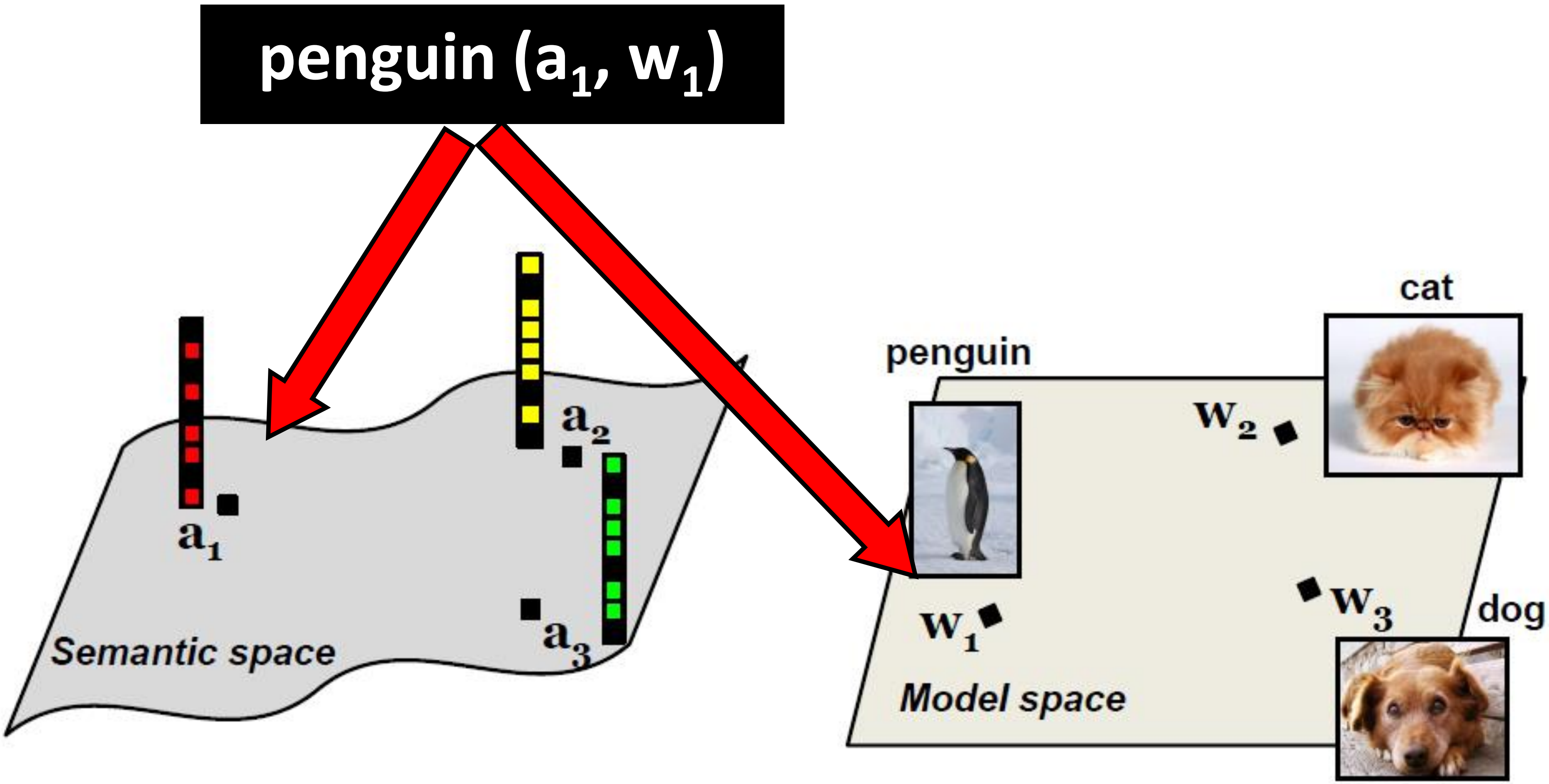
Our Approach: Manifold Learning

Model

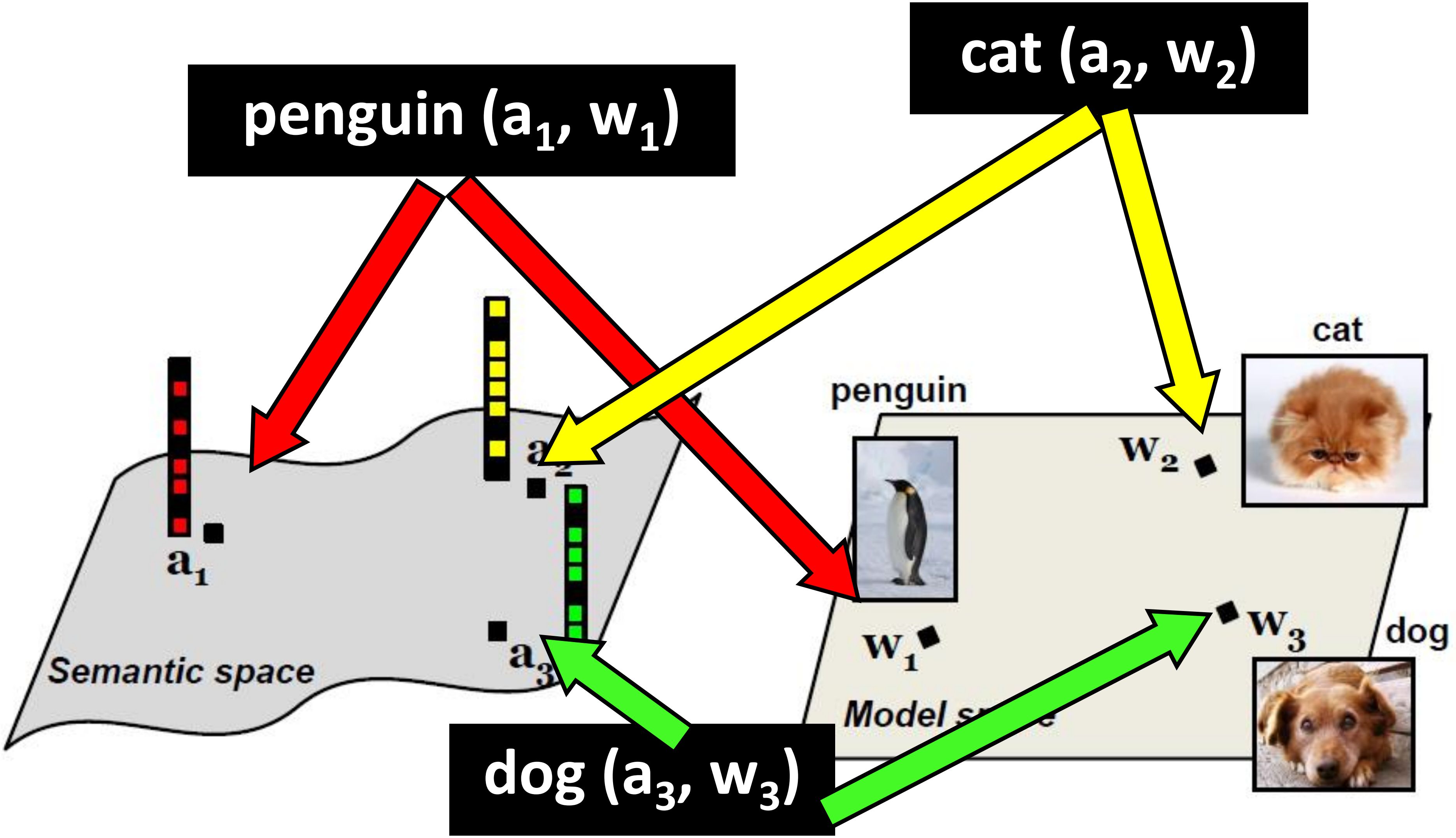




Our Approach: Manifold Learning



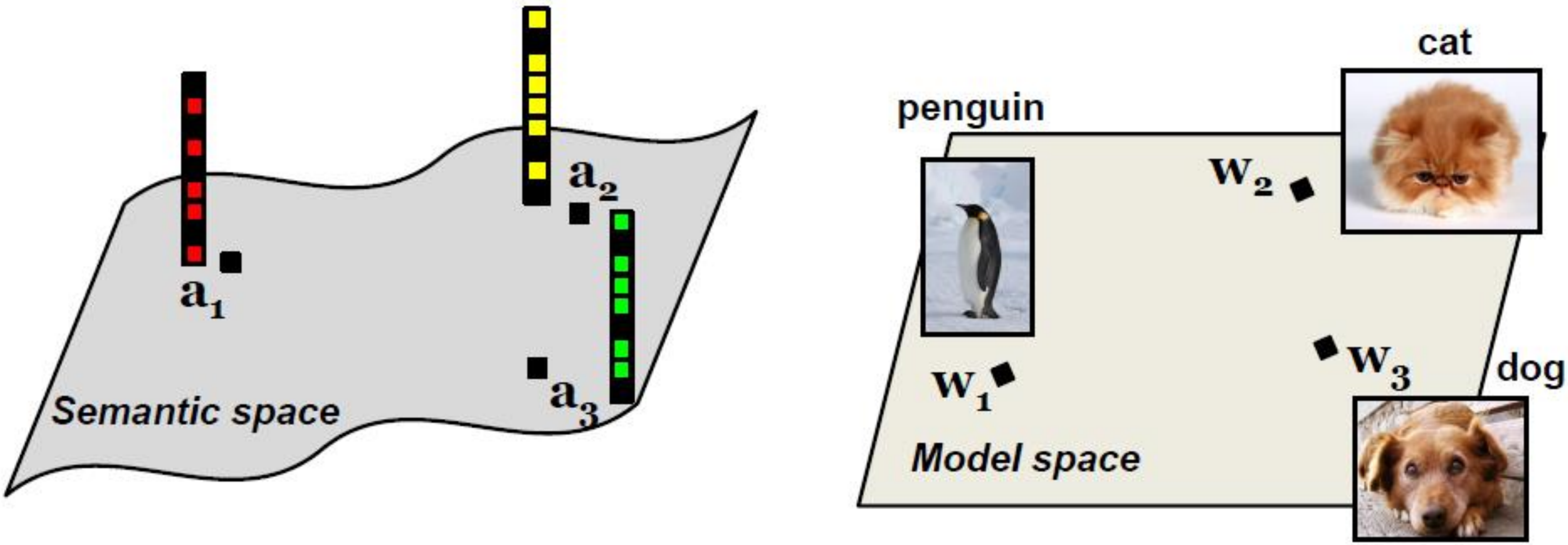
Our Approach: Manifold Learning



Our Approach: Manifold Learning

Main Idea

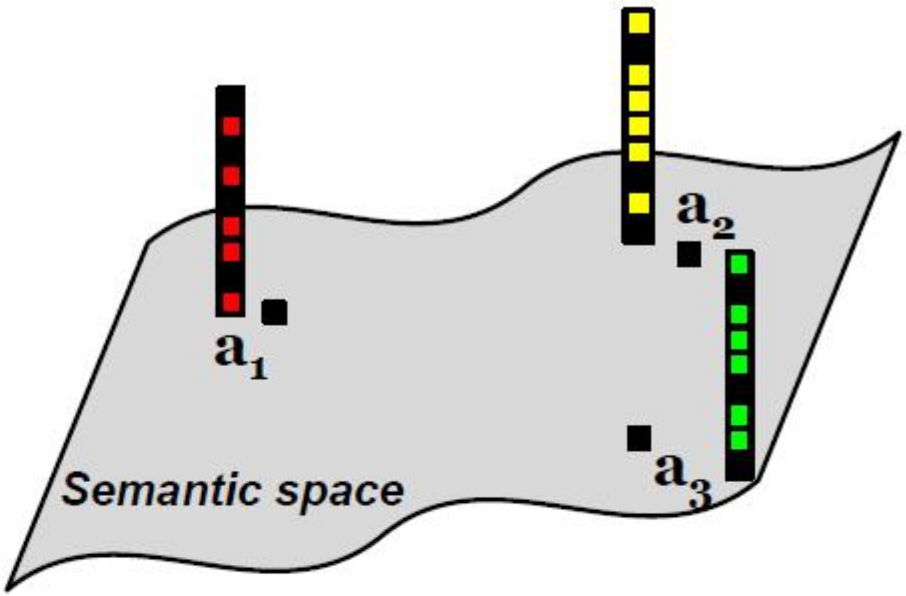
Align the two manifolds

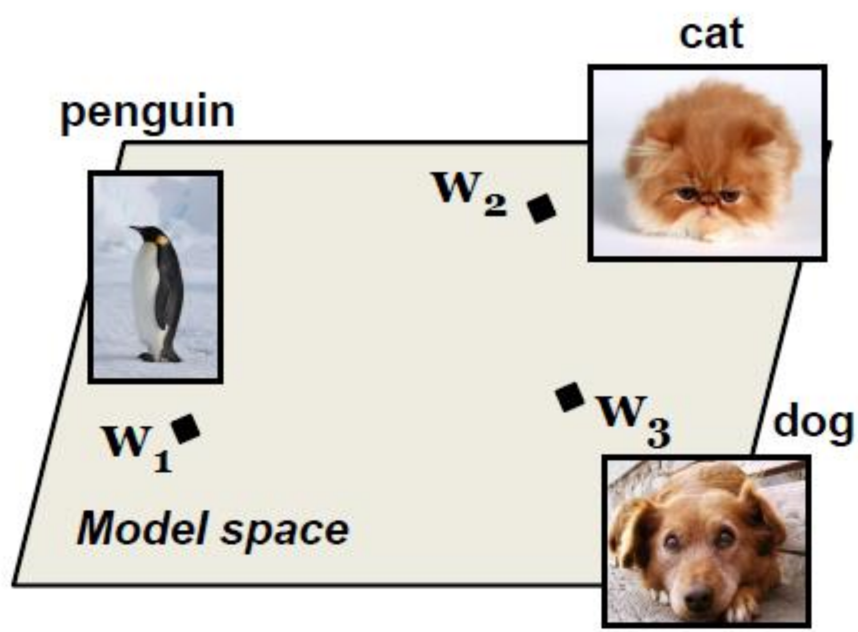


Our Approach: Manifold Learning

If we can align the two manifolds…

We can construct classifiers for ANY classes according to their semantic information.

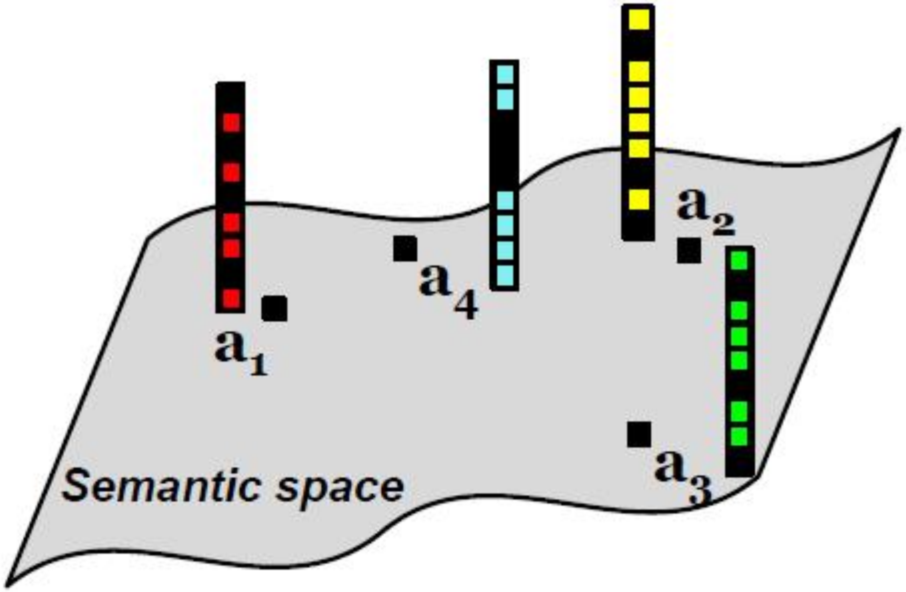


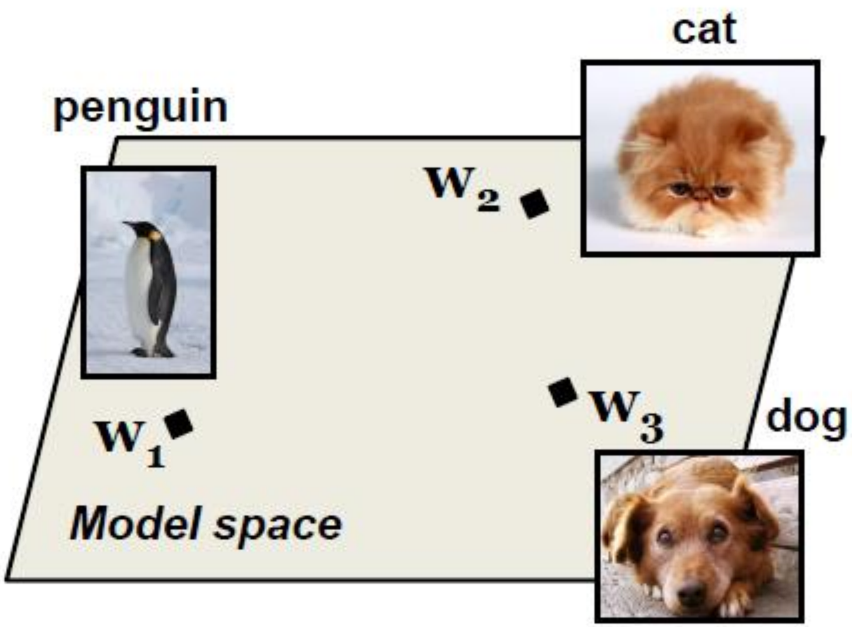


Our Approach: Manifold Learning

If we can align the two manifolds…

We can construct classifiers for ANY classes according to their semantic information.

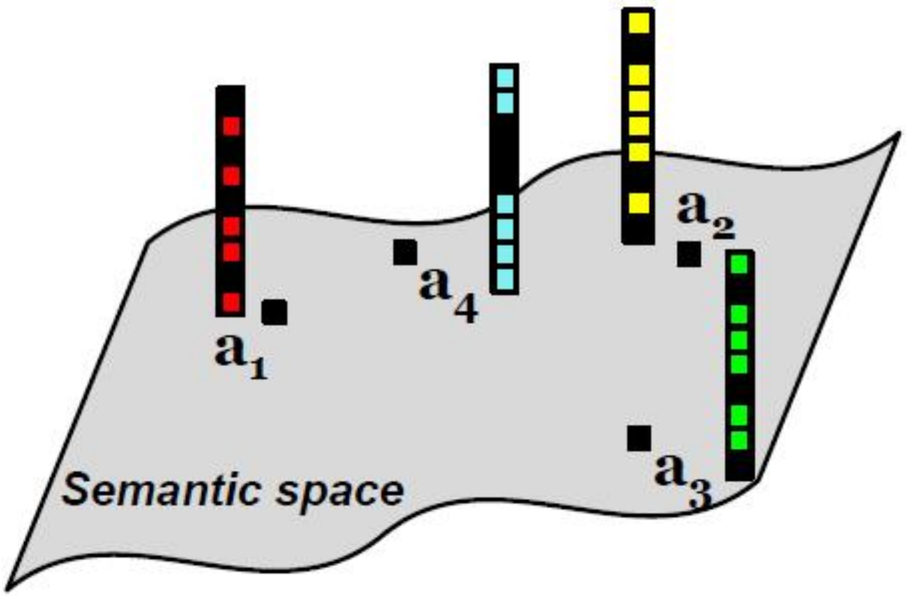


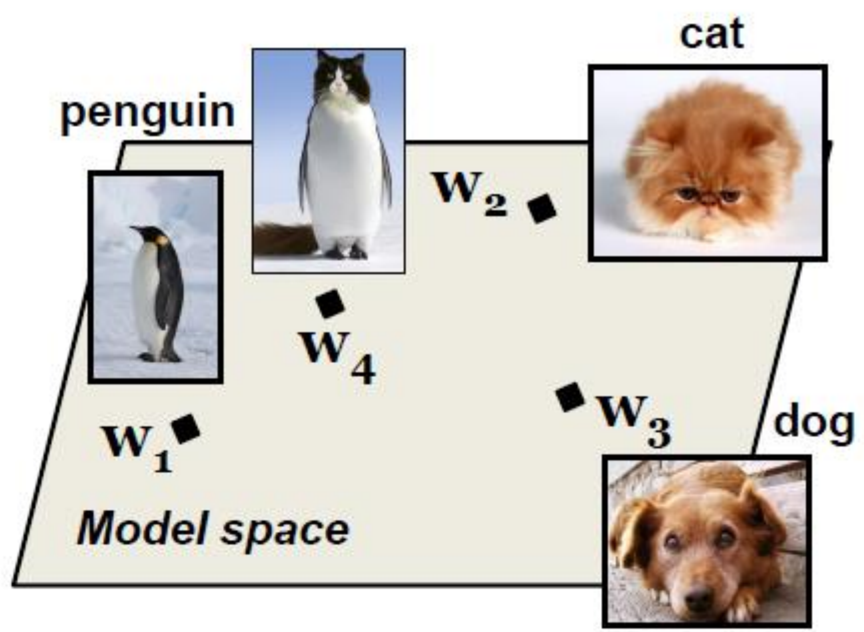


Our Approach: Manifold Learning

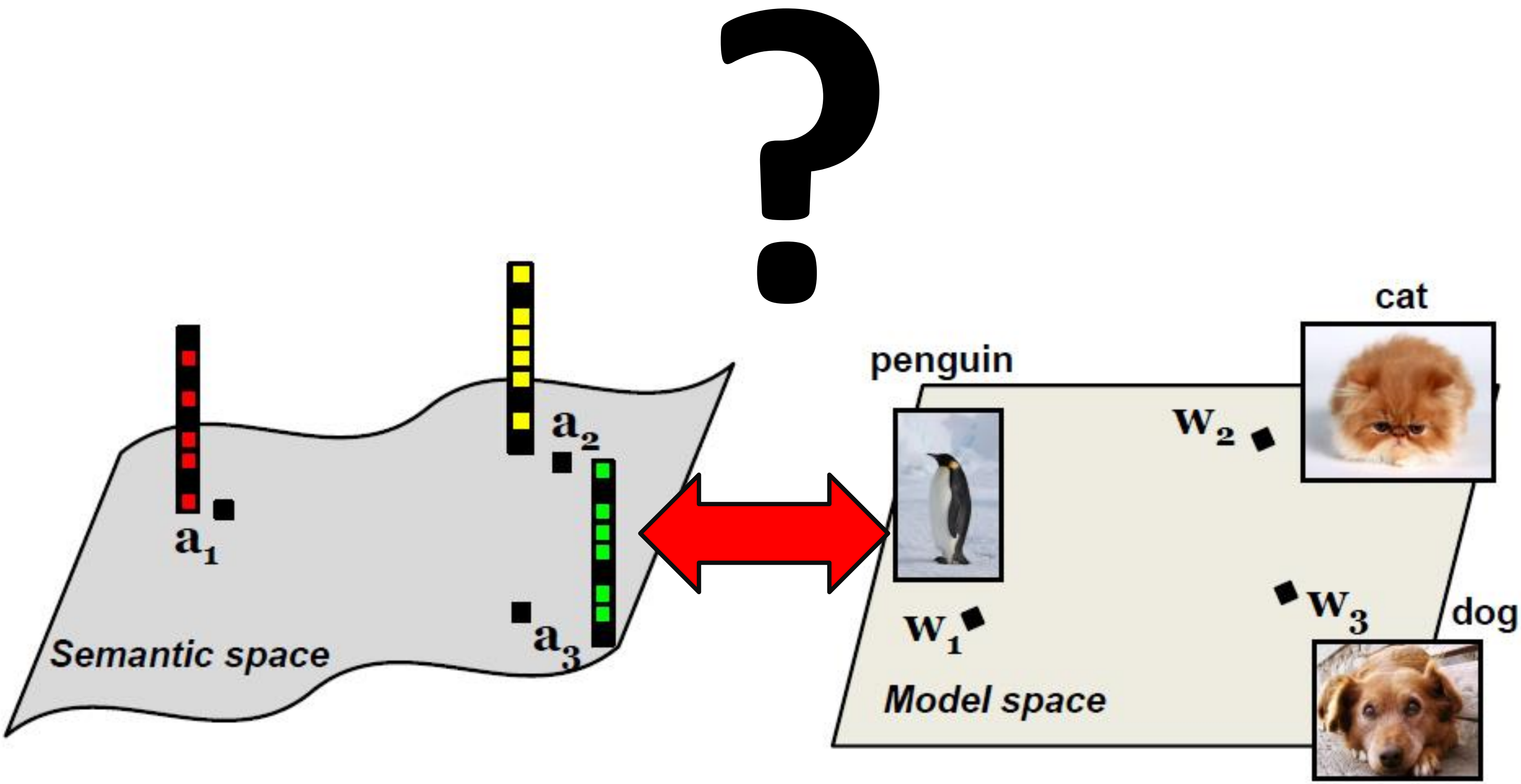
If we can align the two manifolds…

We can construct classifiers for ANY classes according to their semantic information.





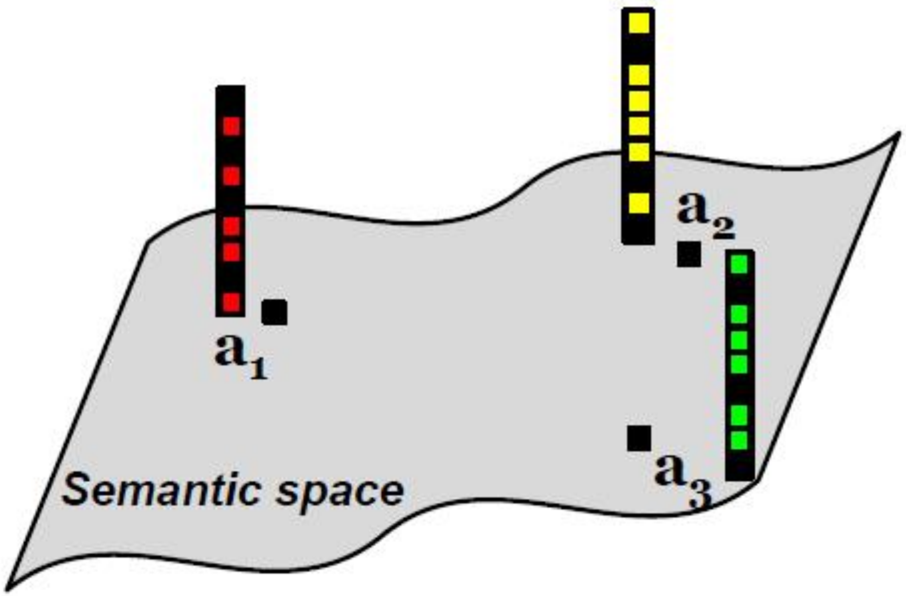
Aligning Manifolds

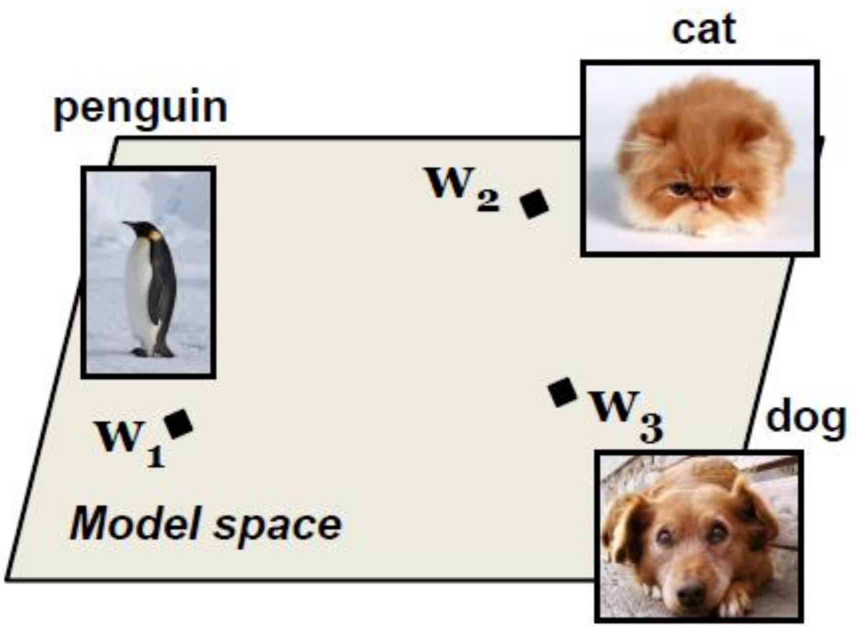


Aligning Manifolds

phantom classes

not corresponding to any objects in the real world

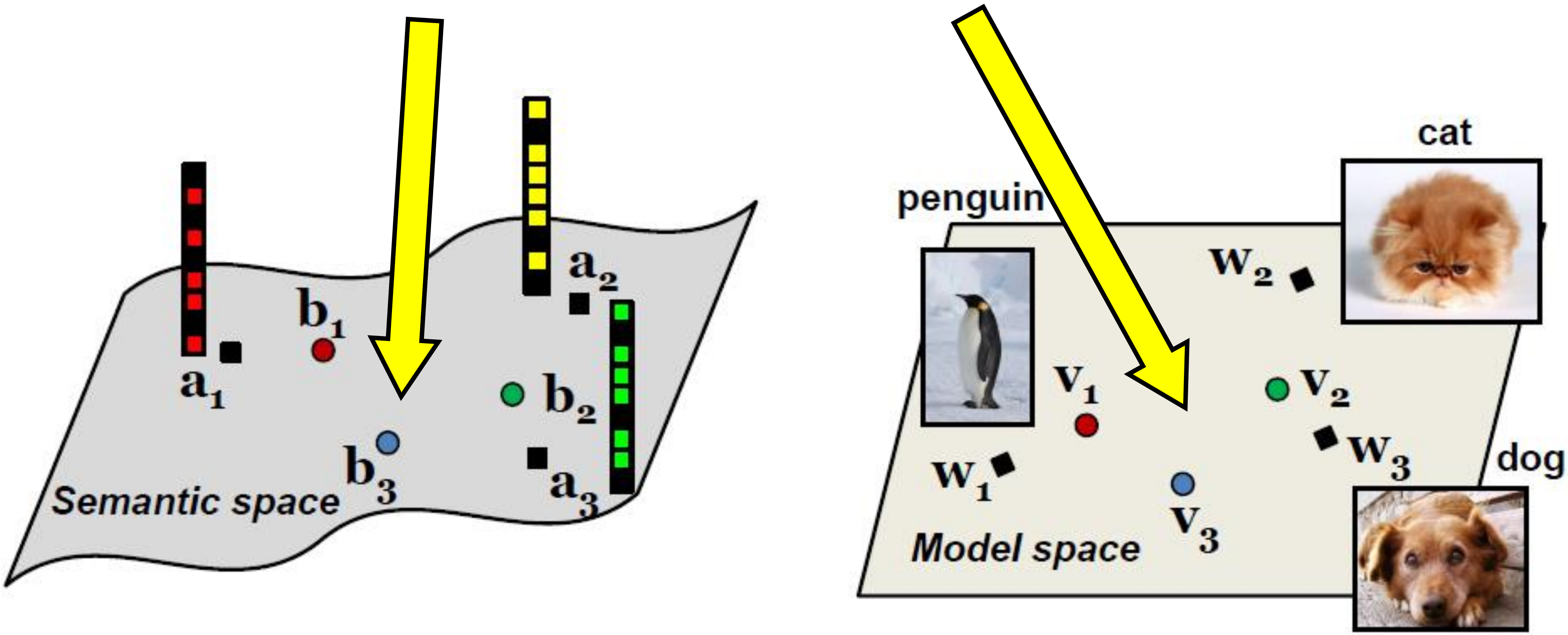




Aligning Manifolds

***phantom* classes**

b_r (semantic) and v_r (model)

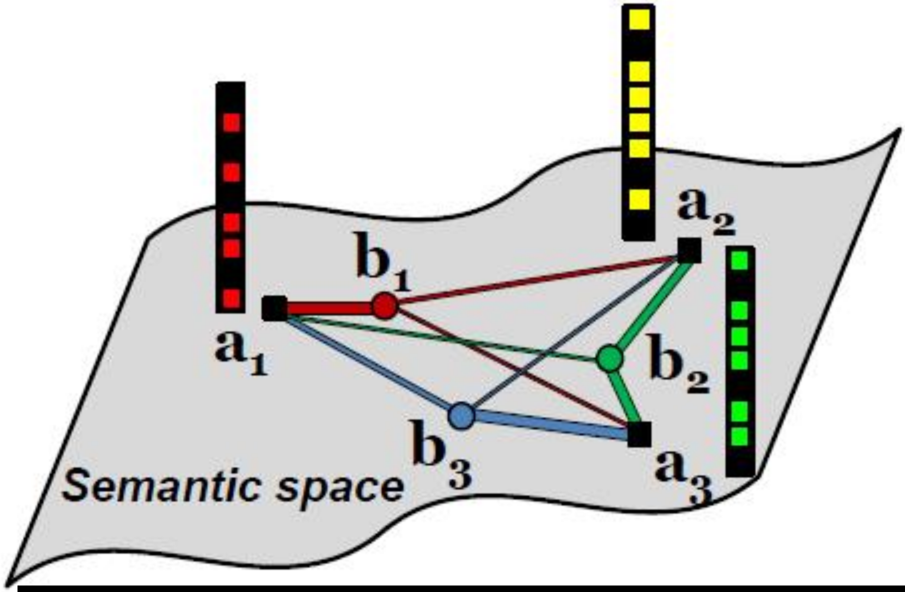


Aligning Manifolds

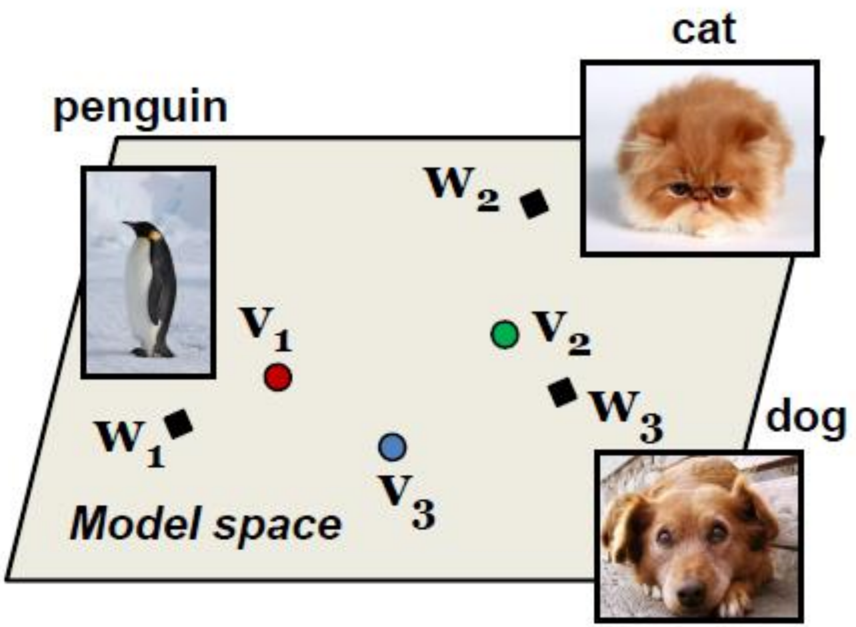
$$s_{cr} = \frac{\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}{\sum_{r=1}^{R} \exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}$$

$$d(\boldsymbol{a}_c, \boldsymbol{b}_r) = (\boldsymbol{a}_c - \boldsymbol{b}_r)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{a}_c - \boldsymbol{b}_r)$$

Define relationships s_{cr} between actual class c and phantom class r in the semantic space



Semantic weighted graph

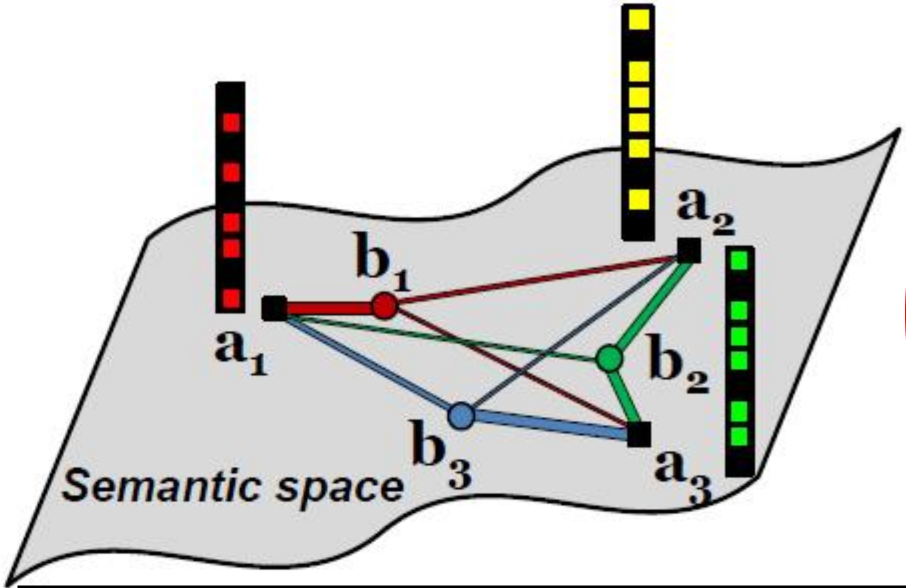


Aligning Manifolds

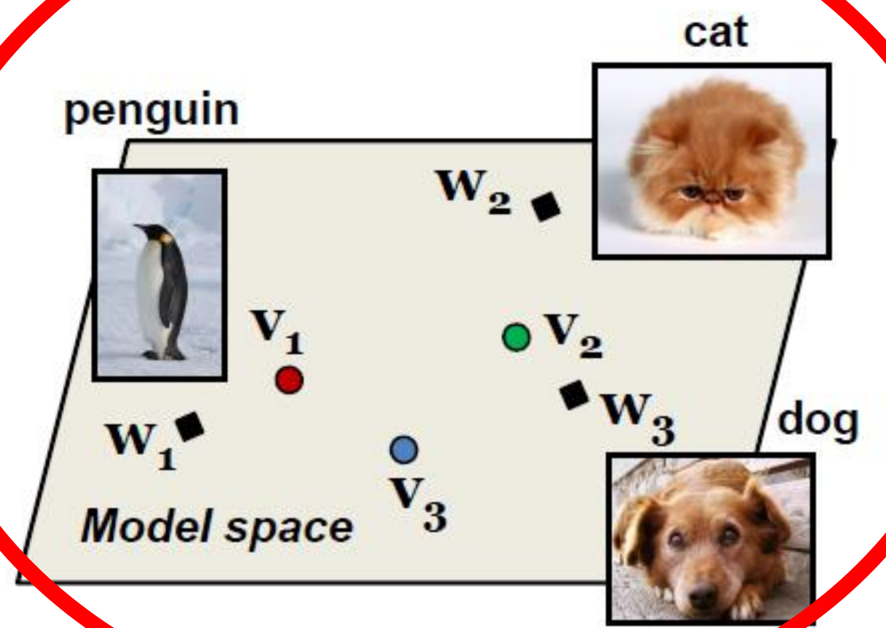
$$s_{cr} = \frac{\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}{\sum_{r=1}^{\mathsf{R}} \exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}$$

$$d(\boldsymbol{a}_c, \boldsymbol{b}_r) = (\boldsymbol{a}_c - \boldsymbol{b}_r)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{a}_c - \boldsymbol{b}_r)$$

View this as the embedding of the semantic weighted graph



Semantic weighted graph

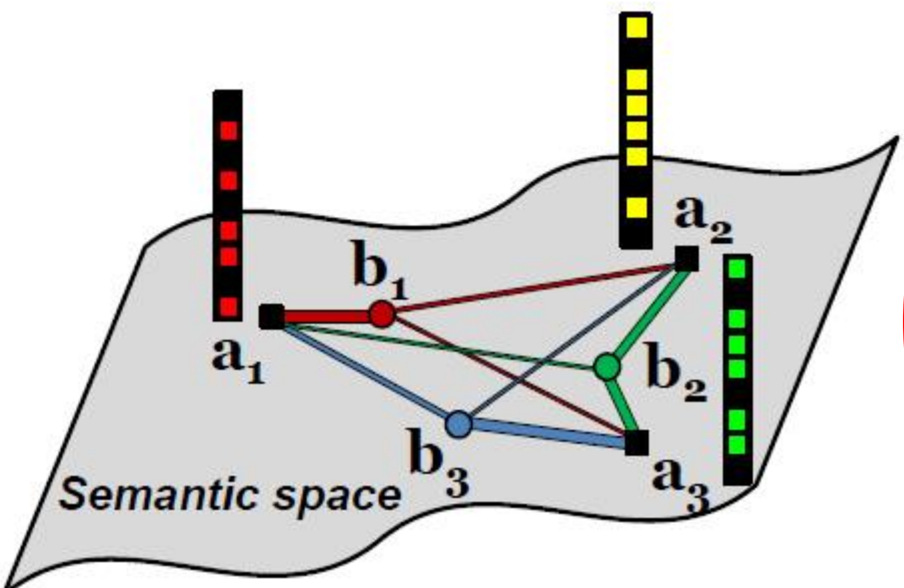


Aligning Manifolds

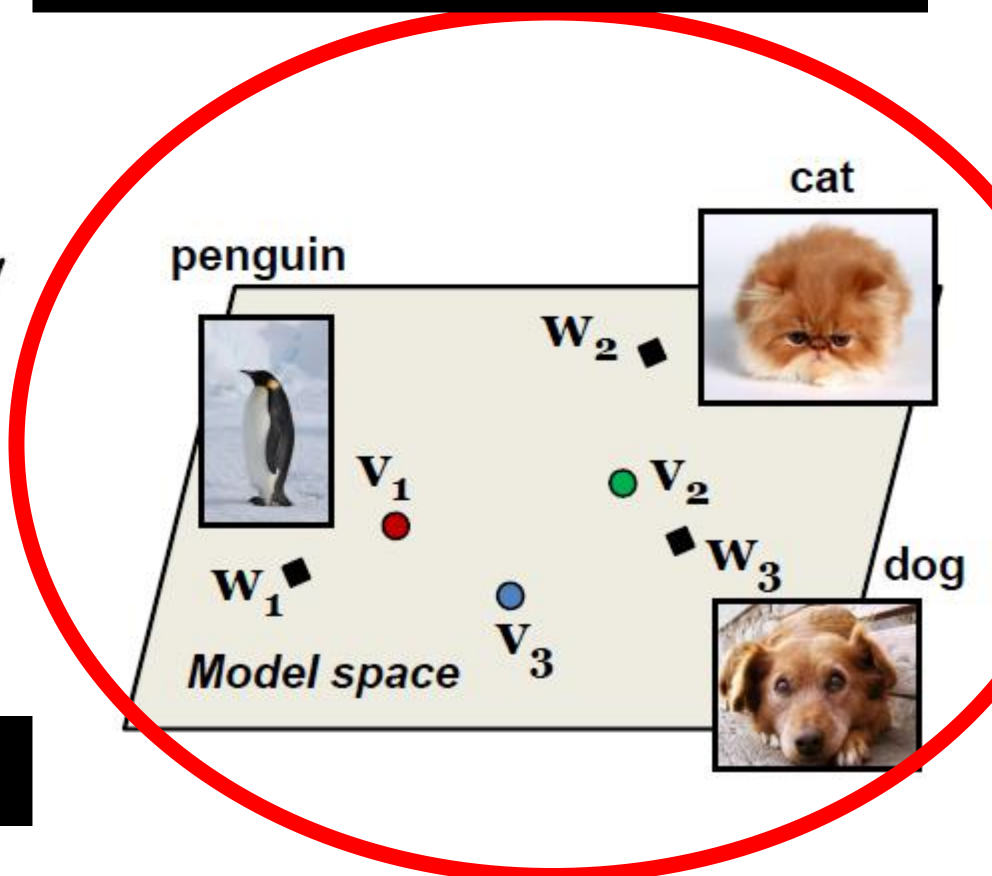
$$s_{cr} = \frac{\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}{\sum_{r=1}^{R} \exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}$$

$$d(\boldsymbol{a}_c, \boldsymbol{b}_r) = (\boldsymbol{a}_c - \boldsymbol{b}_r)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{a}_c - \boldsymbol{b}_r)$$

Let's preserve the structure of the semantic graph here as much as possible



Semantic weighted graph



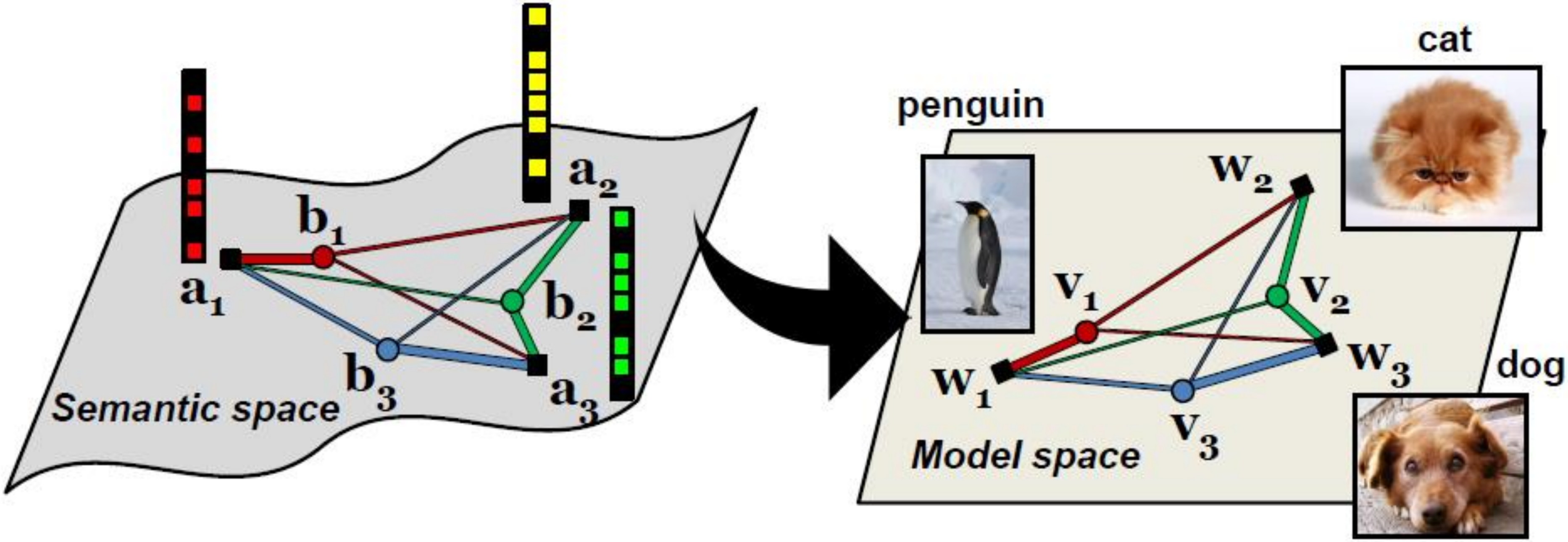
Aligning Manifolds

$$s_{cr} = \frac{\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}{\sum_{r=1}^{\mathrm{R}} \exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}$$

$$d(\boldsymbol{a}_c, \boldsymbol{b}_r) = (\boldsymbol{a}_c - \boldsymbol{b}_r)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{a}_c - \boldsymbol{b}_r)$$

$$\min_{\boldsymbol{w}_c, \boldsymbol{v}_r} \|\boldsymbol{w}_c - \sum_{r=1}^{\mathrm{R}} s_{cr} \boldsymbol{v}_r\|_2^2$$

$$\boxed{\boldsymbol{w}_c = \sum_{r=1}^{\mathrm{R}} s_{cr} \boldsymbol{v}_r}$$

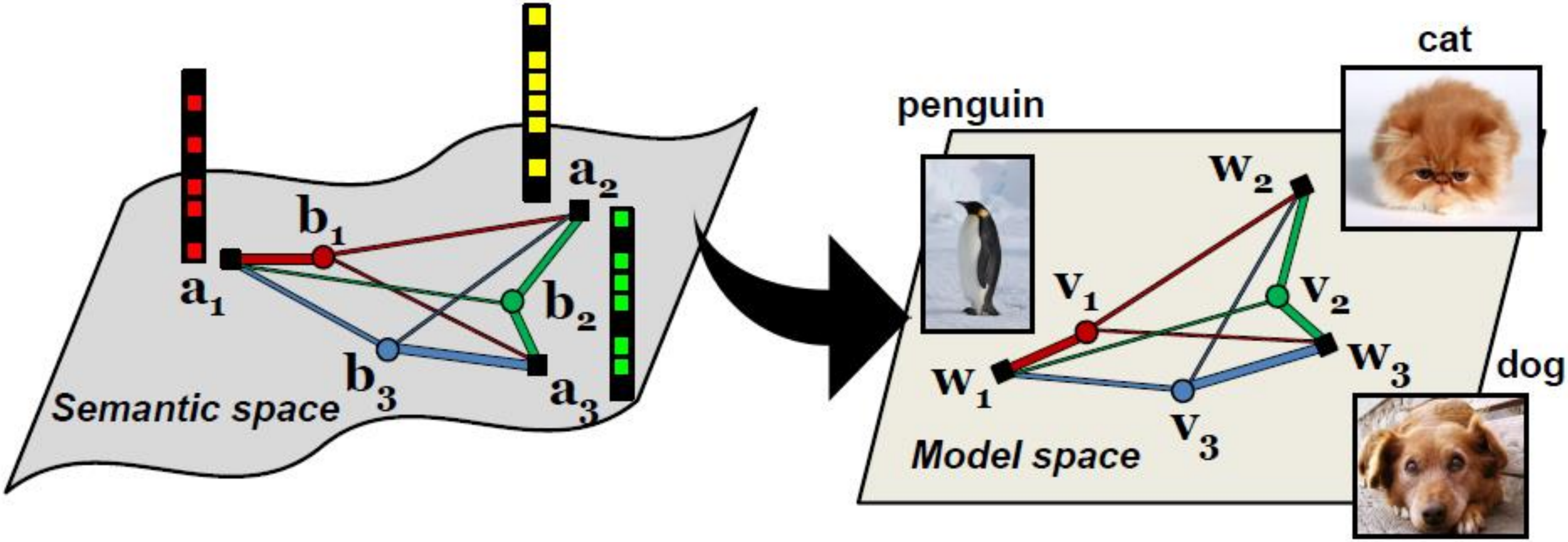


Aligning Manifolds

Formula for classifier synthesis!

$$\boxed{\boldsymbol{w}_c = \sum_{r=1}^{R} s_{cr}\boldsymbol{v}_r}$$

$$s_{cr} = \frac{\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}{\sum_{r=1}^{R}\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}$$

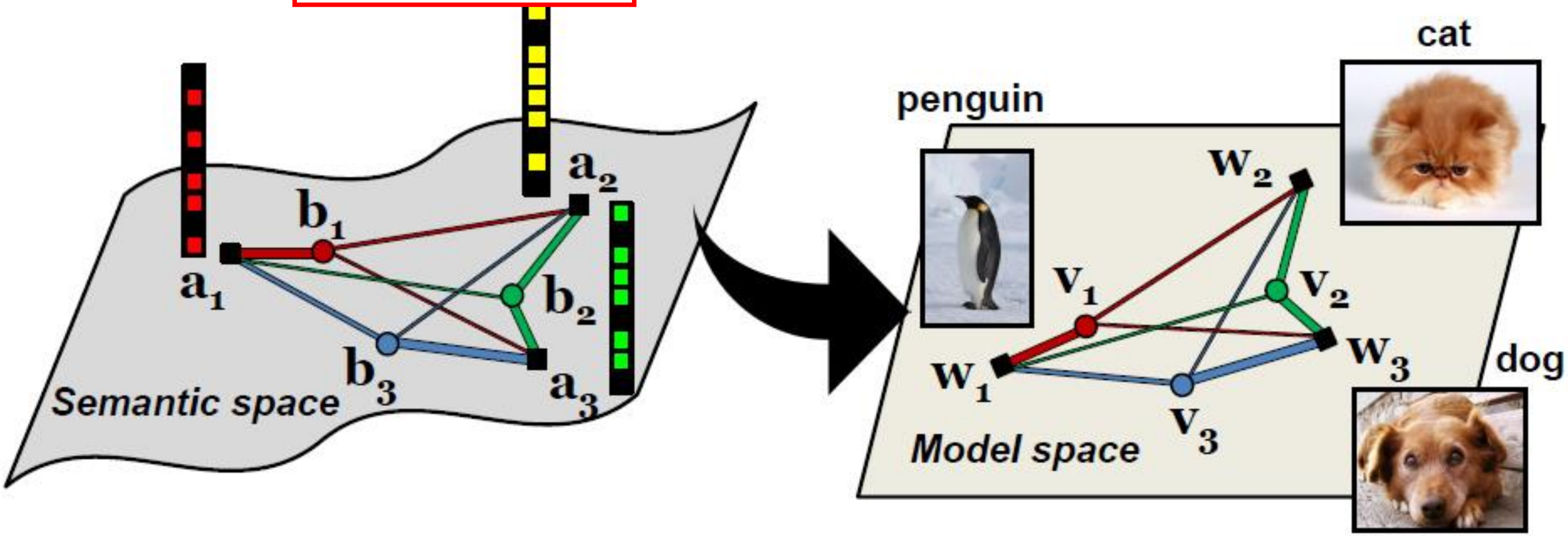


Learning Problem

Learn phantom coordinates v and b for optimal discrimination and generalization performance

$$\boldsymbol{w}_c = \sum_{r=1}^{R} s_{cr}\boldsymbol{v}_r$$

$$s_{cr} = \frac{\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}{\sum_{r=1}^{R}\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}$$



Experiments: Setup

- **Datasets**

	AwA (animals)	**CUB (birds)**	**SUN (scenes)**	**ImageNet**
# of seen classes	40	150	645/646	1,000
# of unseen classes	10	50	72/71	20,842
Total # of images	30,475	11,788	14,340	14,197,122
Semantic embeddings	attributes	attributes	attributes	word vectors

- **Visual features**: GoogLeNet
- **Evaluation**
  - Test images from **unseen classes only**
  - Accuracy of classifying them into **one of the unseen classes**

Experiments: AwA, CUB, SUN

Methods	AwA	CUB	SUN
DAP [*Lampert et al. 09 and 14*]	60.5	39.1	44.5
SJE [*Akata et al. 15*]	66.7	50.1	56.1
ESZSL [*Romera-Paredes et a. 15*]	64.5	44.0	18.7
ConSE [*Norouzi et al. 14*]	63.3	36.2	51.9
COSTA [*Mensink et al. 14*]	61.8	40.8	47.9
Sync$^{\text{o-vs-o}}$ (R, b_r fixed)	69.7	53.4	62.8
Sync$^{\text{struct}}$ (R, b_r fixed)	**72.9**	**54.5**	**62.7**
Sync$^{\text{o-vs-o}}$ (R fixed, b_r learned)	**71.1**	**54.2**	**63.3**

o-vs-o (one-versus-all), struct (Crammer-Singer with l_2 structure loss)
R: the number of phantom classes (fixed to the number of seen classes)
b_r: the semantic embeddings of phantom classes

Experiments: Setup on Full ImageNet

- **3 types of unseen classes**
  - ***2-hop**** **from seen classes 1509 classes**
  - ***3-hop**** **from seen classes 7678 classes**
  - ***All*** **20345 classes**

Harder

- **Metric**
  - ***Flat hit@K***

    Do top K predictions contain the true label?

*** Based on WordNet hierarchy**

Experiments: ImageNet (22K)

Flat Hit@K

2-hop

Methods	**1**	**2**	**5**	**10**	**20**
ConSE [*Norouzi et al. 14*]	9.4	15.1	24.7	32.7	41.8
SynC$^{\text{o-vs-o}}$	**10.5**	**16.7**	**28.6**	**40.1**	**52.0**
SynC$^{\text{struct}}$	9.8	15.3	25.8	35.8	46.5

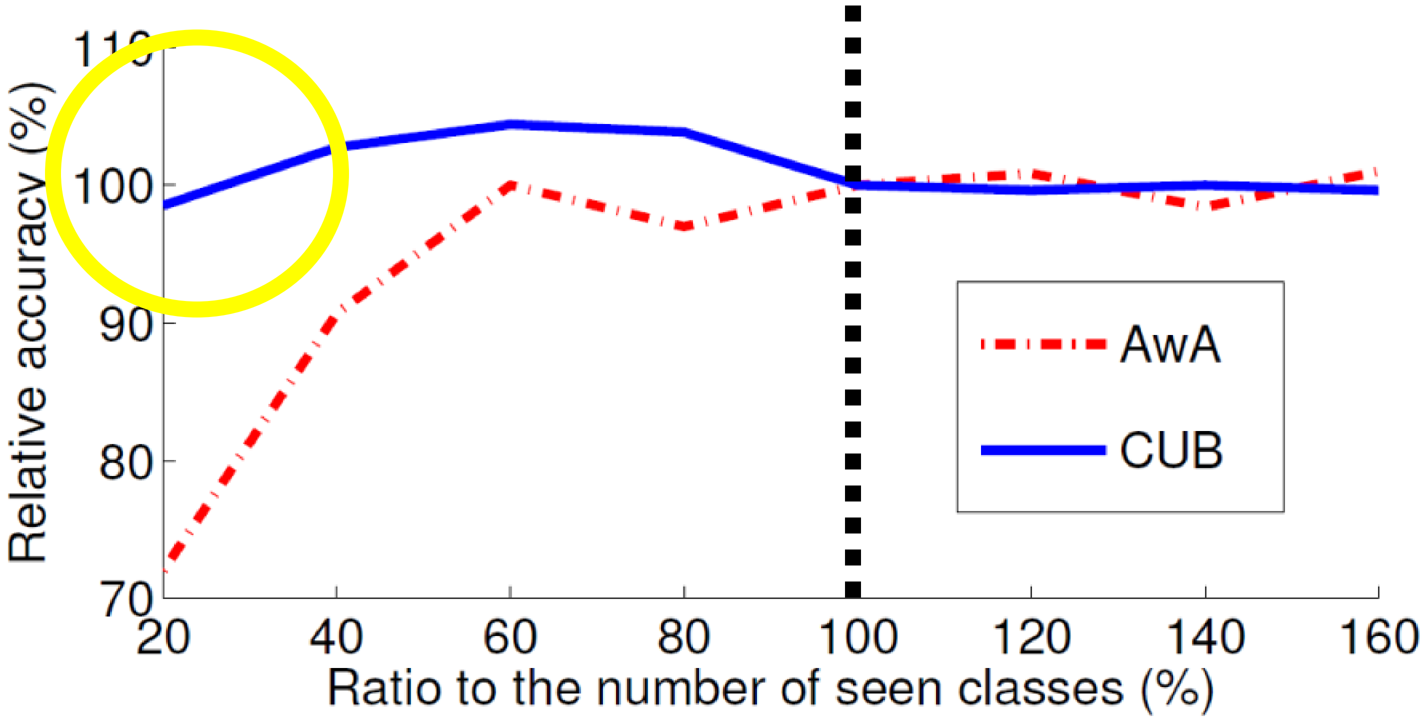
3-hop

Methods	**1**	**2**	**5**	**10**	**20**
ConSE [*Norouzi et al. 14*]	2.7	4.4	7.8	11.5	16.1
SynC$^{\text{o-vs-o}}$	**2.9**	**4.9**	**9.2**	**14.2**	**20.9**
SynC$^{\text{struct}}$	2.9	4.7	8.7	13.0	18.6

All

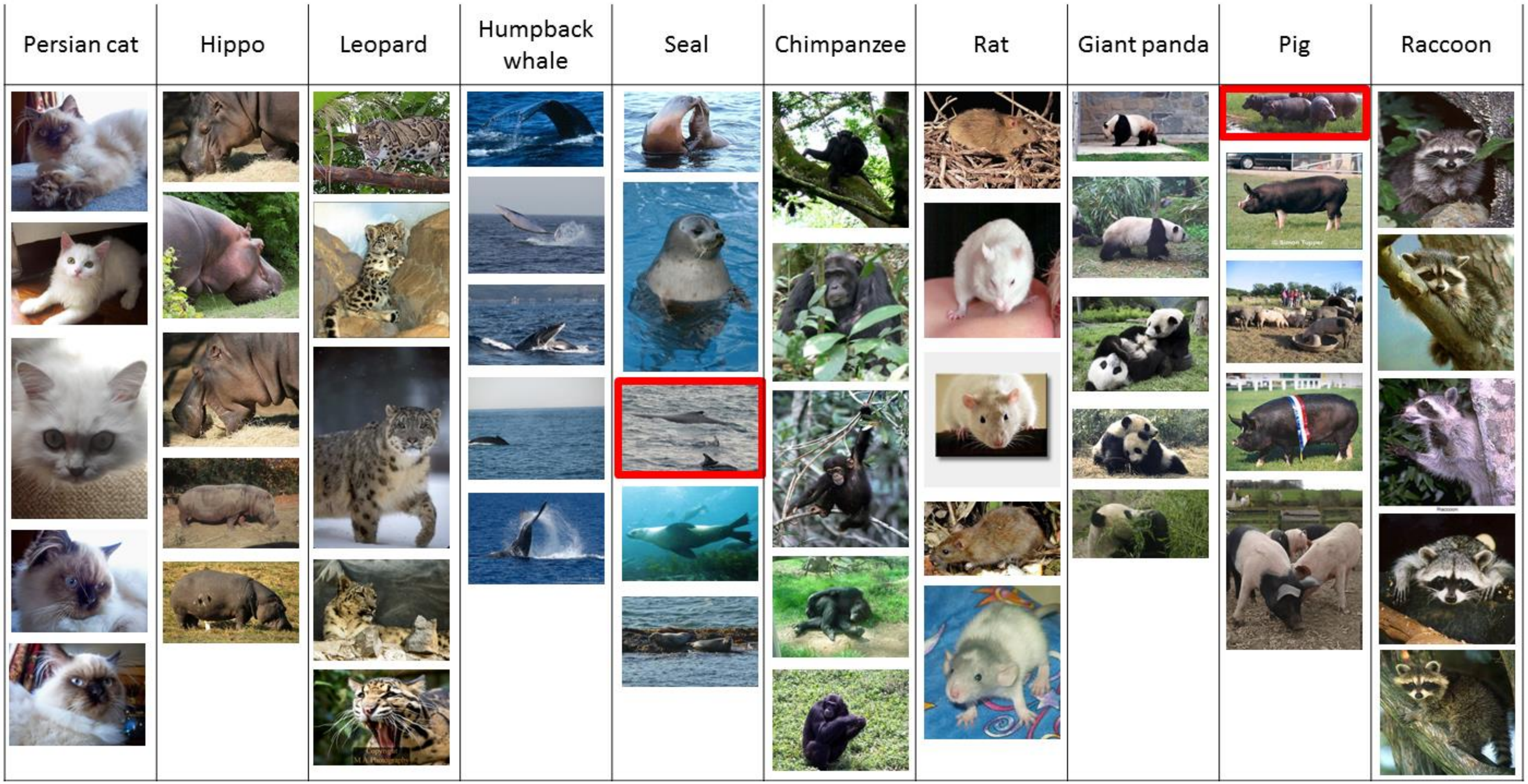
Methods	**1**	**2**	**5**	**10**	**20**
ConSE [*Norouzi et al. 14*]	1.4	2.2	3.9	5.8	8.3
SynC$^{\text{o-vs-o}}$	1.4	**2.4**	**4.5**	**7.1**	**10.9**
SynC$^{\text{struct}}$	**1.5**	**2.4**	4.4	6.7	10.0

Experiments: Number of phantom classes



Top 5 images

AwA dataset



Poster ID 4

Conclusion

Summary

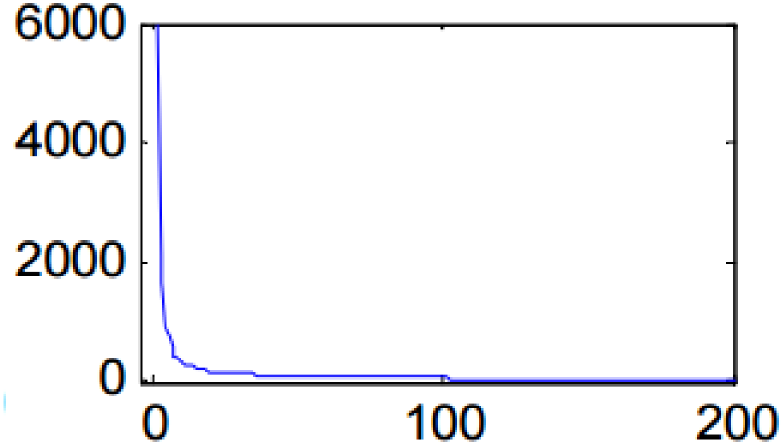
- ✓ Novel **classifier synthesis mechanism** with the state-of-the-art performance on zero-shot learning
- ✓ More results and analysis in the paper

Future work

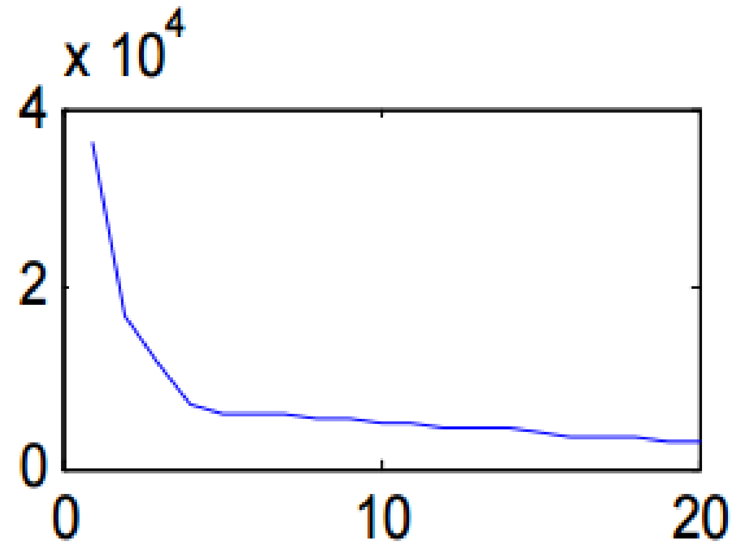
- ✓ **New challenging problem**: we cannot assume future objects only come from unseen classes.
  https://arxiv.org/abs/1605.04253

Thanks!

The Long Tail Phenomena



Objects in ImageNet detection task



Objects in VOC07 detection task

Ouyang et al. CVPR 2016

Current Approaches

- **Embedding based**
  - **Two-stage** *(Lampert et al. 09, Frome et al. 13, Norouzi et al. 14, …)*
    Features → Semantic embeddings → Labels
  - **Unified** *(Akata et al. 13 and 15, Romera-Paredes et al. 15, …)*
    Learning scoring function between features and semantic embeddings of labels
- **Similarity based**
  - **Semantic embeddings define how to combine seen classes' classifiers** *(Mensink et al. 14, …)*

We propose a unified approach that offers richer flexibility in constructing new classifiers than previous approaches.

Learning phantom coordinates

Phantom coordinates in both spaces are **optimized** for optimal discrimination and generalization performance.

$$\min_{\{\boldsymbol{v}_r\}_{r=1}^{\mathrm{R}},\{\beta_{rc}\}_{r,c=1}^{\mathrm{R,S}}} \boxed{\sum_{c=1}^{\mathrm{S}}\sum_{n=1}^{\mathrm{N}} \ell(\boldsymbol{x}_n, \mathbb{I}_{y_n,c}; \boldsymbol{w}_c) + \frac{\lambda}{2}\sum_{c=1}^{\mathrm{S}} \|\boldsymbol{w}_c\|_2^2}$$

Classification loss + Regularizer on classifier weights

$$+\,\eta \sum_{r,c=1}^{\mathrm{R,S}} |\beta_{rc}| + \frac{\gamma}{2}\sum_{r=1}^{\mathrm{R}} (\|\boldsymbol{b}_r\|_2^2 - h^2)^2,$$

$$\text{s.t.} \quad \boxed{\boldsymbol{w}_c = \sum_{r=1}^{\mathrm{R}} s_{cr}\boldsymbol{v}_r, \; s_{cr} = \frac{\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}{\sum_{r=1}^{\mathrm{R}} \exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}}$$

Synthesis mechanism

$$\boldsymbol{b}_r = \sum_{c=1}^{\mathrm{S}} \beta_{rc}\boldsymbol{a}_c, \forall r \in \{1, \cdots, \mathrm{R}\}$$

Learning phantom coordinates

Phantom coordinates in both spaces are **optimized** for optimal discrimination and generalization performance.

$$\min_{\{\boldsymbol{v}_r\}_{r=1}^{\mathsf{R}},\{\beta_{rc}\}_{r,c=1}^{\mathsf{R,S}}} \sum_{c=1}^{\mathsf{S}}\sum_{n=1}^{\mathsf{N}} \ell(\boldsymbol{x}_n, \mathbb{I}_{y_n,c}; \boldsymbol{w}_c) + \frac{\lambda}{2}\sum_{c=1}^{\mathsf{S}} \|\boldsymbol{w}_c\|_2^2$$

$$+ \boxed{\eta \sum_{r,c=1}^{\mathsf{R,S}} |\beta_{rc}| + \frac{\gamma}{2}\sum_{r=1}^{\mathsf{R}} (\|\boldsymbol{b}_r\|_2^2 - h^2)^2,}$$

Regularizers on phantom classes

$$\text{s.t.} \quad \boldsymbol{w}_c = \sum_{r=1}^{\mathsf{R}} s_{cr}\boldsymbol{v}_r, \; s_{cr} = \frac{\exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}{\sum_{r=1}^{\mathsf{R}} \exp\{-d(\boldsymbol{a}_c, \boldsymbol{b}_r)\}}$$

$$\boxed{\boldsymbol{b}_r = \sum_{c=1}^{\mathsf{S}} \beta_{rc}\boldsymbol{a}_c, \forall r \in \{1, \cdots, \mathsf{R}\}}$$

Phantom semantic embedding is a sparse combination of real semantic coordinates

Experiments: Setup on Full ImageNet

- **3 types of unseen classes**
  - ***2-hop**** **from seen classes 1509 classes**
  - ***3-hop**** **from seen classes 7678 classes**
  - ***All*** **20345 classes**

Harder

- **2 types of metric**
  - ***Flat hit@K***

    Do top K predictions contain the true label?
  - ***Hierarchical precision@K***

    How much do top K predictions contain *similar** class to the true label?

More flexible

*** Based on WordNet hierarchy**

Experiments: ImageNet (22K)

Hierarchical Precision@K x 100

2-hop

Methods	2	5	10	20
ConSE [*Norouzi et al. 14*]	21.4	24.7	26.9	28.4
SynC$^{\text{o-vs-o}}$	**25.1**	**27.7**	**30.3**	**32.1**
SynC$^{\text{struct}}$	23.8	25.8	28.2	29.6

3-hop

Methods	2	5	10	20
ConSE [*Norouzi et al. 14*]	5.3	20.2	22.4	24.7
SynC$^{\text{o-vs-o}}$	7.4	**23.7**	**26.4**	**28.6**
SynC$^{\text{struct}}$	**8.0**	22.8	25.0	26.7

All

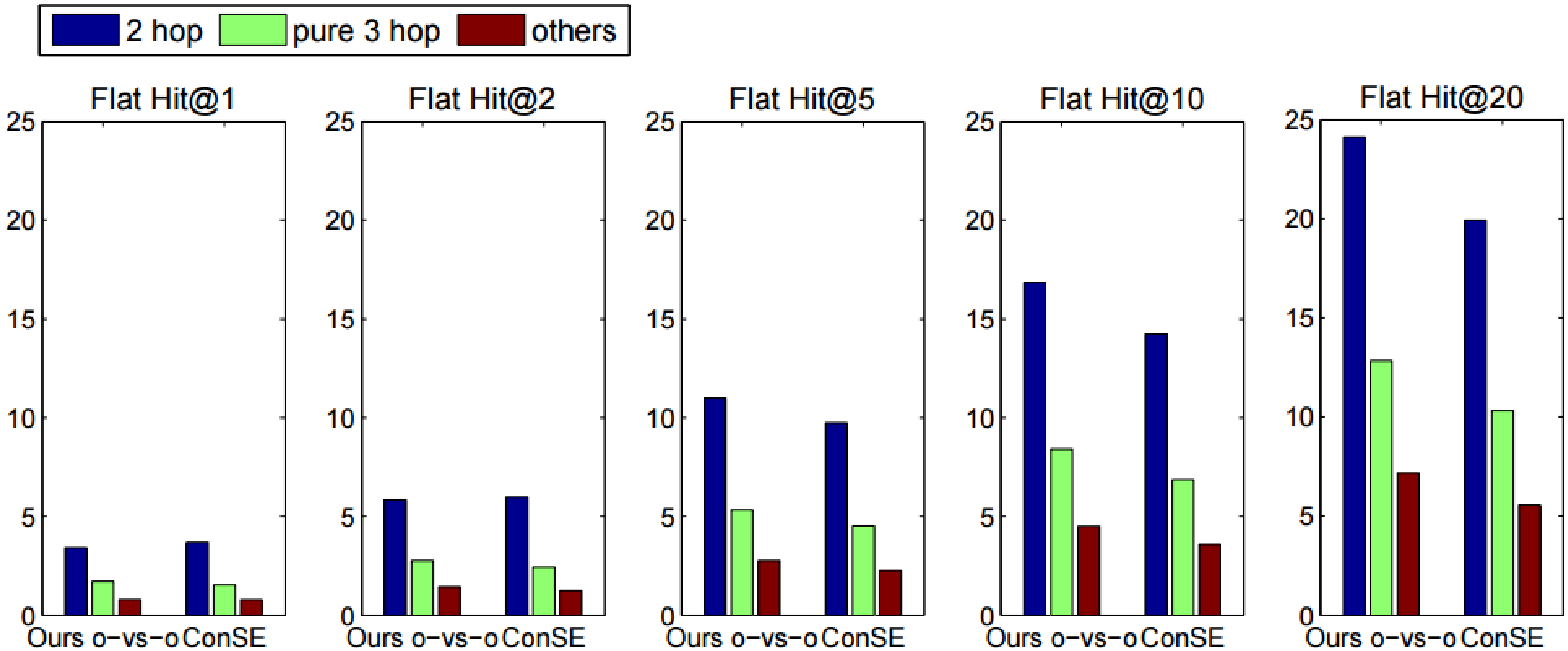
Methods	2	5	10	20
ConSE [*Norouzi et al. 14*]	2.5	7.8	9.2	10.4
SynC$^{\text{o-vs-o}}$	3.1	9.0	10.9	**12.5**
SynC$^{\text{struct}}$	**3.6**	**9.6**	**11.0**	12.2

Experiments: ImageNet (22K)

Scenarios	Methods K=	Flat Hit@K 1	2	5	10	20	Hierarchical precision@K 2	5	10	20
2-hop	ConSE[25]	9.4	15.1	24.7	32.7	41.8	21.4	24.7	26.9	28.4
	ConSE by us	8.3	12.9	21.8	30.9	41.7	21.5	23.8	27.5	31.3
	Ours$^{\text{o-vs-o}}$	10.5	16.7	28.6	40.1	52.0	25.1	27.7	30.3	32.1
	Ours$^{\text{struct}}$	9.8	15.3	25.8	35.8	46.5	23.8	25.8	28.2	29.6
3-hop	ConSE [25]	2.7	4.4	7.8	11.5	16.1	5.3	20.2	22.4	24.7
	ConSE by us	2.6	4.1	7.3	11.1	16.4	6.7	21.4	23.8	26.3
	Ours$^{\text{o-vs-o}}$	2.9	4.9	9.2	14.2	20.9	7.4	23.7	26.4	28.6
	Ours$^{\text{struct}}$	2.9	4.7	8.7	13.0	18.6	8.0	22.8	25.0	26.7
All	ConSE [25]	1.4	2.2	3.9	5.8	8.3	2.5	7.8	9.2	10.4
	ConSE by us	1.3	2.1	3.8	5.8	8.7	3.2	9.2	10.7	12.0
	Ours$^{\text{o-vs-o}}$	1.4	2.4	4.5	7.1	10.9	3.1	9.0	10.9	12.5
	Ours$^{\text{struct}}$	1.5	2.4	4.4	6.7	10.0	3.6	9.6	11.0	12.2

- **2-hop/3-hop/All: further from seen classes = harder**
- **Hierarchical precision: relax the definition of “correct”**

Experiments: ImageNet All (22K)



Accuracy for each type of classes **in All**

Experiments: Attribute v.s. Word vectors

Semantic embedding	Dimensions	Accuracy (%)
word vectors	100	42.2
word vectors	1000	57.5
attributes	85	69.7
attributes + word vectors	185	73.2
attributes + word vectors	1085	**76.3**

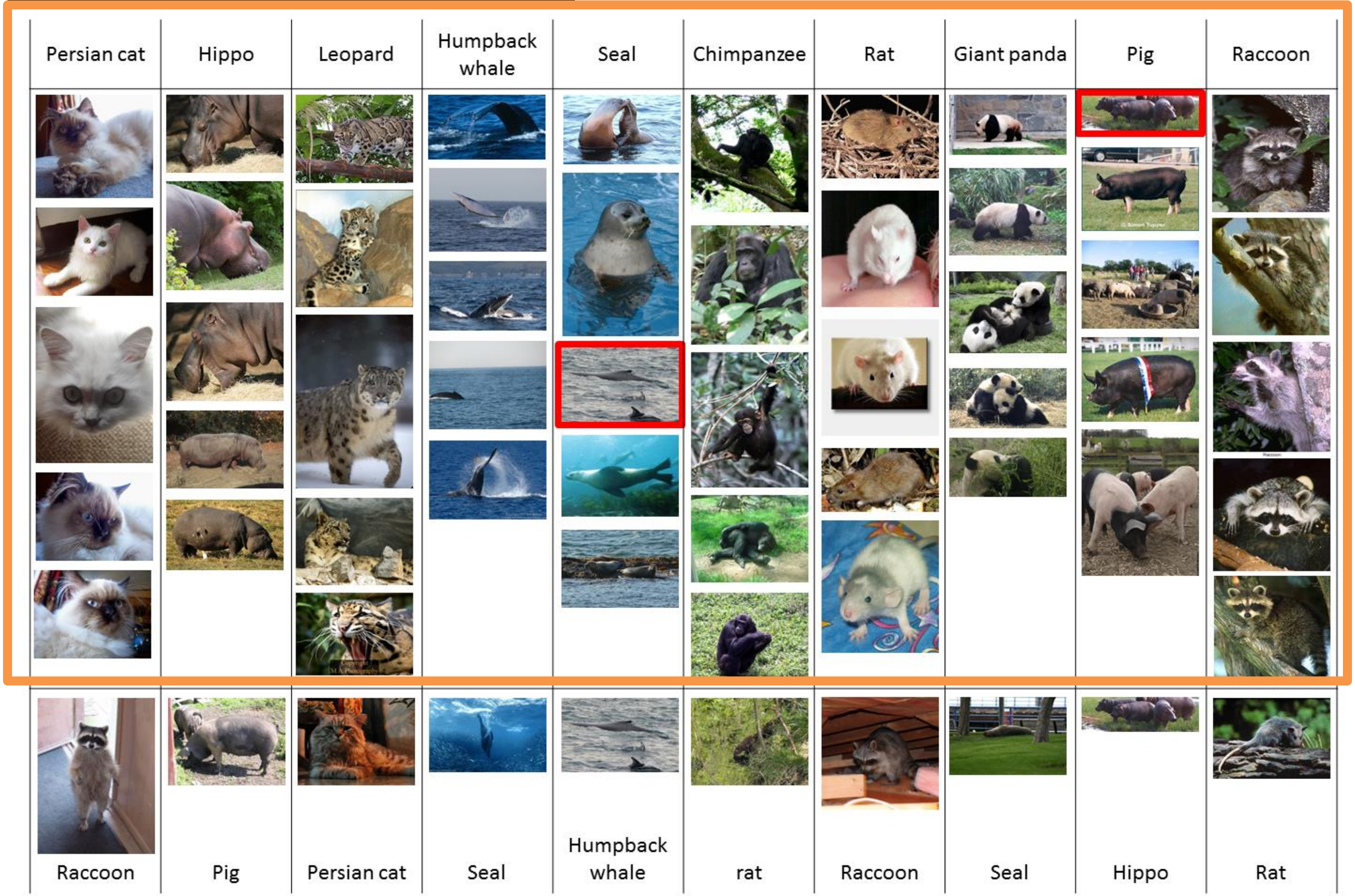
AwA dataset

Experiments: With vs. Without Learning Phantom Classes' Semantic Embeddings

Datasets	Types of embeddings	w/o learning	w/ learning
AwA	attributes	69.7%	71.1%
	100-d word vectors	42.2%	42.5%
	1000-d word vectors	57.6%	56.6%
CUB	attributes	53.4%	54.2%
SUN	attributes	62.8%	63.3%

Top: Top 5 images

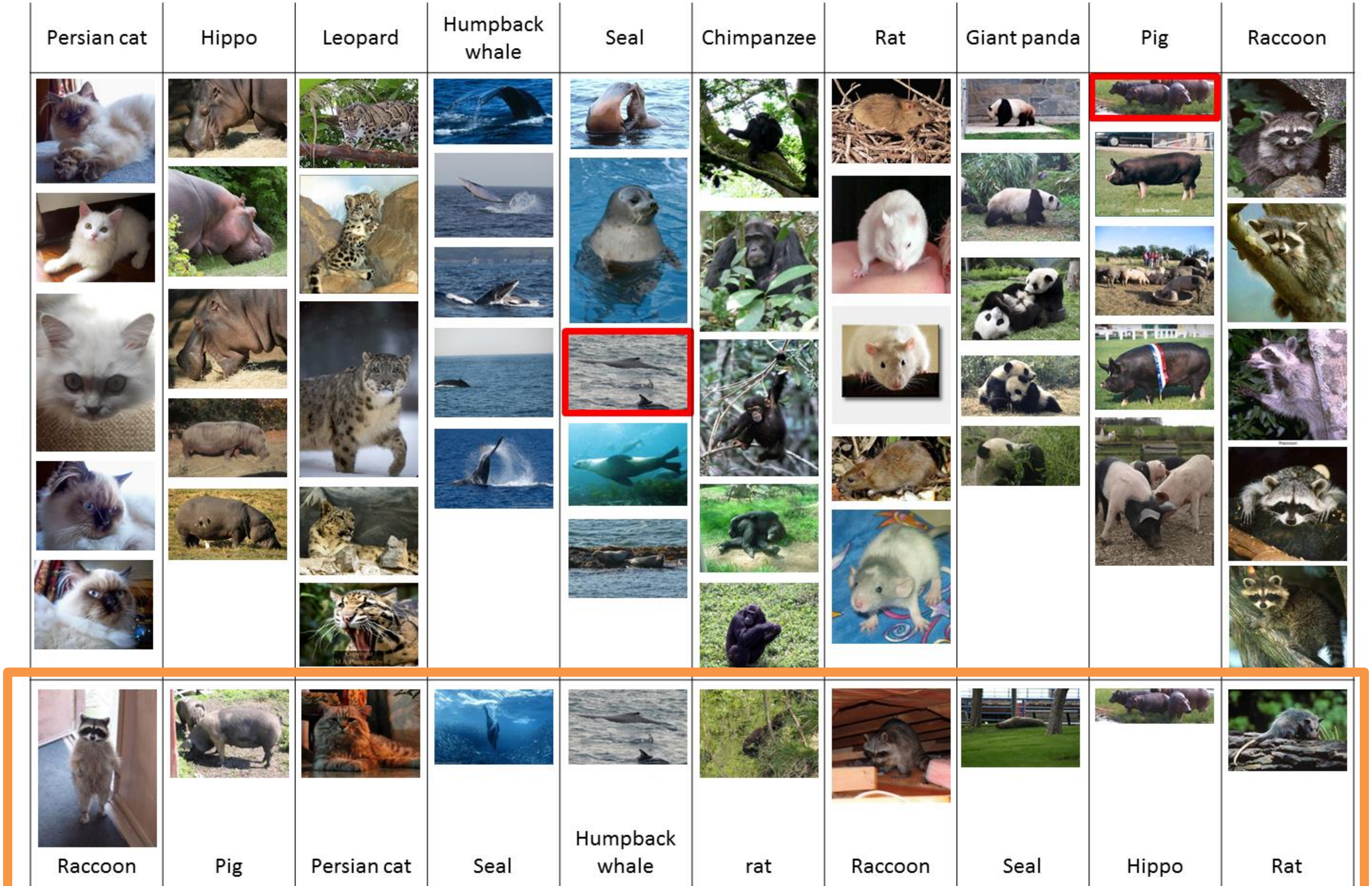
AwA dataset



Bottom: First misclassified image

Top: Top 5 images

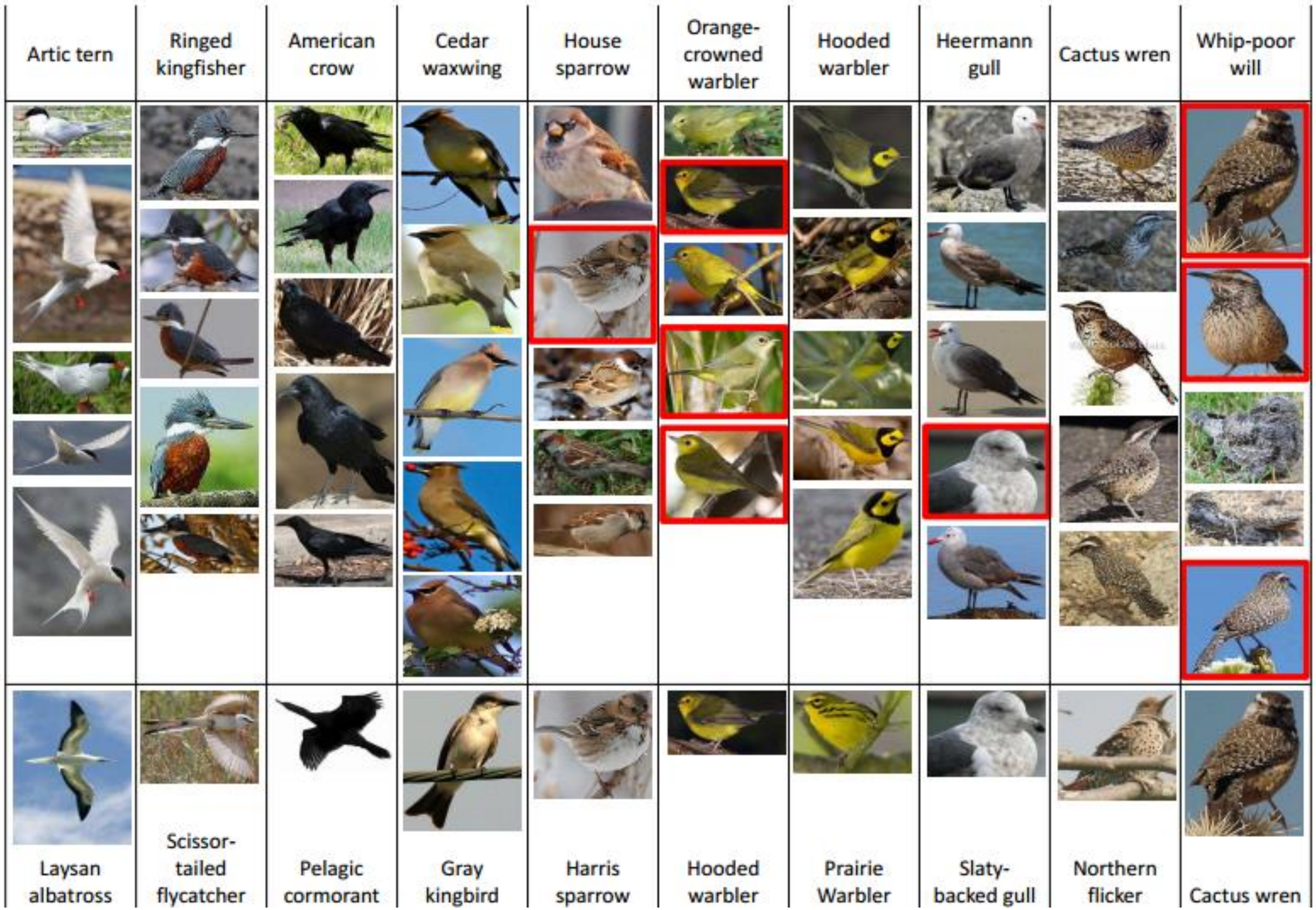
AwA dataset



Bottom: First misclassified image

Top: Top 5 predictions

CUB dataset

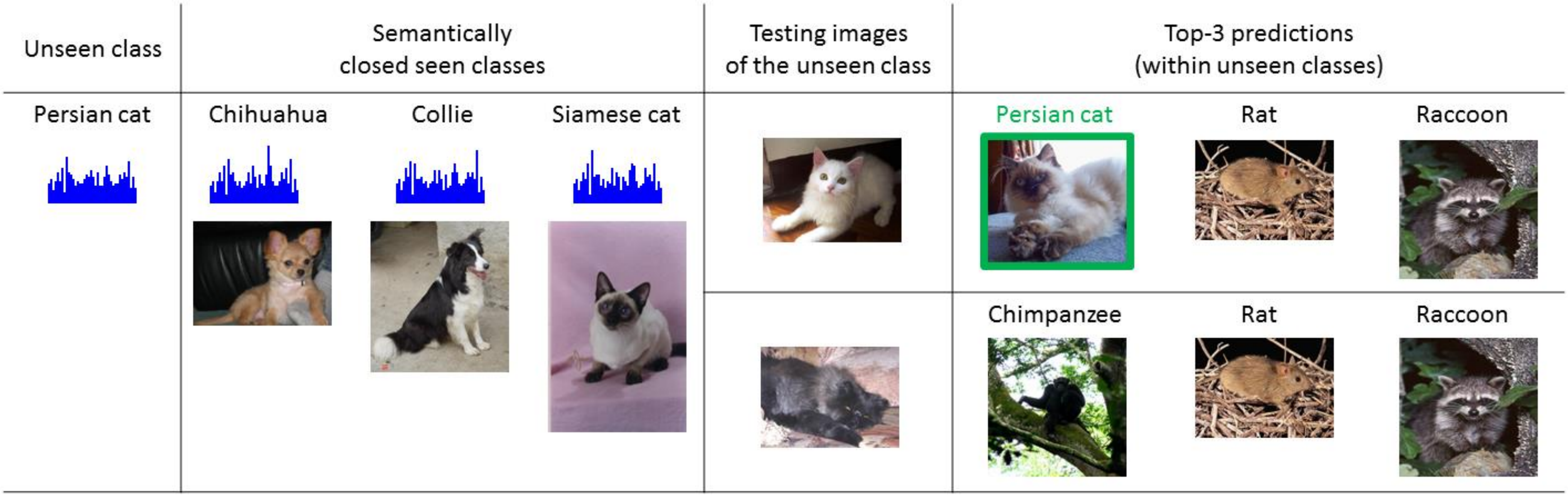


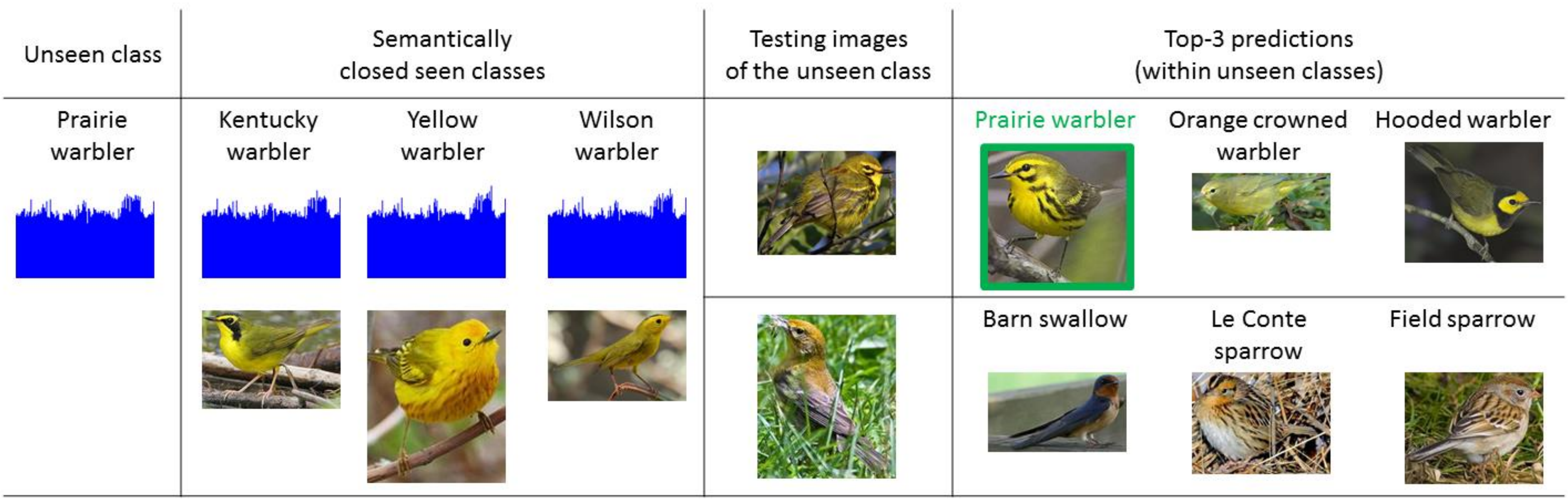
Bottom: First misclassified image

Top: Top 5 predictions

SUN dataset

Bottom: First misclassified image


Unseen class
Semantically closed seen classes
Testing images of the unseen class
Top-3 predictions (within unseen classes)
Persian cat
Chihuahua
Collie
Siamese cat
Persian cat
Rat
Raccoon
Chimpanzee
Rat
Raccoon


Unseen class
Semantically closed seen classes
Testing images of the unseen class
Top-3 predictions (within unseen classes)
Prairie warbler
Kentucky warbler
Yellow warbler
Wilson warbler
Prairie warbler
Orange crowned warbler
Hooded warbler
Barn swallow
Le Conte sparrow
Field sparrow