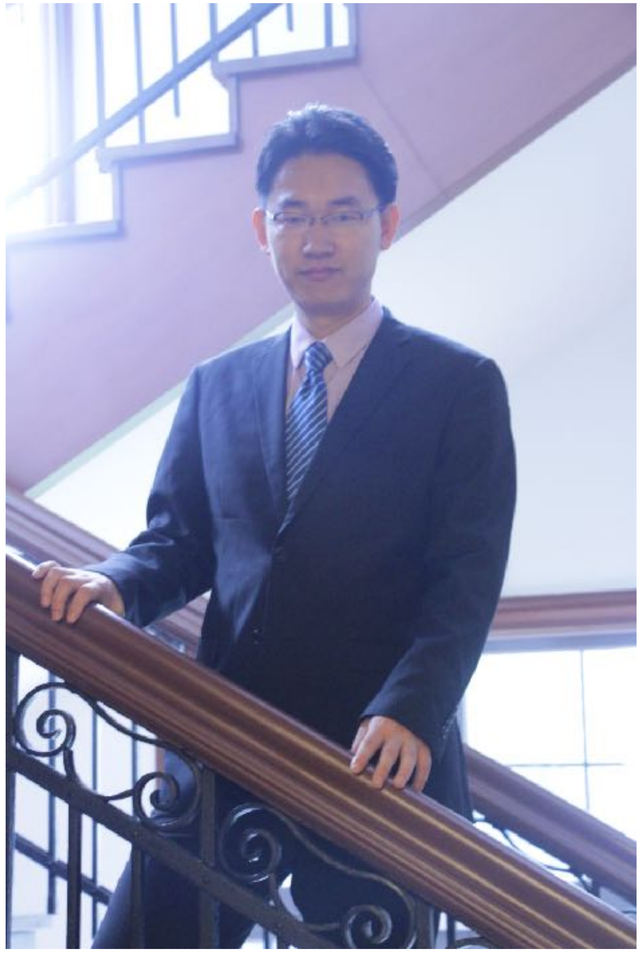
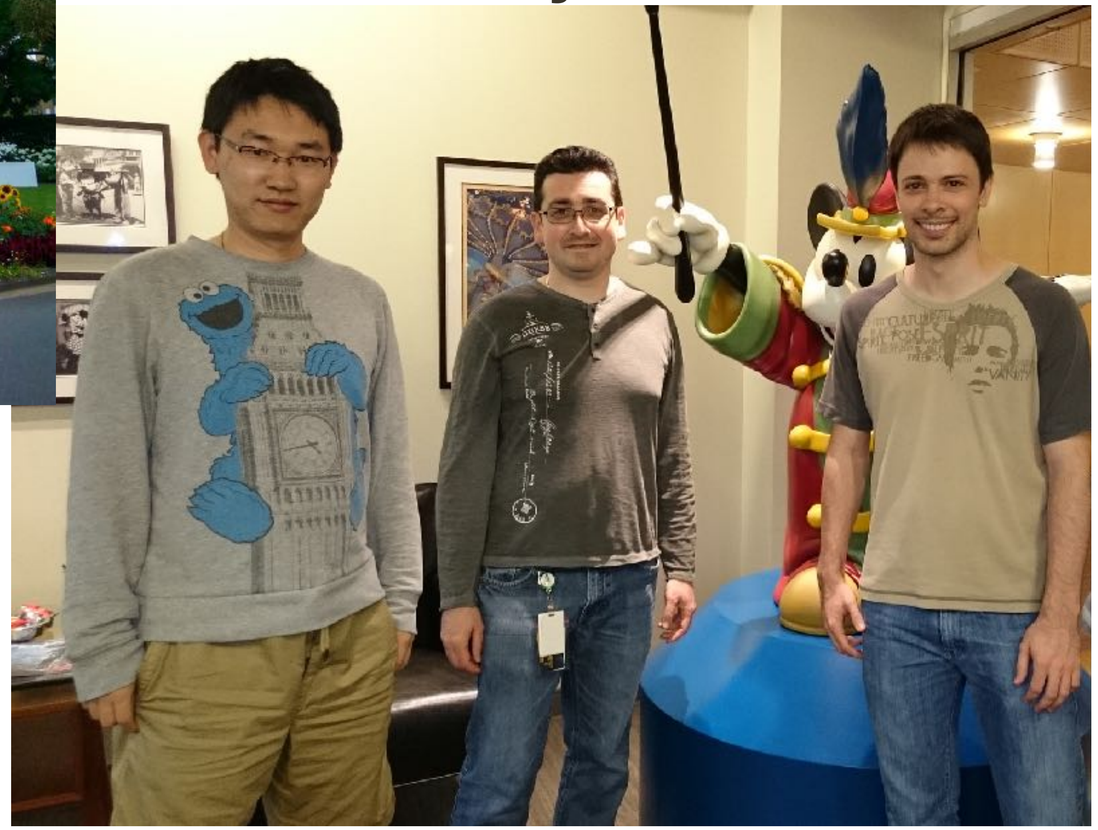
One-shot, Zero-shot and Open-set Recognition of Social Media Analysis

付彦伟
复旦大学大数据学院

About me



Disney Research

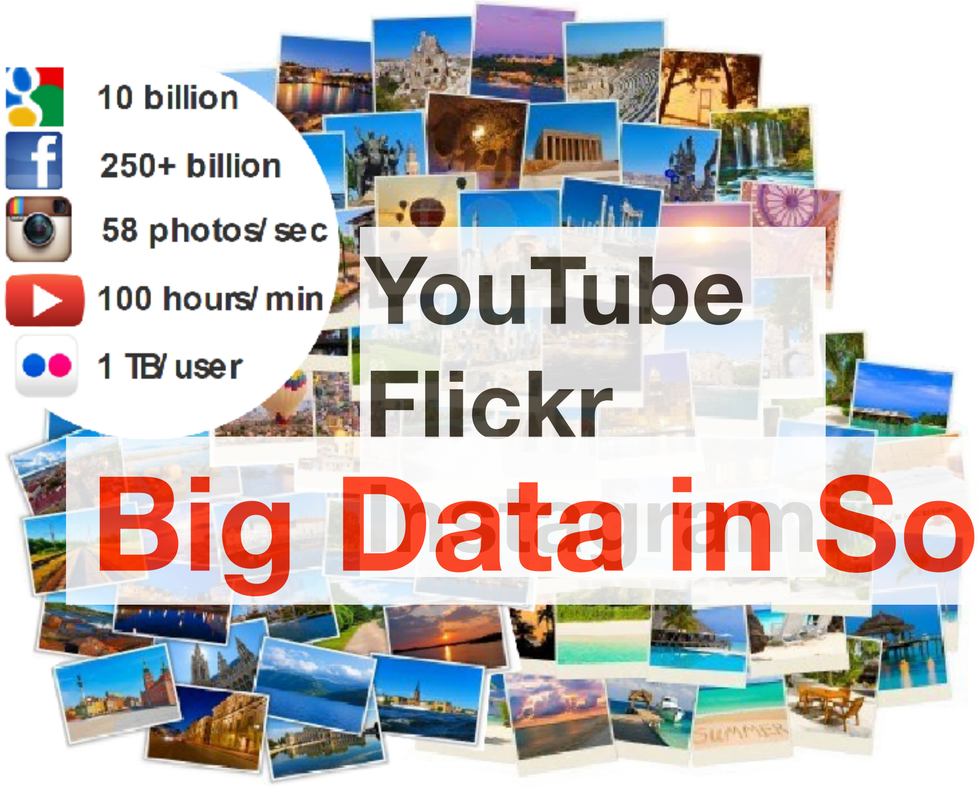


Content

1. Overview
2. Definition
3. Embedding
4. More

Content

1, Overview,
of our works on
Social Media Analysis



Big Data in Social Media Analysis

YouTube vimeo Dailymotion twitch ...

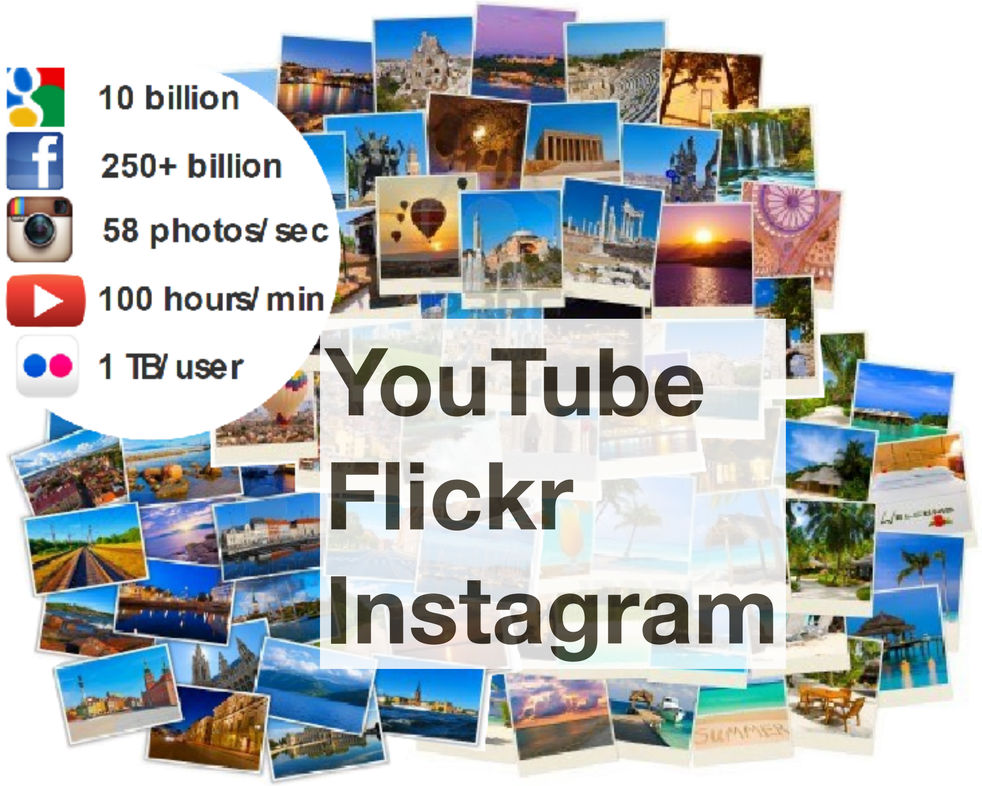


Image Classes

Hockey

Football

Running

Action

Cello Performance

Skiing

Parade

Activity

Sadness

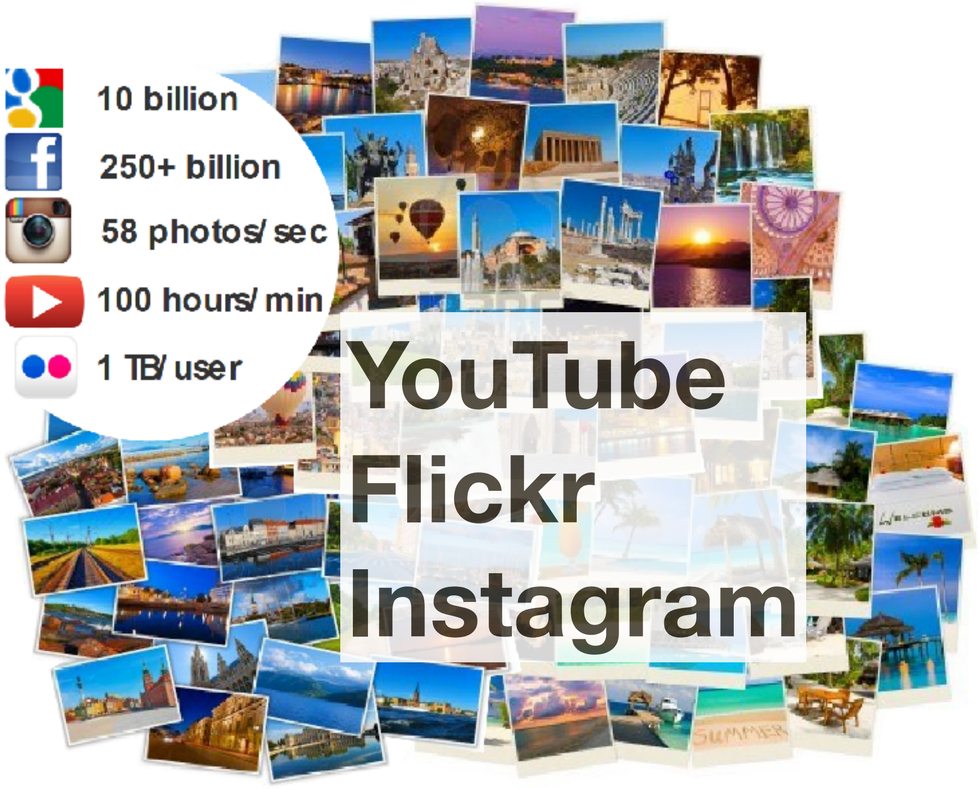
Fear

Anger

Joy

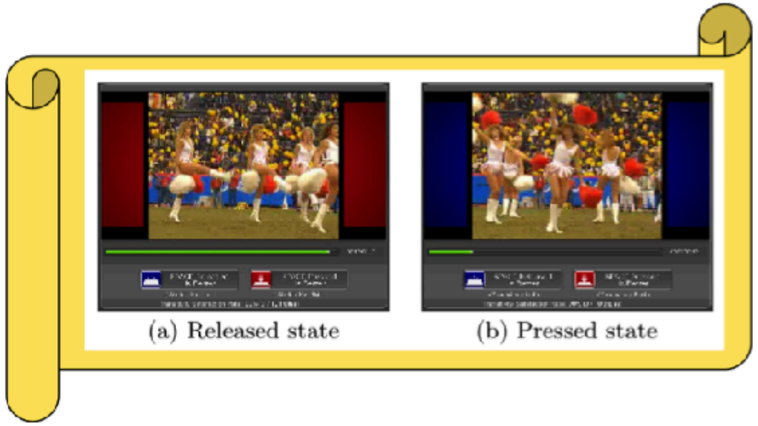
Disgust

Emotion



1. **Yanwei Fu**, Timothy M. Hospedales, Tao Xiang and Shaogang Gong. "*Learning Multi-modal Latent Attributes*" **IEEE TPAMI 2014;**
2. **Yanwei Fu**, Timothy M. Hospedales, Tao Xiang and Shaogang Gong. "*Transductive Multi-view Zero-Shot Learning*" **IEEE TPAMI**, **2015**;
3. **Yanwei Fu**, Timothy M. Hospedales, Tao Xiang and Shaogang Gong. "*Attribute Learning for Understanding Unstructured Social Activity*", **ECCV 2012**;
4. **Yanwei Fu**, Timothy M. Hospedales, Zhenyong Fu, Tao Xiang and Shaogang Gong. "*Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation*" **ECCV 2014;**
5. **Yanwei Fu**, Leonid Sigal. "*Semi-supervised Vocabulary-informed learning*", **CVPR 2016**;
6. ZuXuan Wu, **Yanwei Fu**, Yu-gang Jiang, Leonid Sigal, *Harnessing Object and Scene Semantics for Large-Scale Video Understanding*, **CVPR 2016**;
7. *Heterogeneous Knowledge Transfer in Video Emotion Recognition, Attribution and Summarization*, Baohan Xu, **Yanwei Fu**, Yu-Gang Jiang, Boyang Li and Leonid Sigal. **IEEE Transactions on Affective Computing 2017**.

Crowdsourcing Ranking on Internet



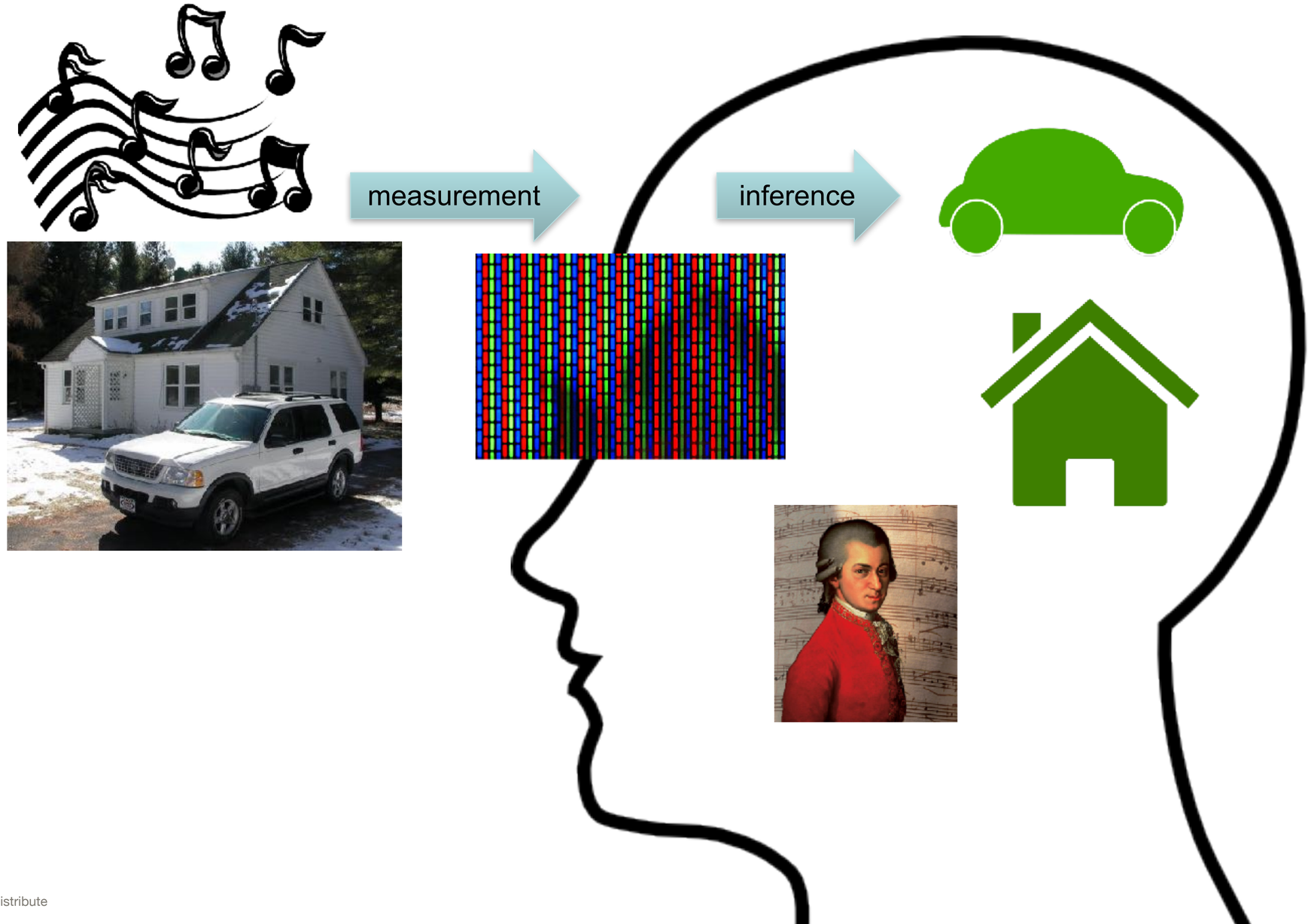
Our Story Start from a movie – The Social Network

1. **Yanwei Fu**, Timothy M. Hospedales, Tao Xiang, Jiechao Xiong, Shaogang Gong, Yizhou Wang and Yuan Yao. "*Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels*" **IEEE TPAMI 2016;**
2. Yu-ting Qiang, **Yanwei Fu**, Yanwen Guo, Zhi-hua Zhou and Leonid Sigal. "*Learning to Generate Posters of Scientific Papers*", **AAAI 2016;**
3. **Yanwei Fu**, Timothy M. Hospedales, Tao Xiang and Shaogang Gong and Yuan Yao. "*Interestingness Prediction by Robust Learning to Rank*" **ECCV 2014**;
4. **Yanwei Fu**, Yanwen Guo,Yanshu Zhu, Feng Liu, Chuanming Song and Zhi-Hua Zhou. *Multi-view Video Summarization*, **IEEE TMM 2010**;
5. **Yanwei Fu**, De-an Huang, Leonid Sigal, Robust Classification by Pre-conditioned LASSO and Transductive Diffusion Component Analysis, http://arxiv.org/abs/1511.06340.

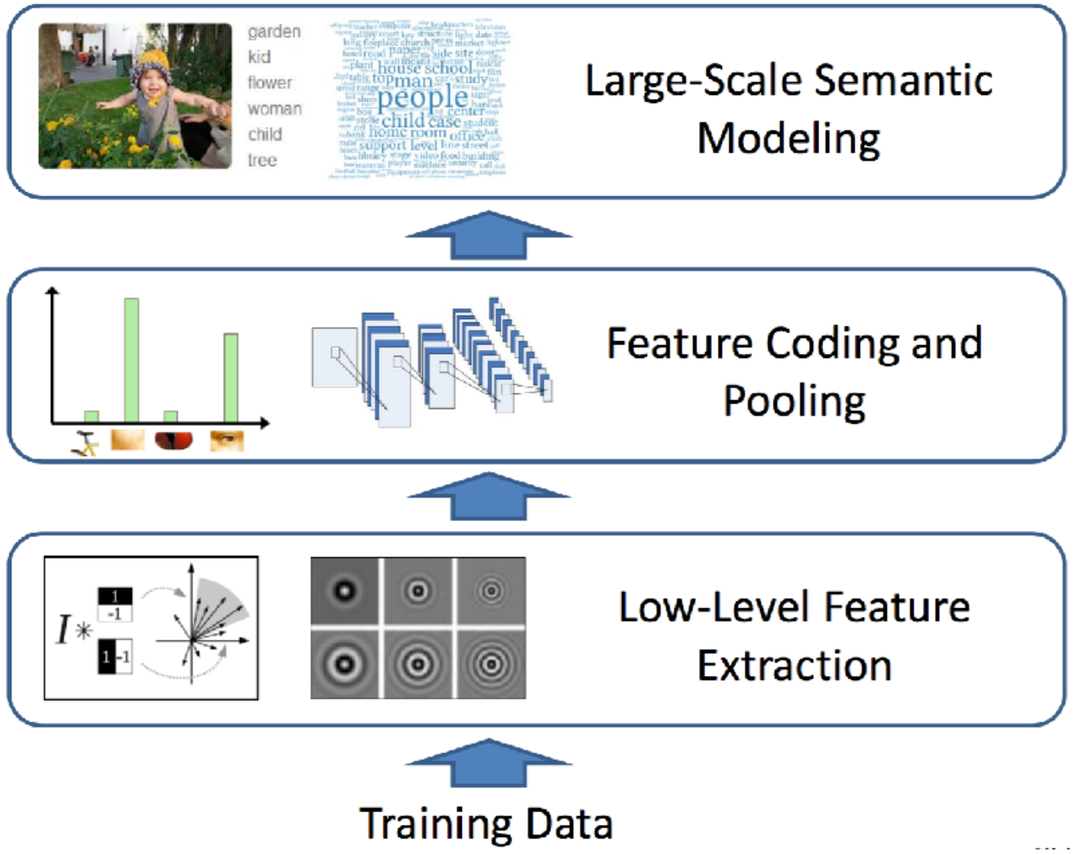
Content

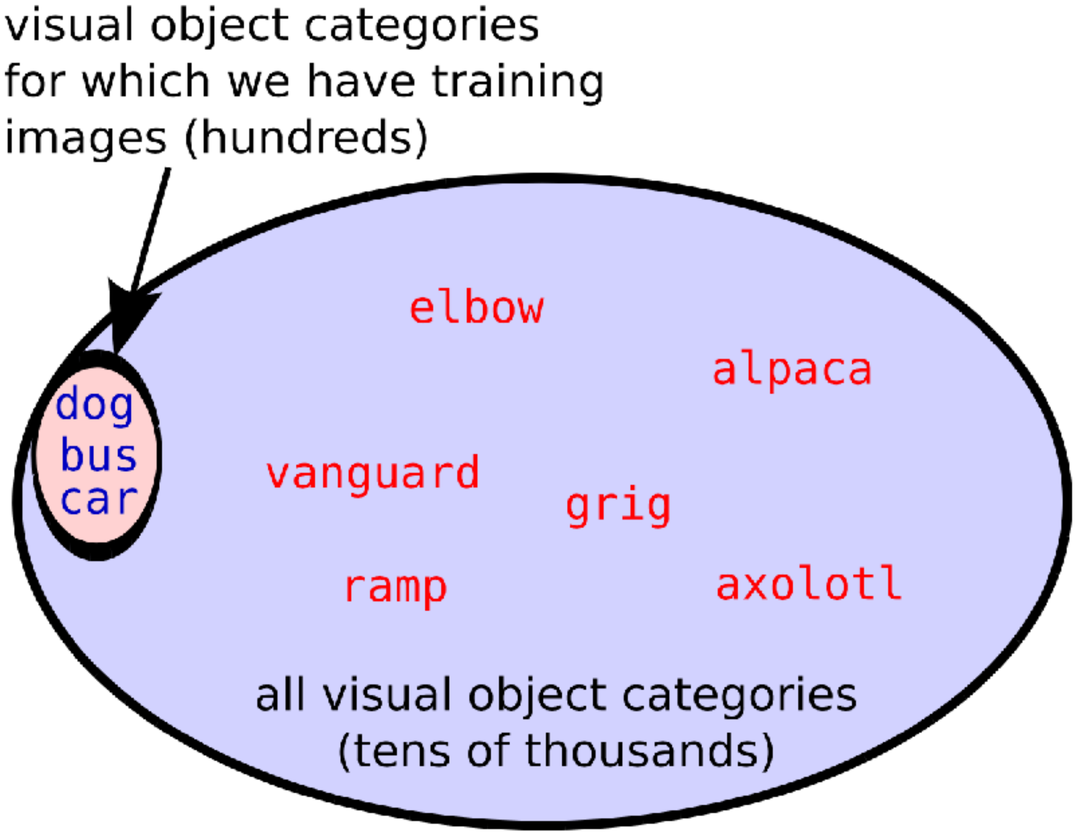
2, Definition,
of one-shot, zero-shot, and
open-set recognition

Recognition in a nutshell



Supervised Recognition

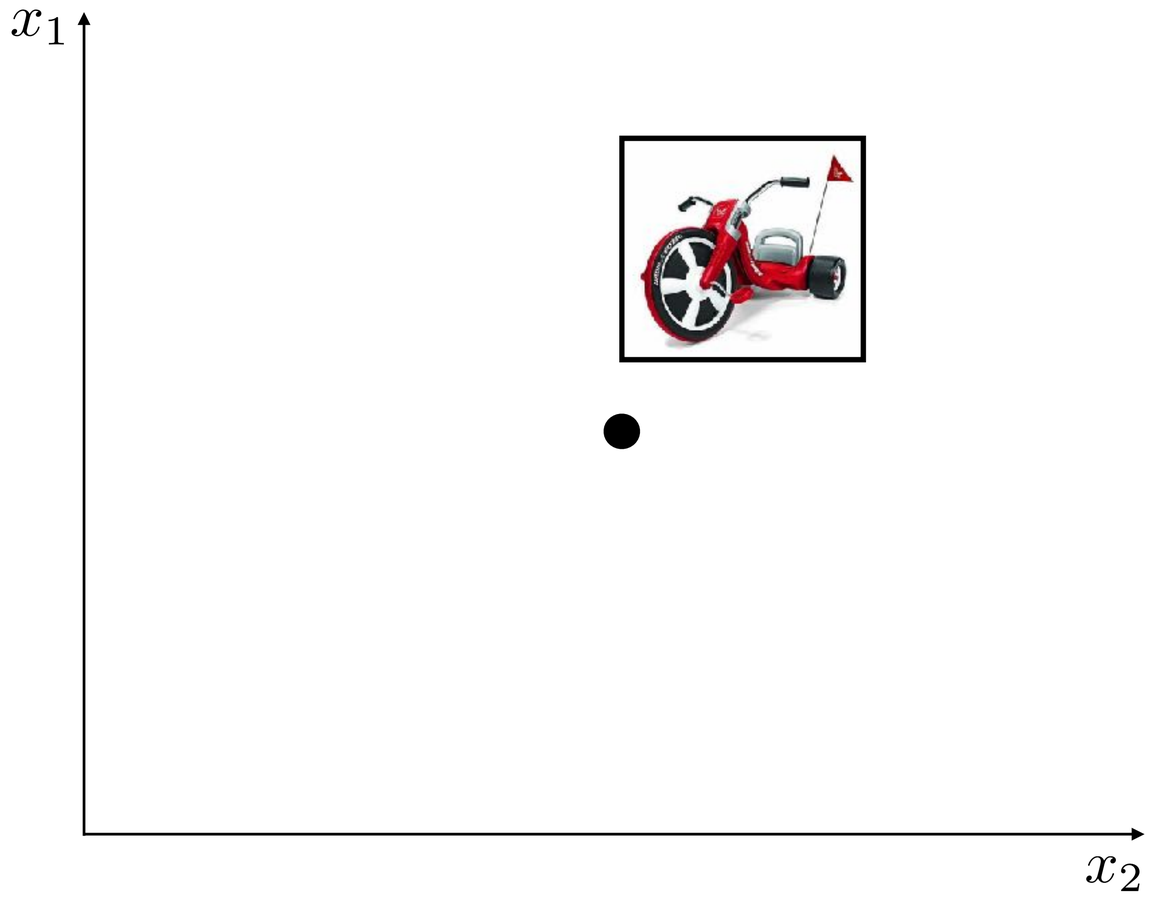




One-shot Learning

Problem Definition

Semantic labels → Visual feature space

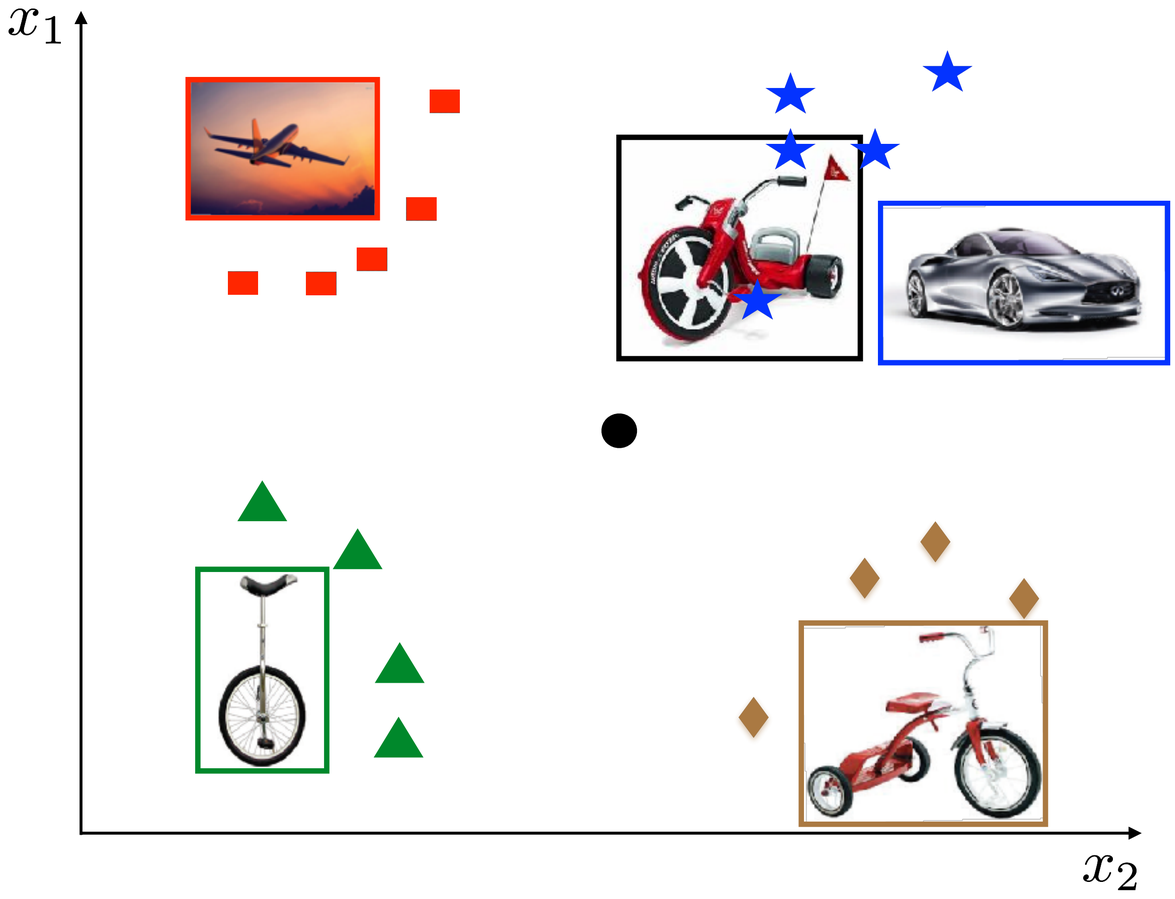


One-shot Learning

Learning

airplane **car**

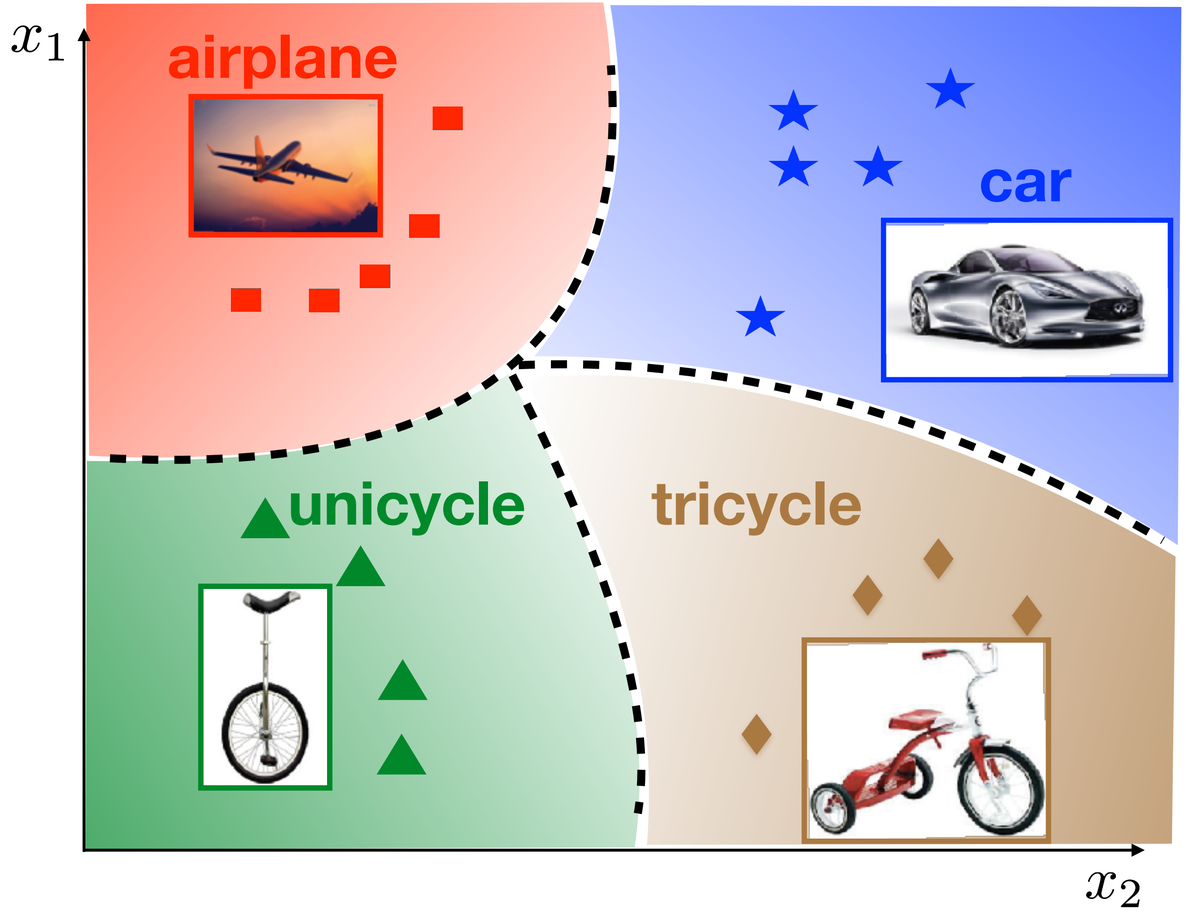
unicycle **tricycle**



Supervised Learning

Learning

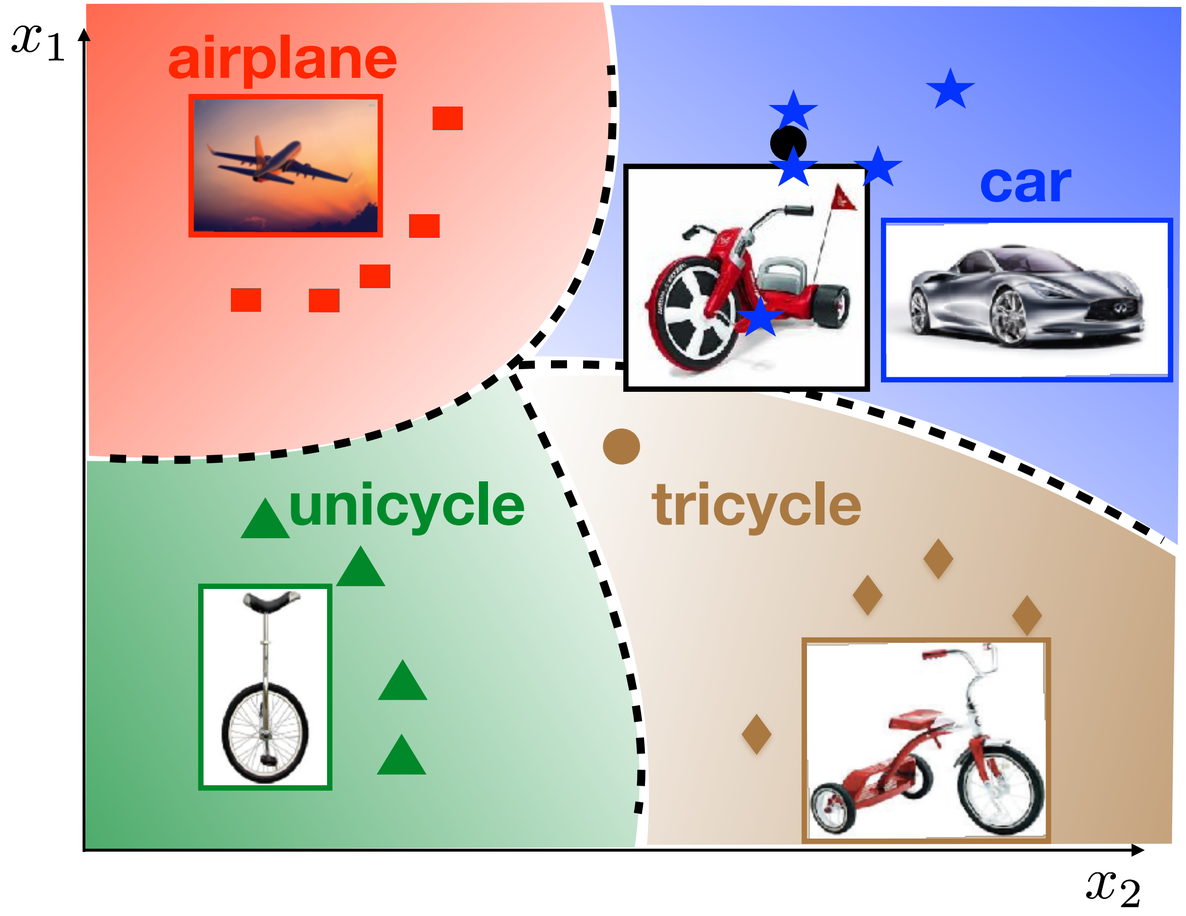
Semantic labels → Visual feature space



Supervised Learning

Inference

Semantic labels → Visual feature space



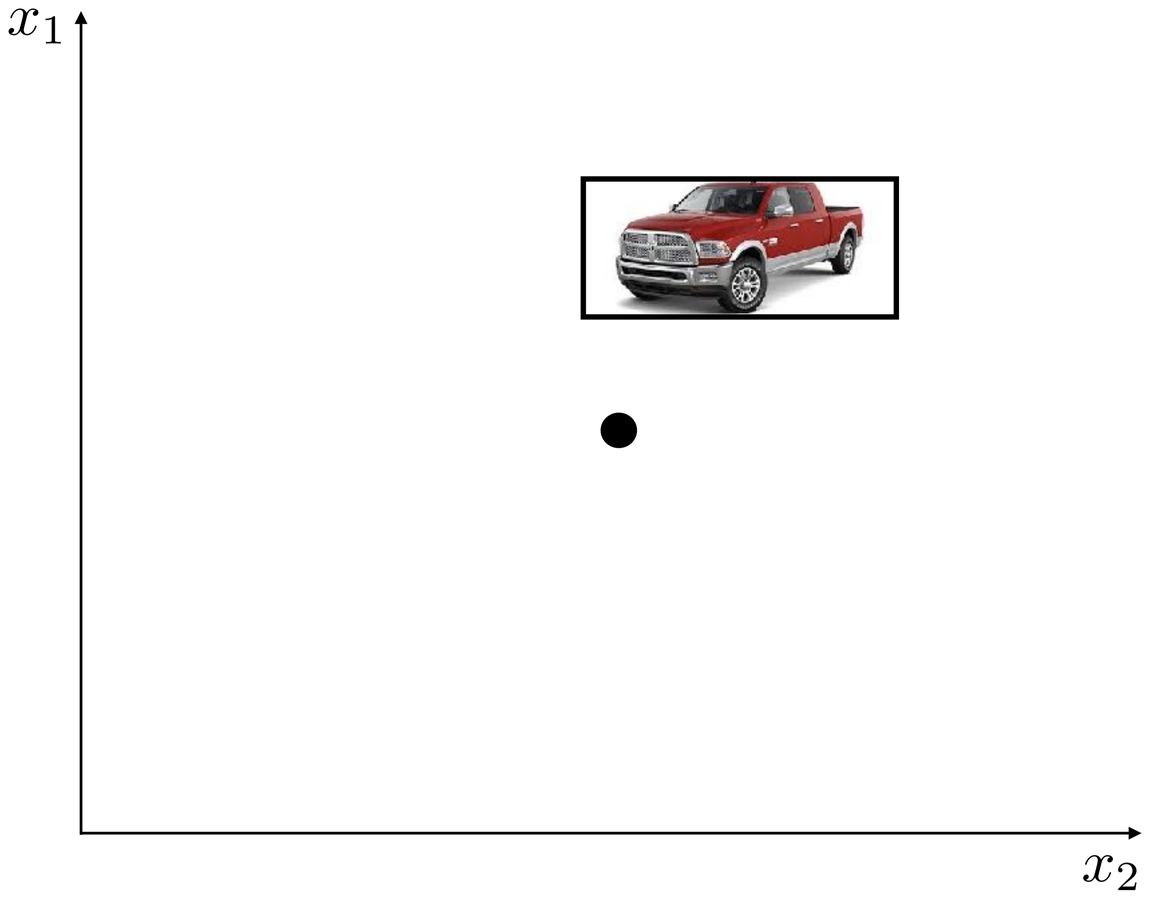
Zero-shot Learning

Problem Definition

We do not have any visually labeled instances of what these look like

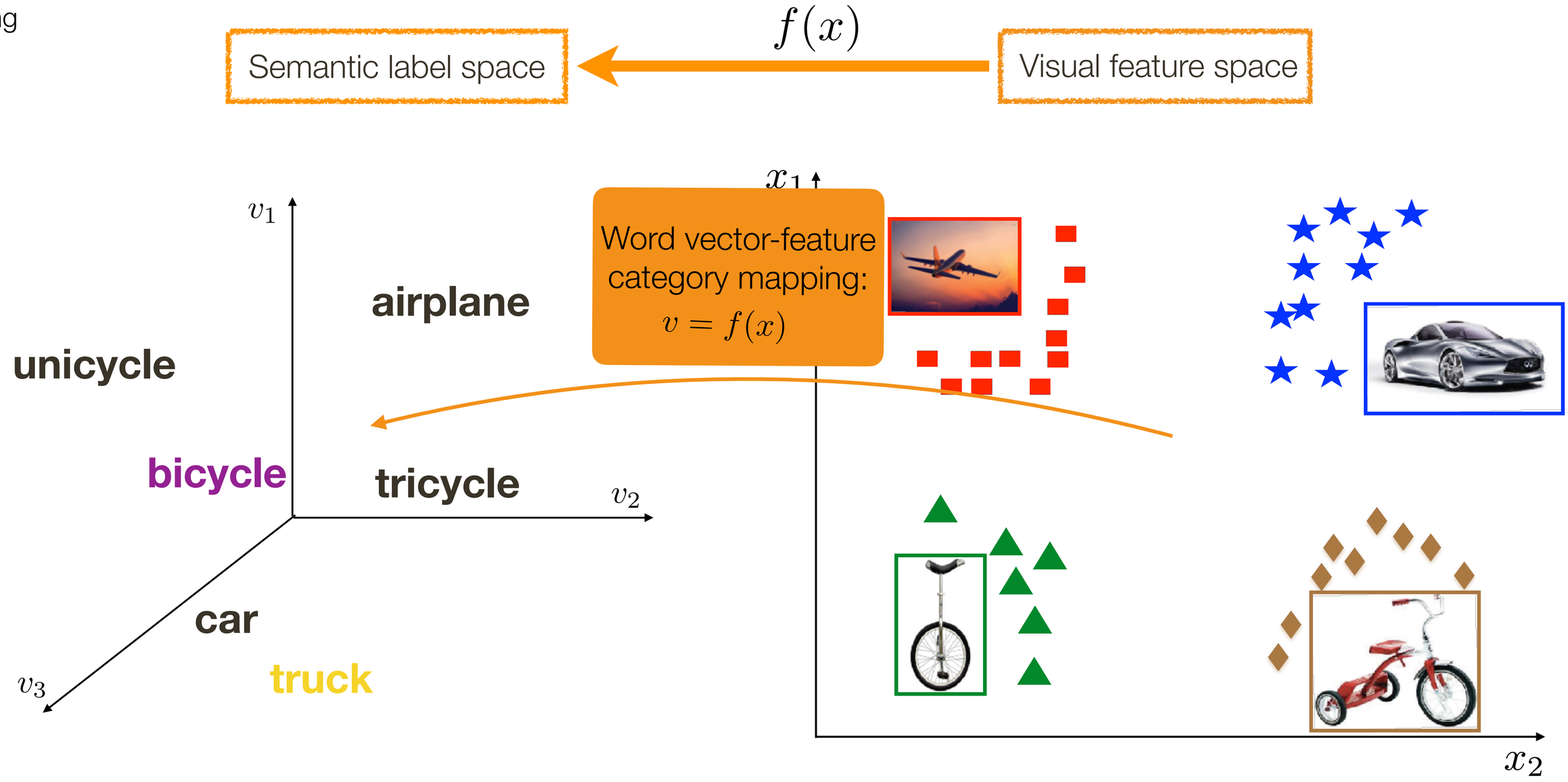
bicycle

truck



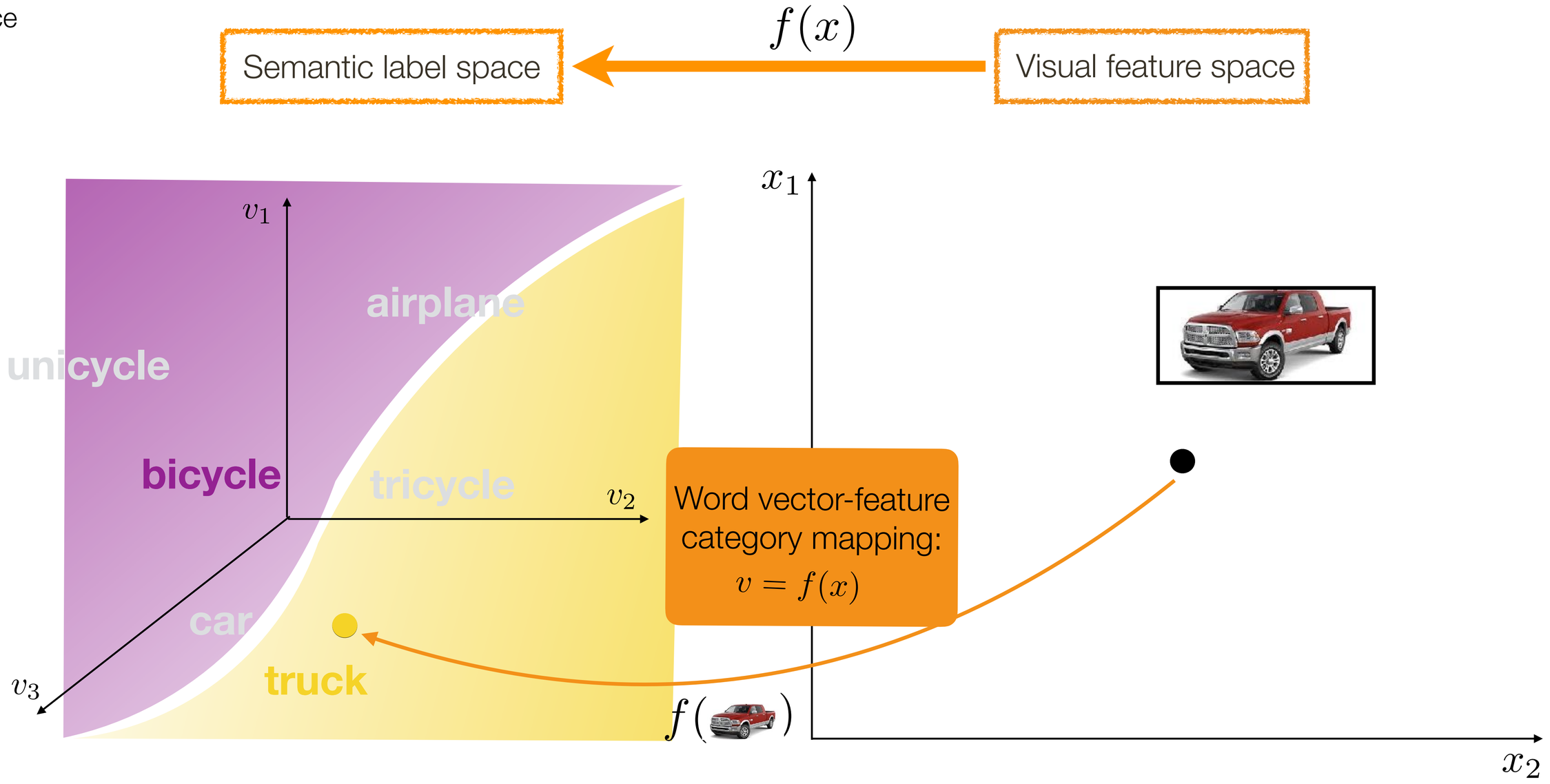
Zero-shot Learning

Learning



Zero-shot Learning

Inference



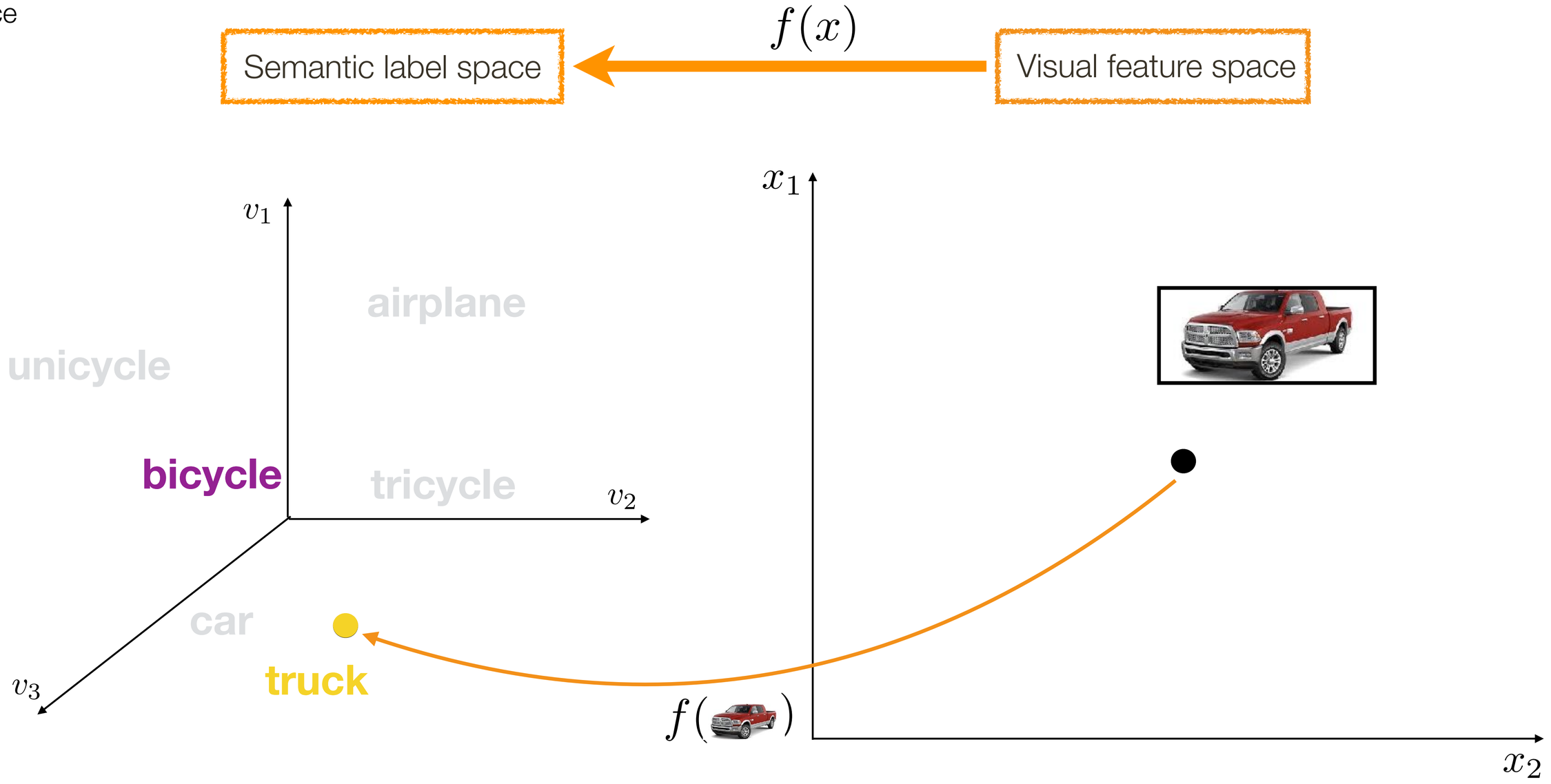
Key Question: How do we define semantic space?

Semantic Label Vector Spaces

Spaces	Type	Advantages	Disadvantages
Semantic Attributes	Supervised	Good interpretability of each dimension: $\text{airplane} := \text{fixed\_wing}, \text{propelled}, \text{has\_pilot}$	Manual annotation Limited vocabulary
Semantic Word Vectors (e.g. word2vec)	Unsupervised	Good vector representation for millions of vocabulary $v\,(\text{Berlin}) - v\,(\text{Germany}) = v\,(\text{Paris}) - v\,(\text{France})$	Limited interpretability of each dimension

Zero-shot Learning

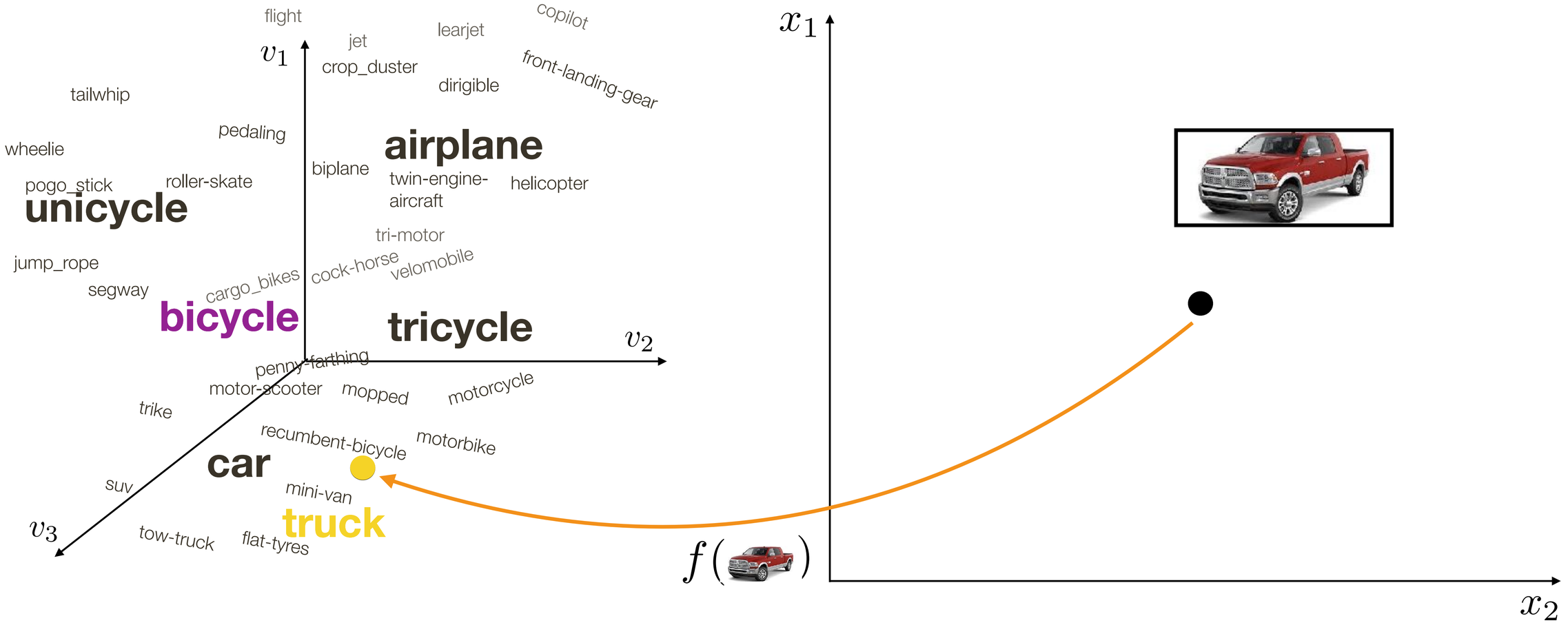
Inference



Open-set Recognition

Problem Definition

Semantic label space ← $f(x)$ ← Visual feature space

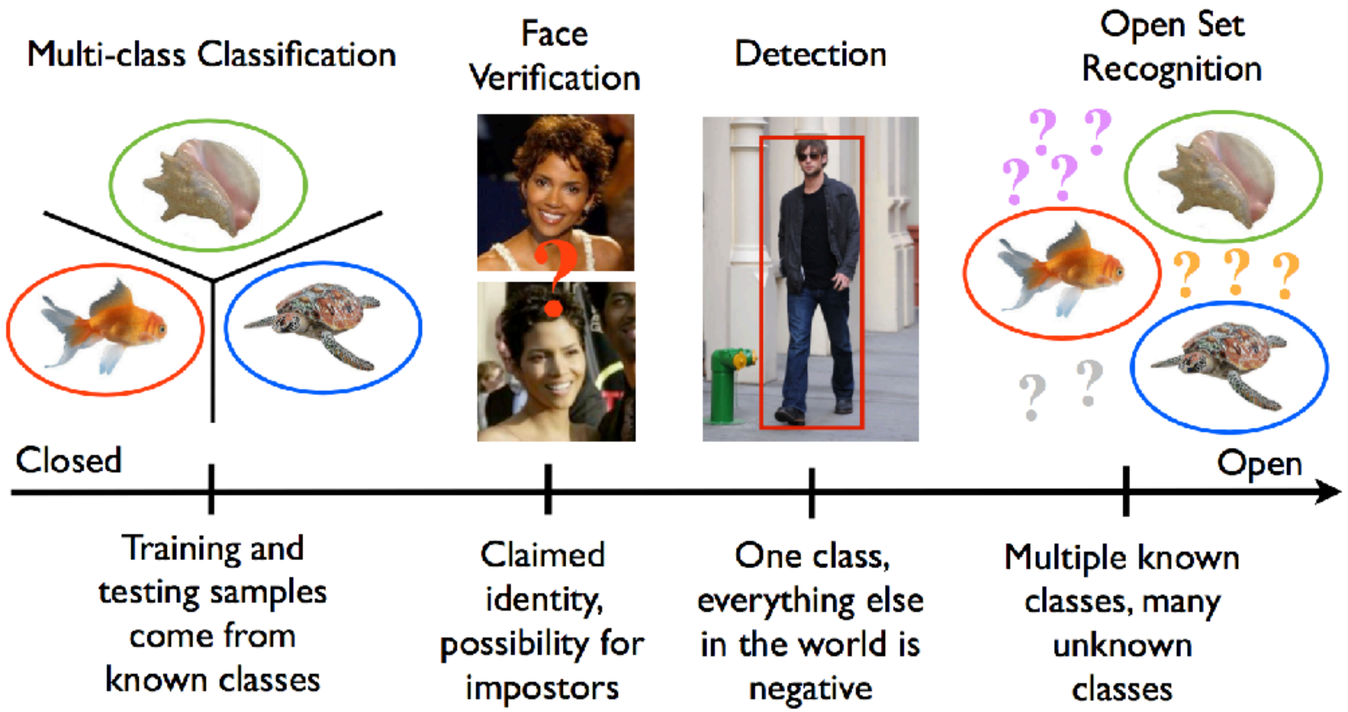


Fu et al. *Semi-supervised Vocabulary-informed learning, CVPR 2016 (Oral)*

Supervised Learning:

Pros: Very good quantitative performance

Cons: Relatively small vocabulary (~1,000 classes)
Requires ***manual labeling*** of all the data



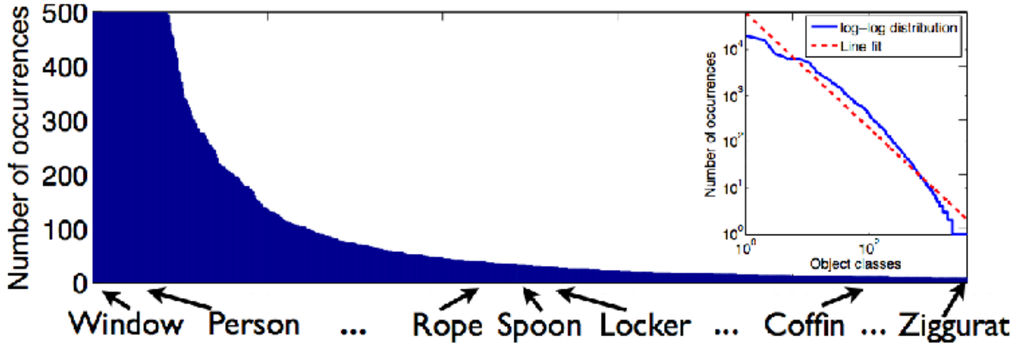
One-shot (N-shot, Few-shot) Learning

Pros: Only require few instances per class
Cons: Relatively small vocabulary (~1,000 classes)
Requires ***manual labeling*** of all the data

Zero-shot Learning:

Pros: Does not require instance labeling for target classes

Cons: Typically limited to recognition with target classes only
Relatively ***small vocabulary*** (~50-10K classes typically)



(a) The number of examples by object class in SUN dataset

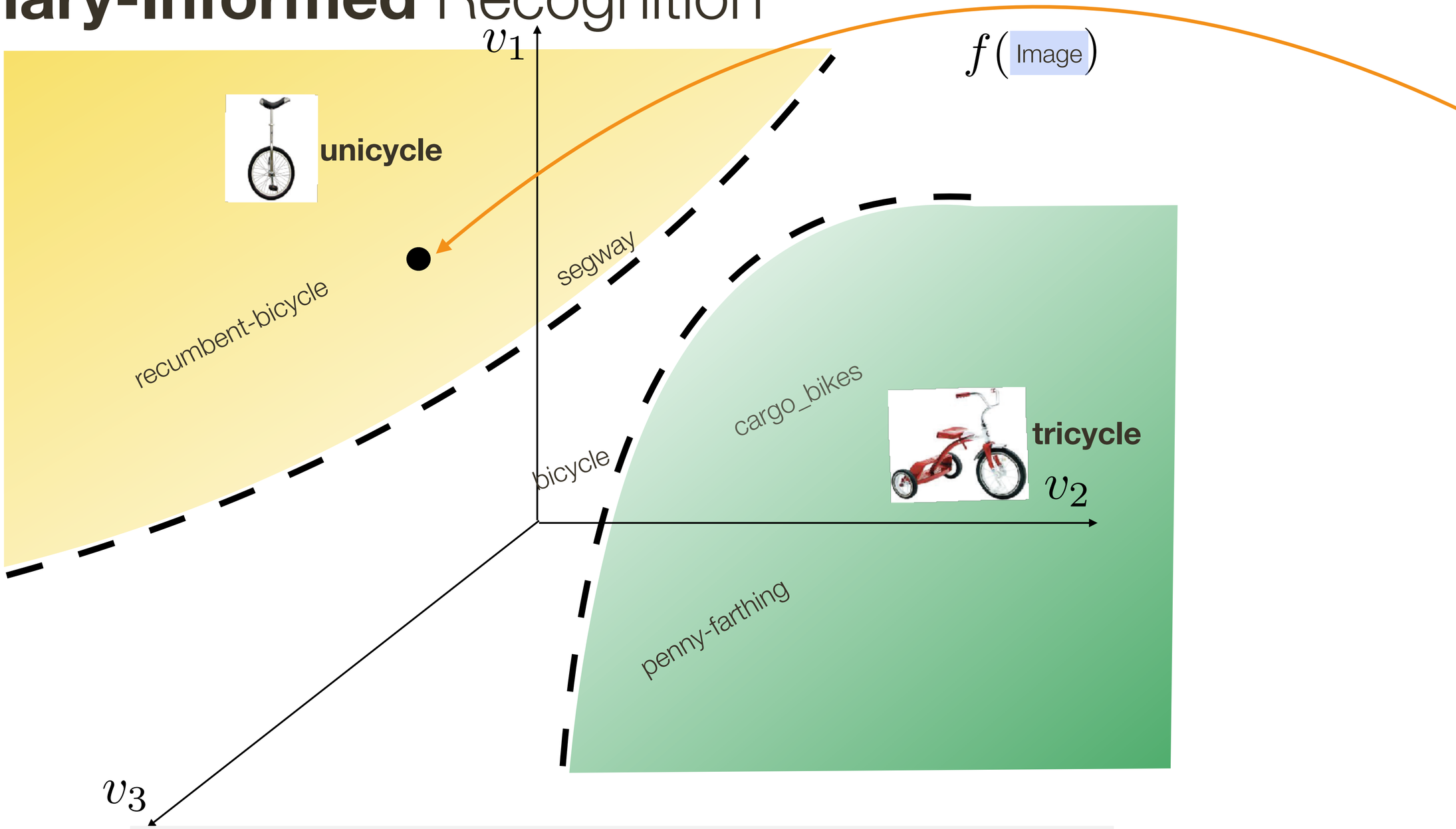
Long-tailed Distribution of object categories in the world

Open-set Learning

Pros: Does not require instance labeling for target classes
Large vocabulary (e.g. 300K classes)

Fu et al. *Semi-supervised Vocabulary-informed learning, CVPR 2016 (Oral)*

Vocabulary-Informed Recognition

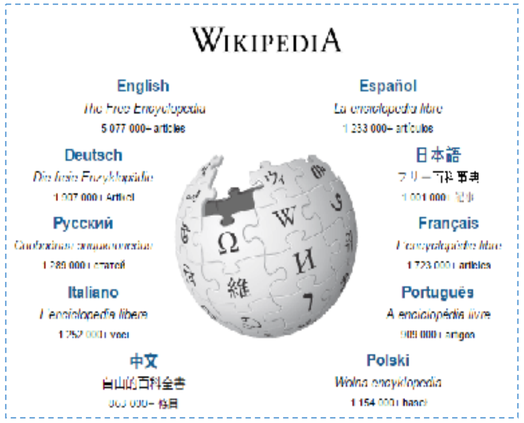


Fu et al. *Semi-supervised Vocabulary-informed learning, CVPR 2016 (Oral)*

Content

3. Embedding
in Multi-view or Multi-modal

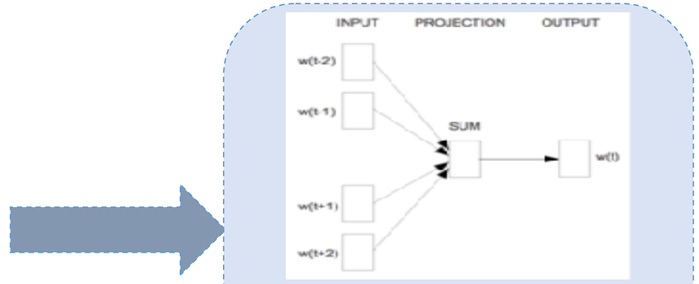
Multi-view/Multi-modal Embedding



Wikipedia, Articles

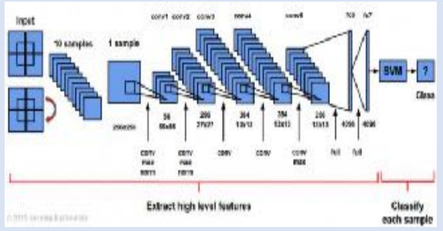
...

ImageNet



Language Models

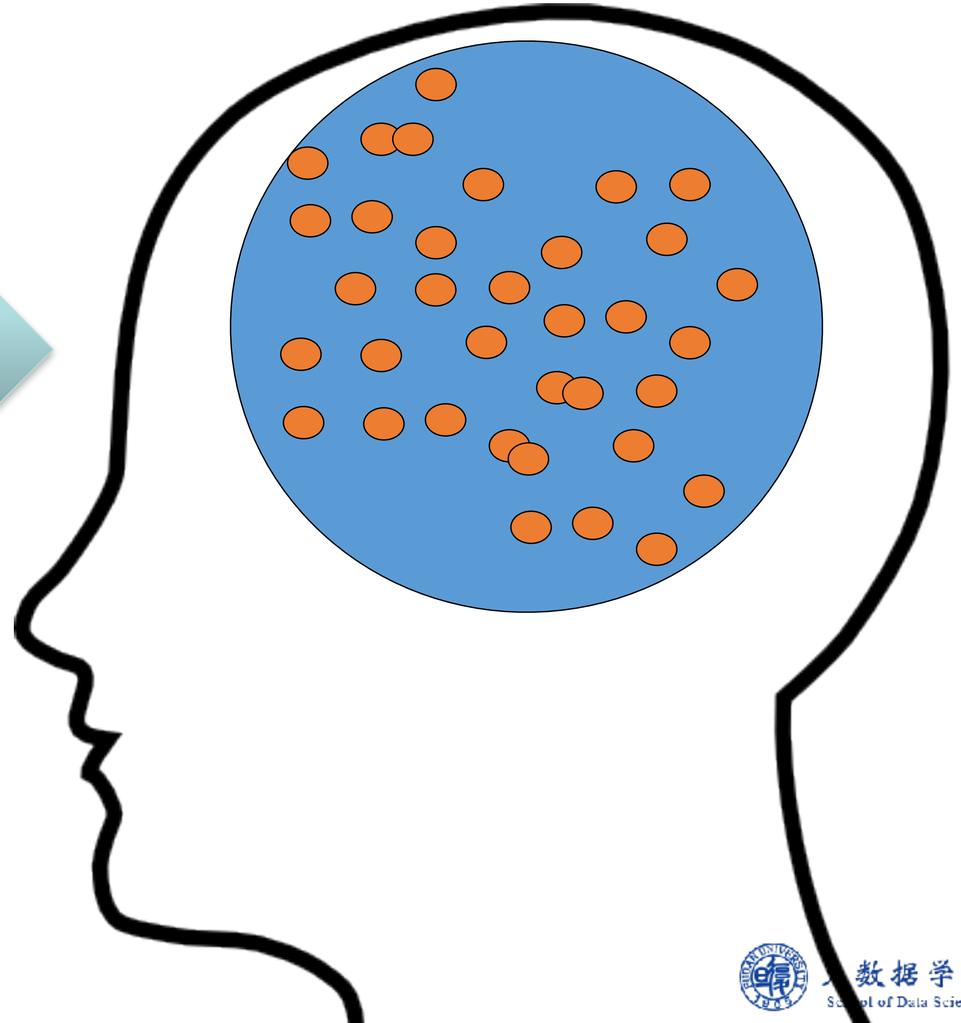
...



Deep Networks

?

A unified embedding space



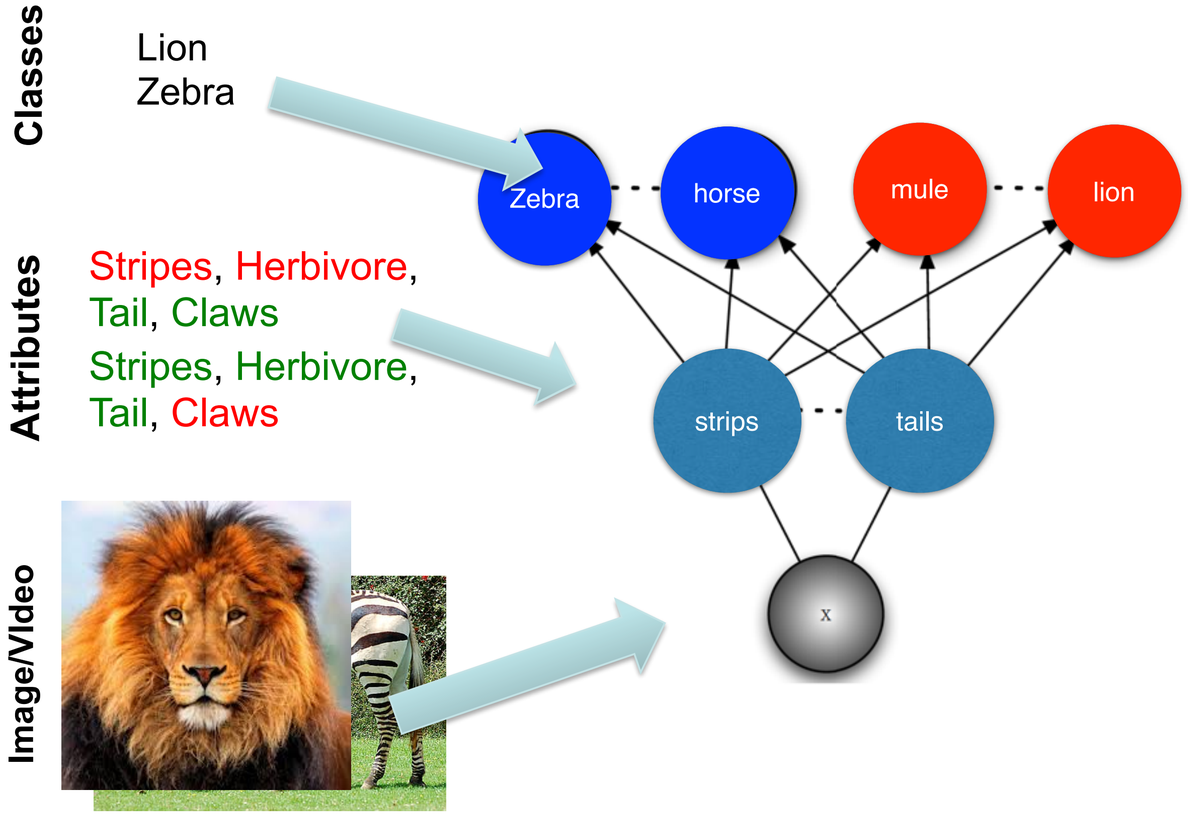
Embedding

Learning Multi-modal Latent Attributes

Fu et al. *Attribute Learning for Understanding Unstructured Social Activity*", ECCV 2014
Fu et al. *Learning Multi-modal Latent Attributes*, IEEE TPAMI 2014

Multi-modal Embedding

Existing Attribute Learning Pipeline

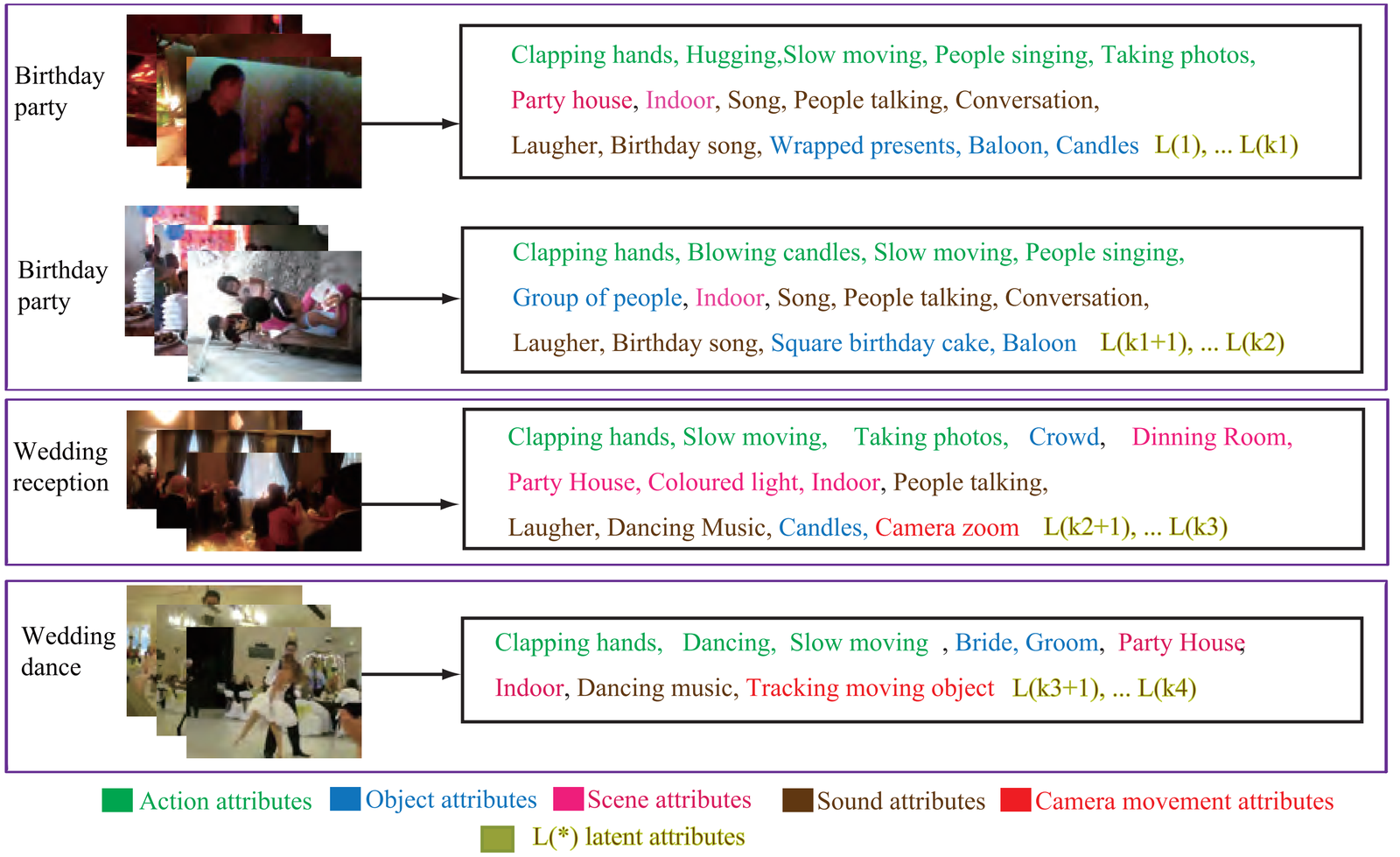


Challenges of attribute annotations

- Incomplete user-defined attributes;
- Sparse user-defined attributes;
- Ambiguous user-defined attributes;

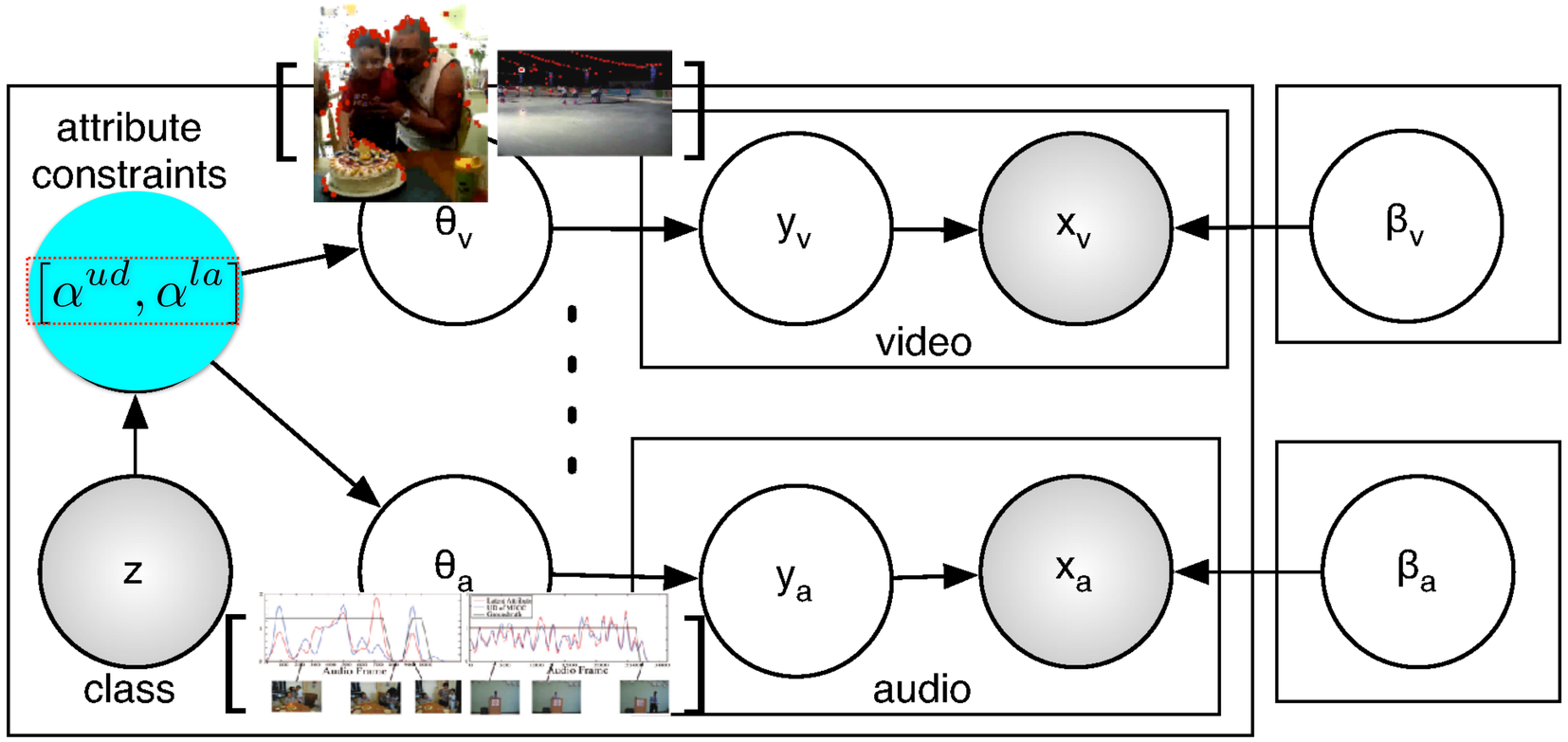
Fu et al. *Attribute Learning for Understanding Unstructured Social Activity*", ECCV 2014
Fu et al. *Learning Multi-modal Latent Attributes*, IEEE TPAMI 2014

Multi-modal Embedding



Fu et al. *Attribute Learning for Understanding Unstructured Social Activity*", ECCV 2014
Fu et al. *Learning Multi-modal Latent Attributes*, IEEE TPAMI 2014

Multi-modal Latent Attribute Topic Model(M2LATM)



$$\theta_m \sim Dir(\alpha) \qquad y_m \sim Multi(\theta_m) \quad m \in \{v, a, \cdots\}$$

Fu et al. *Attribute Learning for Understanding Unstructured Social Activity*", ECCV 2014
Fu et al. *Learning Multi-modal Latent Attributes*, IEEE TPAMI 2014

Experimental Results

Dataset & Settings:

USAA dataset (8 video classes, e.g. birthday party, graduation party.);

Animal with Attributes (AwA) dataset (40 auxiliary cls; 10 testing cls);

Comparisons

Direct: KNN/SVM of features to classes;

DAP: Direct Attribute Prediction [Lampert et al. CVPR 2009];

SVM-UD: a SVM generalization of DAP;

SCA: Topic models in [Wang et al CVPR 2009];

ST: Synthetic Transfer in [Yu et al ECCV 2010];

Multi-task Learning

	Direct	SVM-DAP	SCA	M2LATM
100l, A/69	**66.0**	65.7	44.0	65.6
10l, A/69	26.8	40.2	32.2	**40.6**
10l, R/7	26.8	26.4	25.6	**38.3**
10l, N/0	26.8	-	17.3	**40.4**
10l, T/7	26.8	32.4	26.0	**38.3**
10l, B/7	26.8	18.2	26.0	**38.9**

Supervised classification performance for USAA. 8 classes, chance =12.5%. Row labels are l: number of training instances per class, A: all attributes, R: random subset of attributes, N: no attributes, T: top attributes, B: bottom attributes.

Transfer Learning — ZSL

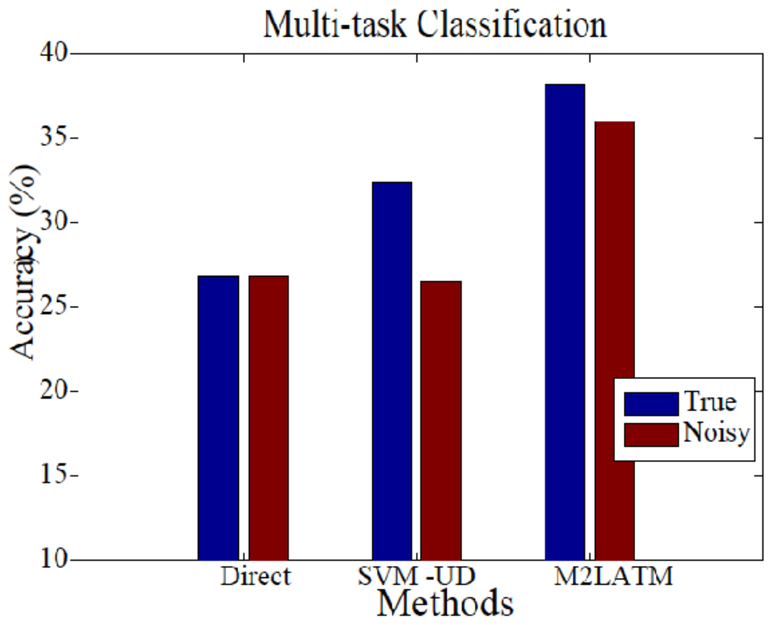
USAA dataset: 4 auxiliary classes; 4 testing classes

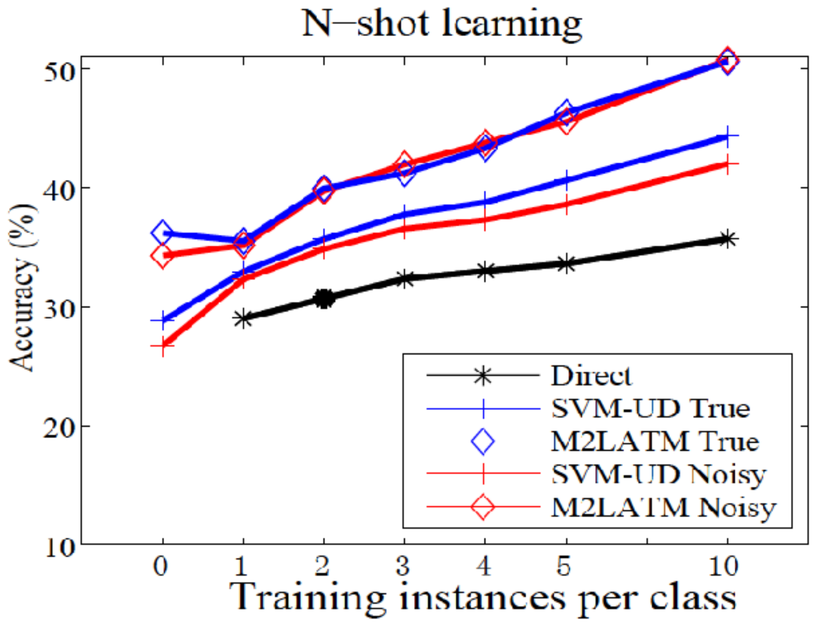
	SVM-UD	ST[34]	M2LATM
R/7	27.1	18.1	**33.8**
O/15	31.3	36.9	**39.4**
R/34	36.7	30.9	**39.2**
A/69	33.2	31.0	**41.9**

AwA dataset: 40 auxiliary classes; 10 testing classes;

		No Attrib Prior		Attrib Prior	
	[39]/[26]/[17]	DAP	M2LATM	DAP	M2LATM
R/9	-	26.3	26.9	27.8	**29.2**
R/42	-	34.4	38.2	36.0	**39.7**
A/85	33.0/33.0/32.7	37.0	39.2	39.2	**41.3**

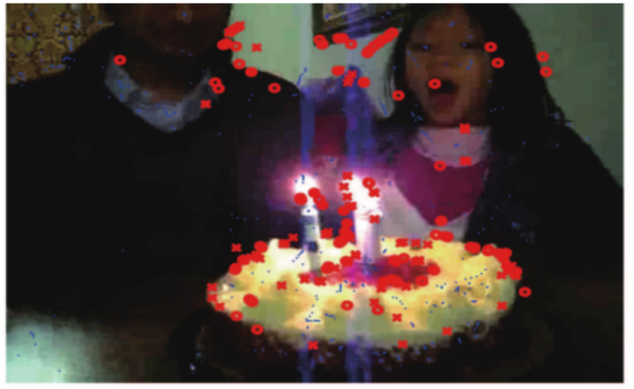
Robustness to Label Noise





To simulate label noise, we randomly flipped 50% of attribute annotations on 50% of the training videos (so 25% wrong annotations).

Qualitative Results

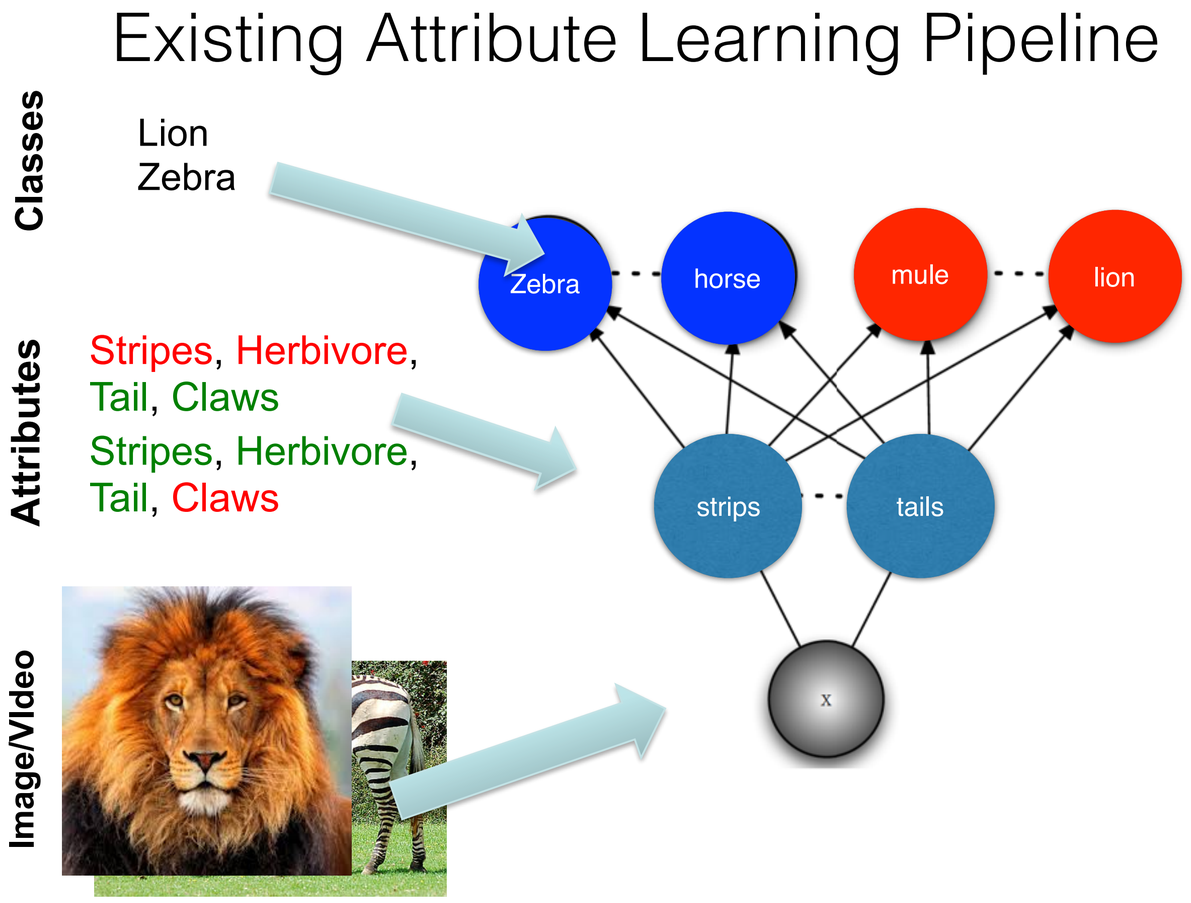



Red illustrates interest-points from UD or latent attributes;
Blue illustrates interest-points not related to attributes concerned.

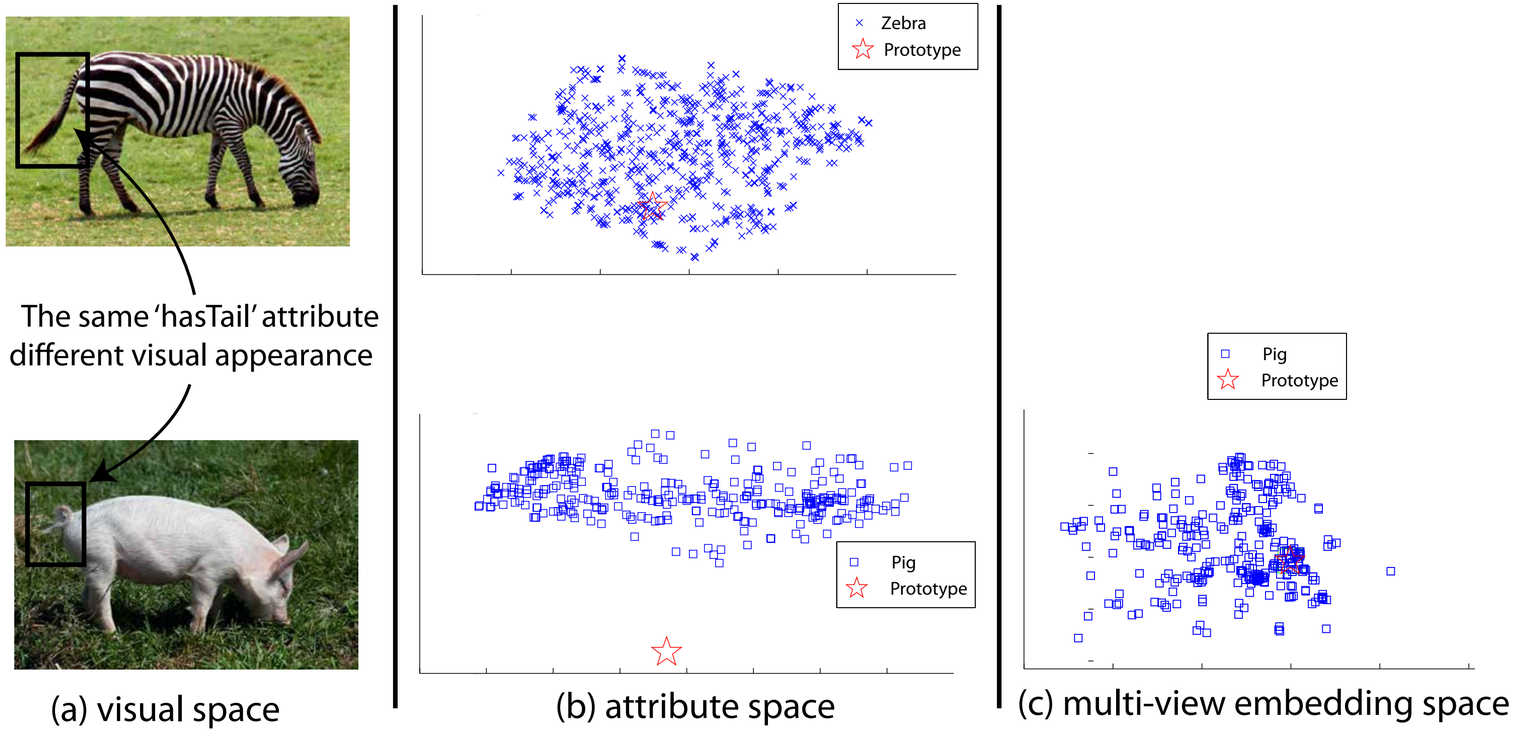
Embedding

Multi-view Embedding

Multi-view Embedding



Projection Domain Shift Problem



- Zebra (one of auxiliary classes) and Pig (one of target classes) share 'hasTail' attributes and visual appearance differs greatly;
- Projection of low-level features to attributes learned from auxiliary data is shifted when used in target data.

Fu et al. *Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation*, ECCV 2014
Fu et al. *Transductive Multi-View Zero-Shot Learning*, IEEE TPAMI 2015

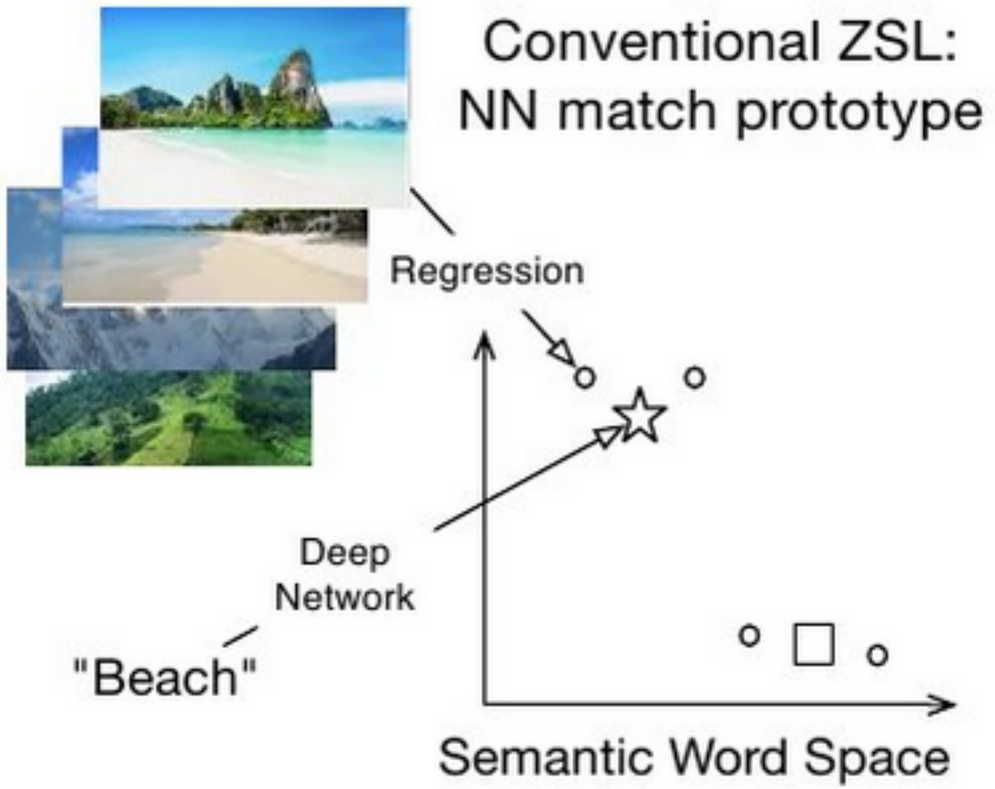
Prototype Sparsity and Multiple Representation Embedding

Prototype sparsity problem,

- Each class only has a single attribute prototype which is insufficient to fully represent what that class looks like;
- Semantic representations (e.g. word vectors) have manifold structure.

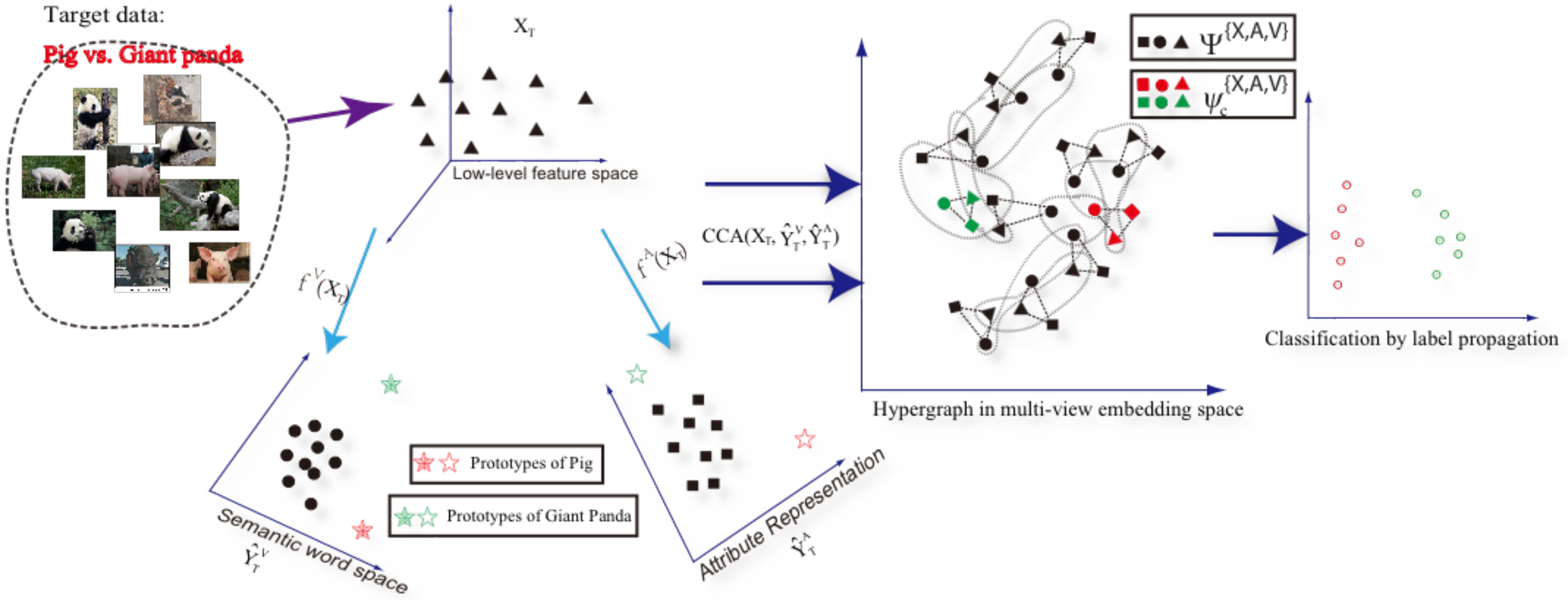
Embedding multiple representations,

- Distributed word vectors (word2vec, GloVec);
- Non-distributed word vectors (tf–idf); and WordNet;
- Each may contain complementary information of different classes.



Fu et al. *Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation*, ECCV 2014
Fu et al. *Transductive Multi-View Zero-Shot Learning*, IEEE TPAMI 2015

Tranductive Multi-vew CCA Embedding



Fu et al. *Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation*, ECCV 2014
Fu et al. *Transductive Multi-View Zero-Shot Learning*, IEEE TPAMI 2015

Experimental Results

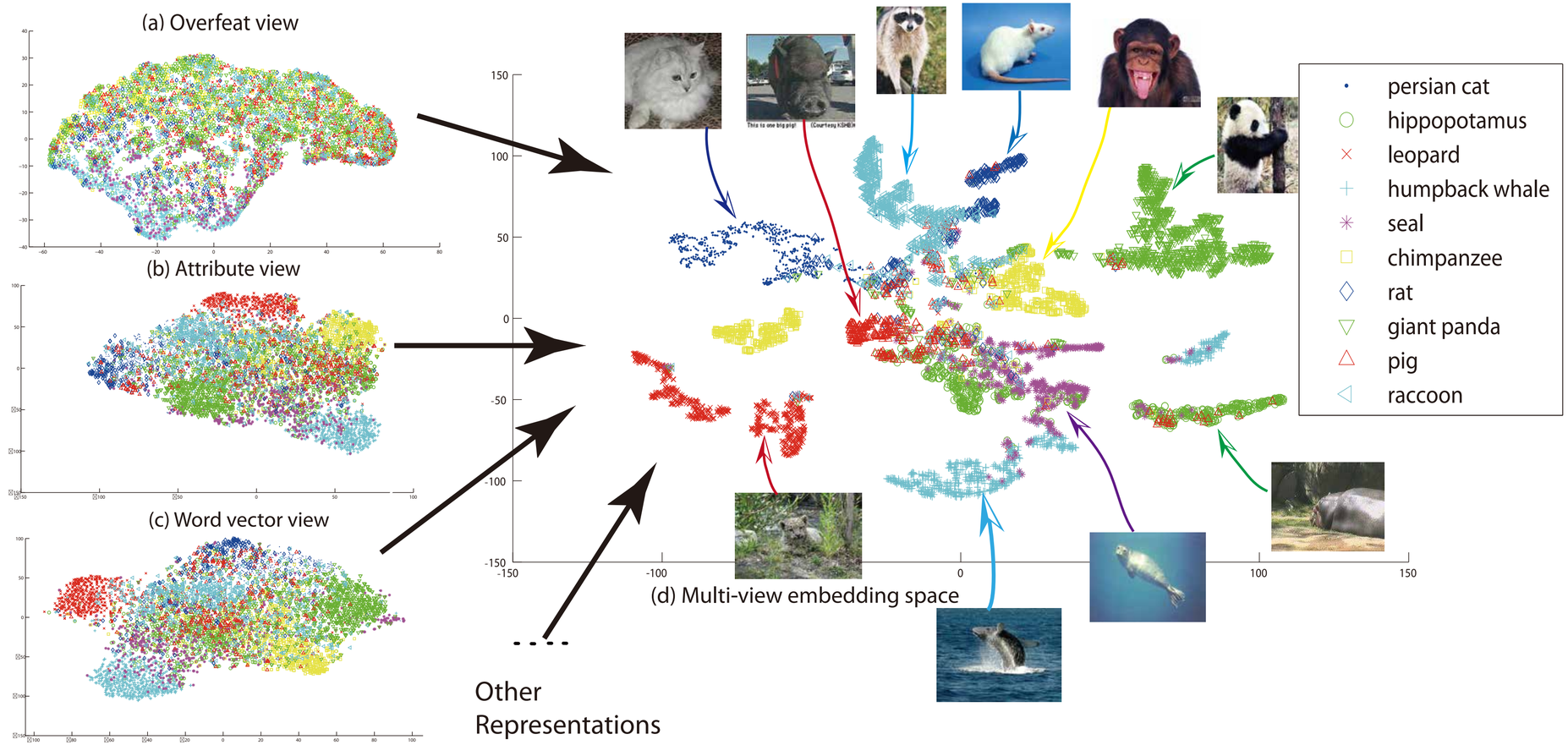
Approach	AwA ($\mathcal{H}$ [27])	AwA ($\mathcal{O}$)	AwA ($\mathcal{O}, \mathcal{D}$)	USAA	CUB ($\mathcal{O}$)	CUB ($\mathcal{F}$)
DAP	40.5([27])/41.4([28])/38.4*	51.0*	57.1*	33.2([15])/35.2*	26.2*	9.1*
IAP	27.8([27])/42.2([28])	–	–	–	–	–
M2LATM [15]***	41.3	–	–	41.9	–	–
ALE/HLE/AHLE [1]	37.4/39.0/43.5	–	–	–	–	18.0*
Mo/Ma/O/D [38]	27.0/23.6/33.0/35.7	–	–	–	–	–
PST [36]***	42.7	54.1*	62.9*	36.2*	38.3*	13.2*
[54]	48.3**	–	–	–	–	–
TMV-BLP [14]***	47.7	69.9	77.8	48.2	45.2	16.3
TMV-HLP***	**49.0**	**73.5**	**80.5**	**50.4**	**47.9**	**19.5**

Comparison with State-of-the-Art of ZSL on AwA, USAA and CUB

1. $\mathcal{H}$, $\mathcal{O}$, $\mathcal{D}$ and $\mathcal{F}$ represent hand-crafted, OverFeat, DeCAF, and Fisher Vector.
2. *: our implementation; **: requires additional human annotations; ***: requires unlabelled data, i.e. a transductive setting.
3. '–': no result reported;

Fu et al. *Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation*, ECCV 2014
Fu et al. *Transductive Multi-View Zero-Shot Learning*, IEEE TPAMI 2015

Visualization of AwA Dataset



t-SNE Visualisation of (a) OverFeat view, (b) attribute view, (c) word vector view, (d) transition probability of pairwise nodes by our framework. The unlabelled target classes are much more separable in (d).

Fu et al. *Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation*, ECCV 2014
Fu et al. *Transductive Multi-View Zero-Shot Learning*, IEEE TPAMI 2015

Embedding

Pairwise Graph Embedding

Subjective Visual Properties

Richer semantic representations;

- Less ambiguity: *who is smiling more*? Vs. *how much is smiling*?
- Better for transfer learning & active learning;
- Better for interactive image retrieval;

Subjective visual properties,

- Image/video interestingness & aesthetics;
- Image memorability & image/video quality.

Who is more beautiful?

Fu et al. *Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels*", IEEE TPAMI 2016;
Fu et al. *Interestingness Prediction by Robust Learning to Rank*, ECCV 2014

Crowdsourced Paired Comparisons

Advantages:

1. Cheap and easier for annotators (AMT);
2. We can annotate large-scale datasets;

Problems:

Outliers:

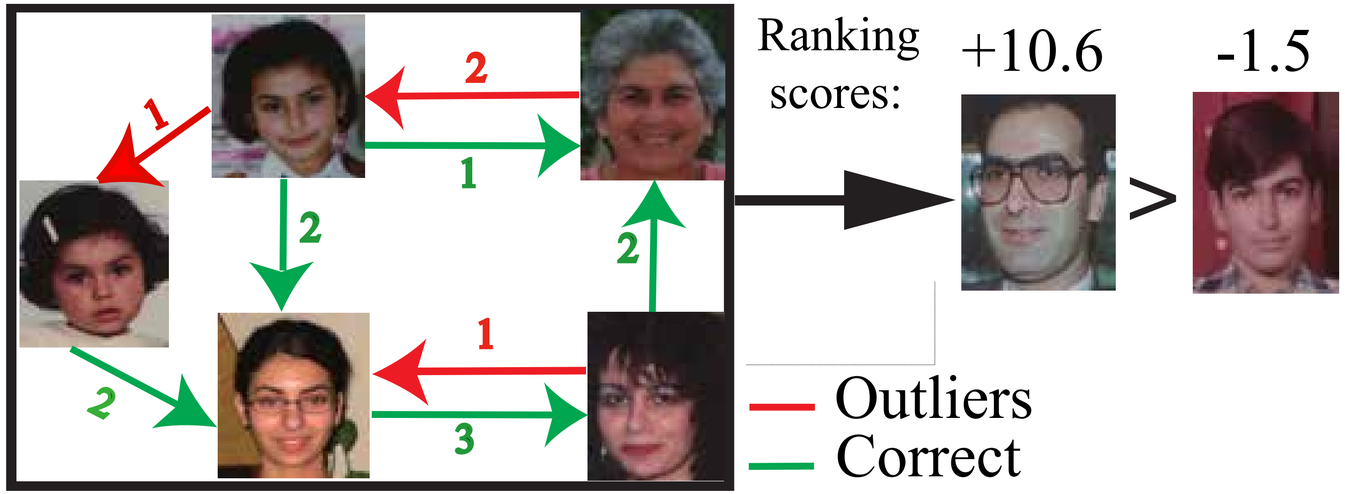
- Ambiguous or Malicious/ Lazy annotators;
- Cultural&Psychological factors for inconsistent comparisons;

Sparsity:

- $\mathbf{O}(n^2)$ for n instances;
- Large-scale dataset even with lots of comparisons, resulting in sparse data.

Fu et al. *Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels*", IEEE TPAMI 2016;
Fu et al. *Interestingness Prediction by Robust Learning to Rank*, ECCV 2014

Overview — Robust Learning to Rank



Synthetic experiment: Human Age Prediction from Face Images

Pairs data are naturally directed graphs $G = (V, E)$

$$E = \{i \to j | Y_{ij} > 0\}$$
$$V = \{i\}_{i=1}^{I}$$

Low-level feature matrix $\Phi = \left[\phi_i^T\right]_{i=1}^{I} \in R^{I \times d}$;

The edge $i \to j$ exists if $Y_{ij} > 0$.

Fu et al. *Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels*", IEEE TPAMI 2016;
Fu et al. *Interestingness Prediction by Robust Learning to Rank*, ECCV 2014

Robust Learning to Rank (URLR) Framework

Each edge $i \to j \in E$, Y_{ij} is modelled as

$$Y_{ij} = \beta^T \phi_i - \beta^T \phi_j + \gamma_{ij}$$

The edges of the whole dataset are

$$Y = C\Phi\beta + \Gamma$$

where C is incidence matrix. γ_{ij} is also known as incidental parameter [*1].

We minimize the discrepancy between annotation and prediction $C\Phi\beta + \Gamma$ as well as keeping sparse outliers Γ,

$$\min_{\beta,\Gamma} \frac{1}{2}\|Y - C\Phi\beta - \Gamma\|_2^2 + \lambda\|\Gamma\|_1$$

$$:= \sum_{i \to j \in E} \left[\frac{1}{2}(Y_{ij} - \gamma_{ij} - \beta^T \phi_i - \beta^T \phi_j)^2 + \lambda|\gamma_{ij}| \right]$$

[*1] **Jianqing Fan**, Runlong Tang and Xiaofeng Shi, Partial Consistency with Sparse Incidental Parameters, arXiv:1210.6950, 2012.

Robust Learning to Rank (URLR) Framework (cont.)

To solve it, we rewrite the cost function as Lagaragain,

$$L(\beta, \Gamma) = \frac{1}{2}\|Y - X\beta - \Gamma\|_2^2 + \lambda\|\Gamma\|_1$$

where $X = C\Phi$, $X = U\Sigma A^T$ and $U = [U_1; U_2]$

We have

$$\hat{\beta} = (X^T X)^\dagger X^T (Y - \Gamma)$$

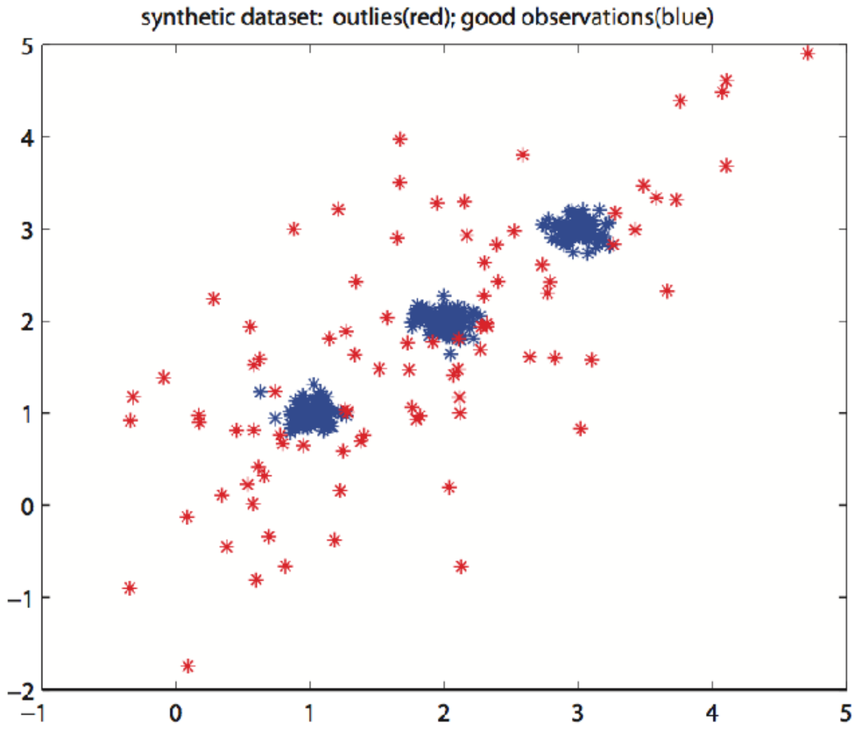
$$\hat{\Gamma} = \arg\min_{\Gamma} \|U_2^T Y - U_2^T \Gamma\| + \lambda\|\Gamma\|_1$$

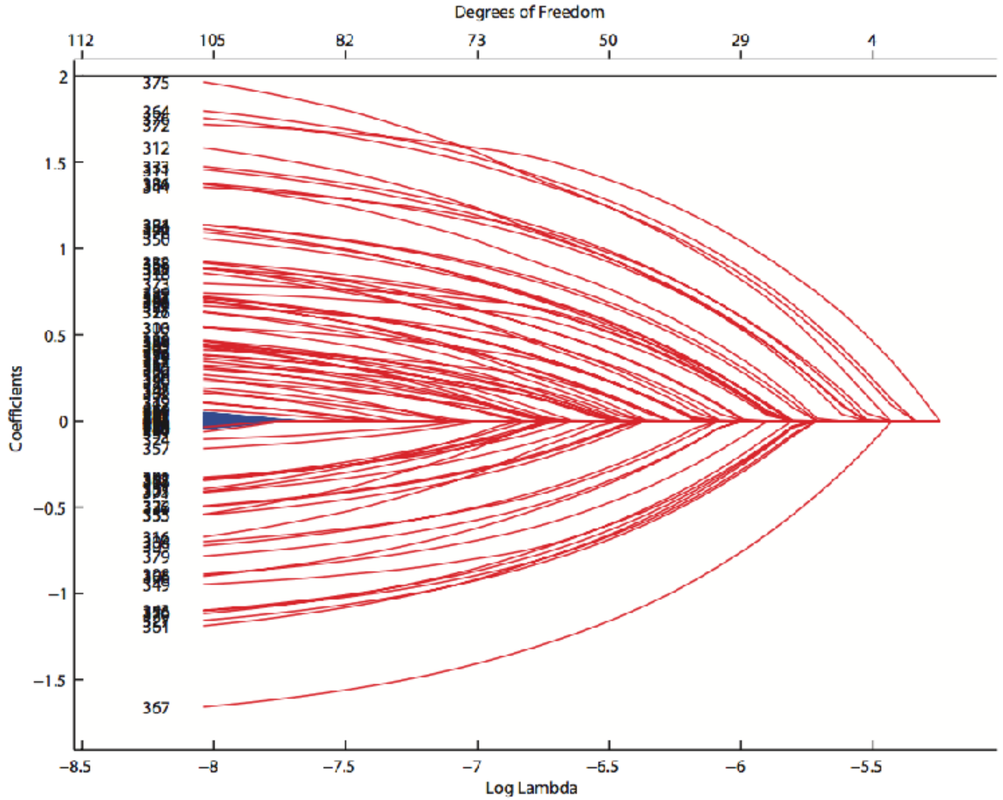
We solve Γ checking regularisation path,

It's one type of Preconditioned Lasso![*2].

[*2] Fabian L. Wauthier, Nebojsa Jojic and Michael I. Jordan, A Comparative Framework for Preconditioned Lasso Algorithms, NIPS 2013.

Checking Regularisation Path





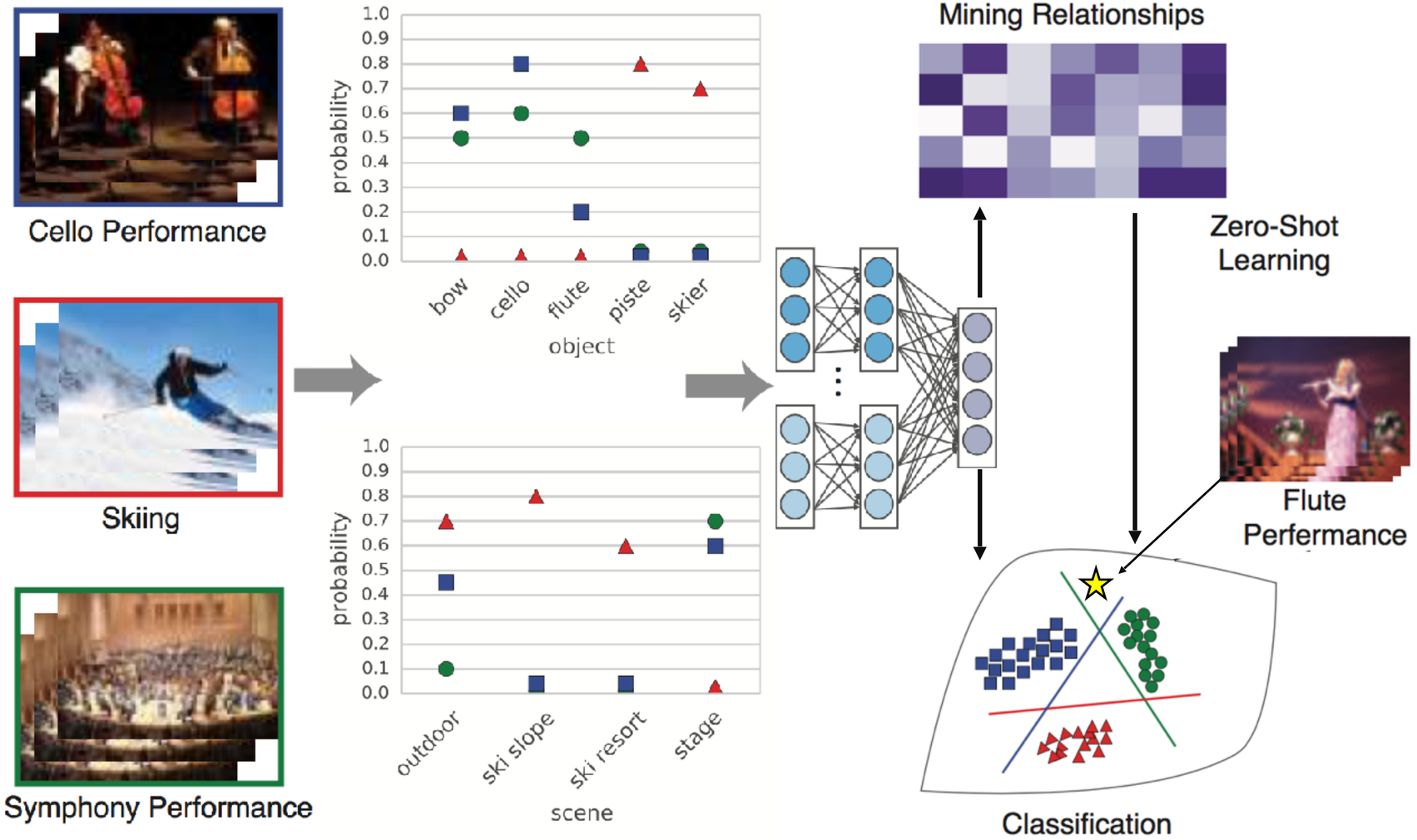
Red lines & red points indicate outliers;
Blue lines & blue points are inliers. Figures from [*3].

[*3] **Yanwei Fu**, De-An Huang, Leonid Sigal, Robust Classification by Pre-conditioned LASSO and Transductive Diffusion Component Analysis, submitted to ICLR 2016: http://arxiv.org/abs/1511.06340

Content

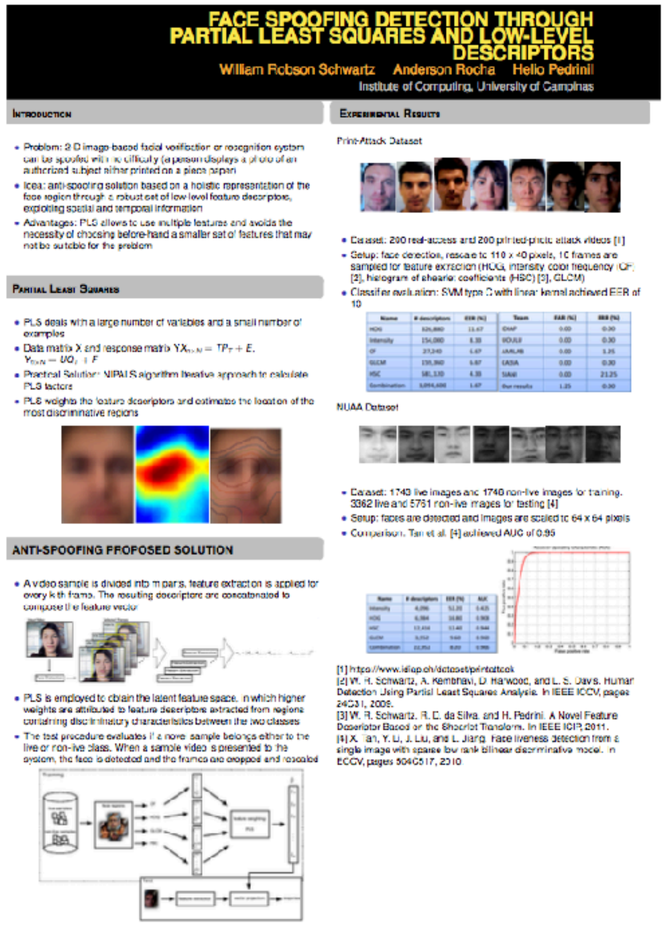
More
of our works

Video Understanding with Objects/Scenes

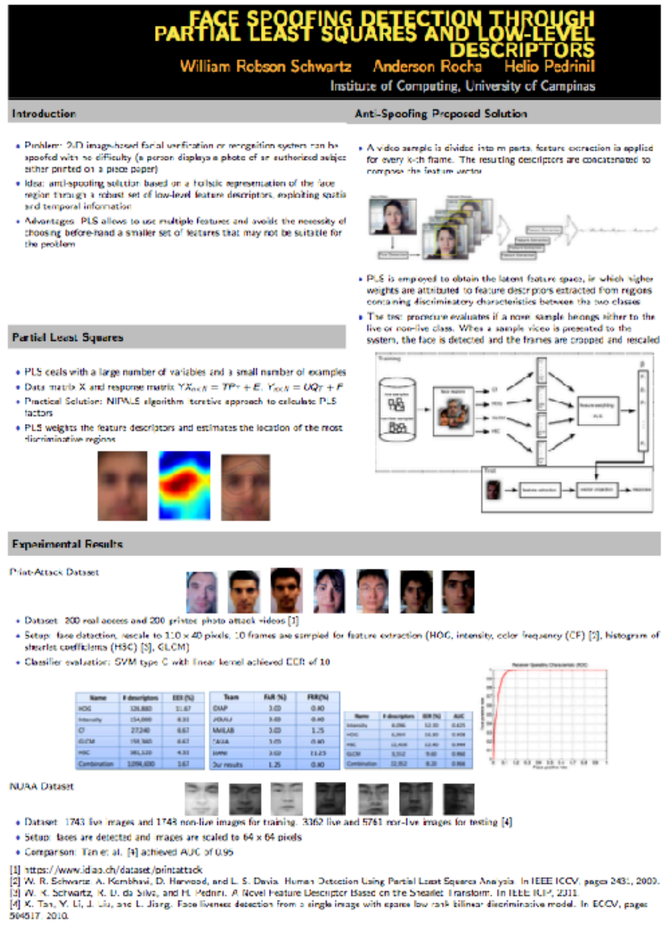


Wu *et al*. *Harnessing Object and Scene Semantics for Large-Scale Video Understanding*, **CVPR 2016**;

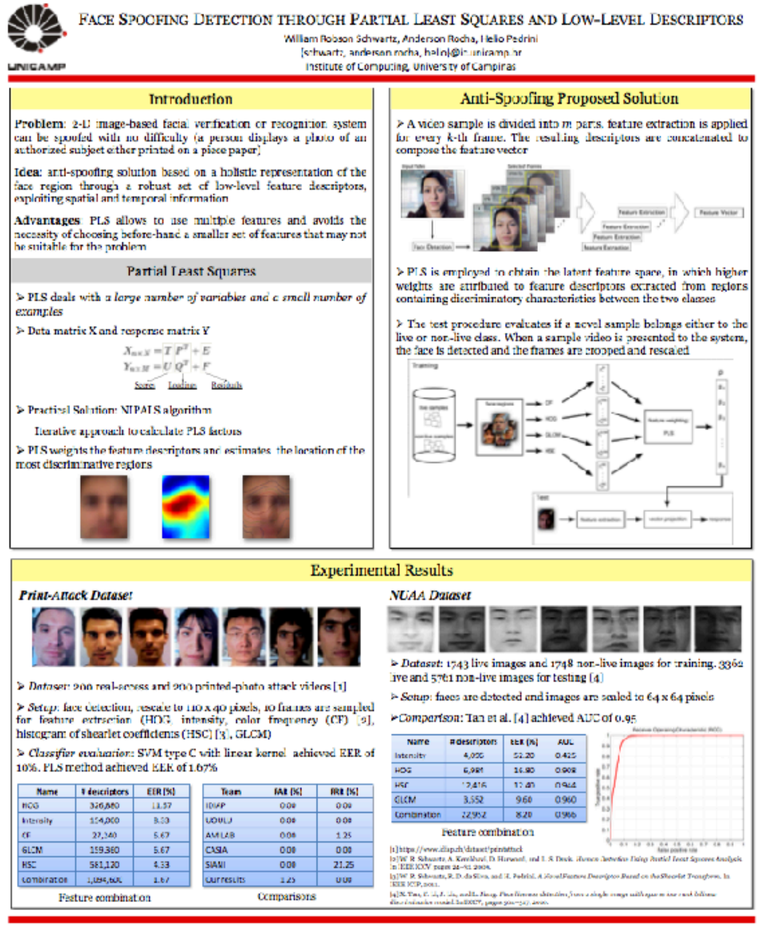
Learning to Generate Posters of Scientific Papers



(a) Designed by novice



(b) Our result



(c) Original poster

Qiang *et al.* "*Learning to Generate Posters of Scientific Papers*", **AAAI 2016;**
Qiang *et al.* "*Learning to Generate Posters of Scientific Papers by Probabilistic Graphical Model*", submitted to a journal;

Thanks!

yanwei.fu@iclouds.com