Semi-supervised Vocabulary-informed Learning

Yanwei Fu*

Leonid Sigal

* School of Data Science, Fudan University

Disney Research

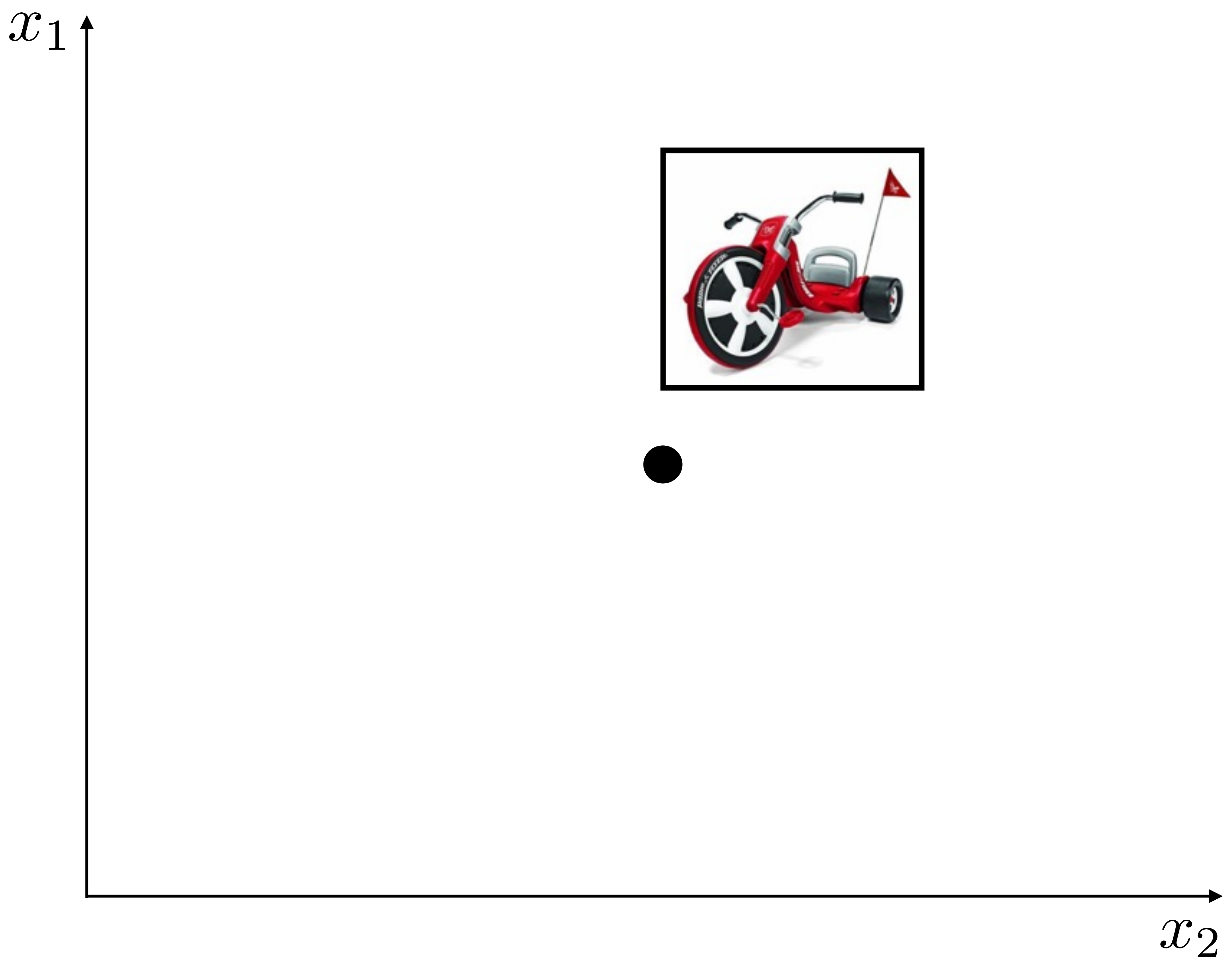
Supervised Learning

Problem Definition

Semantic labels → Visual feature space

airplane car

unicycle tricycle



Supervised Learning

Learning

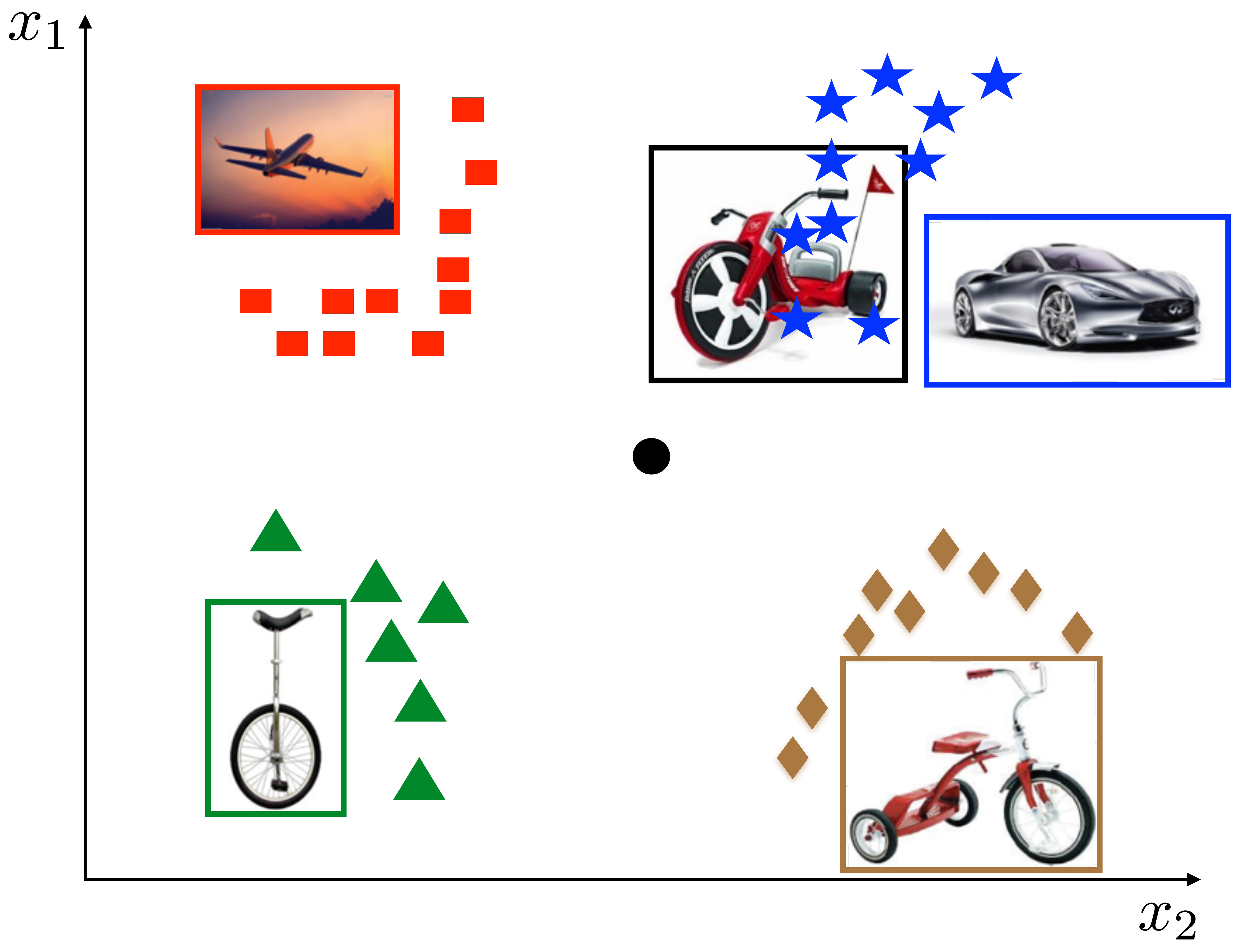
Semantic labels → Visual feature space

airplane

car

unicycle

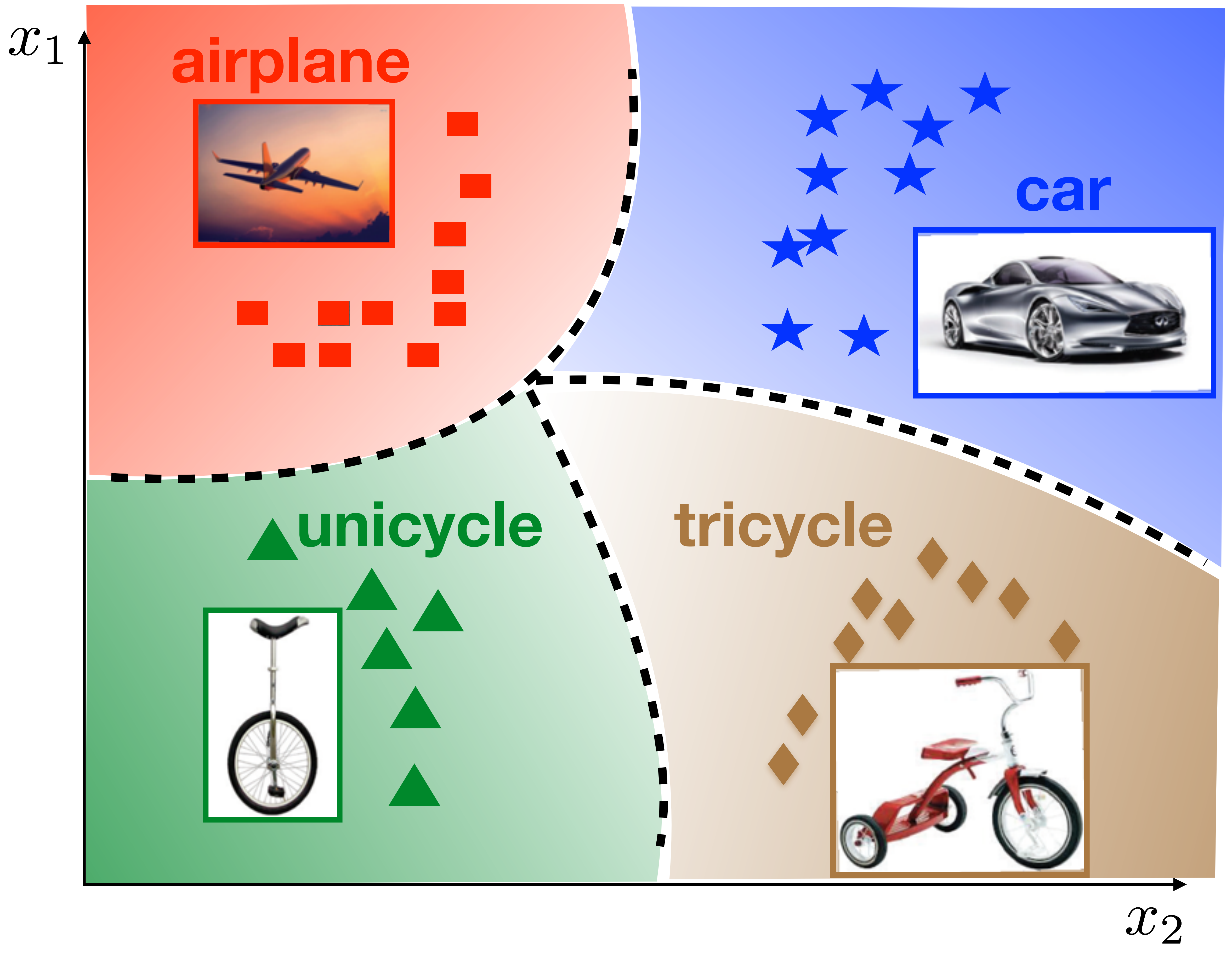
tricycle



Supervised Learning

Learning

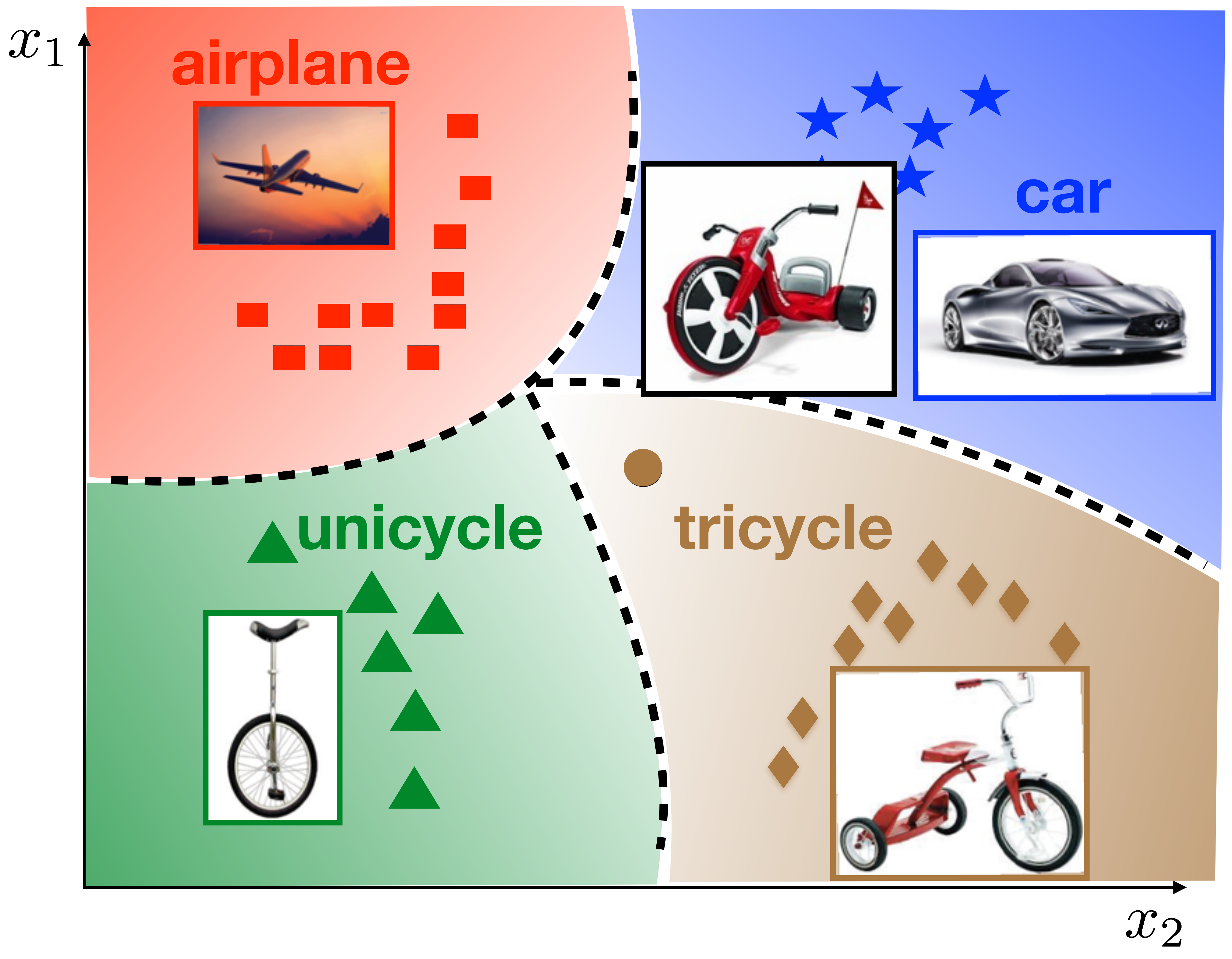
Semantic labels → Visual feature space



Supervised Learning

Inference

Semantic labels → Visual feature space



Zero-shot Learning

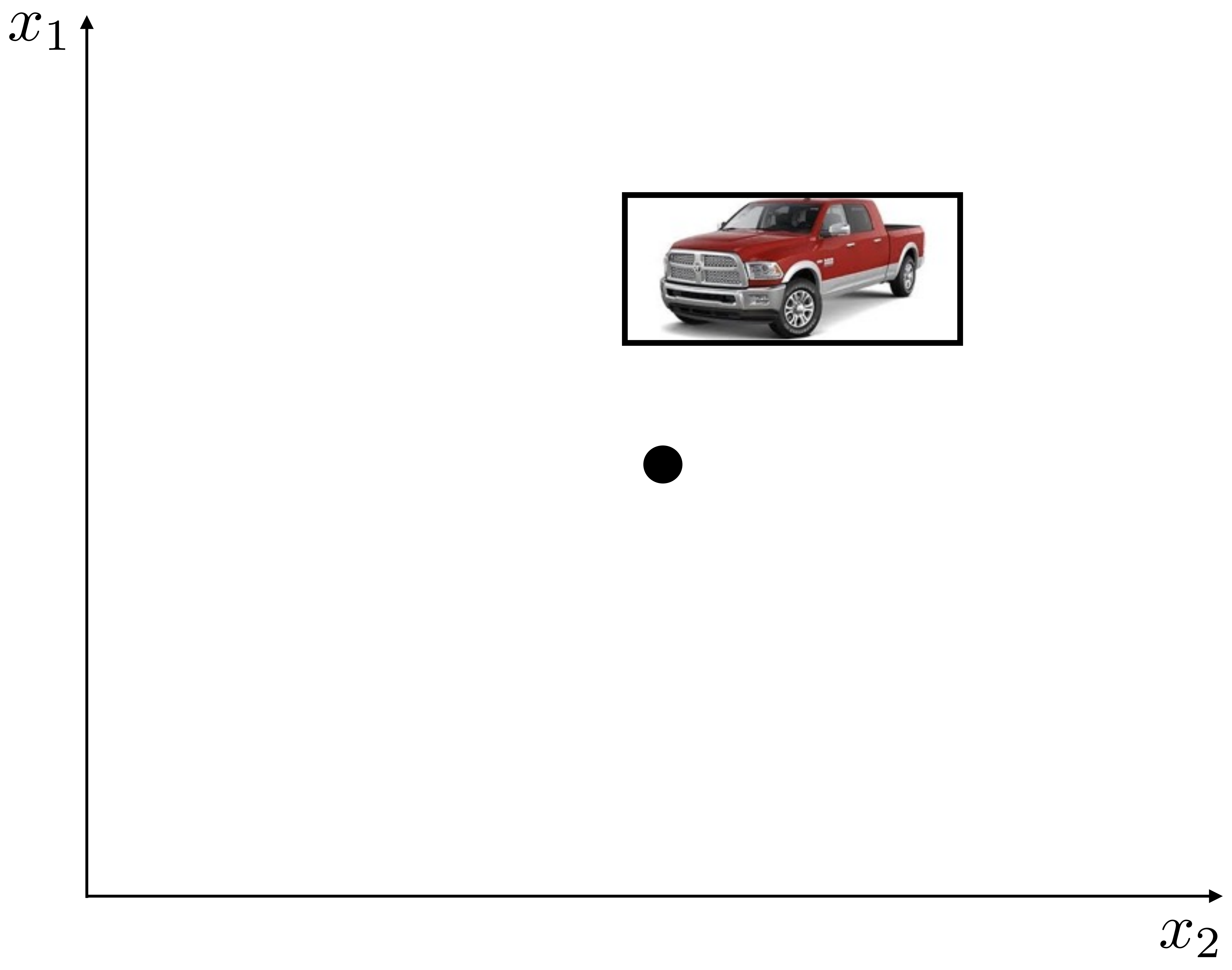
Problem Definition

Semantic labels → Visual feature space

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

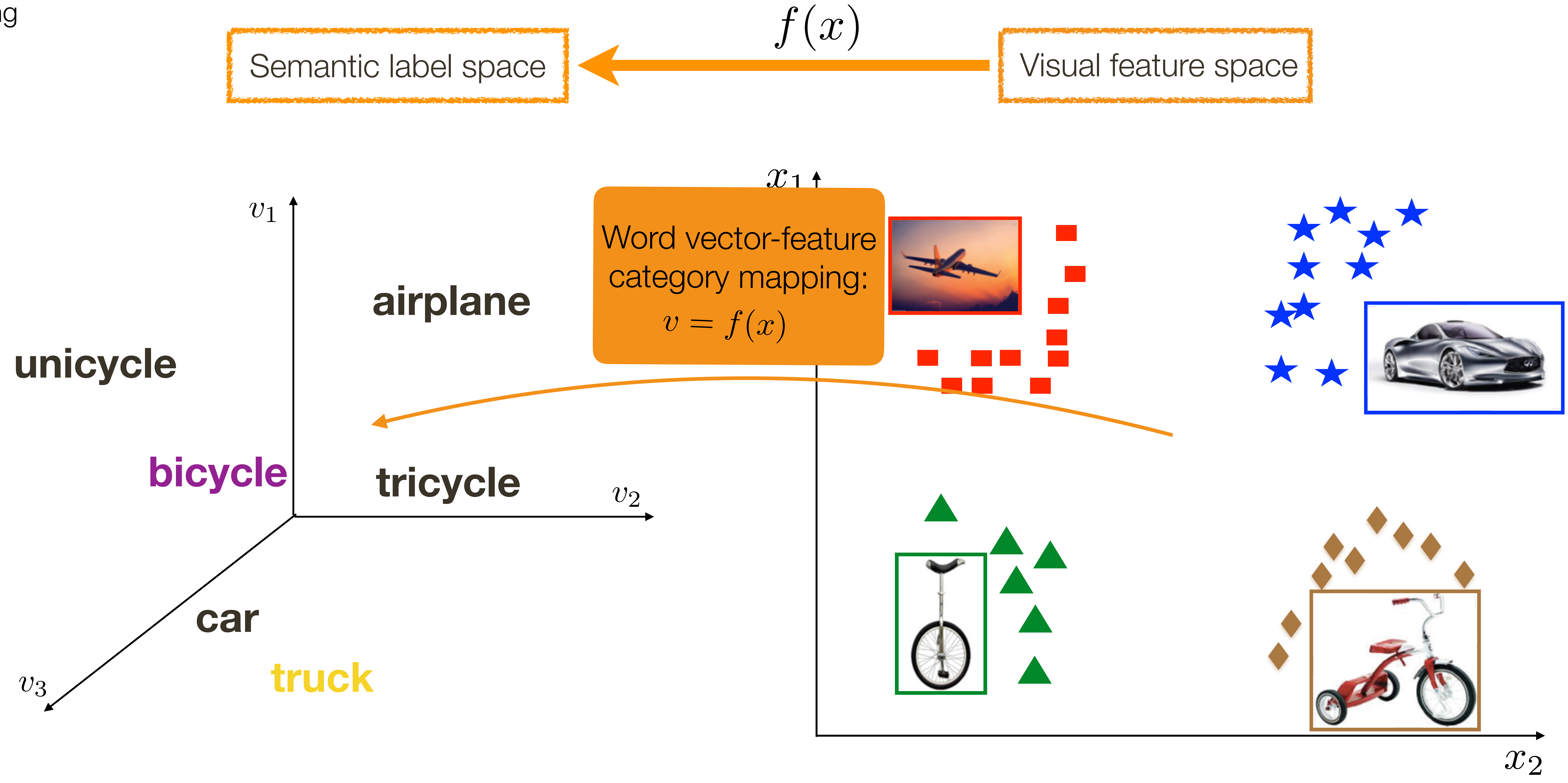
bicycle

truck



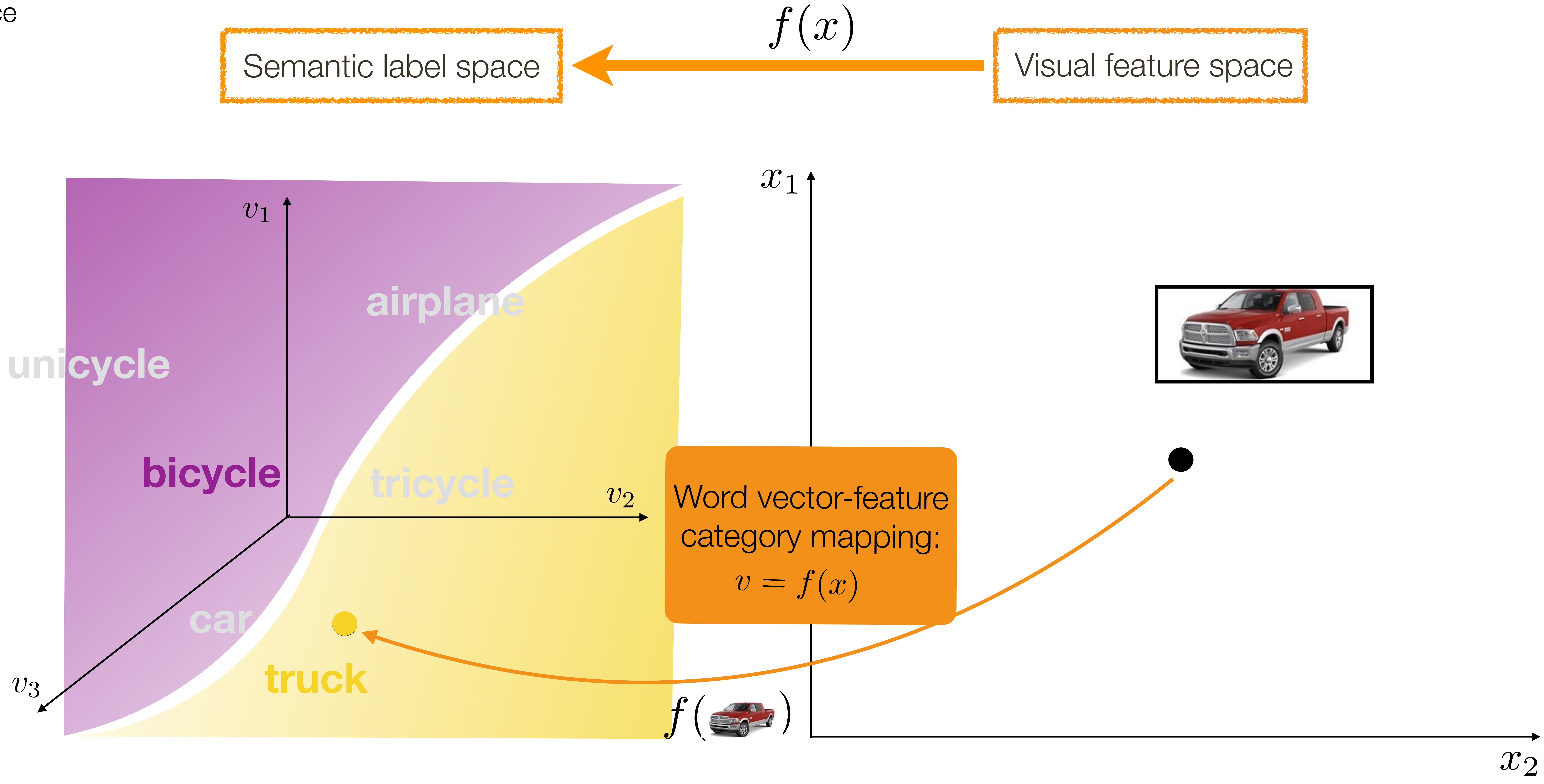
Zero-shot Learning

Learning



Zero-shot Learning

Inference



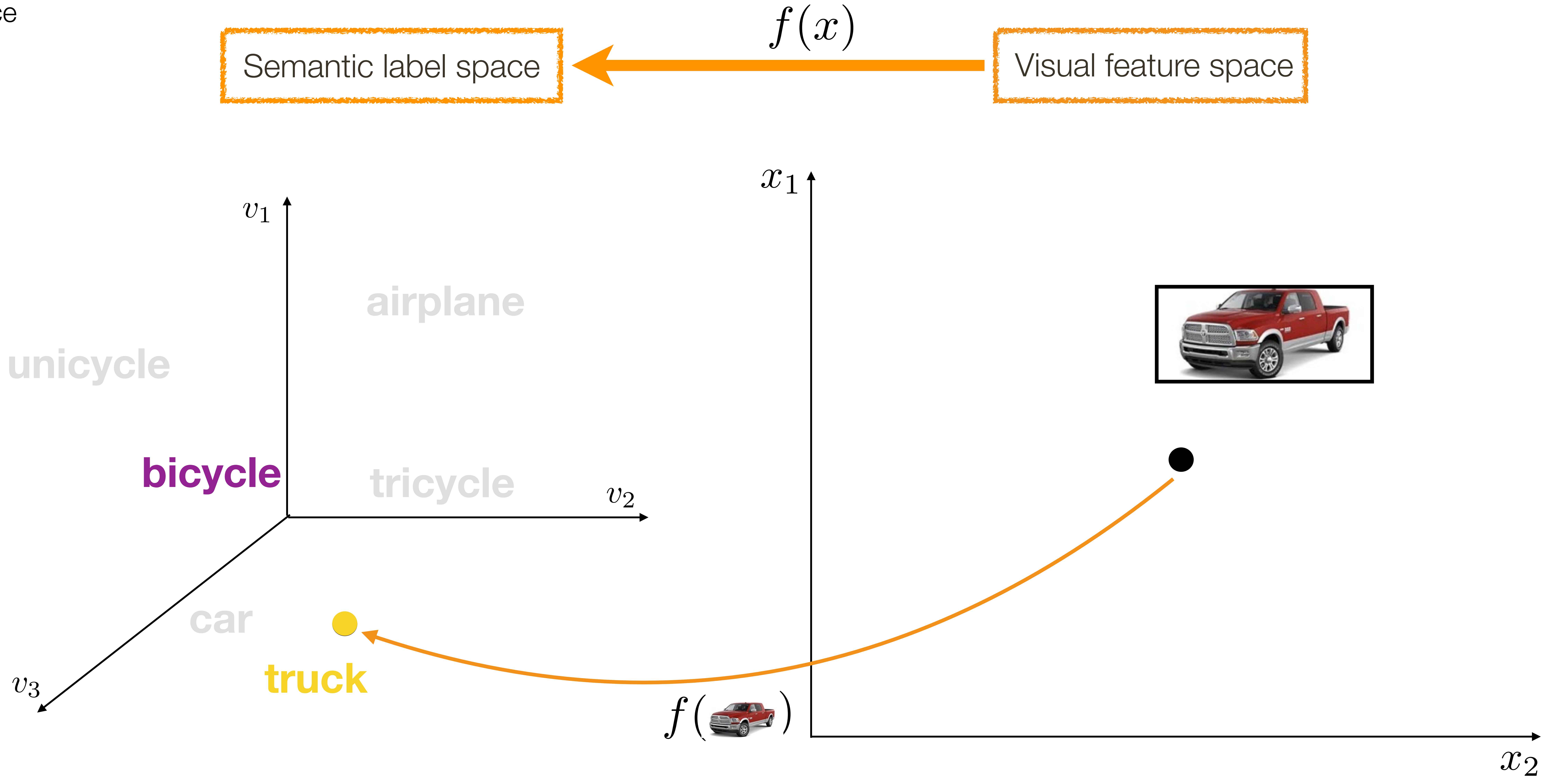
Key Question: How do we define semantic space?

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 $v\left(\text{Berlin}\right) - v\left(\text{Germany}\right) = v\left(\text{Paris}\right) - v\left(\text{France}\right)$	Limited interpretability of each dimension

Zero-shot Learning

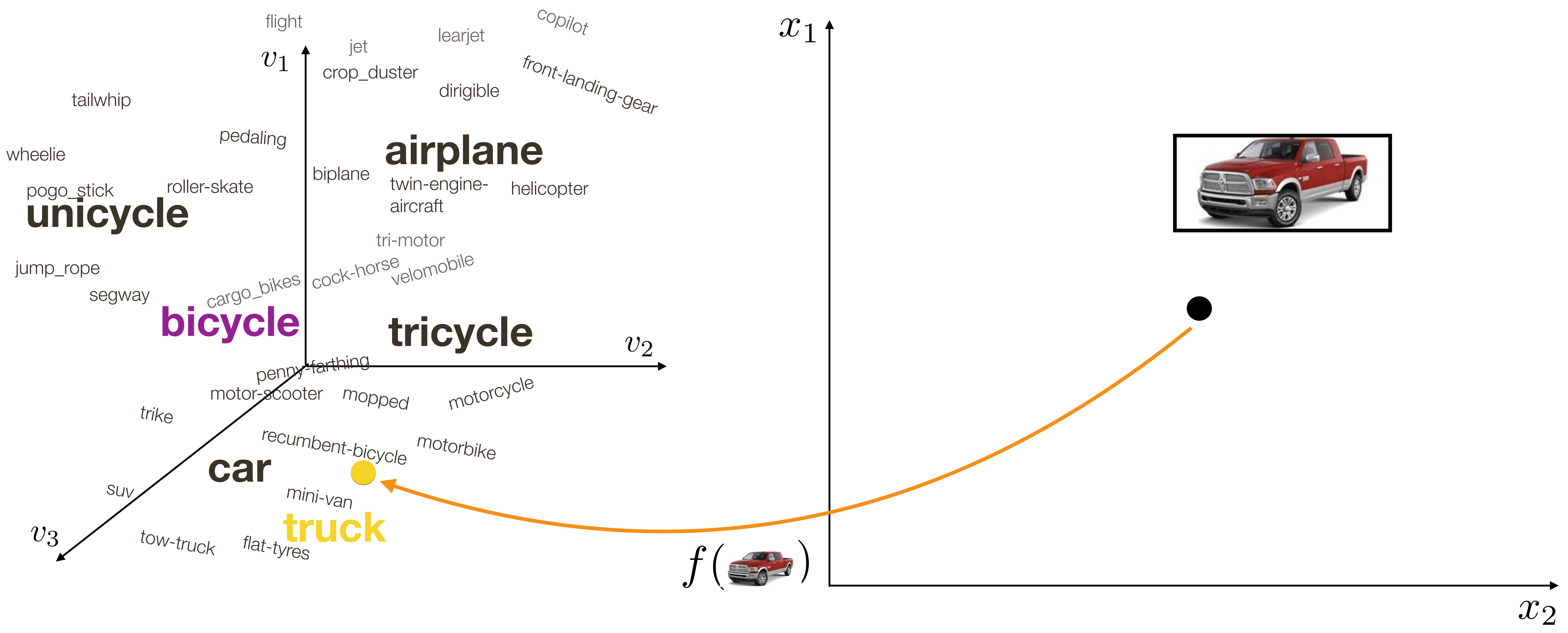
Inference



Open-set Recognition

Problem Definition

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



Summary:

Supervised Learning: [Fei-Fei *et al*. TPAMI’06], [Deng *et al*. ECCV’14], [Torralba *et al.* TPAMI’08], [Weston *et al.* IJCAI’11]

Pros: Very good quantitative performance

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

Zero-shot Learning: [Palatucci *et al*. NIPS’09],[Lampert *et al.* CVPR’09], [Farhadi *et al.* CVPR’09], [Rohrbach *et al.* CVPR’10]

Pros: Does not require instance labeling for target classes

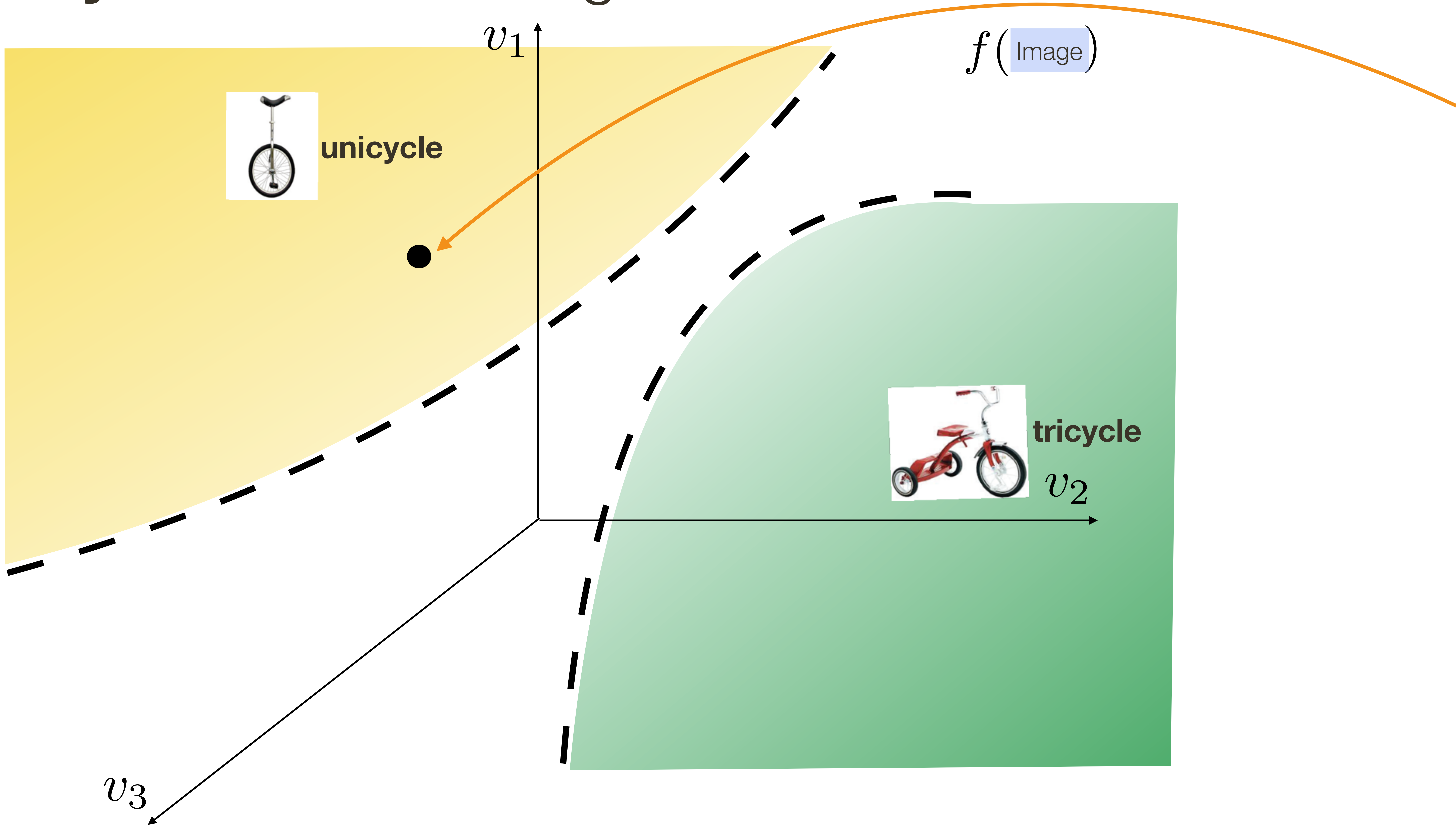
Cons: Typically limited to recognition with target classes only
Relatively ***small vocabulary*** (~50-200 classes typically)

Open-set Learning: [Scheirer *et al.* TPAMI’13], [Sattar *et al.* CVPR’15], [Bendale *et al.* CVPR’15] [Guadarrama *et al*. RSS’14]

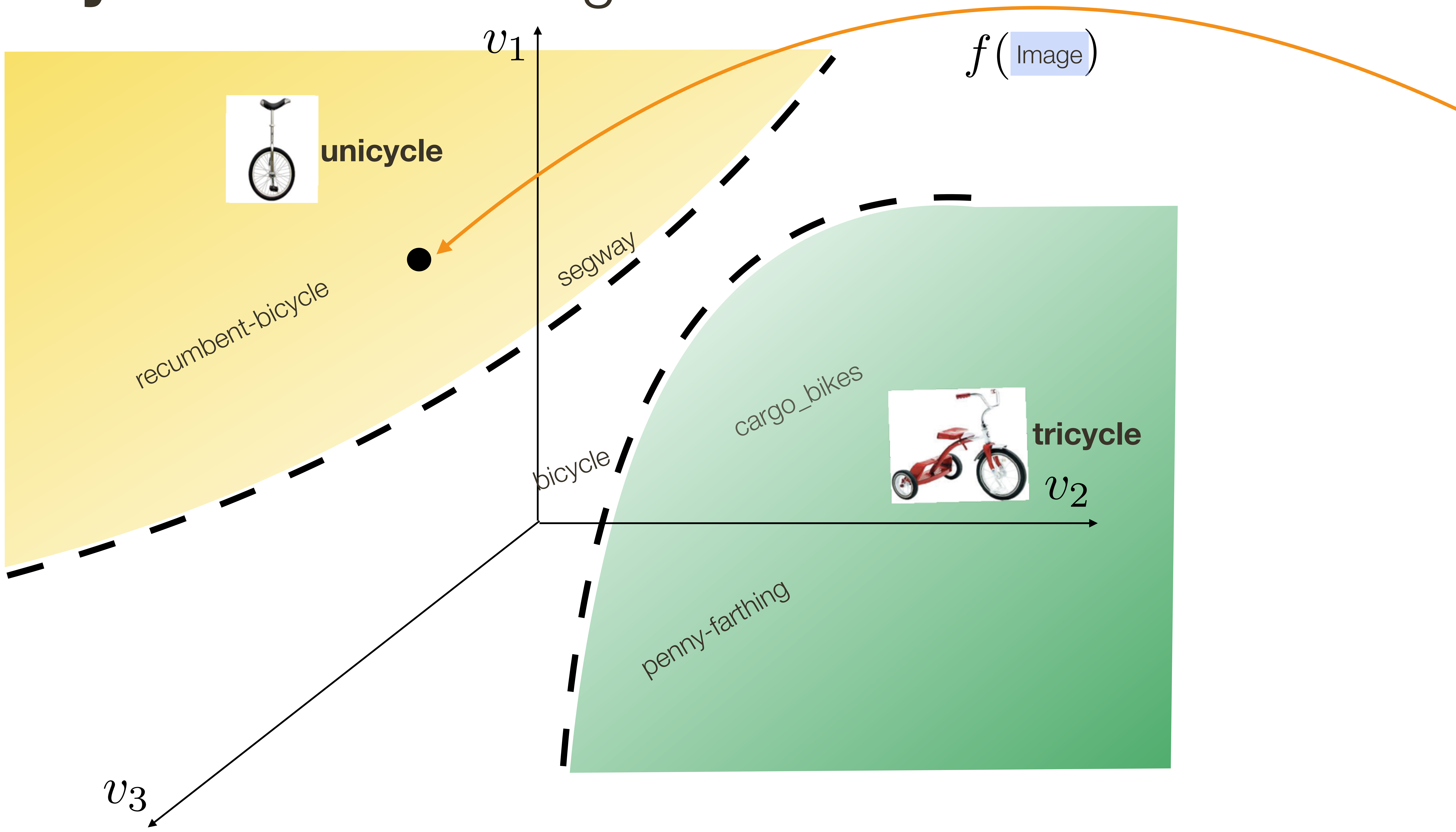
Pros: Does not require instance labeling for target classes
Large vocabulary (up to 310K classes)

Key Question: Can this large vocabulary be actually useful for recognition?

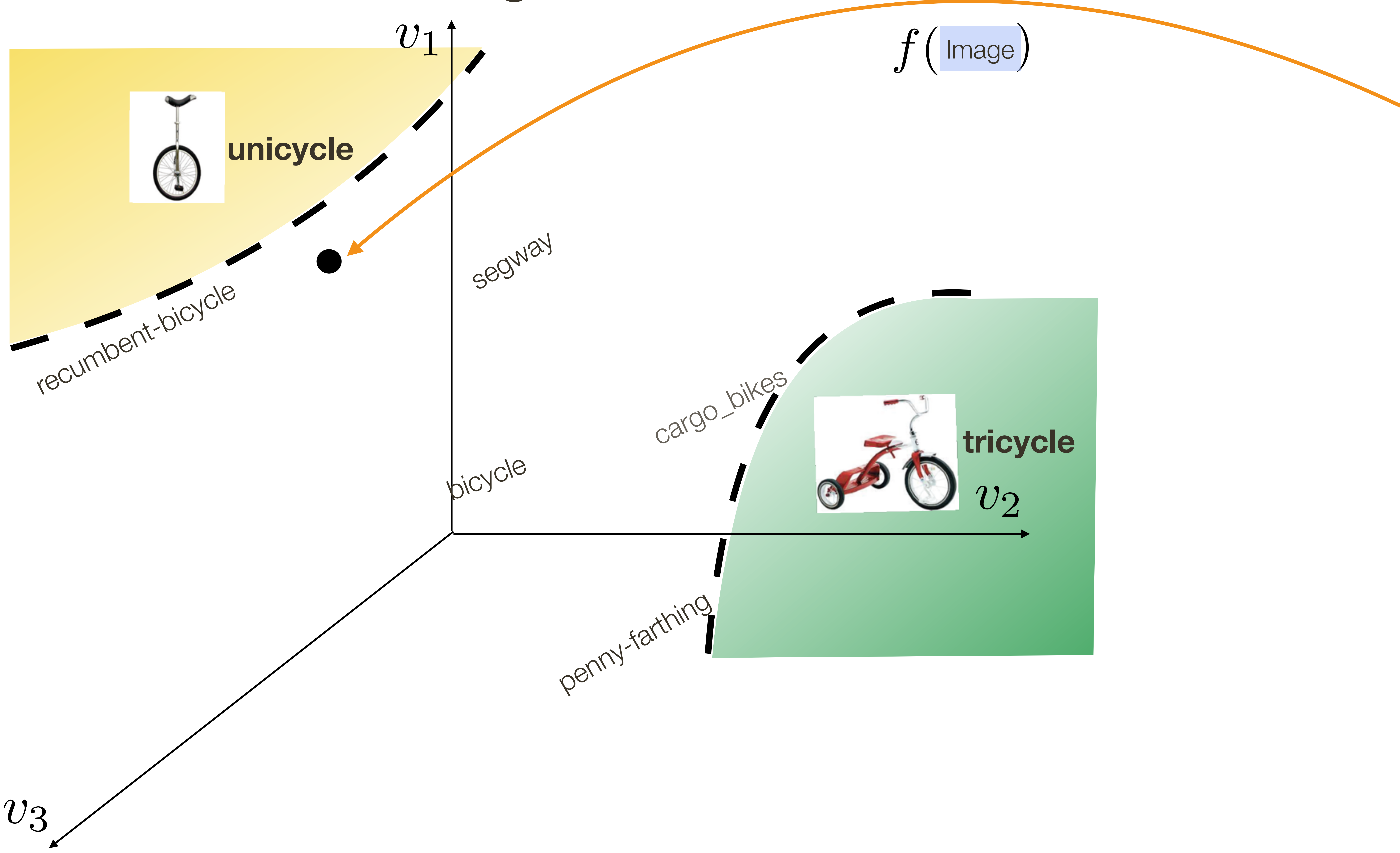
Vocabulary-Informed Recognition



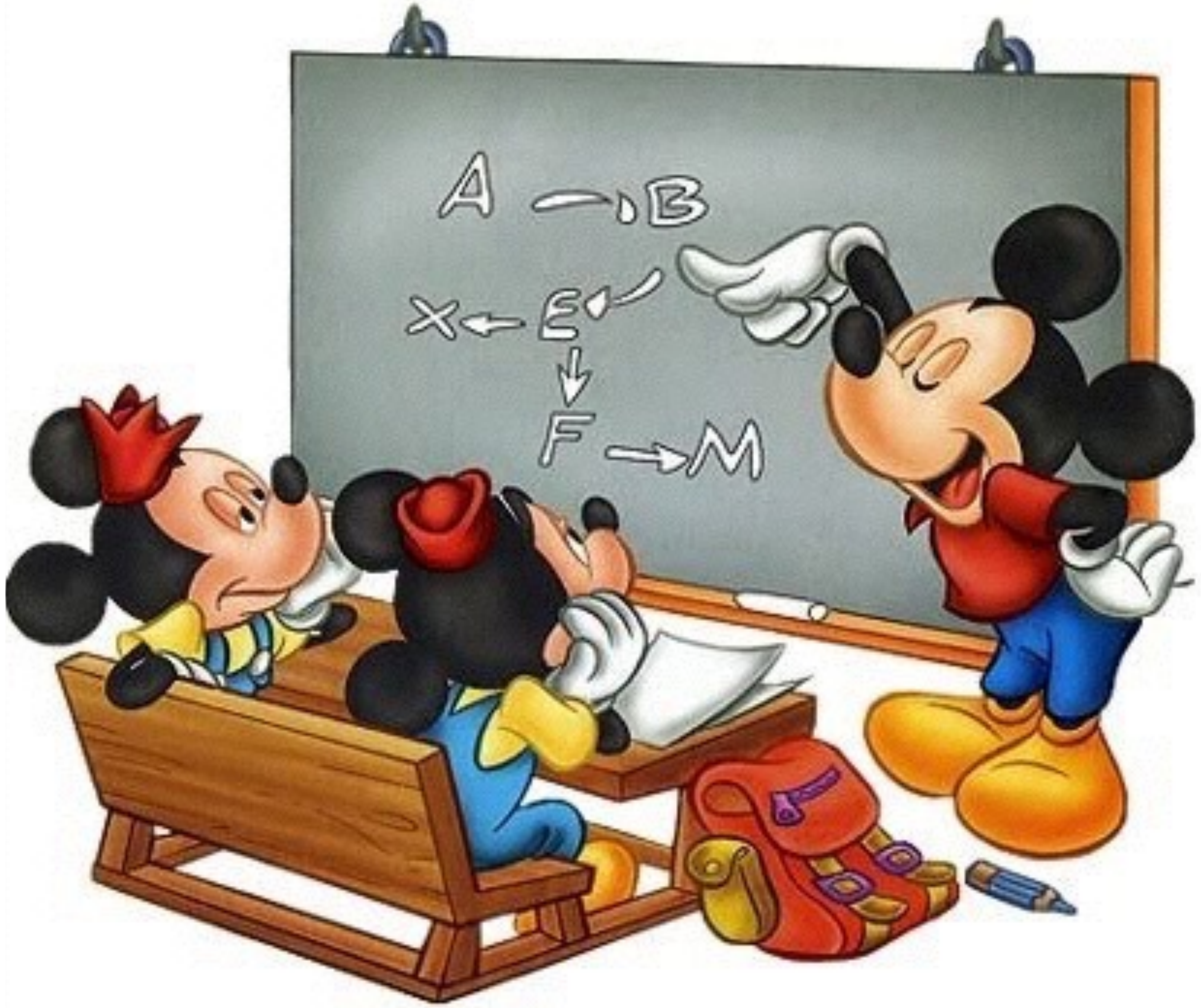
Vocabulary-Informed Recognition



Vocabulary-Informed Recognition



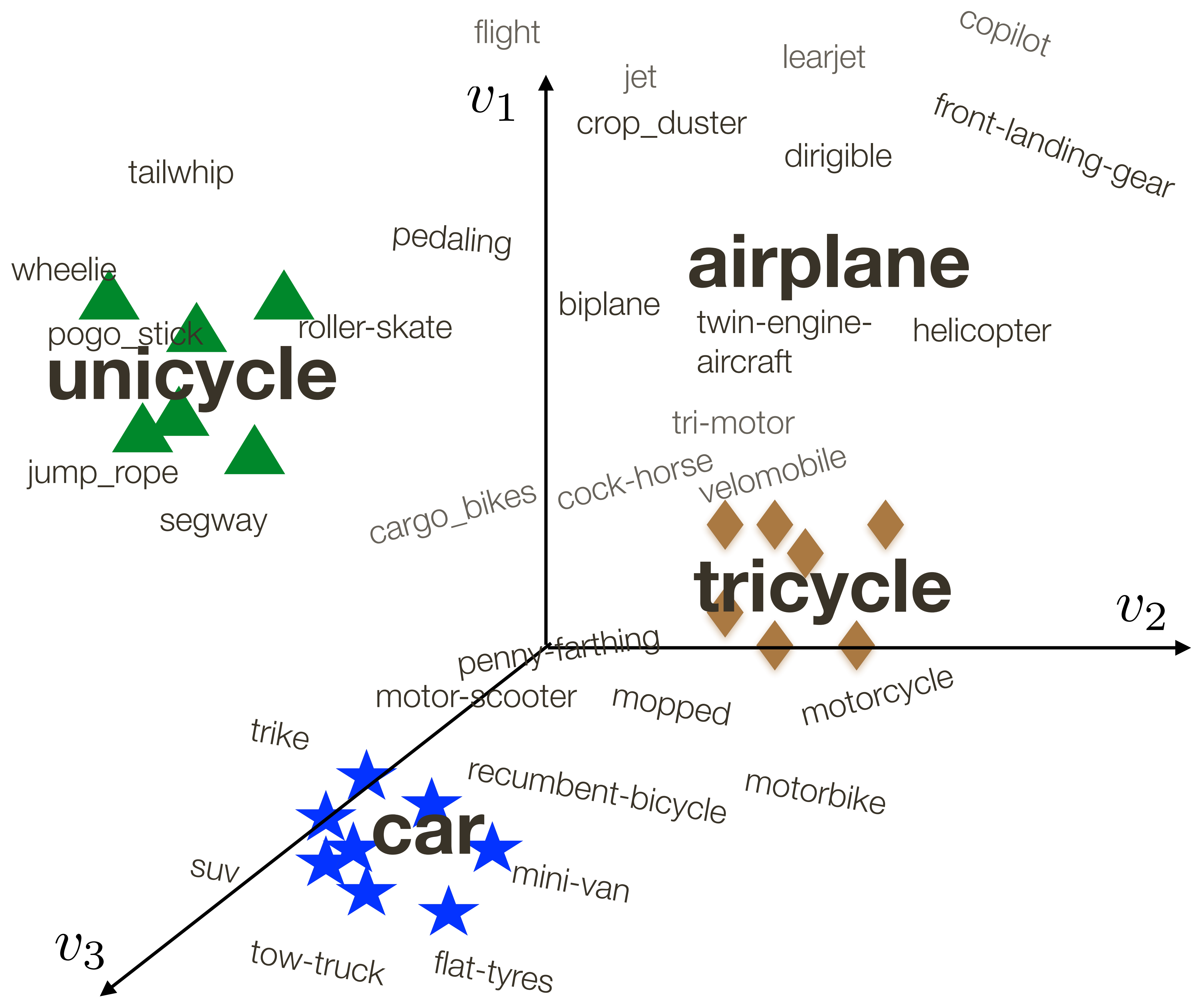
Formulation

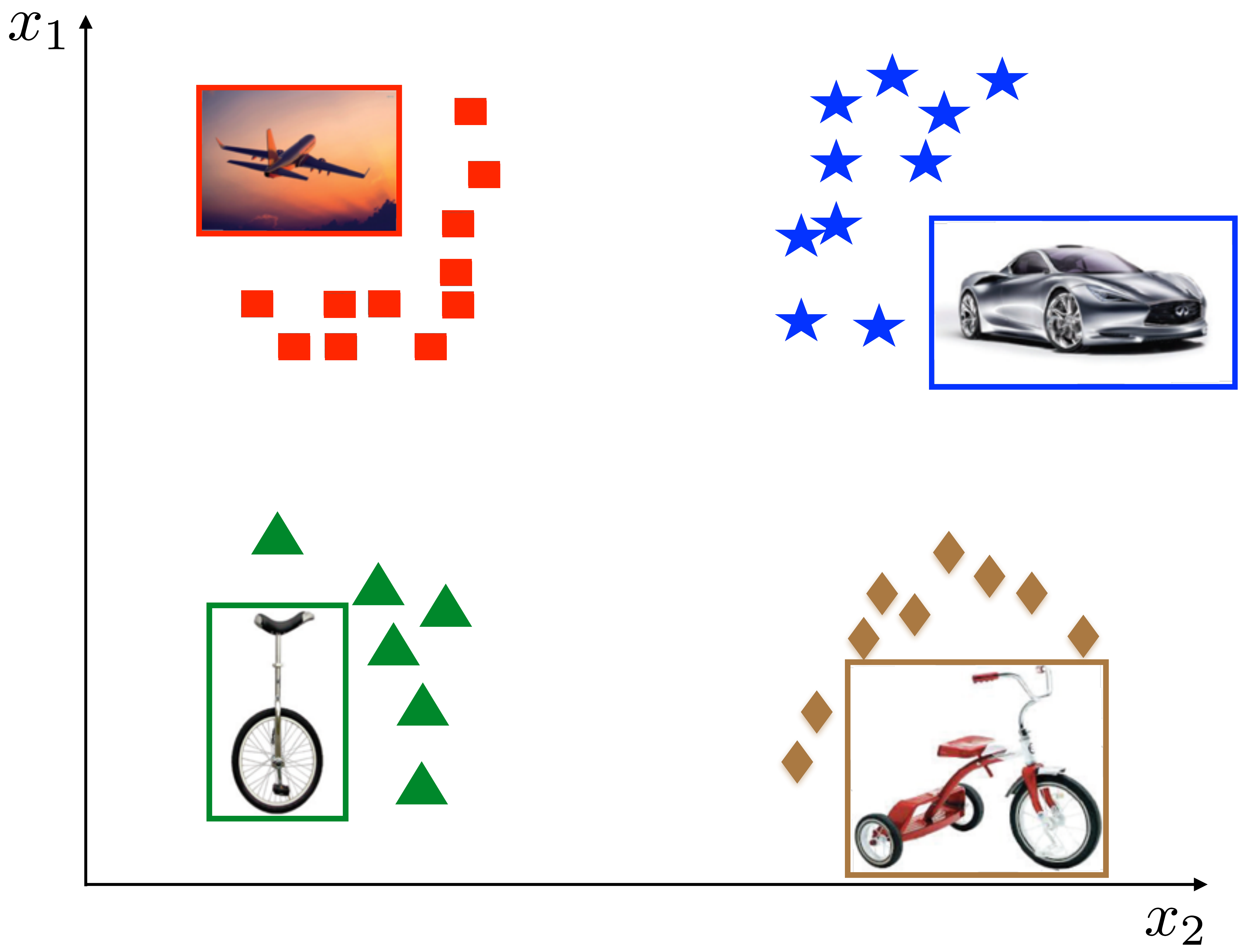


Semi-Supervised Vocabulary-Informed Learning (SS-Voc)

Regression term

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



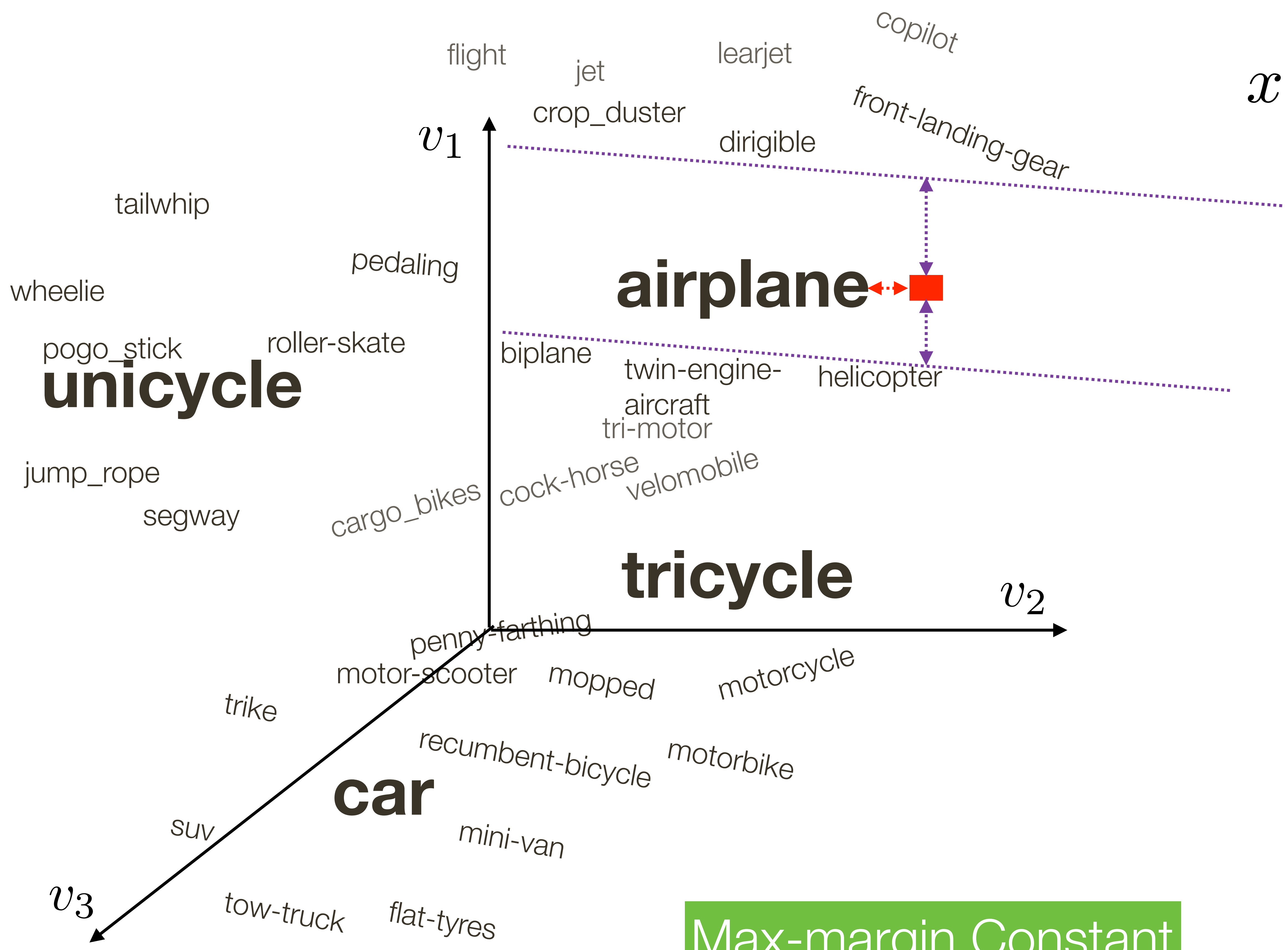


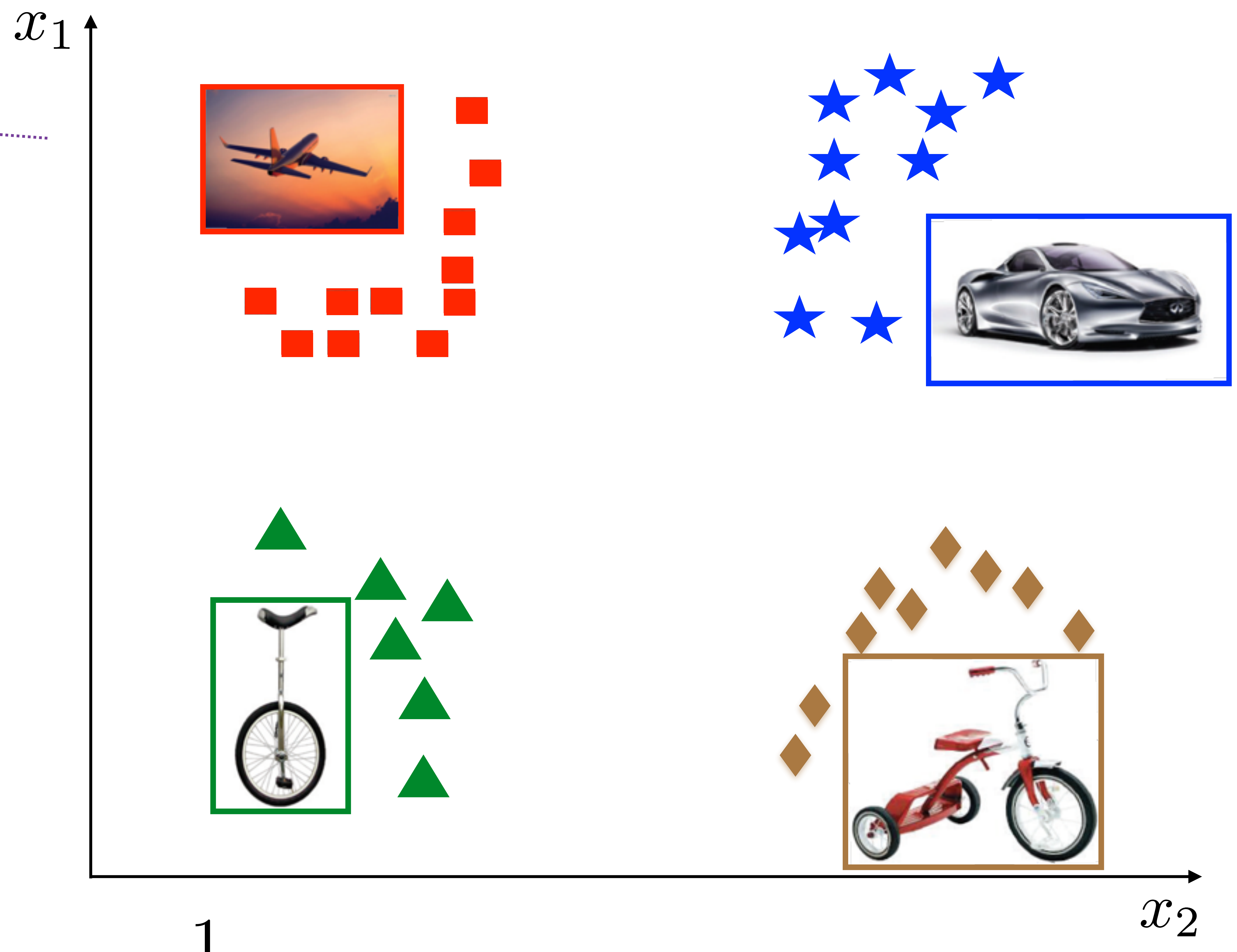
$$\sum \mathcal{L}_{\epsilon}(f(\text{[image]}, W), \mathbf{u}_{\text{airplane}})$$

Semi-Supervised Vocabulary-Informed Learning (SS-Voc)

Max-margin term

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



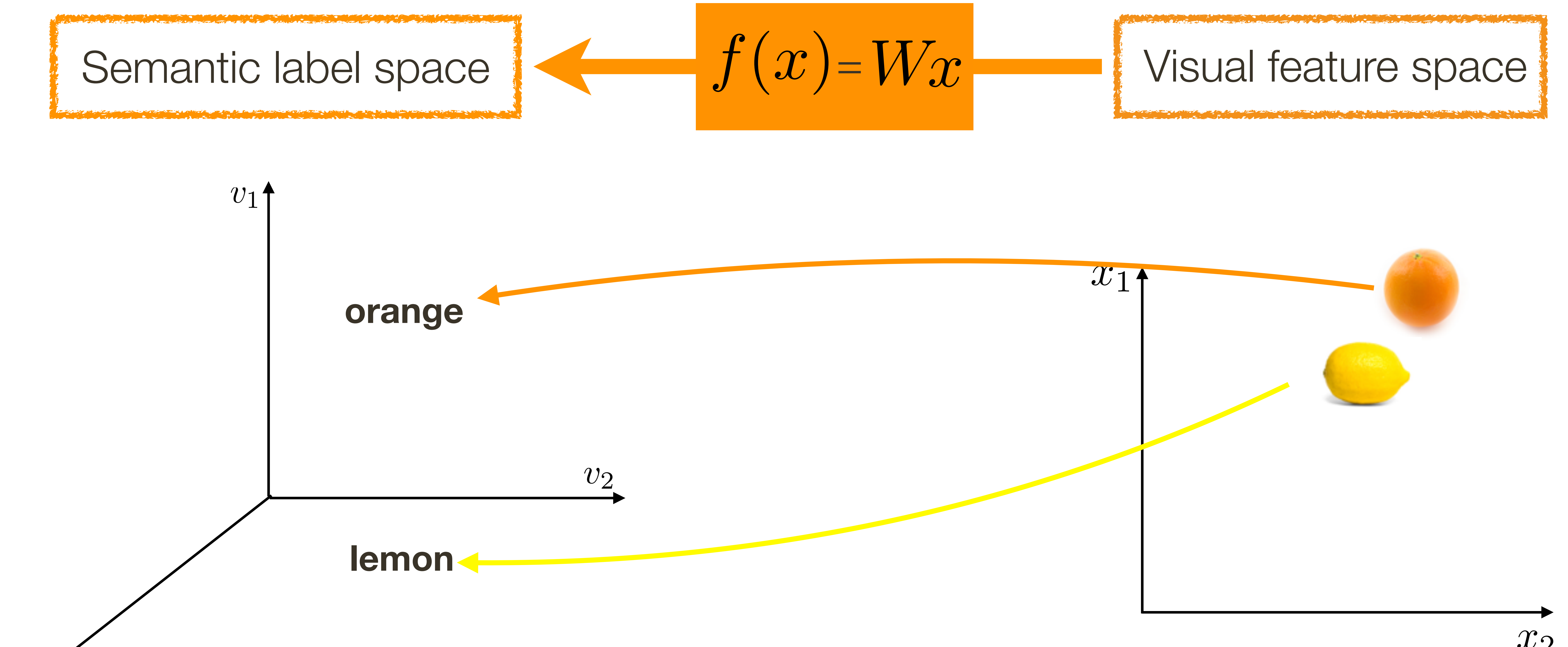


Max-margin Constant

$$\sum \left[\underline{C} - \frac{1}{2}\underline{\mathcal{L}_\epsilon(f(\;), \mathbf{u}_{\text{airplane}})} + \frac{1}{2}\underline{\mathcal{L}_\epsilon(f(\;), \mathbf{u}_{\text{helicopter}})} \right]$$

Semi-Supervised Vocabulary-Informed Learning (SS-Voc)

Final full objective



Allows fine-tuning of the word-vector space

$$\sum \mathcal{L}_\epsilon(f(\;, W), \mathbf{u}_{\text{airplane}} V) + \sum \left[C - \frac{1}{2}\mathcal{L}_\epsilon(f(\;), \mathbf{u}_{\text{airplane}} V) + \frac{1}{2}\mathcal{L}_\epsilon(f(\;), \mathbf{u}_{\text{helicopter}} V)\right]$$

Regression Term

Word-vector sp Max-margin Term mation

Advantages of the Approach

- A new paradigm for learning informed by very large vocabulary
- A unified framework for supervised, zero-shot learning
- Competitive quantitative performance
- Our framework can even scale up to open set image recognition with 310,000 vocabulary entities

Evaluation

Datasets

Animals with Attributes(AwA) [Lampert *et al.* CVPR 2009]:

40 auxiliary classes (24295 images), 10 target classes (6180 images);

We use 5 instances per auxiliary class for learning;

ImageNet 2012/2010 [Deng *et al.* CVPR 2009]:

1000 auxiliary classes (from ImageNet 2012);

360 target classes (from ImageNet 2010).

We use 3 instance per auxiliary class for learning;

Recognition Tasks

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

AwA/ImageNet	No. Testing Classes			No. Testing Words	
	Auxiliary	Target	Total	Vocabulary	Chance(%)
SUPERVISED	✓		40/1000	40/1000	2.5/0.1
ZERO-SHOT		✓	10/360	10/360	10/0.28
OPEN-SET	✓	✓	50/1360	310K/310K	$\boxed{\text{3.2E-04}}$

The tasks are only separated in **evaluation**;
We train **one unified SS-Voc model** for all the settings

Baselines

SUPERVISED LEARNING	SVM: Input Features —> Semantic Labels SVR-Map:Input Features —> Semantic Word Vectors

One-shot Learning: Bart *et al*. CVPR 2005; Fei-Fei *et al*. TPAMI 2006; Mensink *et al.* ECCV 2012; Fu *et al*. TPAMI 2013;

Baselines

SUPERVISED LEARNING	SVM: Input Features —> Semantic Labels SVR-Map:Input Features —> Semantic Word Vectors
ZERO-SHOT LEARNING	DAP/IAP, DeViSE, ConSE, AMP, PST, HEX, TMV-BLP

DAP/IAP(Lampert *et al.* TPAMI 2013); DeViSE(Frome *et al.* NIPS 2013); ConSE(Norouzi *et al.* ICLR 2014); AMP(Fu et al. CVPR 2015), PST(Rohrbach et al. NIPS 2013).

Baselines

SUPERVISED LEARNING	SVM: Input Features —> Semantic Labels SVR-Map:Input Features —> Semantic Word Vectors
ZERO-SHOT LEARNING	DAP/IAP, DeViSE, ConSE, AMP, PST, HEX, TMV-BLP
OPEN-SET IMAGE RECOGNITION	SVR-Map

Schemer *et al*. TPAMI 2013,TPAMI 2014; Sattar et al. CVPR 2015; Bendale et al. CVPR 2015;

Variants of Our Model

SS-Voc(W,V):

$\sum \mathcal{L}_\epsilon(f($  $, W), \mathbf{u}_{\text{airplane}} V) + \sum [C - \frac{1}{2}\mathcal{L}_\epsilon(f($  $), \mathbf{u}_{\text{airplane}} V) + \frac{1}{2}\mathcal{L}_\epsilon(f($  $), \mathbf{u}_{\text{helicopter}} V)]$

Word-vector space global transformation

Regression Term

Max-margin Term

SS-Voc(W):

$\sum \mathcal{L}_\epsilon(f($  $, W), \mathbf{u}_{\text{airplane}}) + \sum [C - \frac{1}{2}\mathcal{L}_\epsilon(f($  $), \mathbf{u}_{\text{airplane}}) + \frac{1}{2}\mathcal{L}_\epsilon(f($  $), \mathbf{u}_{\text{helicopter}})]$

Regression Term

Max-margin Term

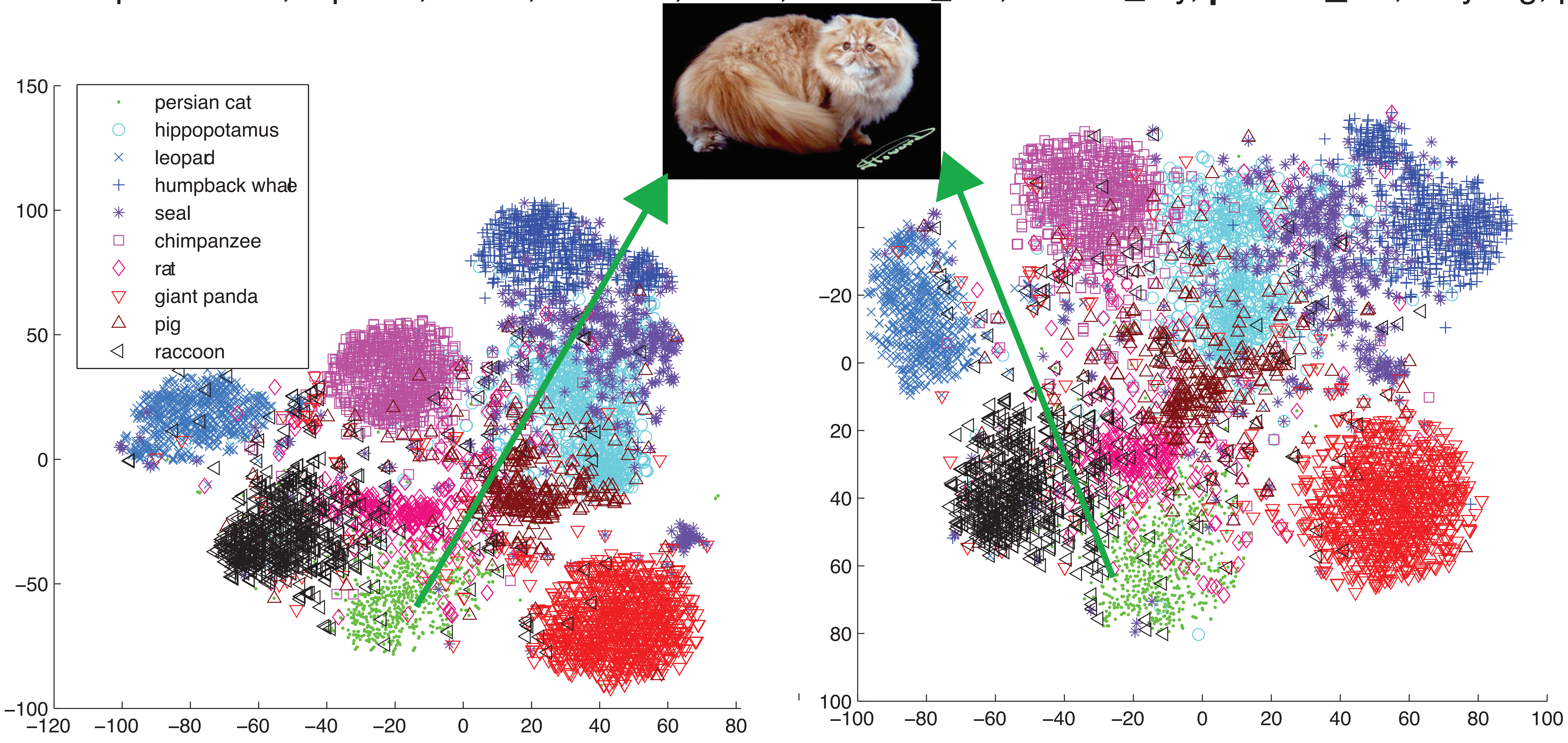
SVR-Map: $\sum \mathcal{L}_\epsilon(f($  $W)\ \mathbf{u}_{\text{airplane}})$

Regression Term

t-SNE Visualization of AwA 10 Testing Classes

SS-Voc: **persian_cat,** siamese_cat, hamster, weasel, rabbit, monkey, zebra, owl, anthropomorphized, cat

SVR-Map: hamster, squirrel, rabbit, raccoon, kitten, siamese_cat, stuffed_toy, **persian_cat**, ladybug, puppy



Semi-Supervised Vocabulary-informed Learning (SS-Voc)

Support Vector Regression (SVR)

Supervised Results

AwA dataset

Method	Accuracy
SS-Voc (W,V)	59.1
SS-Voc (W)	58.6
SVM	52.1
SVR-Map	57.1

ImageNet dataset

Method	Accuracy
SS-Voc (W,V)	37.1
SS-Voc (W)	36.3
SVM	33.8
SVR-Map	25.6

Zero-shot Results—AwA dataset

3.3%

0.82%

Method	Features	Accuracy
SS-Voc: full instances	$CNN_{OverFeat}$	78.3
800 instances (20 inst*40 class);	$CNN_{OverFeat}$	74.4
200 instances (5 inst*40 class);	$CNN_{OverFeat}$	68.9
Akata *et al*. CVPR 2015	$CNN_{GoogLeNet}$	73.9
TMV-BLP (Fu *et al*. ECCV 2014)	$CNN_{OverFeat}$	69.9
AMP (SR+SE) (Fu *et al*. CVPR 2015)	$CNN_{OverFeat}$	66.0
DAP (Lampert *et al*. TPAMI 2013)	CNN_{VGG19}	57.5
PST (Rohrbach *et al*. NIPS 2013)	$CNN_{OverFeat}$	53.2
DS (Rohrbach *et al*. CVPR 2010)	$CNN_{OverFeat}$	52.7
IAP (Lampert *et al.* TPAMI 2013)	$CNN_{OverFeat}$	44.5
HEX (Deng *et al*. ECCV 2014)	CNN_{DECAF}	44.2

Zero-shot Results—ImageNet

Method	Features	T-1 Accuracy (full instances)	T-5 Accuracy (full instances)	T-1 Accuracy (3000 instances)	T-5 Accuracy (3000 instances)
SS-Voc	CNN_{VGG-19}	9.5	16.8	8.9	14.9
ConSE	CNN_{VGG-19}	7.8	15.5	5.5	13.1
DeViSE	CNN_{VGG-19}	5.2	12.8	3.7	11.8
AMP	CNN_{VGG-19}	6.1	13.1	3.5	10.5

Open-Set Image Recognition — AwA dataset

AwA	Testing Classes			Vocabulary
	Auxiliary	Target	Total	
$\textsc{Open-Set}_{1K-NN}$	(LEFT)	(RIGHT)	40/10	1000*
$\textsc{Open-Set}_{1K-RND}$			40/10	1000†
$\textsc{Open-Set}_{310K}$			40/10	310K

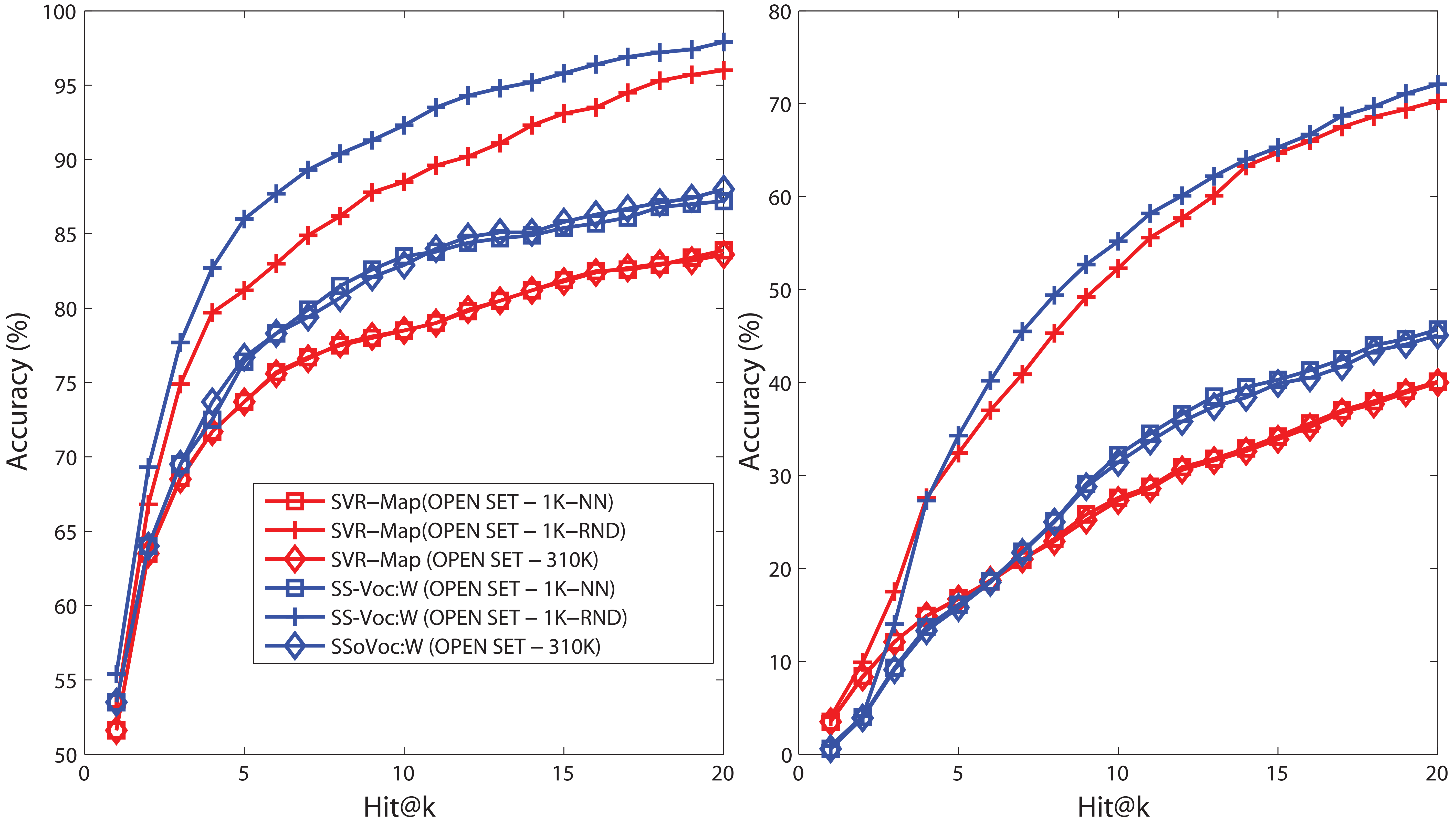
\textsc{\small Open-Set}$_{1K-NN}$

$\textsc{Open-Set}_{1K-NN}$	1000 candidate labels (of 310K labels) sampled from nearest neighbor set of ground-truth class prototypes
$\textsc{Open-Set}_{1K-RND}$	1000 candidate labels randomly sampled from 310K vocabulary set.
$\textsc{Open-Set}_{310K}$	the large vocabulary of approximately 310K entities

Open-Set Image Recognition — AwA dataset

AwA	Testing Classes			Vocabulary
	Auxiliary	Target	Total	
OPEN-SET$_{1K-NN}$	(LEFT)	(RIGHT)	40/10	1000*
OPEN-SET$_{1K-RND}$			40/10	1000†
OPEN-SET$_{310K}$			40/10	310K

\textsc{\small Open-Set}$_{1K-NN}$



Disney Research

Take-home

1. A new learning paradigm — vocabulary-informed learning
2. A unified semantic embedding framework for supervised, zero-shot and open-set image recognition

Thanks! Q&A