From Large Scale Image Categorization to Entry-Level Categories

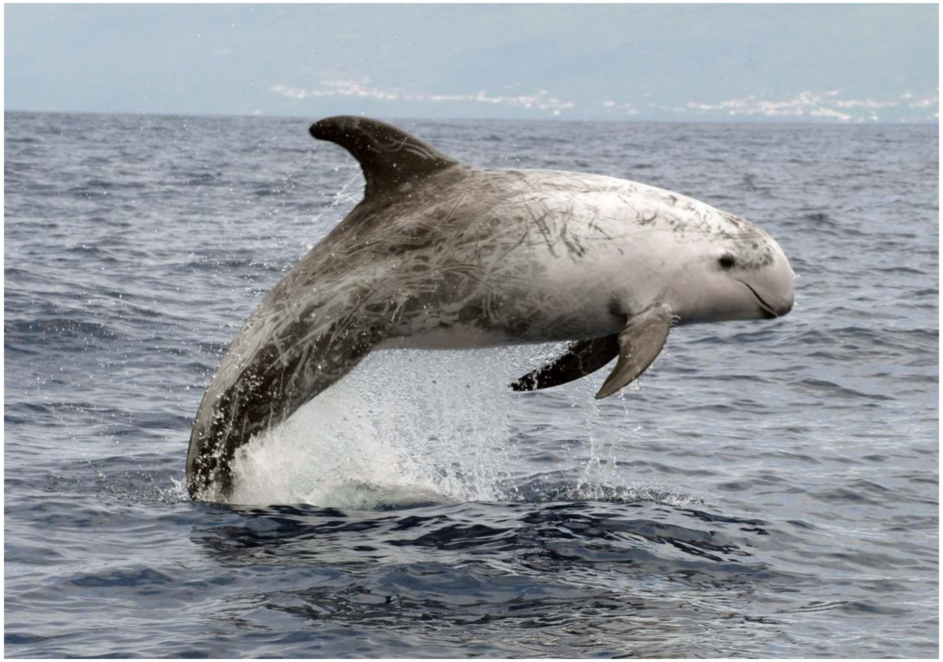
Vicente Ordonez, Jia Deng, Yejin Choi,
Alexander C. Berg, Tamara L. Berg

Stony Brook University

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

Stanford University

What would you call this?



Grampus griseus

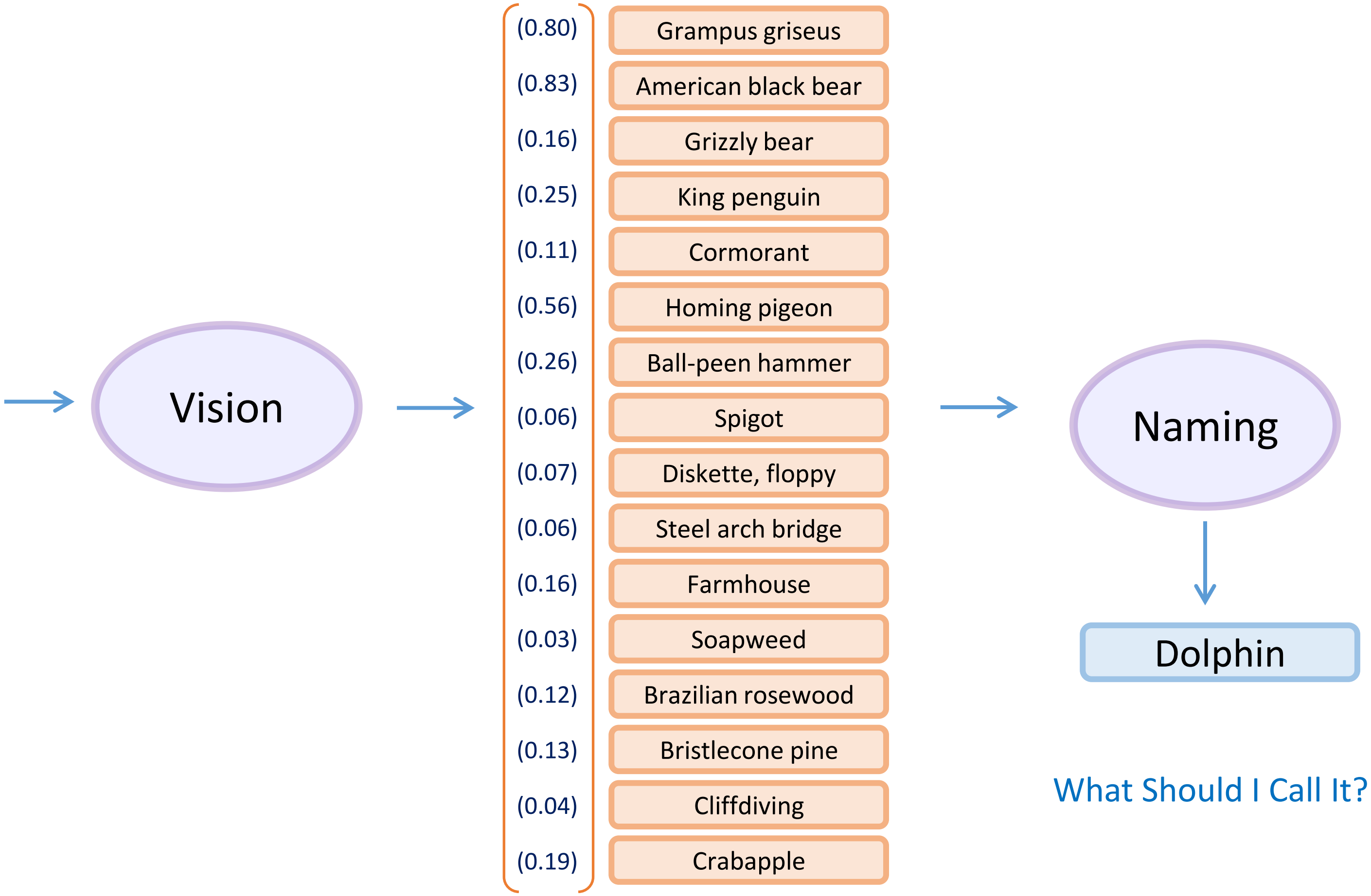
[✓ Dolphin]

What would you call this?

Object
Organism
Animal
Chordate
Vertebrate
Bird
Aquatic bird
[✓ Swan]
Whistling swan
Cygnus Colombianus

Naming Image Content

Input Image



Vision

(0.80)	Grampus griseus
(0.83)	American black bear
(0.16)	Grizzly bear
(0.25)	King penguin
(0.11)	Cormorant
(0.56)	Homing pigeon
(0.26)	Ball-peen hammer
(0.06)	Spigot
(0.07)	Diskette, floppy
(0.06)	Steel arch bridge
(0.16)	Farmhouse
(0.03)	Soapweed
(0.12)	Brazilian rosewood
(0.13)	Bristlecone pine
(0.04)	Cliffdiving
(0.19)	Crabapple

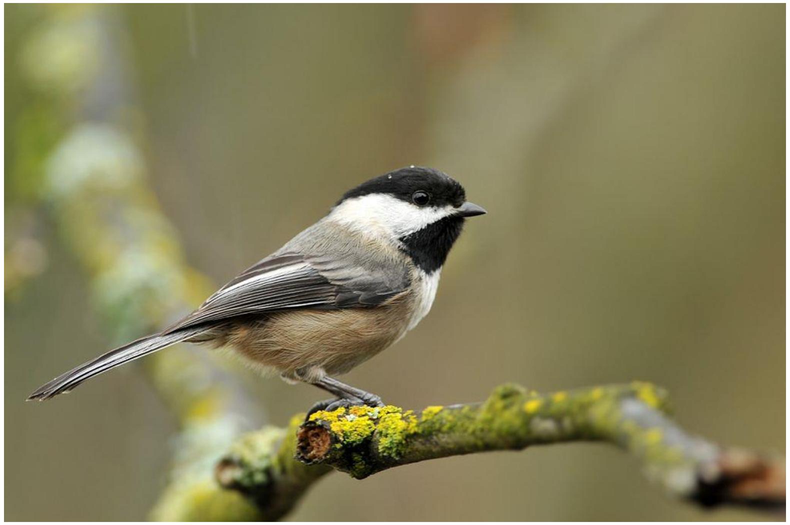
Thousands of Noisy Category Predictions

Naming

Dolphin

What Should I Call It?

Entry-Level Category

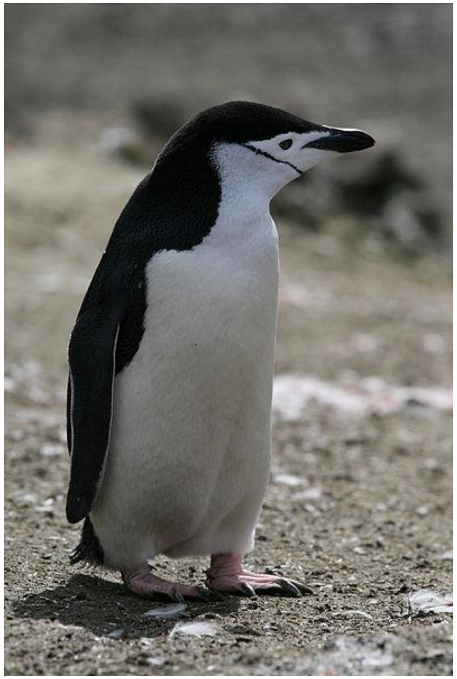


The category that people are likely to name when presented with a depiction of an object.

Rosch et al, 1976
Jolicoeur, Gluck & Kosslyn, 1984

Superordinates: animal, vertebrate
Entry Level: bird
Subordinates: Black-capped chickadee

Entry-Level Category



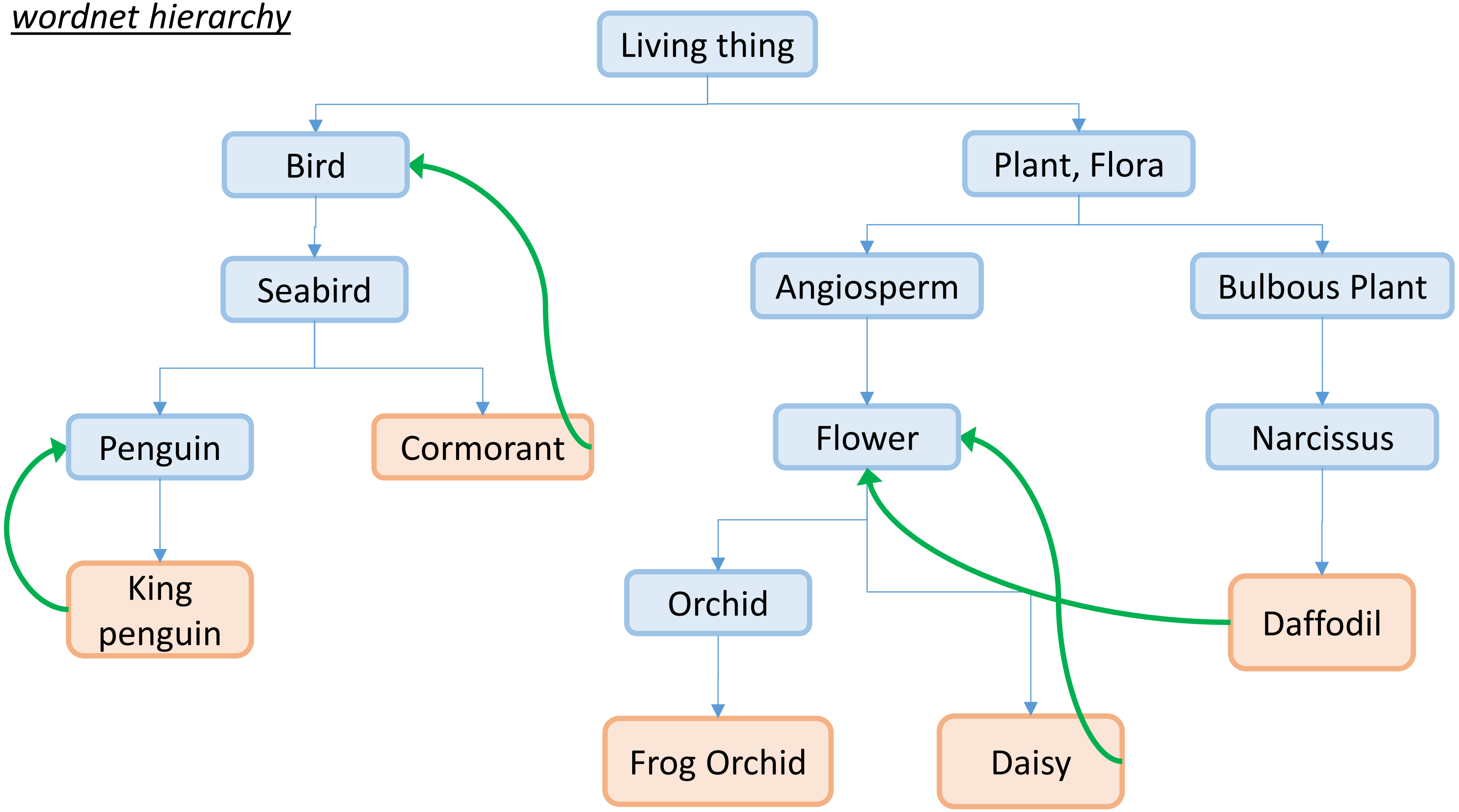
The category that people are likely to name when presented with a depiction of an object.

Rosch et al, 1976
Jolicoeur, Gluck & Kosslyn, 1984

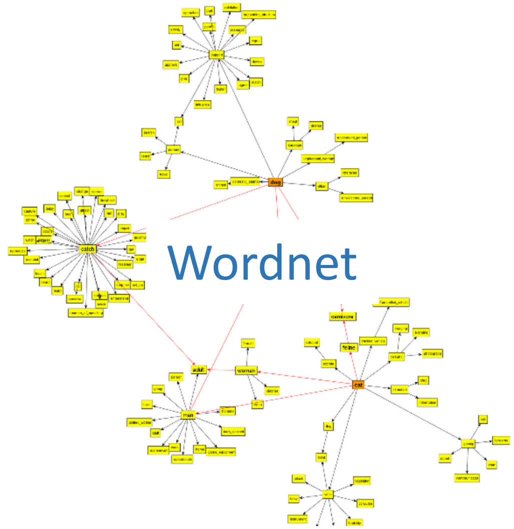
Superordinates: animal, bird
Entry Level: penguin
Subordinates: Chinstrap penguin

Is this hard?

wordnet hierarchy



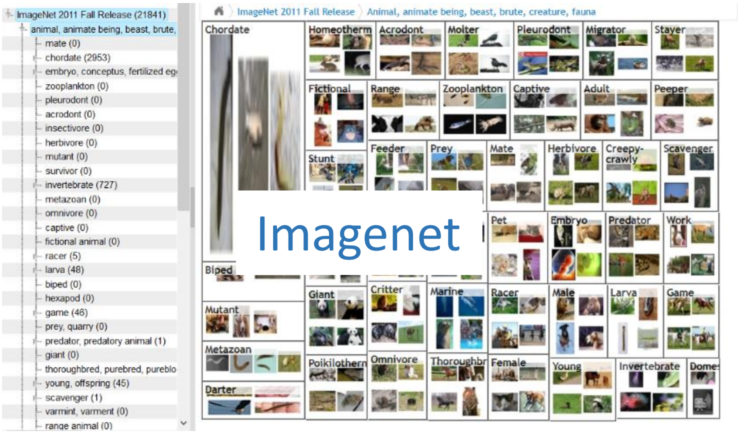
How will we do it?



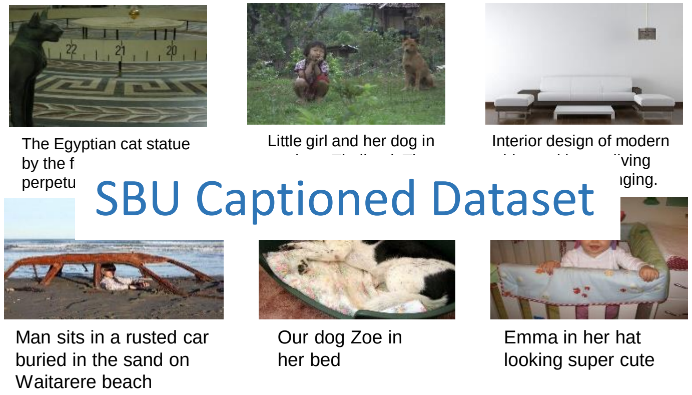
Linguistic resources

Computer Vision

Lots of text



Labeled Images



The Egyptian cat statue by the f perpetu

Little girl and her dog in

Interior design of modern ving ging.

Man sits in a rusted car buried in the sand on Waitarere beach

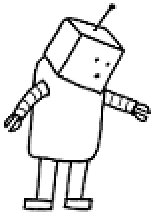
Our dog Zoe in her bed

Emma in her hat looking super cute

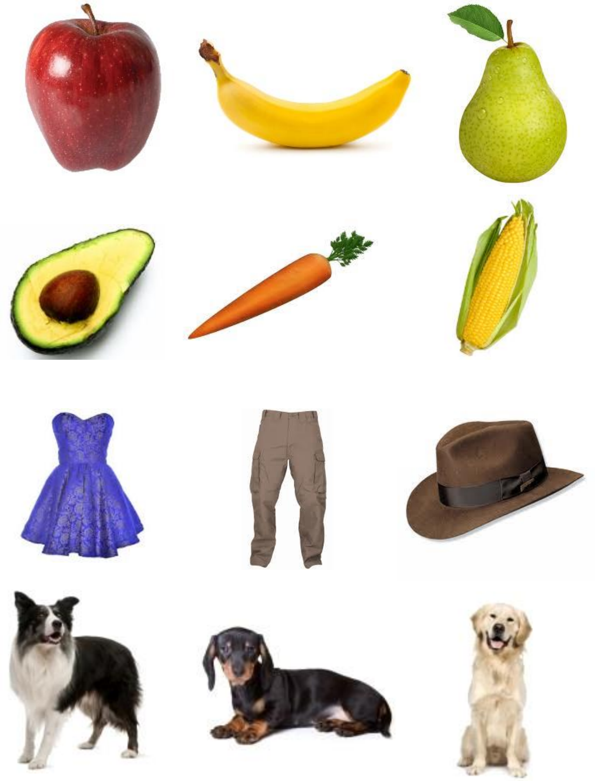
Lots of images with text

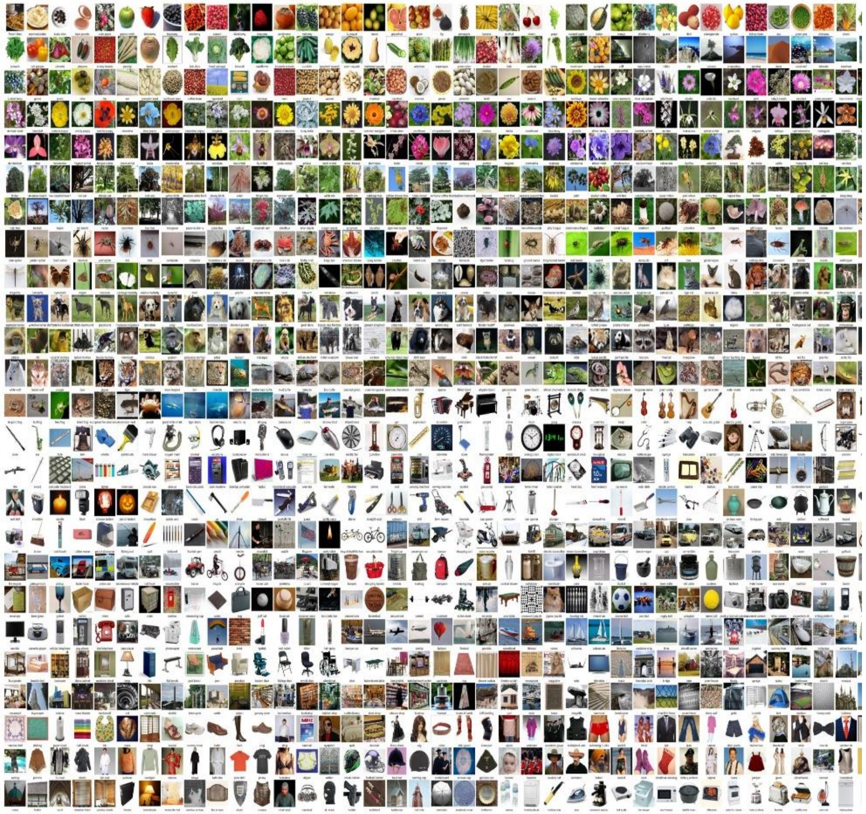
Scaling Naming Tasks!

48 categories



> 7000 categories

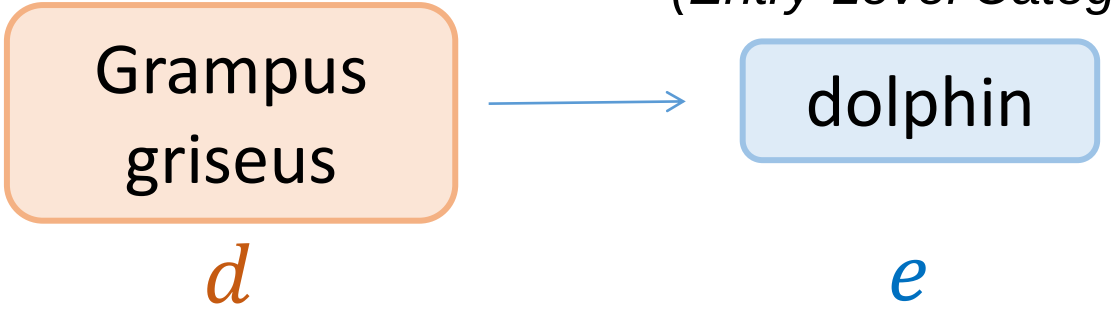




1. Goal: Category Translation

Detailed Category

What should I Call It?
(Entry-Level Category)



Grampus griseus

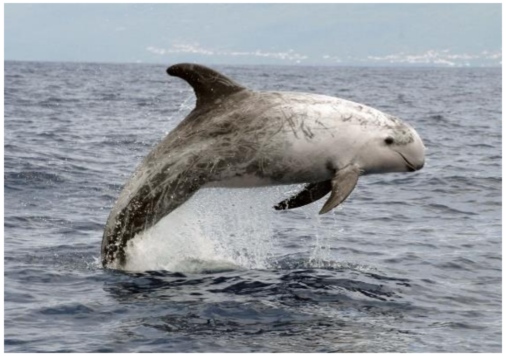
dolphin

d

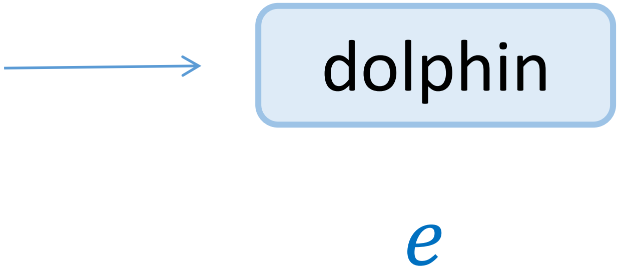
e

2. Goal: Content Naming

Input Image



What should I Call It?
(Entry-Level Category)



dolphin

e

1. Goal: Category Translation

Detailed Category

Grampus griseus → dolphin

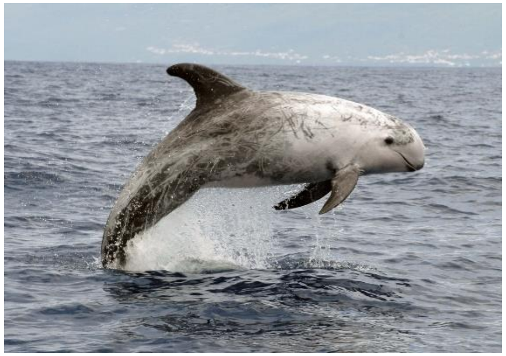
d

What should I Call It?
(Entry-Level Category)

e

2. Goal: Content Naming

Input Image



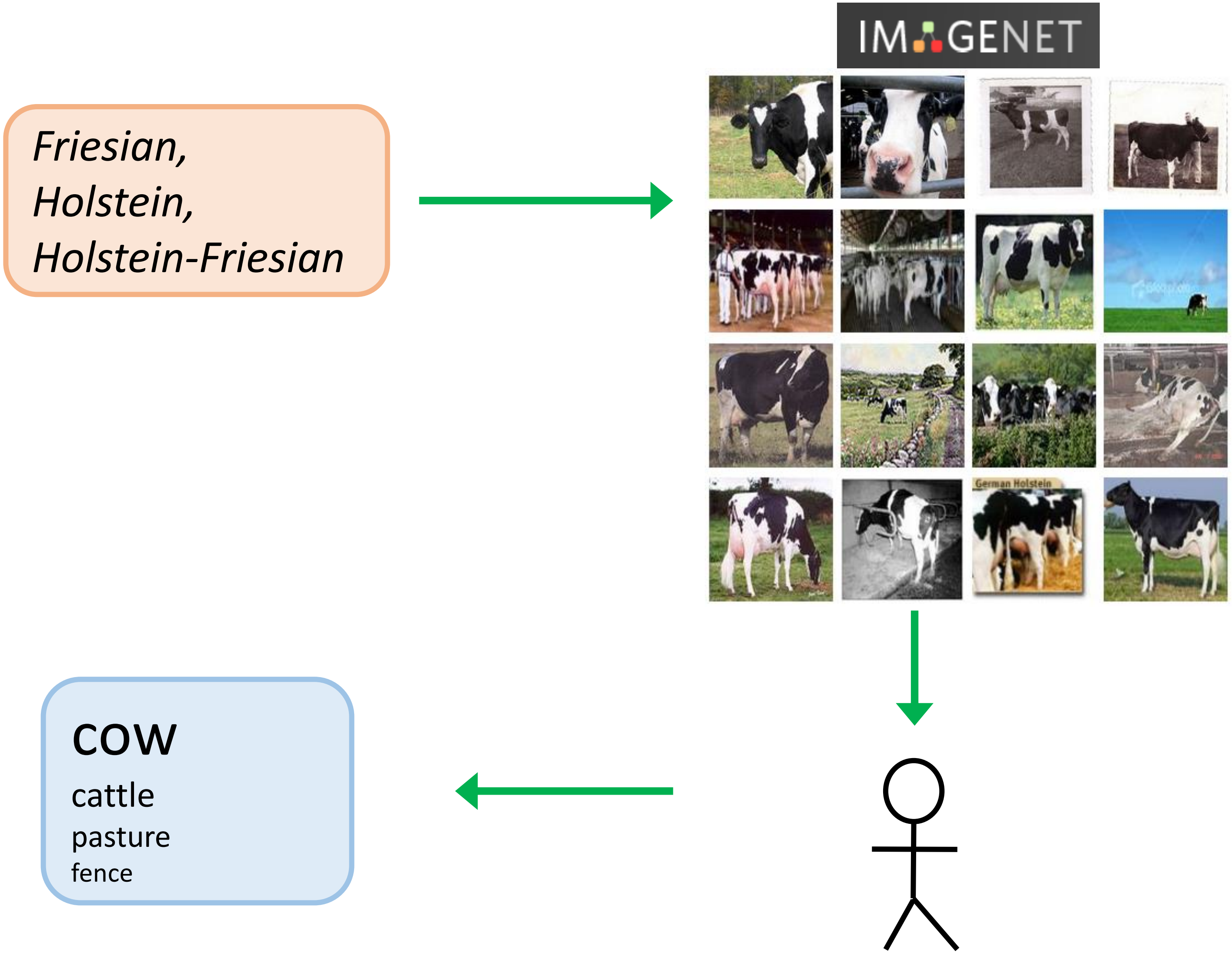
→

What should I Call It?
(Entry-Level Category)

dolphin

e

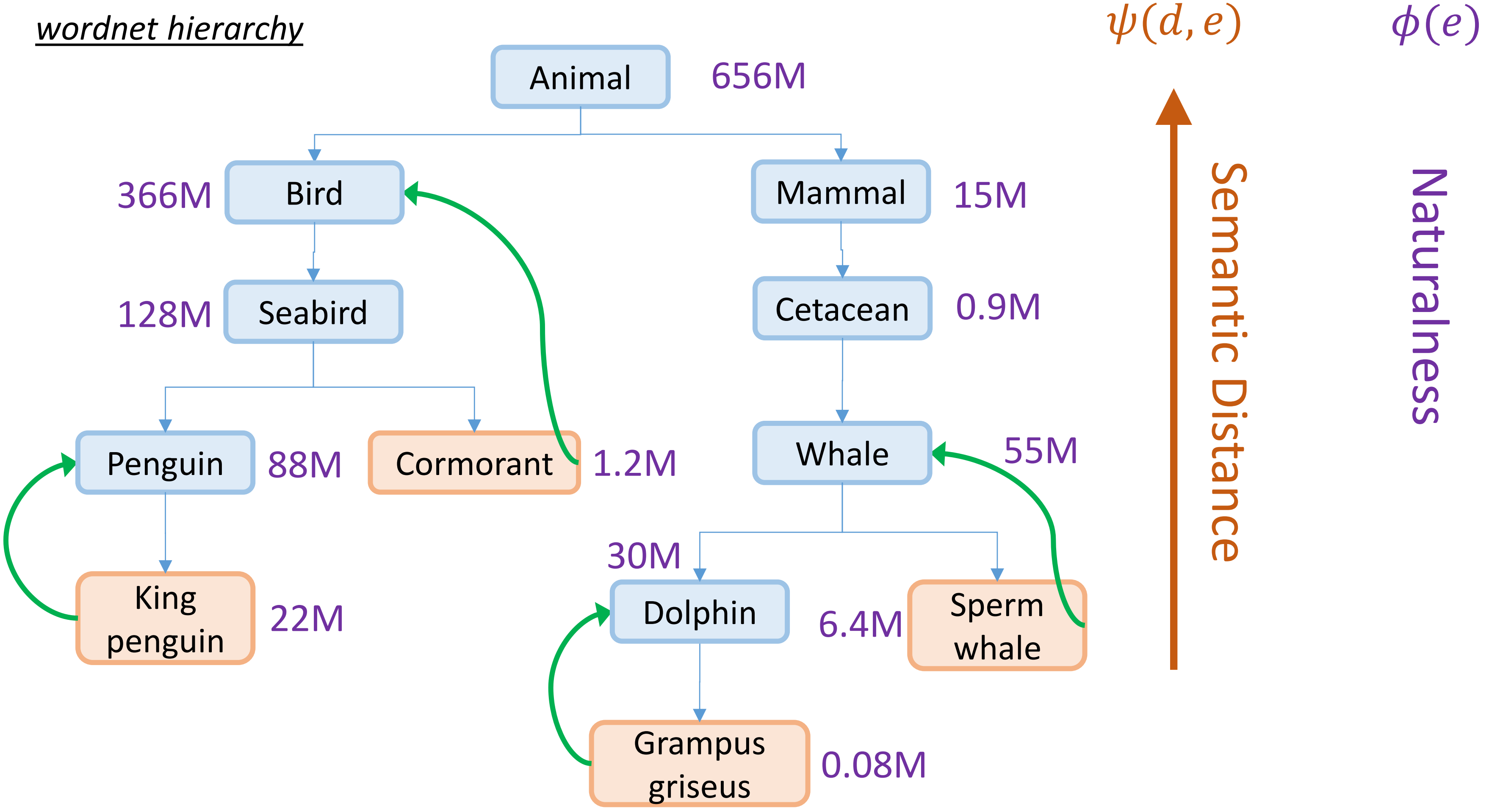
Category Translation by Humans



Friesian,
Holstein,
Holstein-Friesian

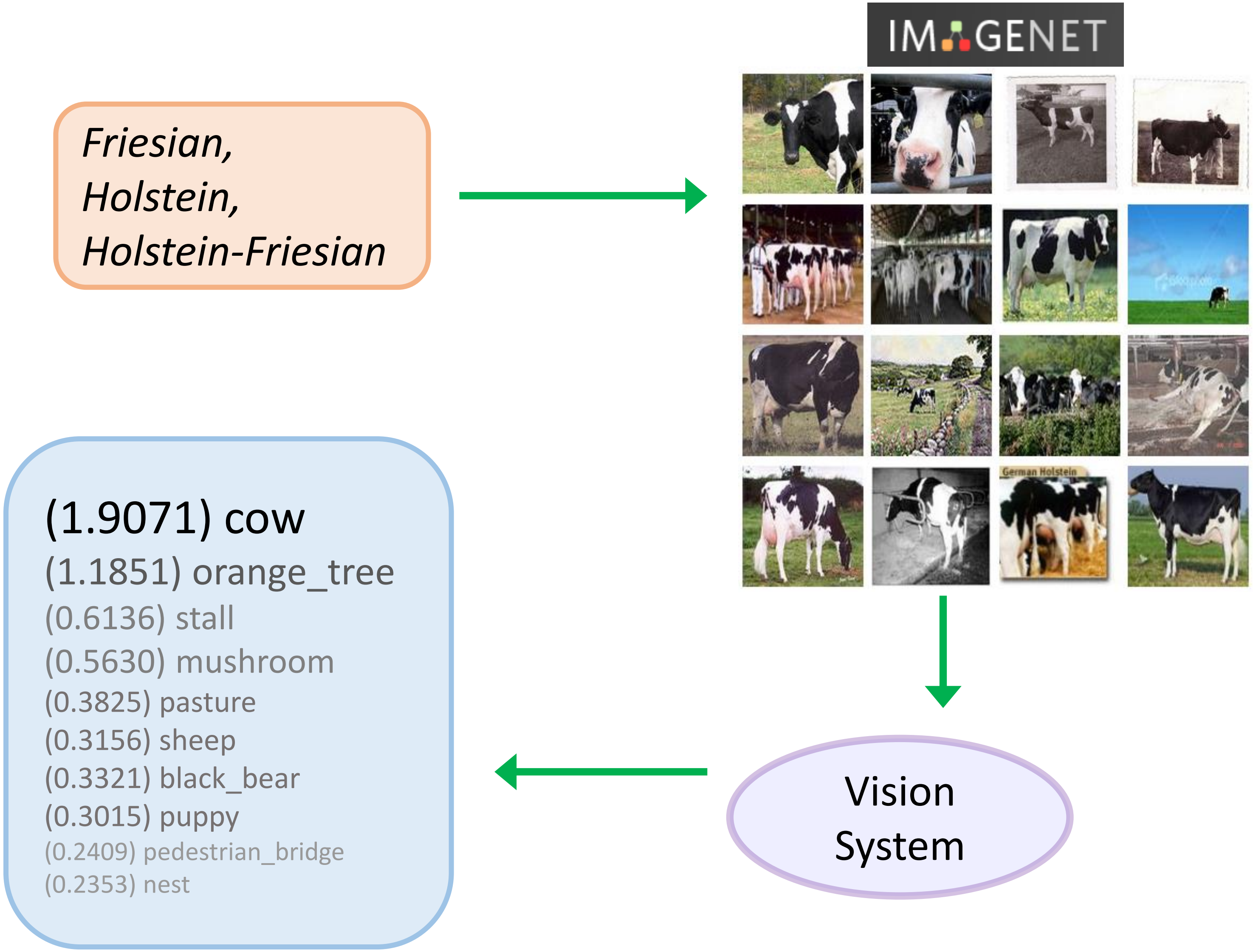
cow
cattle
pasture
fence

1.1 Category Translation: Text-based



$$\tau(d,\lambda) = \underset{w}{\operatorname{argmax}}[\phi(e) - \lambda\psi(d,e)]$$

1.2 Category Translation: Image-based



Friesian,
Holstein,
Holstein-Friesian

Vision System

(1.9071) cow
(1.1851) orange_tree
(0.6136) stall
(0.5630) mushroom
(0.3825) pasture
(0.3156) sheep
(0.3321) black_bear
(0.3015) puppy
(0.2409) pedestrian_bridge
(0.2353) nest

Category Translation: Examples

	HUMANS	TEXT BASED	IMAGE BASED
cactus wren	bird	bird	bird
buzzard, Buteo buteo	hawk	hawk	bird
whinchat, Saxicola rubetra	bird	chat	bird
Weimaraner	dog	dog	dog
numbat, banded anteater, anteater	anteater	anteater	cat
rhea, Rhea americana	ostrich	bird	grass
Europ. black grouse, heathfowl	bird	bird	duck
yellowbelly marmot, rockchuck	Squirrel	marmot	rock

1. Goal: Category Translation

Detailed Category

Grampus griseus →

d

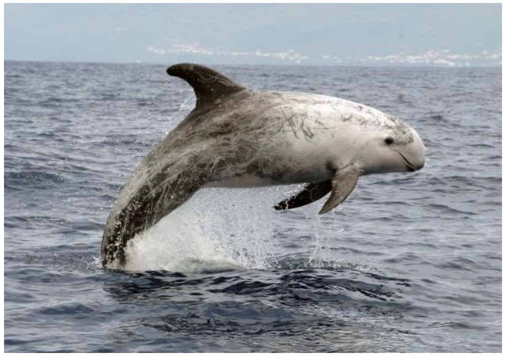
What should I Call It?
(Entry-Level Category)

dolphin

e

2. Goal: Content Naming

Input Image



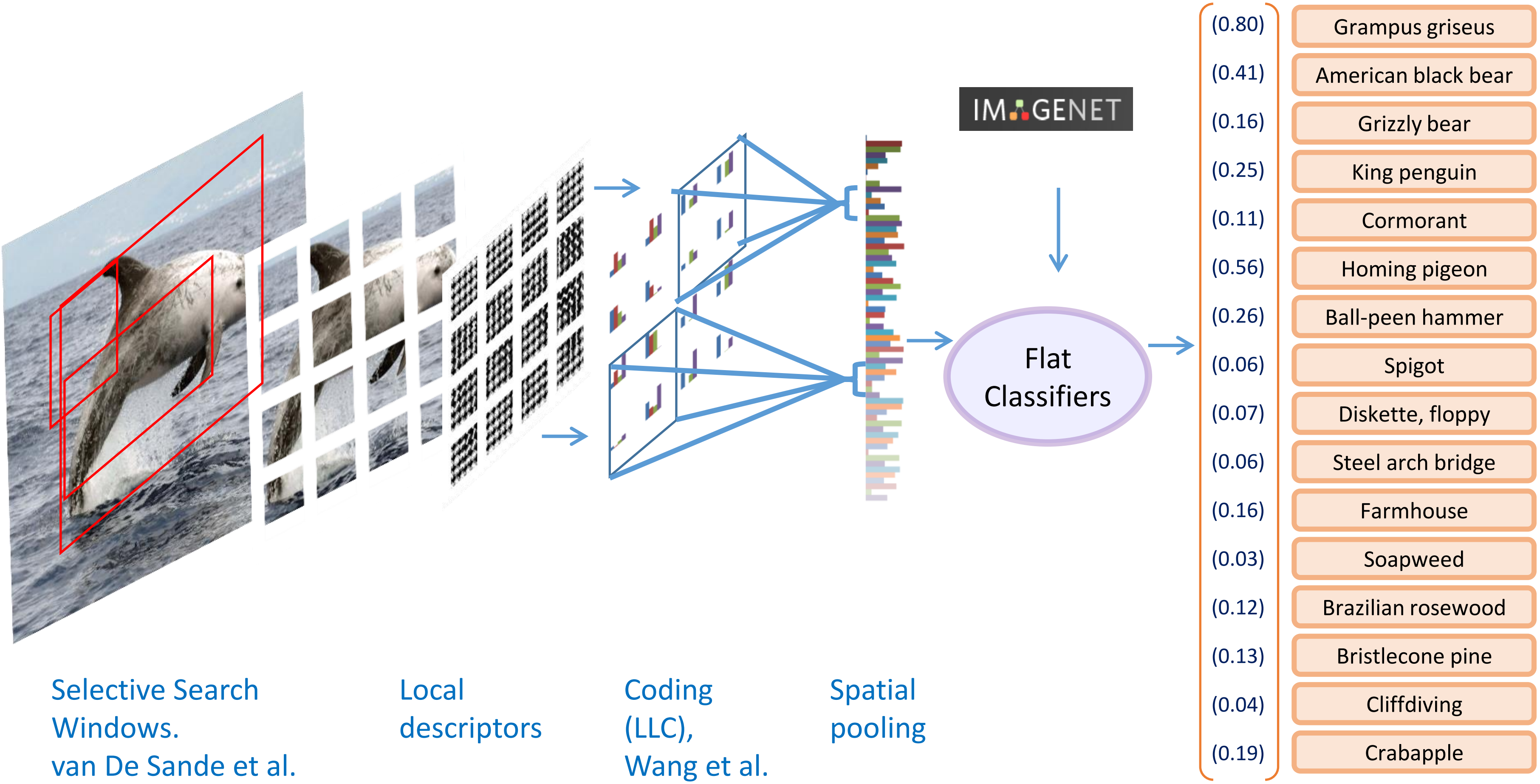
→

What should I Call It?
(Entry-Level Category)

dolphin

e

Large Scale Categorization



(0.80)	Grampus griseus
(0.41)	American black bear
(0.16)	Grizzly bear
(0.25)	King penguin
(0.11)	Cormorant
(0.56)	Homing pigeon
(0.26)	Ball-peen hammer
(0.06)	Spigot
(0.07)	Diskette, floppy
(0.06)	Steel arch bridge
(0.16)	Farmhouse
(0.03)	Soapweed
(0.12)	Brazilian rosewood
(0.13)	Bristlecone pine
(0.04)	Cliffdiving
(0.19)	Crabapple

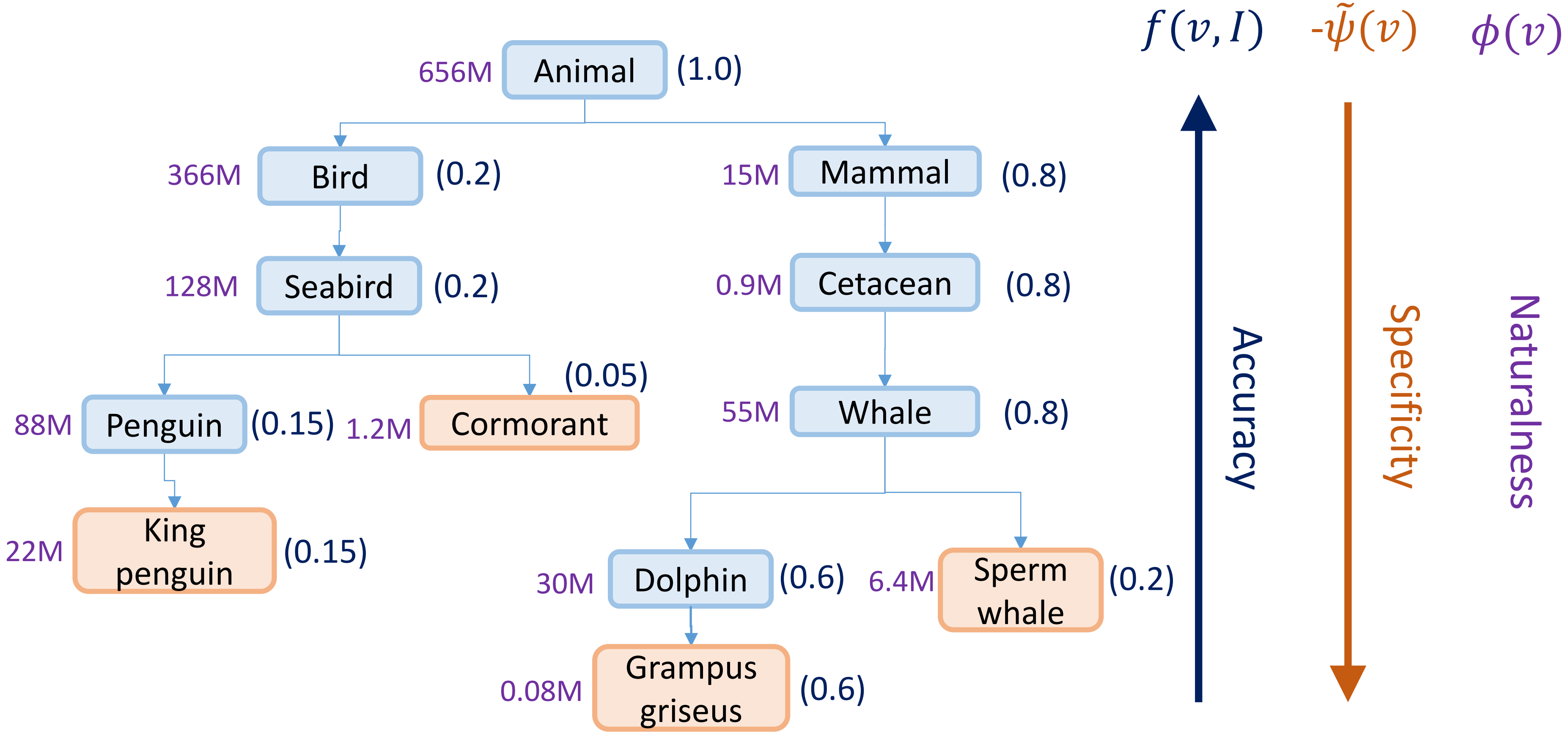
Selective Search Windows.
van De Sande et al.
ICCV 2011

Local descriptors

Coding (LLC),
Wang et al.
CVPR 2010

Spatial pooling

2.1 Propagated Visual Estimates



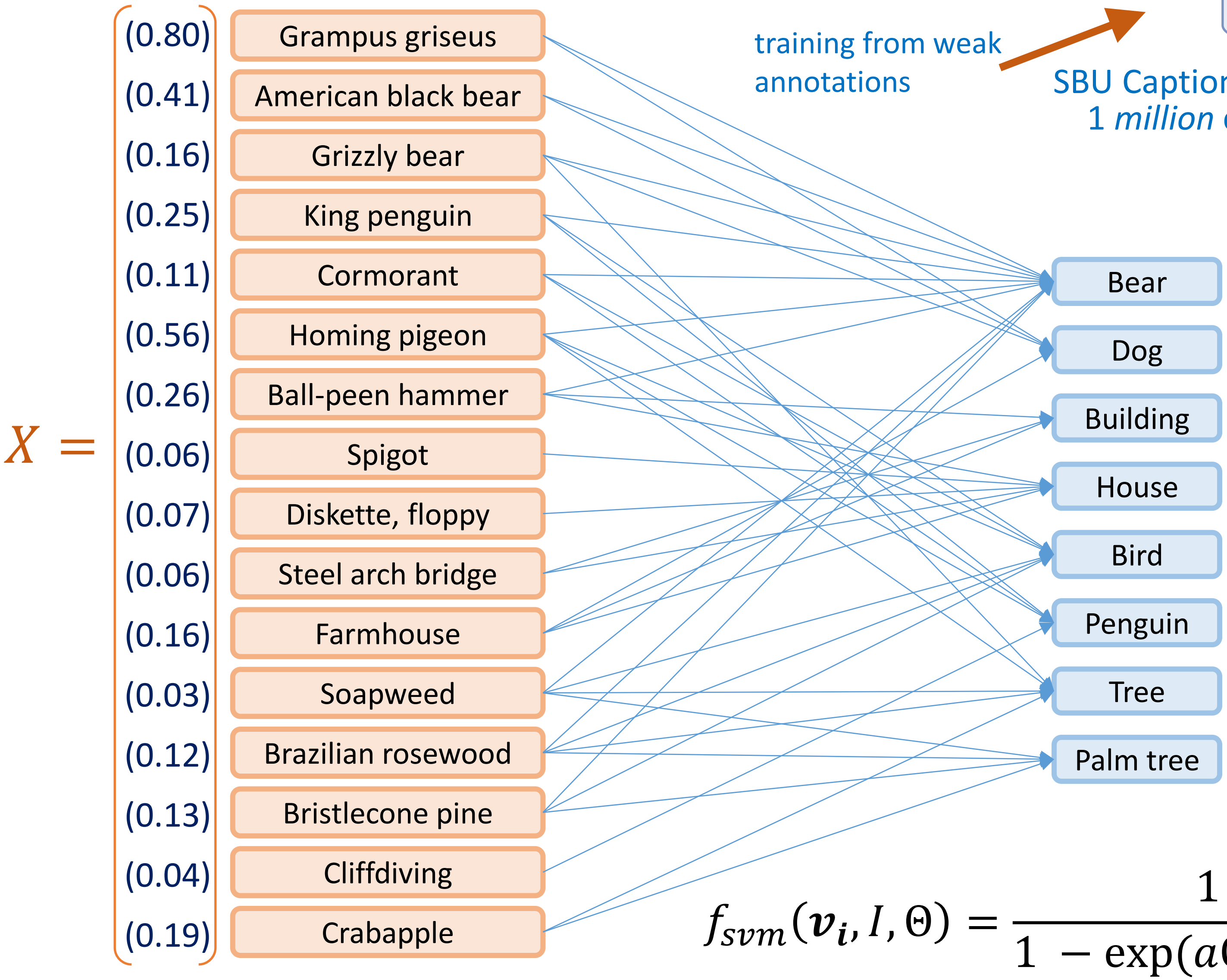
Our work

$$f_{nat}(v,I,\tilde{\lambda}) = f(v,I)\ [\phi(v) - \tilde{\lambda}\tilde{\psi}(v)]$$

2.2 Supervised Learning

training from weak annotations

SBU Captioned Photo Dataset
1 *million* captioned images!



$X =$

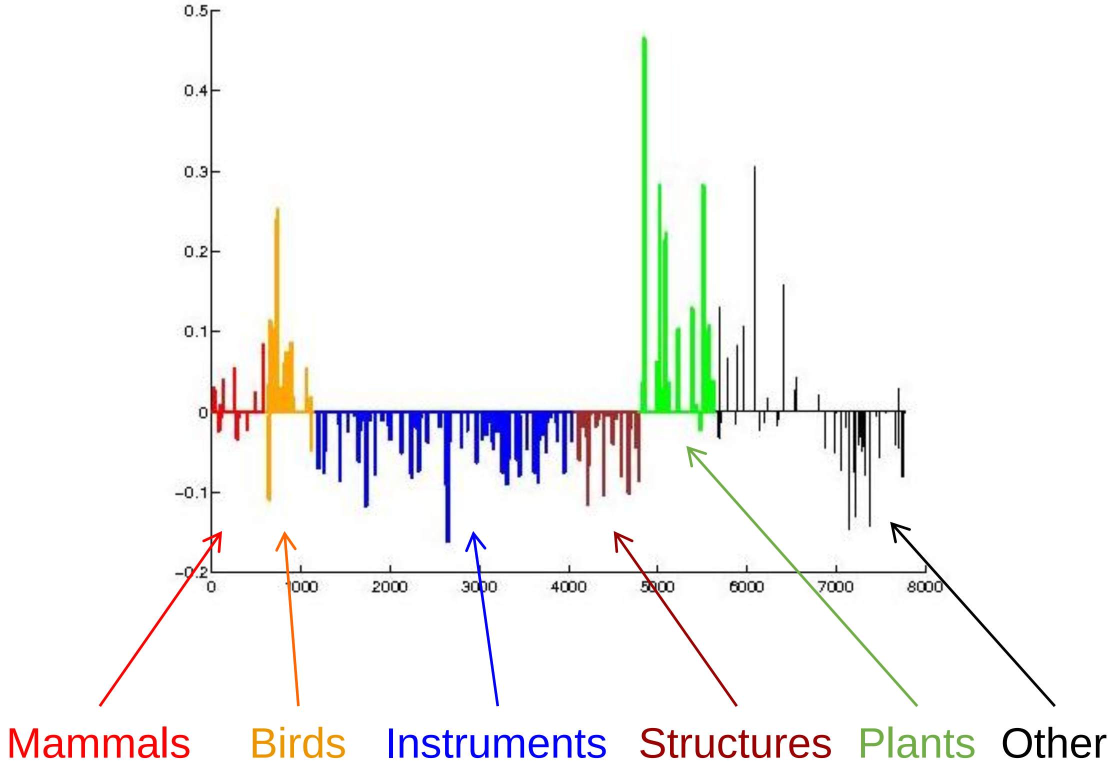
(0.80) Grampus griseus
(0.41) American black bear
(0.16) Grizzly bear
(0.25) King penguin
(0.11) Cormorant
(0.56) Homing pigeon
(0.26) Ball-peen hammer
(0.06) Spigot
(0.07) Diskette, floppy
(0.06) Steel arch bridge
(0.16) Farmhouse
(0.03) Soapweed
(0.12) Brazilian rosewood
(0.13) Bristlecone pine
(0.04) Cliffdiving
(0.19) Crabapple

Bear
Dog
Building
House
Bird
Penguin
Tree
Palm tree

$$f_{svm}(\boldsymbol{v_i}, I, \Theta) = \frac{1}{1 \; - \exp(a\Theta^T X + b)}$$

Extracting Meaning from Data

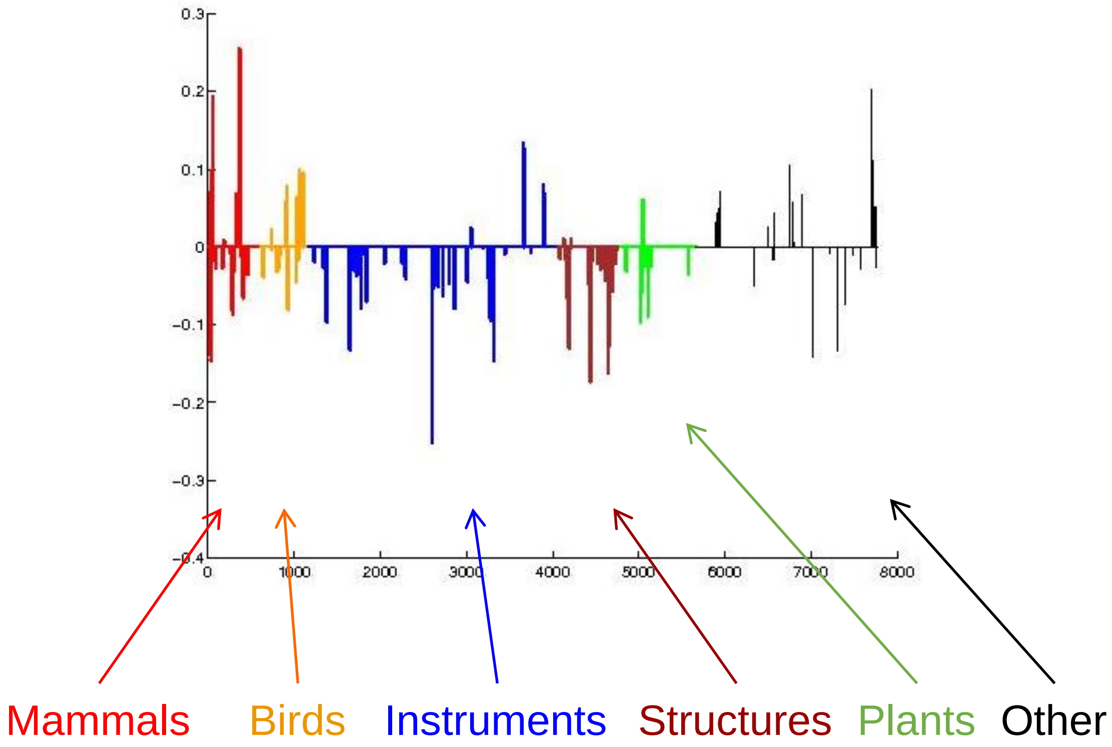
Weights learned to recognize images with **“tree”** in caption



snag
shade tree
bracket fungus, shelf fungus
bristlecone pine, Rocky Mountain bristlecone pine, Pinus aristata
Brazilian rosewood, caviuna wood, jacaranda, Dalbergia nigra
redheaded woodpecker, redhead, Melanerpes erythrocephalus
redbud, Cercis canadensis
mangrove, Rhizophora mangle
chiton, coat-of-mail shell, sea cradle, polyplacophore
crab apple, crabapple
papaya, papaia, pawpaw, papaya tree, melon tree, Carica papaya
frogmouth

Extracting Meaning from Data

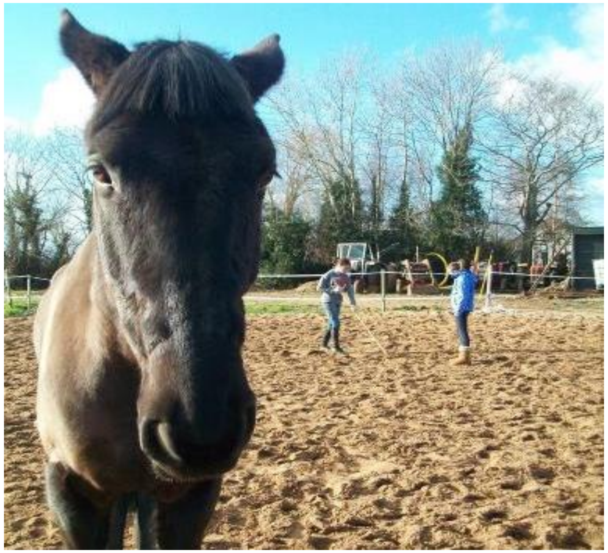
Weights learned to recognize images with **“water”** in caption



Mammals Birds Instruments Structures Plants Other

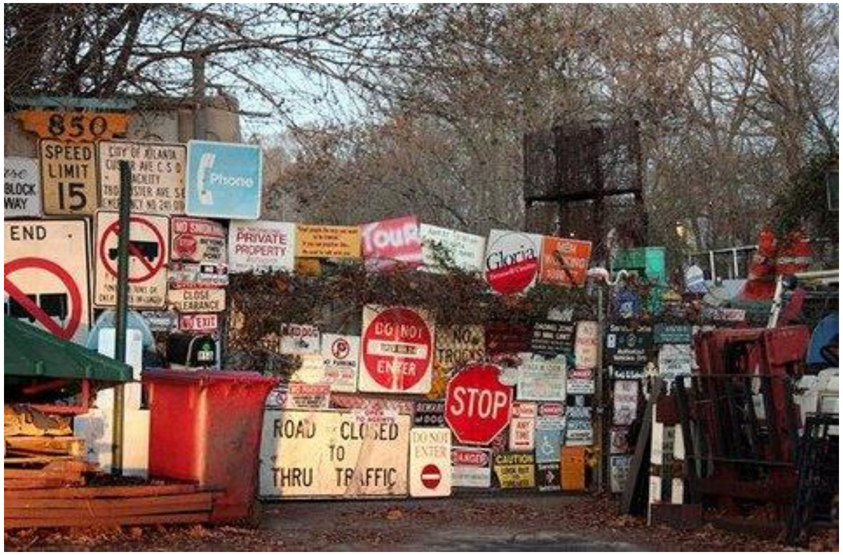
water dog
surfing, surfboarding, surfriding
manatee, Trichechus manatus
punt
dip, plunge
cliff diving
fly-fishing
sockeye, sockeye salmon, red salmon, blueback salmon, Oncorhynchus nerka
sea otter, Enhydra lutris
American coot, marsh hen, mud hen, water hen, Fulica americana
booby
canal boat, narrow boat, narrowboat

Results: Content Naming



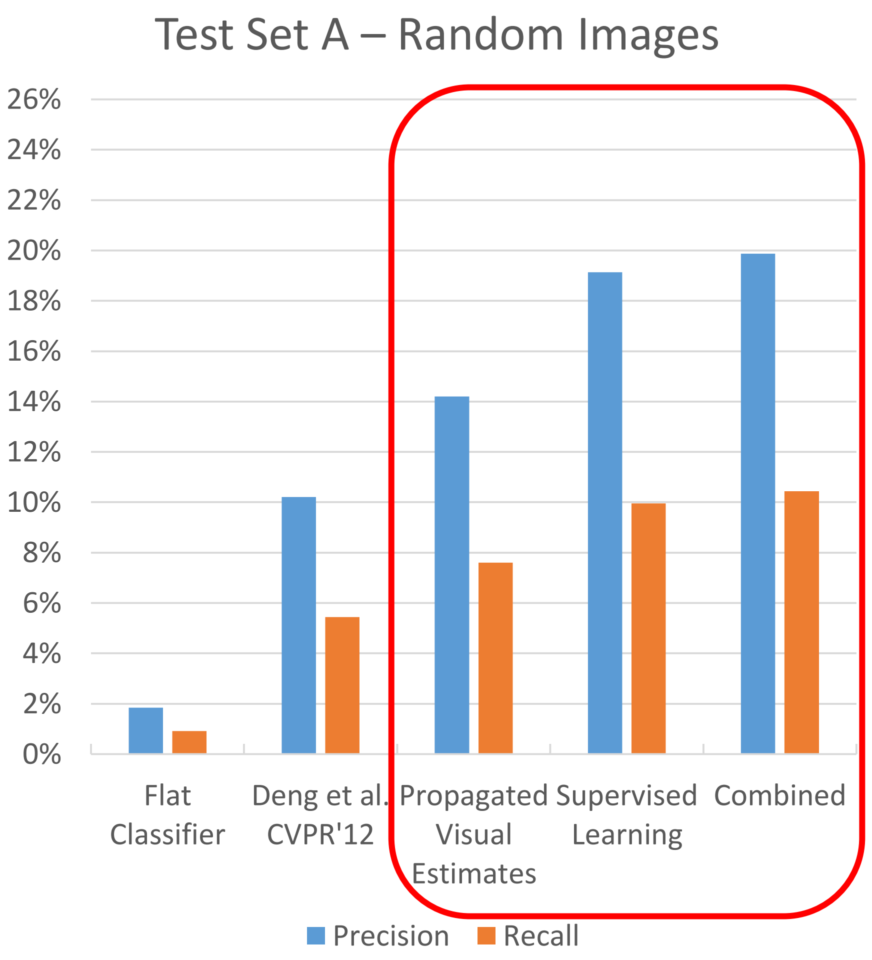
Human Labels	Flat Classifier	Deng et al. CVPR’12	Propagated Visual Estimates	Supervised Learning	Joint
farm, fence	gelding	**horse**	**horse**	**horse**	**horse**
field	yearling	equine	**tree**	pasture	pasture
horse, mule	shire	perissodactyl	equine	**field**	**field**
kite, dirt	yearling	ungulate	male	cow	cow
people	draft	male	gelding	**fence**	**fence**
tree, zoo					

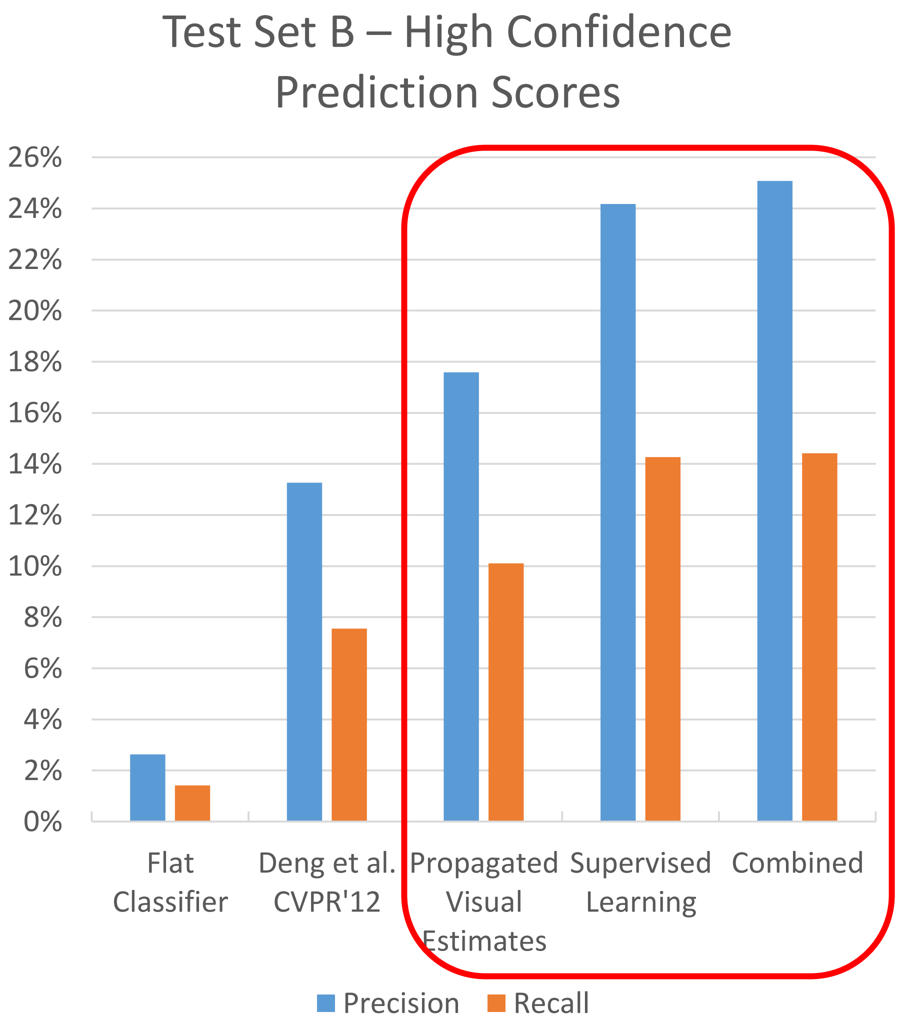
Results: Content Naming



Human Labels	Flat Classifier	Deng et al. CVPR’12	Propagated Visual Estimates	Supervised Learning	Joint
fence, junk	feeder	woody	tree	logo	logo
sign	Hyla	tree	structure	street	street
stop sign	cleaner	structure	building	neighborhood	neighborhood
street sign	box	plant	plant	building	building
trash can	large	vascular	area	office building	office
tree					

Evaluation: Content Naming





Conclusions/Future Work

- We explored different models for content naming in images.

- Results can be used to improve the larger goal of generating human-like image descriptions.

- Go beyond nouns and infer other type of abstractions on action and attribute words.

Questions?