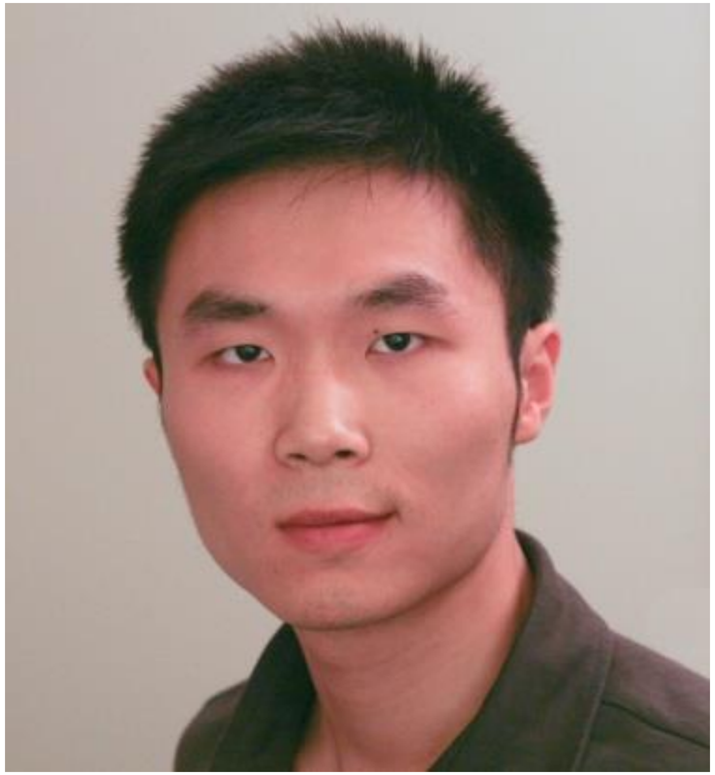
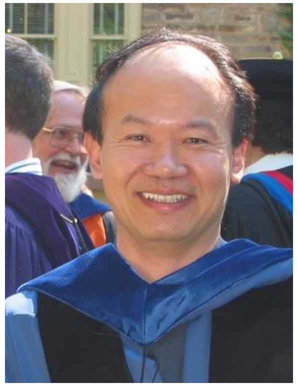
Large-Scale Visual Recognition Powered by Big Data and Big Crowd

Fei-Fei Li
Stanford University



Dr. Jia Deng
Stanford U. -> U. Michigan

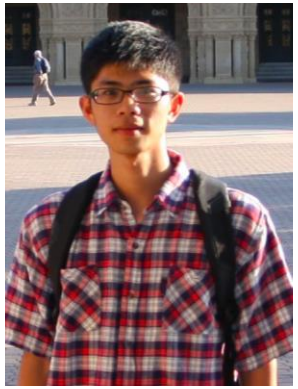


Prof. Kai Li
Princeton U.

Prof. Alex Berg
Stony Brook U.

Sanjeev Satheesh
Stanford U.

Jonathan Krause
Stanford U.



Zhiheng Huang
Stanford U.

Olga Russakovsky
Stanford U.

Build a computer to recognize EVERYTHING

Surveillance

Robotics

Assistive tools

Wearable devices

Recognition Engine

Driverless cars

Smart photo album

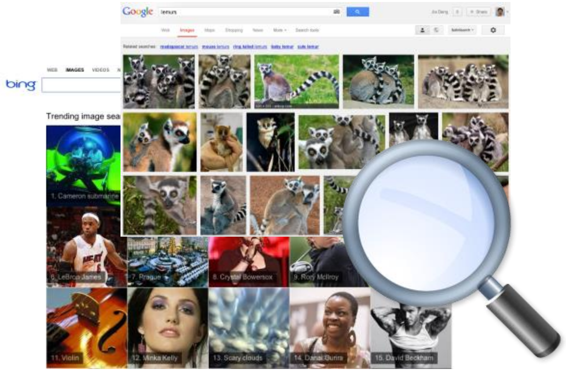


Image search

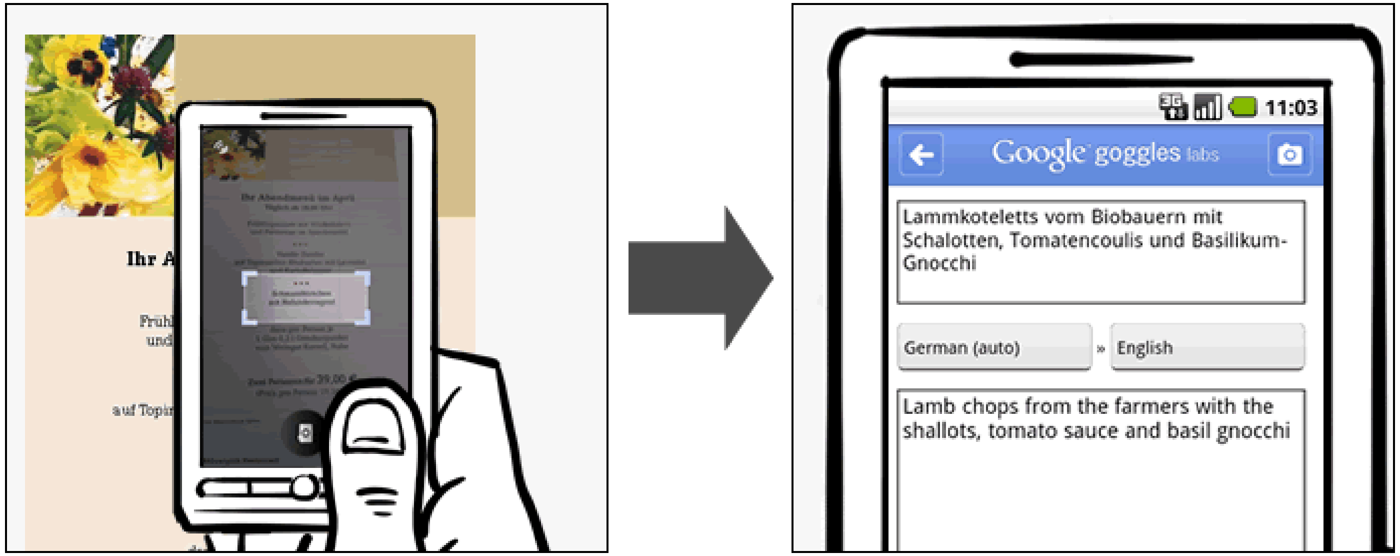
Mining social media

What can computers already recognize?


AUTO
10M
IN
39
Nikon
The Nikon S60. Detects up to 12 faces.

Google Goggles

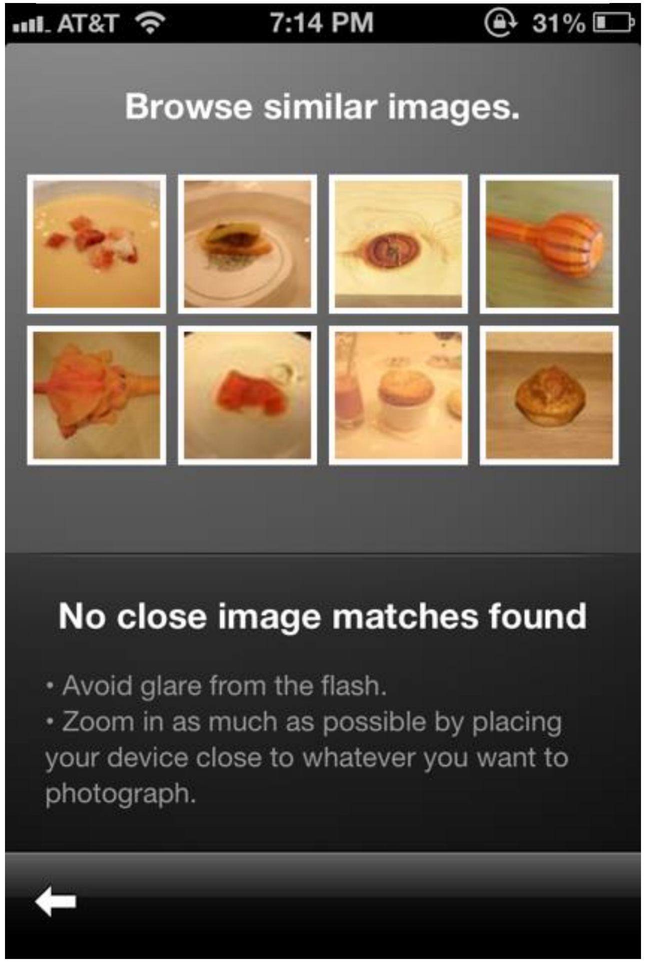
Use pictures to search the web.



But when it comes to generic objects in the world…

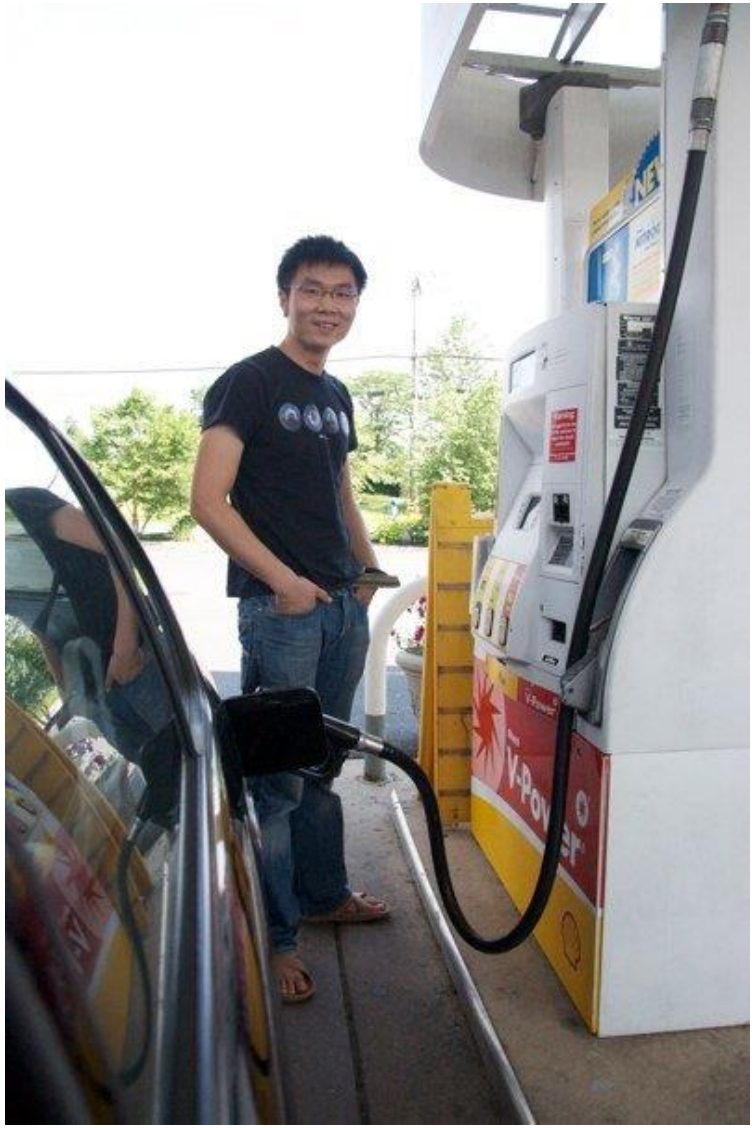
Google Goggles

Use pictures to search the web.



But when it comes to generic objects in the world…

What about Gas Pumps!



But when it comes to generic objects in the world…

20 object classes: PASCAL VOC [Everingham et al. 2006-2012]

Airplane	**Dining table**
Bird	**Dog**
Boat	**Horse**
Bike	**Motorbike**
Bottle	**Person**
Bus	**Potted plant**
Car	**Sheep**
Cat	**Sofa**
Chair	**Train**
Cow	**TV monitor**

How many things are there?

PASCAL VOC

WordNet

20
[Everingham ’06-’12]

60K+
product categories

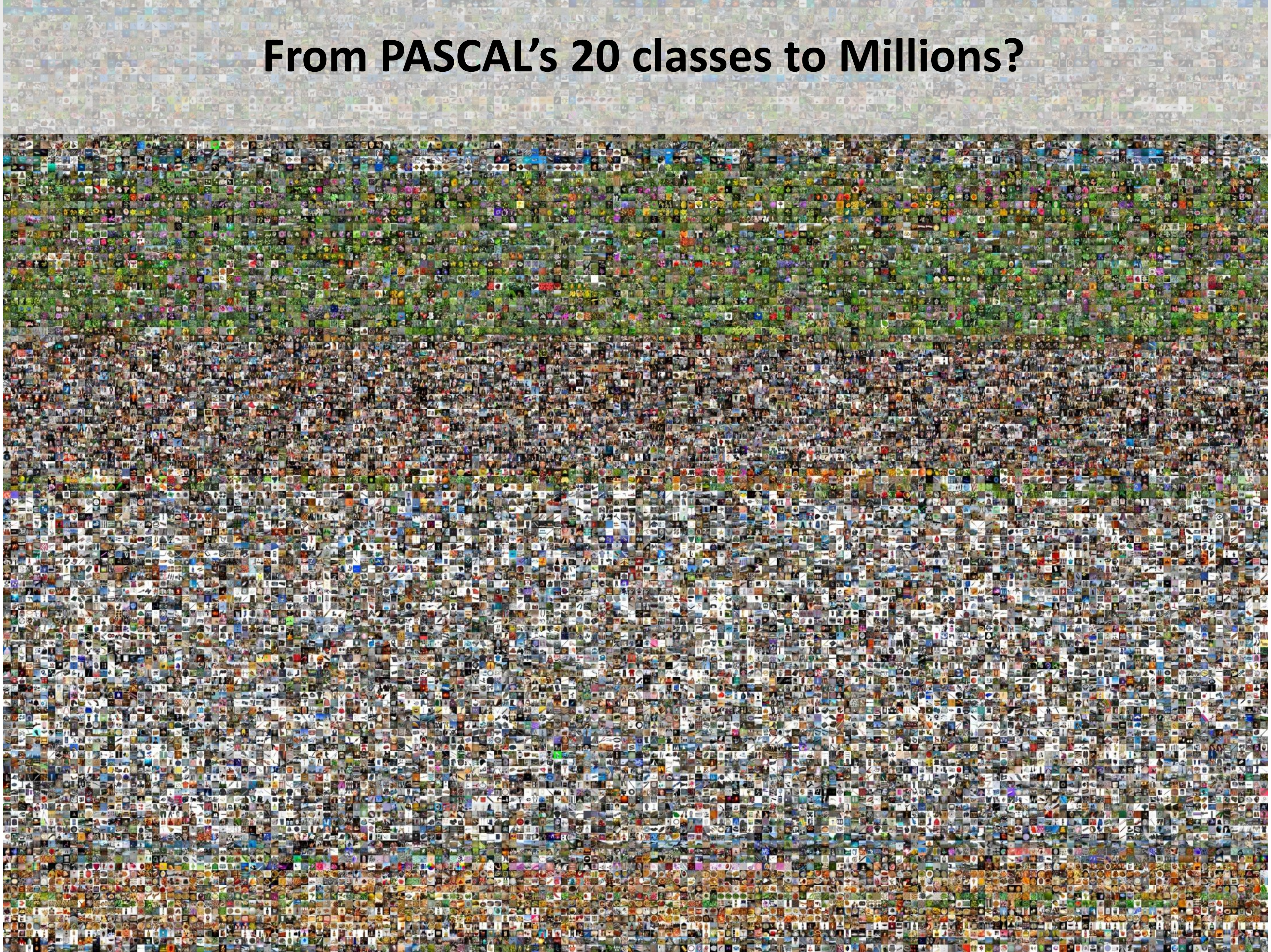
80K+
English nouns
[Miller ’95; Fellbaum ’98]

3.5M+
unique tags
[Sigurbjörnsson & Zwol ’08]

10K+
[Biederman ’87]

4.1M+
articles

From PASCAL's 20 classes to Millions?



EVA

Engine for Visual Annotation

The EVA system, powered by **ImageNet**, can annotate images with guaranteed accuracies. It currently recognizes over **10,000** visual categories. See the project page to find out more.

Paste a URL | Upload an image

ANNOTATE

Agenda

How to build a large-scale recognition engine using big data

- **STEP 1:** ?
- **STEP 2:** ?
- **STEP 3:** ?

Agenda

How to build a large-scale recognition engine using big data

STEP 1: Build a Large Knowledge Base

STEP 2: ?

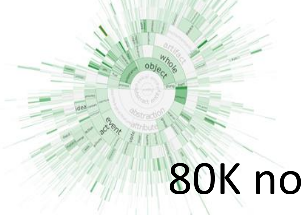
STEP 3: ?

Get a list of everything

↓

Crawl the web

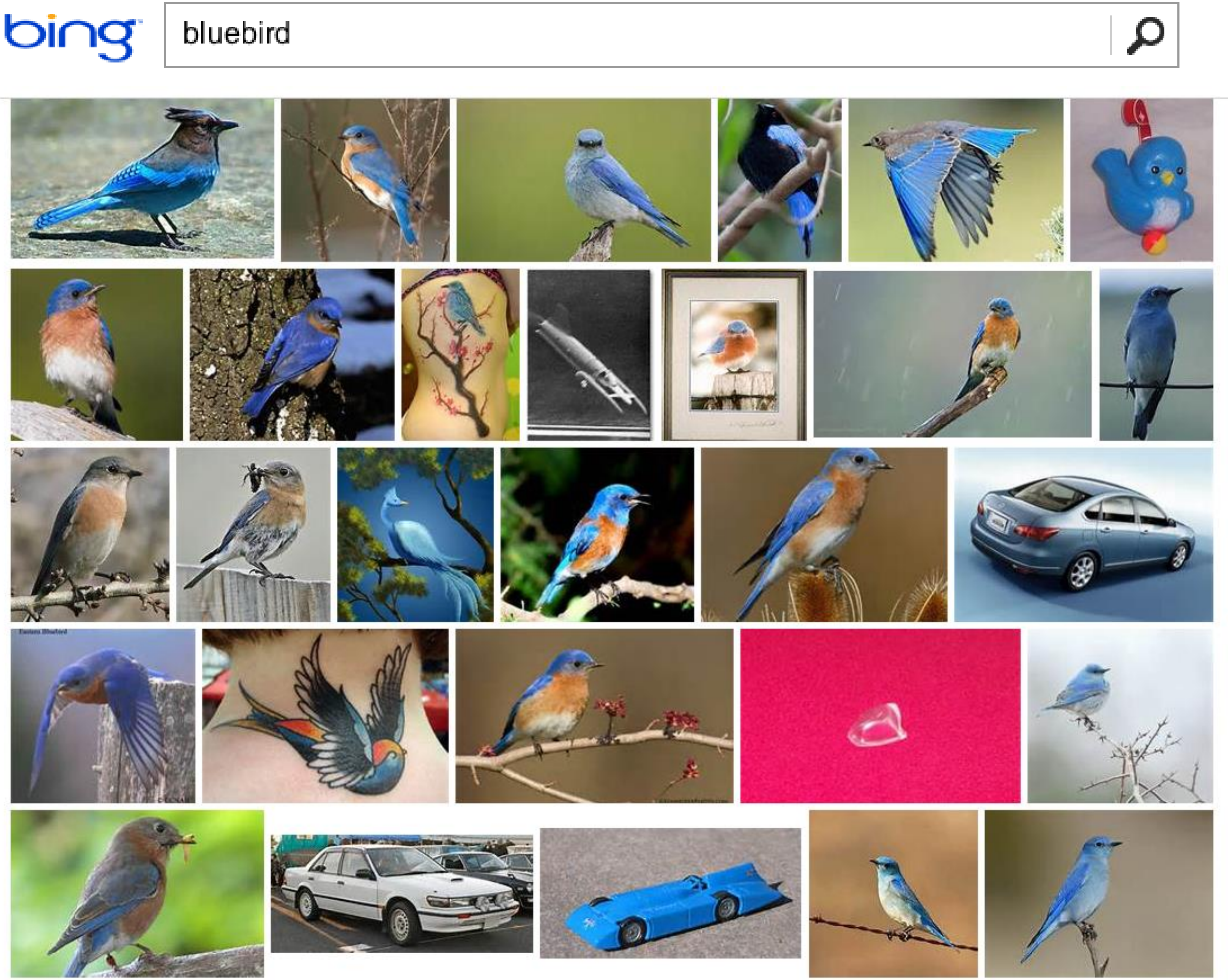
WordNet

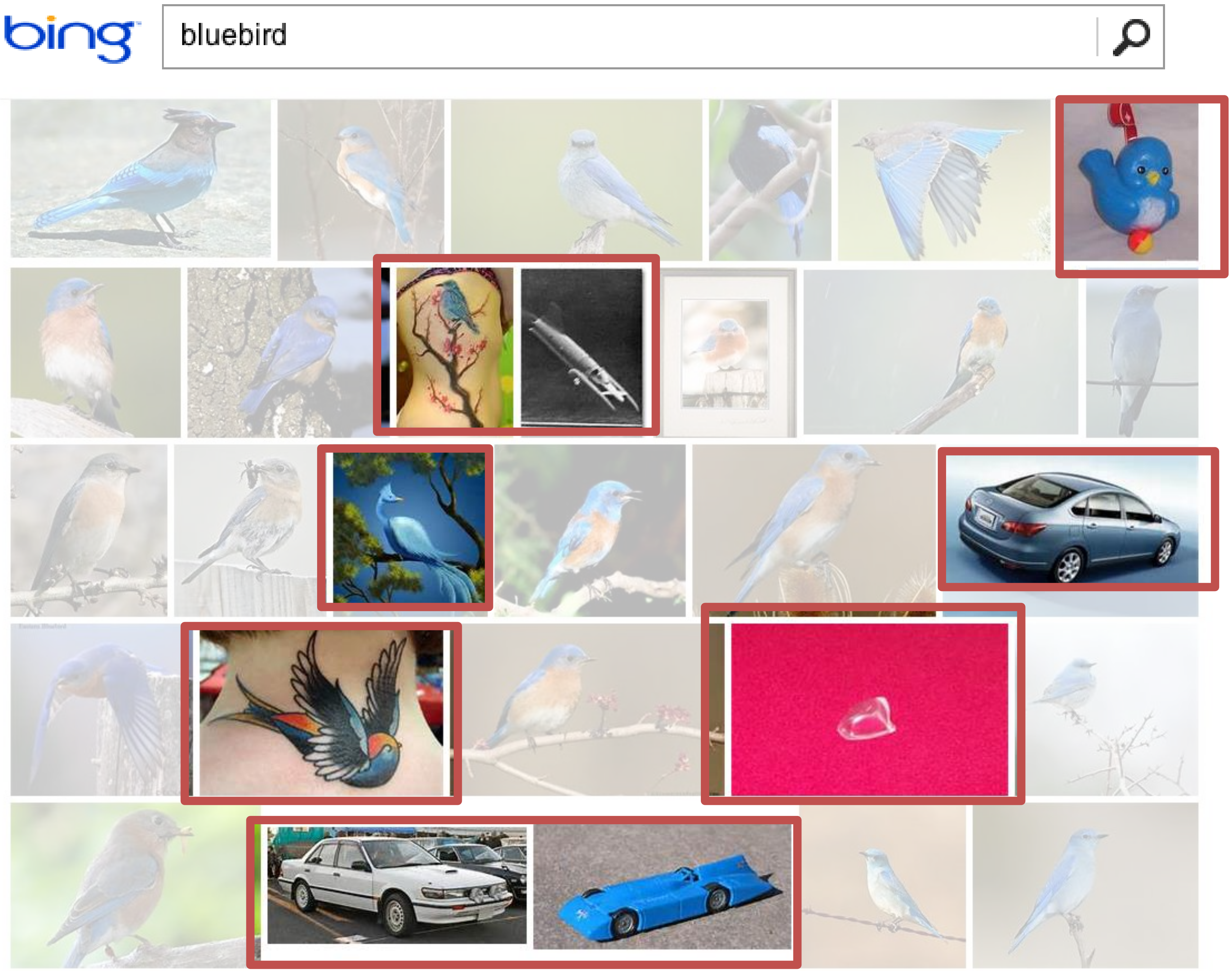


80K nouns

- Expert constructed
- Rich structure
  - Taxonomy, Partonomy
- Widely used

[Torralba, Fergus, Freeman ’08]
[Yao, Yang, Zhu ’07]
[Everingham et al ’06]
[Russell et al ’05]
[Griffin & Perona ’03]
[Fei-Fei, Fergus, Perona ’03]


bing
bluebird


bing
bluebird

Get a list of everything

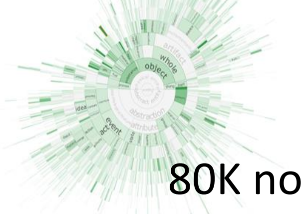
↓

Crawl the web

↓

Clean up

WordNet

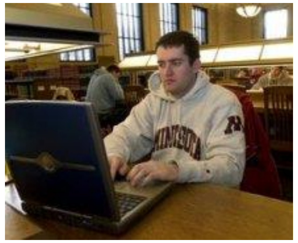


80K nouns

- Expert constructed
- Rich structure
    - Taxonomy, Partonomy
- Senses disambiguated
- Widely used

[Torralba, Fergus, Freeman ’08]
[Yao, Yang, Zhu ’07]
[Everingham et al ’06]
[Russell et al ’05]
[Griffin & Perona ’03]
[Fei-Fei, Fergus, Perona ’03]

Graduate Students



Good at complex tasks

Good quality

Very few of them

High cost

Estimate: 20 Years, $2M+

Graduate Students

Good at complex tasks

Good quality

Very few of them

High cost

The Crowd

amazon mechanical turk

beta

Artificial Artificial Intelligence

<u>Graduate Students</u>

Good at complex tasks

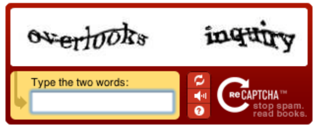
Good quality

Very few of them

High cost

<u>The Crowd</u>

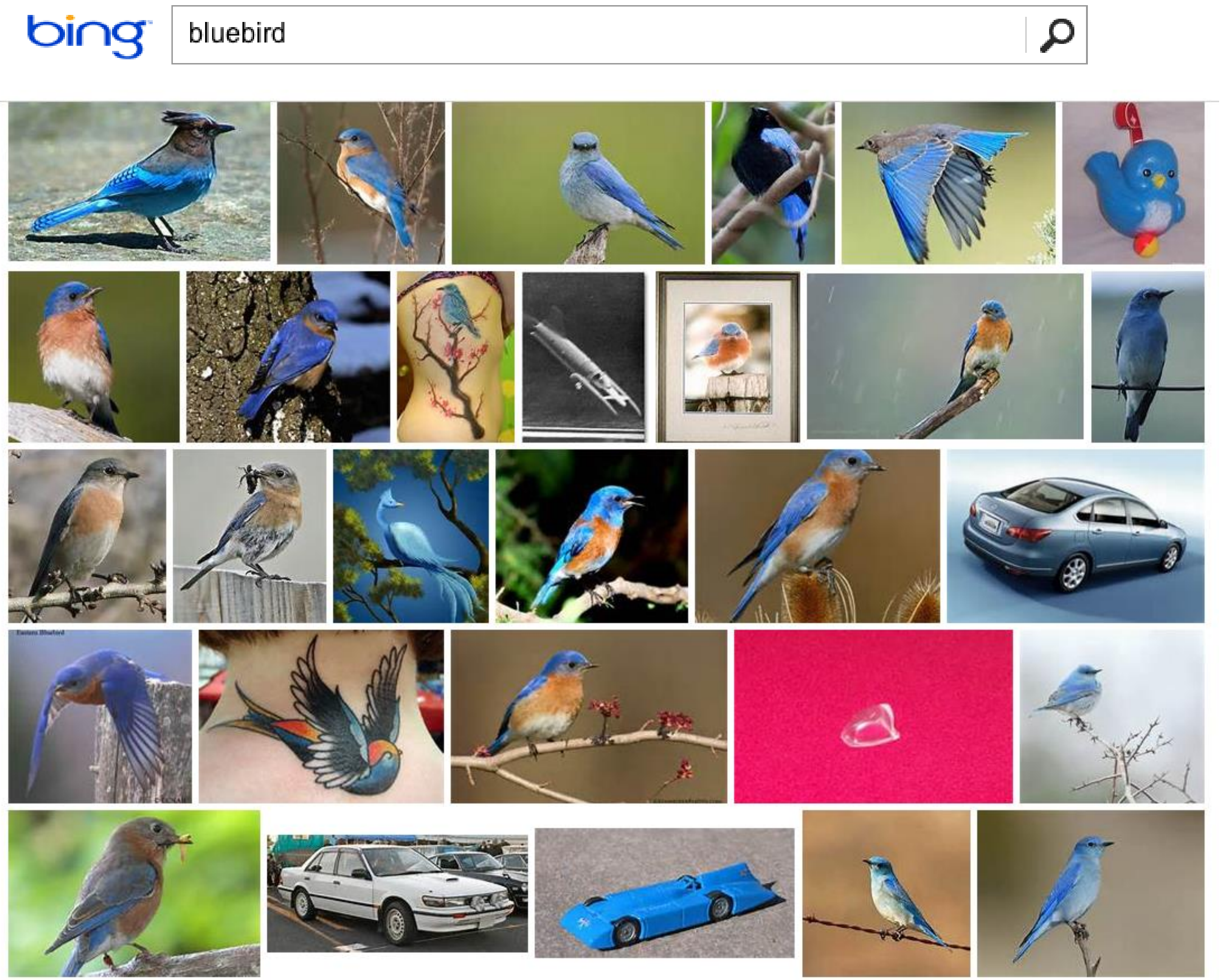
Good at simple tasks



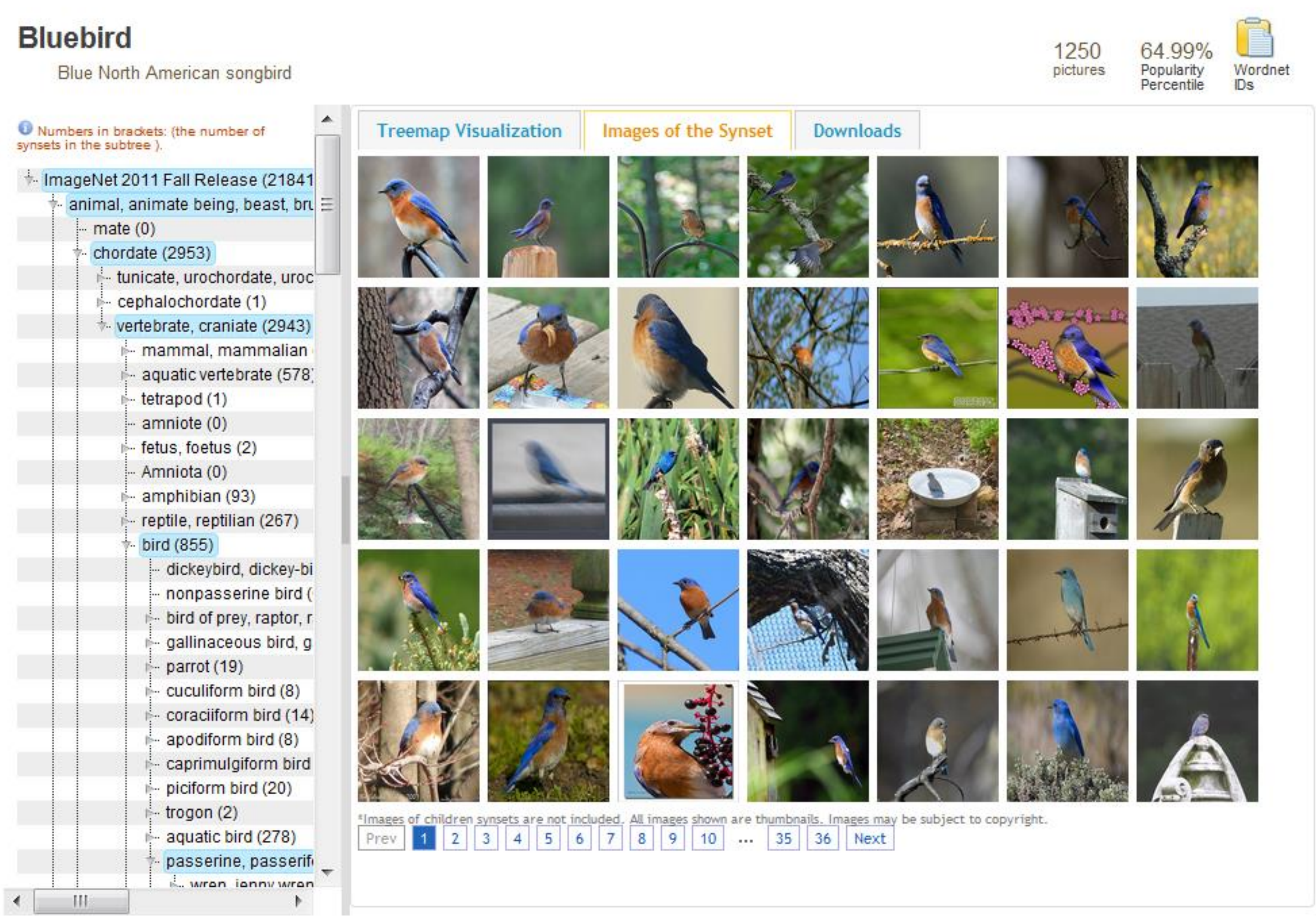
Mixed quality

Many of them

Low cost


bing
bluebird


amazon.com

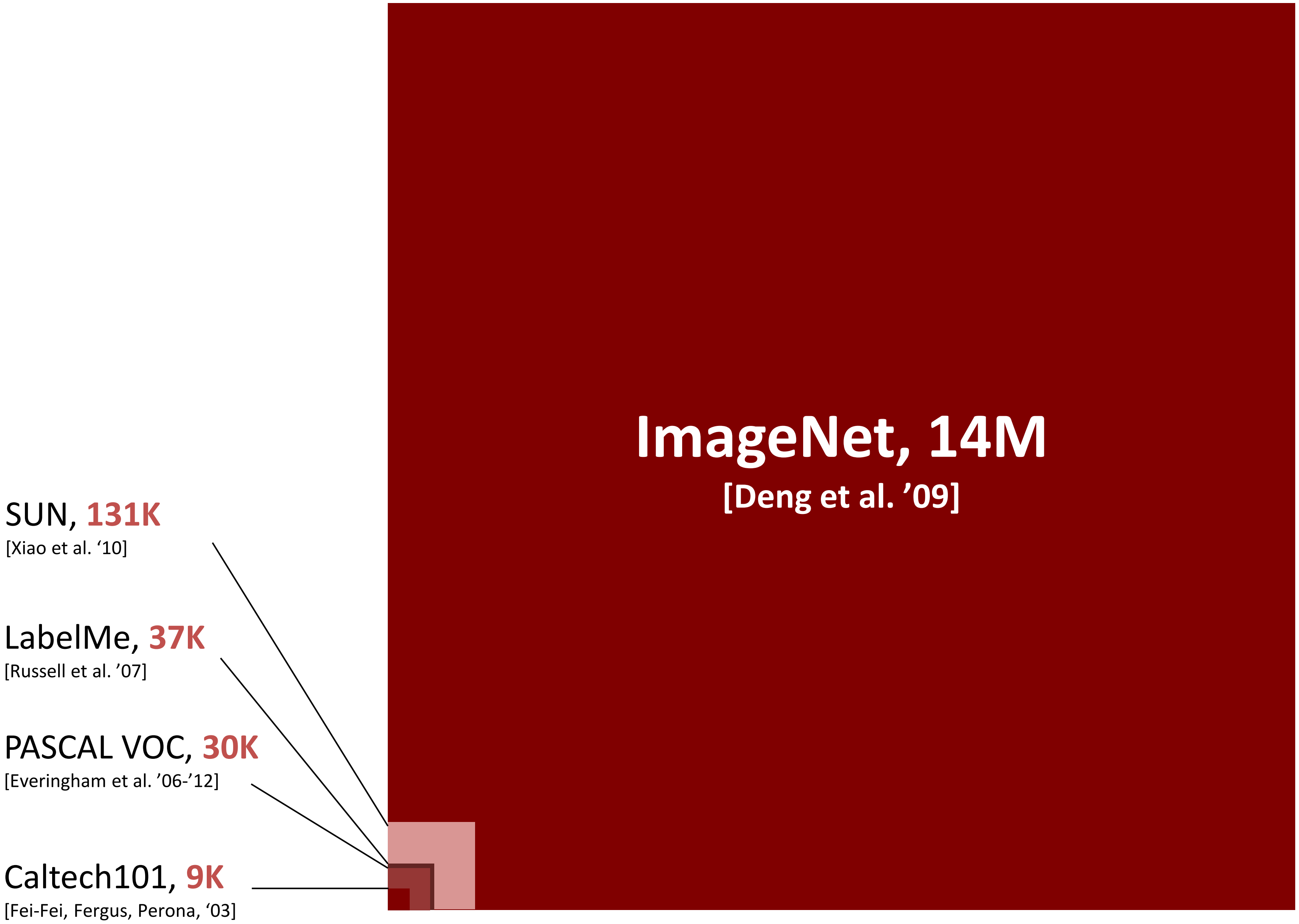

Bluebird
Blue North American songbird
1250
pictures
64.99%
Popularity
Percentile
Wordnet
IDs
Numbers in brackets: (the number of synsets in the subtree).
ImageNet 2011 Fall Release (21841
animal, animate being, beast, bru
mate (0)
chordate (2953)
tunicate, urochordate, uroc
cephalochordate (1)
vertebrate, craniate (2943)
mammal, mammalian
aquatic vertebrate (578)
tetrapod (1)
amniote (0)
fetus, foetus (2)
Amniota (0)
amphibian (93)
reptile, reptilian (267)
bird (855)
dickeybird, dickey-bi
nonpasserine bird (
bird of prey, raptor, r
gallinaceous bird, g
parrot (19)
cuculiform bird (8)
coraciiform bird (14)
apodiform bird (8)
caprimulgiform bird
piciform bird (20)
trogon (2)
aquatic bird (278)
passerine, passerif
Treemap Visualization
Images of the Synset
Downloads
*Images of children synsets are not included. All images shown are thumbnails. Images may be subject to copyright.
Prev
1
2
3
4
5
6
7
8
9
10
...
35
36
Next

IMAGENET [Deng et al. 2009] **www.image-net.org**

22,000 categories and **14,000,000+** images

- Animals
  - Bird
  - Fish
  - Mammal
  - Invertebrate
- Plants
  - Tree
  - Flower
- Food
- Materials
- Structures
- Artifact
  - Tools
  - Appliances
  - Structures
- Person
- Scenes
  - Indoor
  - Geological Formations
- Sport Activities

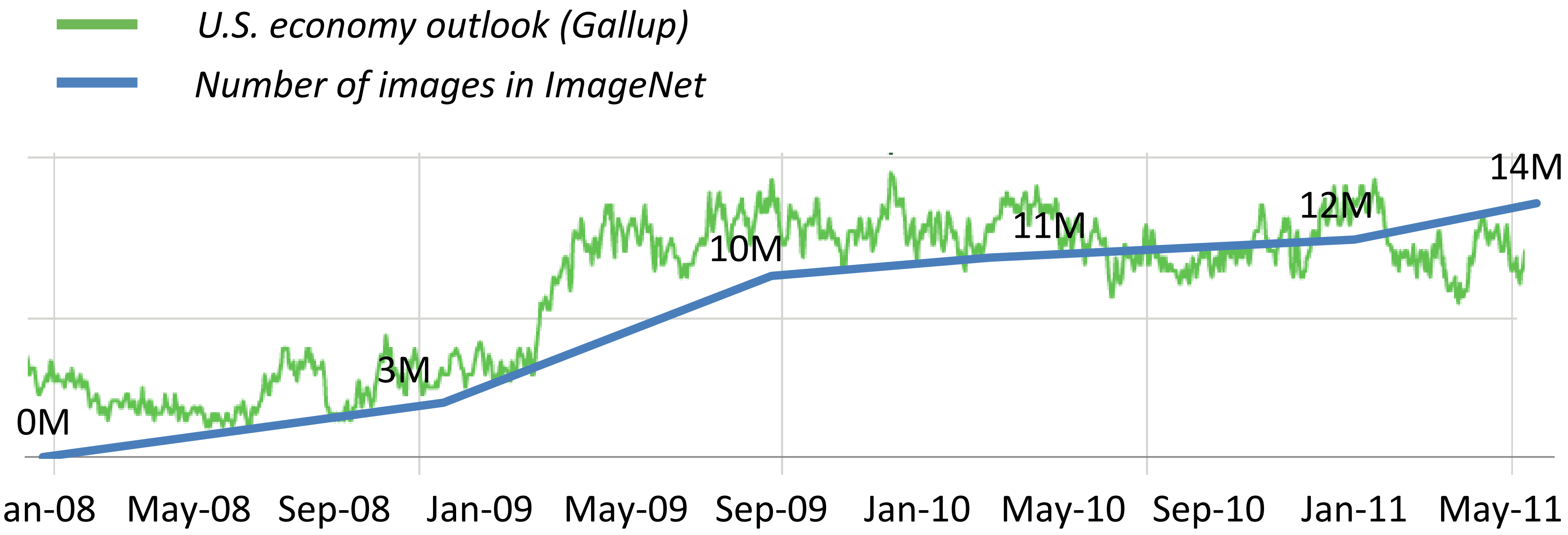
Number of Labeled Images



IMAGENET hired **50K+** AMT workers

who looked at **160M+** images

and made **550M+** binary decisions



ECCV 2012
Best paper Award

Kuettel, Guillaumin, Ferrari. **Segmentation Propagation in ImageNet.** ECCV 2012

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition

The New York Times

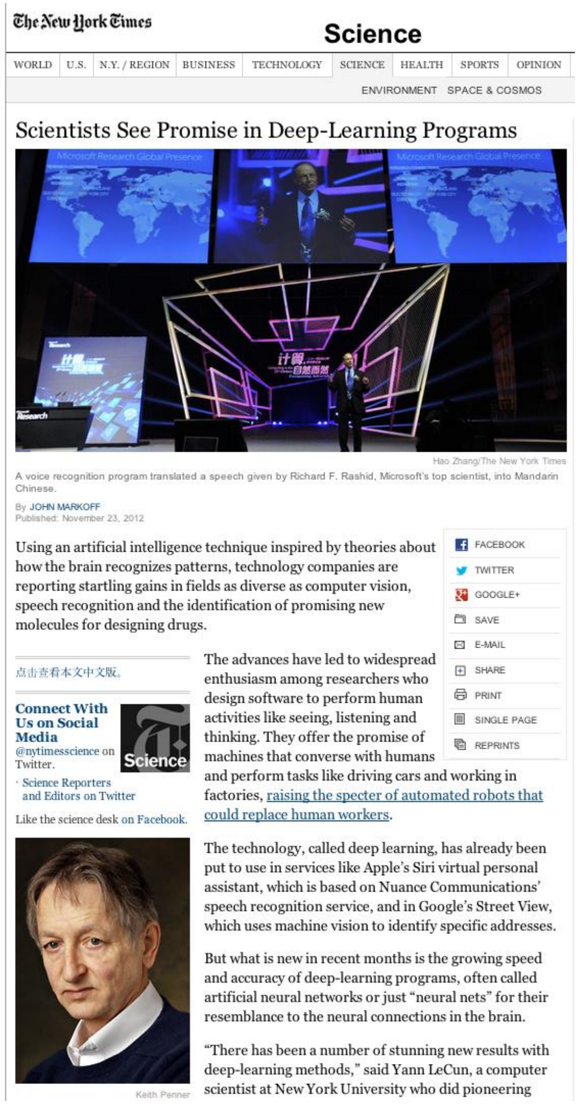
Business Day
Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

How Many Computers to Identify a Cat? 16,000

Jim Wilson/The New York Times

Le et al. **Building high-level features using large scale unsupervised learning**. ICML 2012.



The New York Times

Science

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

ENVIRONMENT SPACE & COSMOS

Scientists See Promise in Deep-Learning Programs

Hao Zhang/The New York Times

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

By JOHN MARKOFF
Published: November 23, 2012

Using an artificial intelligence technique inspired by theories about how the brain recognizes patterns, technology companies are reporting startling gains in fields as diverse as computer vision, speech recognition and the identification of promising new molecules for designing drugs.

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT
SINGLE PAGE
REPRINTS

点击查看本文中文版。

Connect With Us on Social Media
@nytimesscience on Twitter.
· Science Reporters and Editors on Twitter
Like the science desk on Facebook.

The advances have led to widespread enthusiasm among researchers who design software to perform human activities like seeing, listening and thinking. They offer the promise of machines that converse with humans and perform tasks like driving cars and working in factories, raising the specter of automated robots that could replace human workers.

The technology, called deep learning, has already been put to use in services like Apple's Siri virtual personal assistant, which is based on Nuance Communications' speech recognition service, and in Google's Street View, which uses machine vision to identify specific addresses.

But what is new in recent months is the growing speed and accuracy of deep-learning programs, often called artificial neural networks or just "neural nets" for their resemblance to the neural connections in the brain.

"There has been a number of stunning new results with deep-learning methods," said Yann LeCun, a computer scientist at New York University who did pioneering

Keith Penner

Krizhevsky, Sutskever, Hinton. **ImageNet classification with deep convolutional neural networks.** NIPS 2012

Seeking a Better Way to Find Web Images

By JOHN MARKOFF
Published: November 19, 2012

STANFORD, Calif. — You may think you can find almost anything on the Internet.

But even as images and video rapidly come to dominate the Web, search engines can ordinarily find a given image only if the text entered by a searcher matches the text with which it was labeled. And the labels can be unreliable, unhelpful ("fuzzy" instead of "rabbit") or simply nonexistent.

To eliminate those limits, scientists will need to create a new generation of visual search technologies — or else, as the Stanford computer scientist Fei-Fei Li recently put it, the Web will be in danger of "going dark."

Now, along with computer scientists from Princeton, Dr. Li, 36, has built the world's largest visual database in an effort to mimic the human vision system. With more than 14 million labeled objects, from obsidian to orangutans to ocelots, the database has because a vital resource for computer vision researchers.

The labels were created by humans. But now machines can learn from the vast database to recognize similar, unlabeled objects, making possible a striking increase in recognition accuracy.

This summer, for example, two Google computer scientists, Andrew Y. Ng and Jeff Dean, tested the new system, known as ImageNet, on a huge collection of labeled photos.

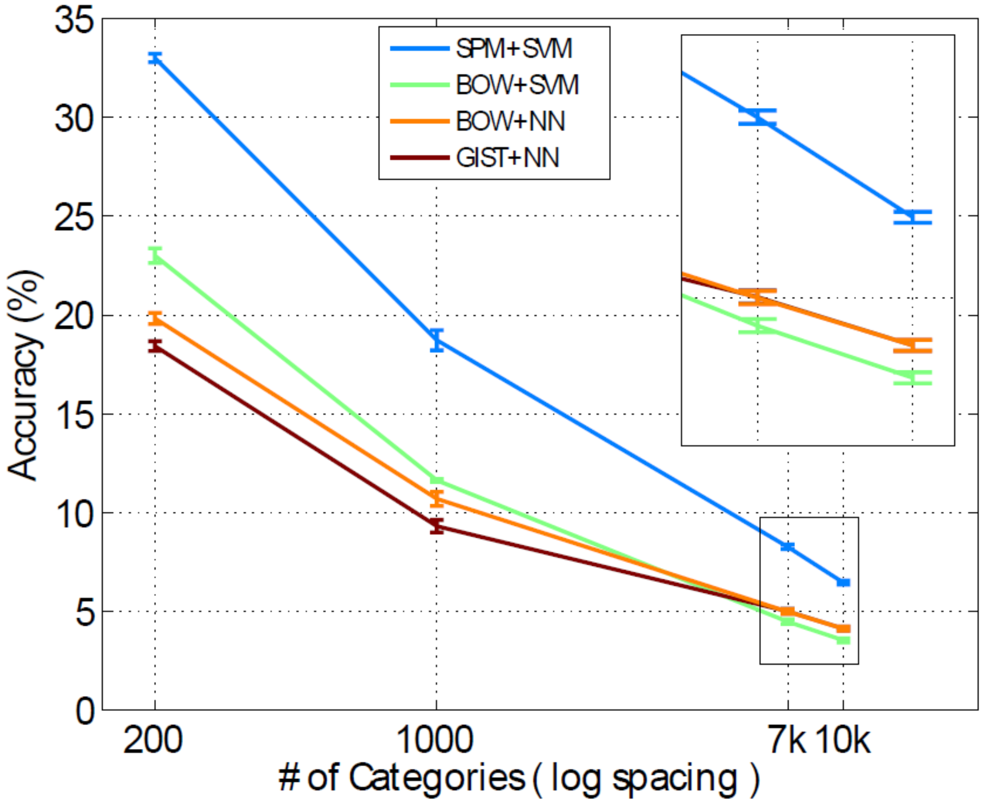
Agenda

How to build a large-scale recognition engine using big data

- **STEP 1:** Build a Large Knowledge Base (ImageNet)
- **STEP 2:** ?
- **STEP 3:** ?

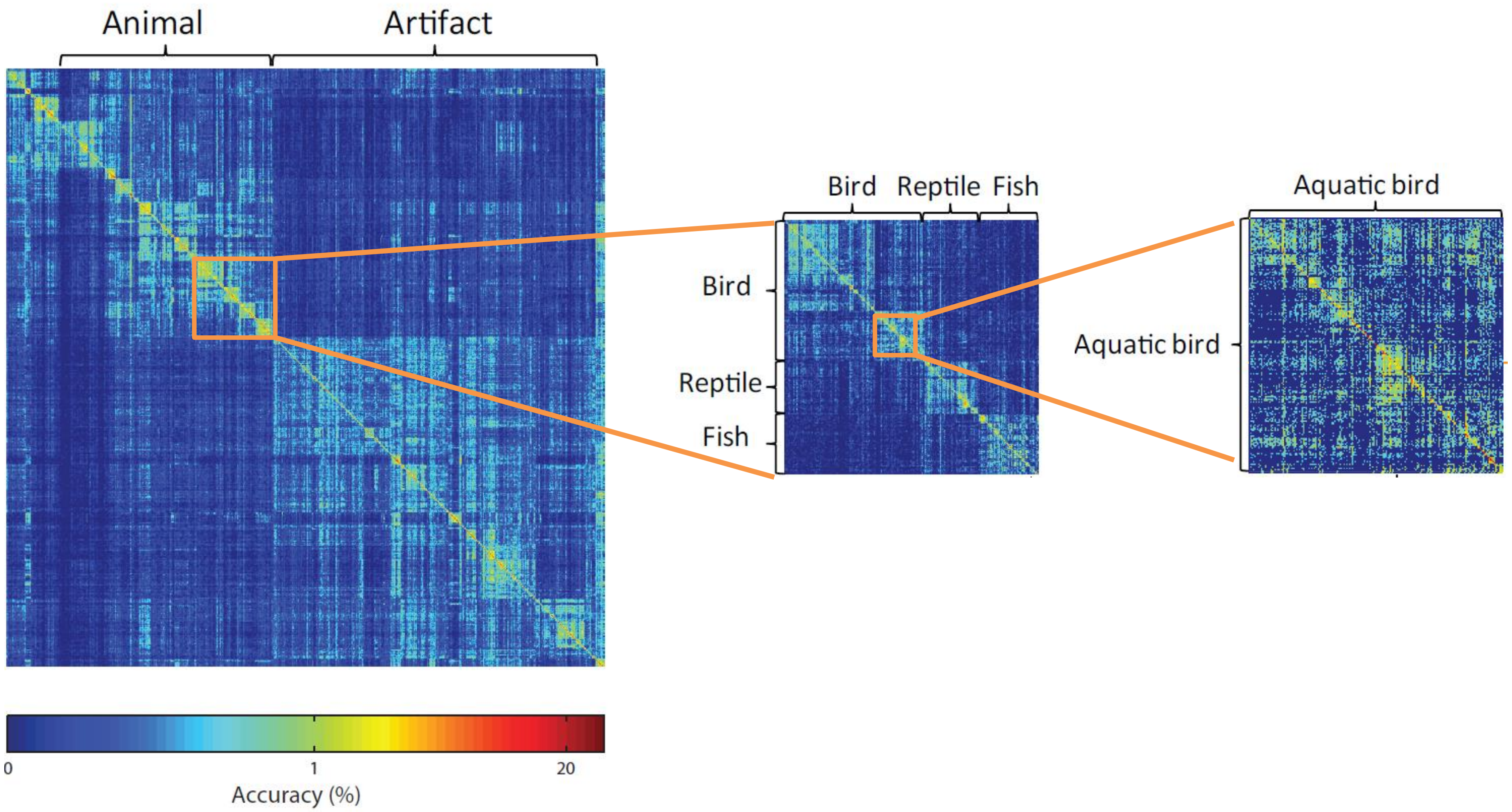
Learn to Classify 10K Classes

- 9 Million images
- 4 methods
  - SPM+SVM [Lazebnik et al. '06]
  - BOW+SVM [Csurka et al. '04]
  - BOW+NN
  - GIST+NN [Oliva et al. '01]
    - 6.4% for 10K categories



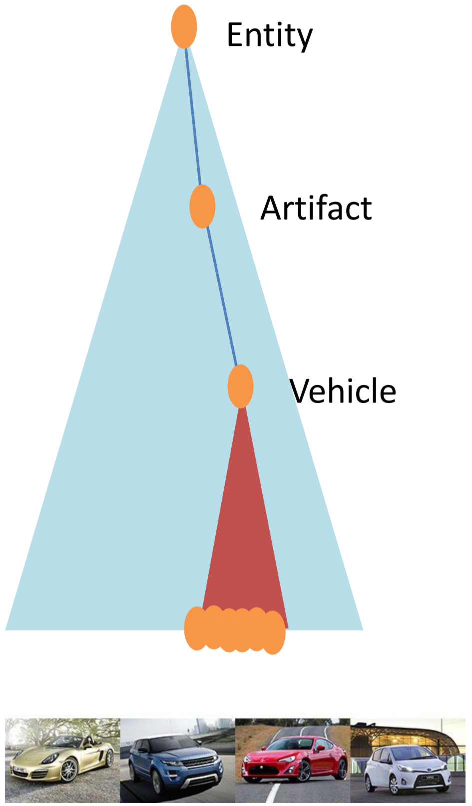
Deng, Berg, Li, & Fei-Fei, *ECCV2010*

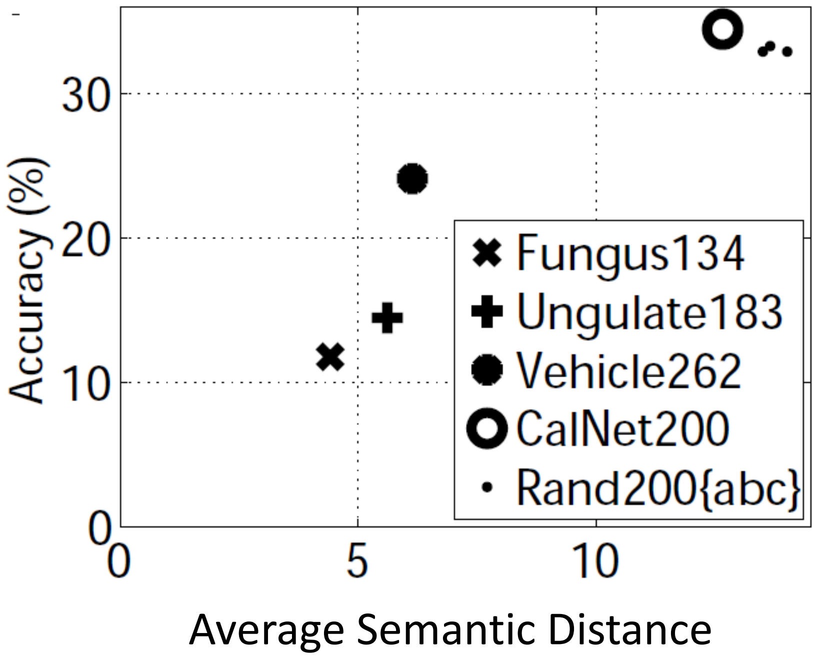
Learn to Classify 10K Classes

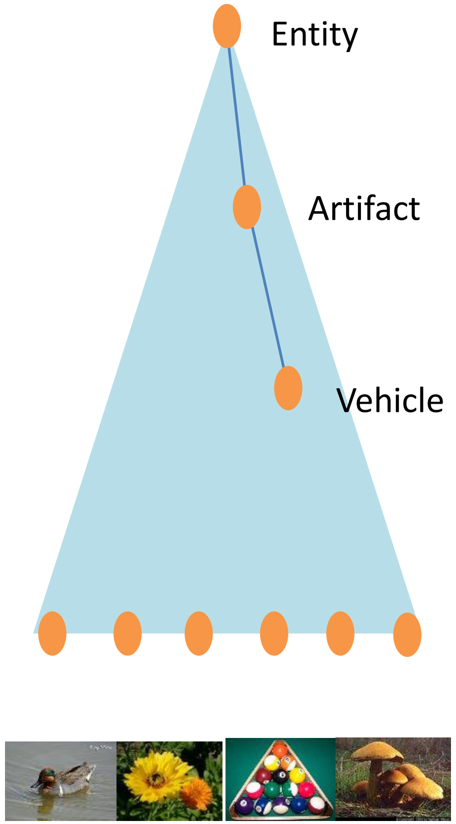


Deng, Berg, Li, & Fei-Fei, *ECCV2010*

Fine-grained categories are a lot harder







Finer **Coarser**

Deng, Berg, Li, & Fei-Fei, *ECCV2010*

Agenda

How to build a large-scale recognition engine using big data

STEP 1: Build a Large Knowledge Base (ImageNet)

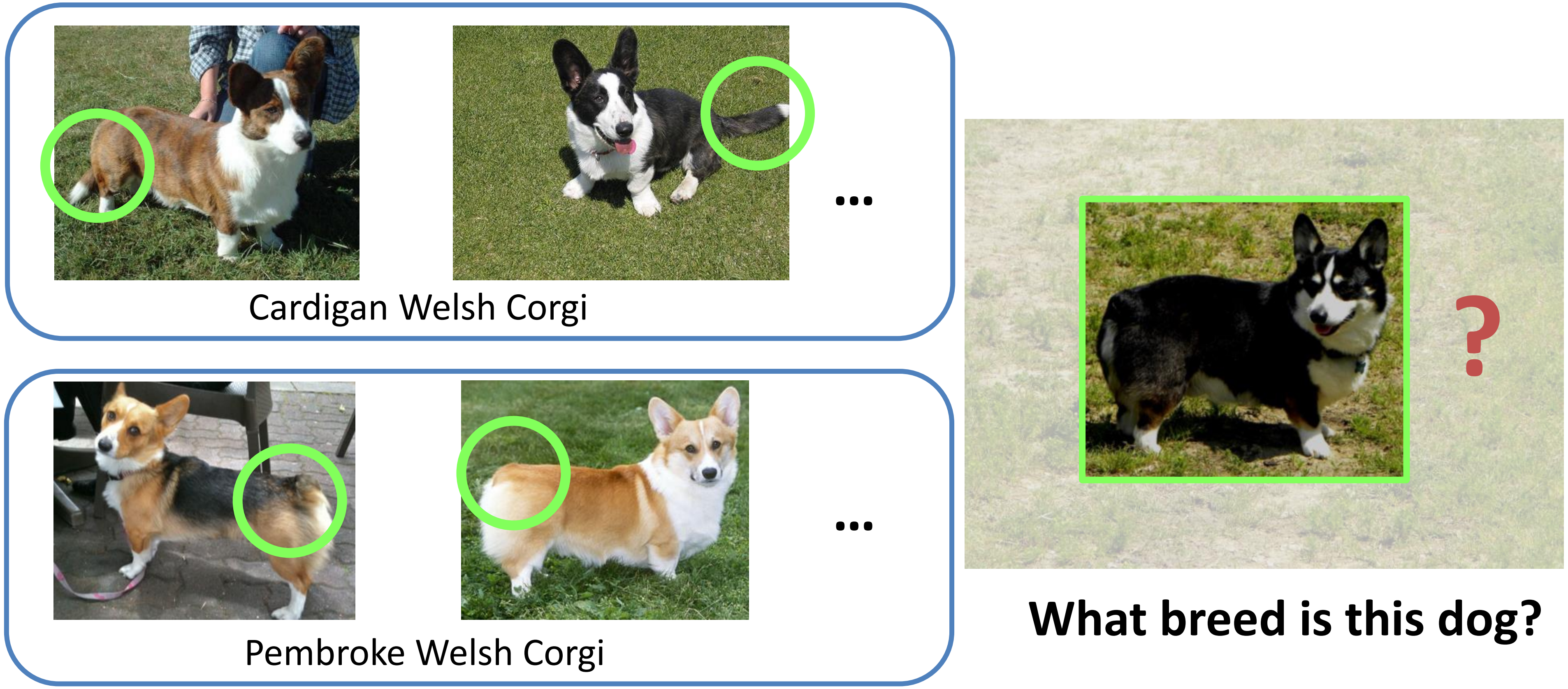
STEP 2: Fine-Grained Recognition

STEP 3: ?

Why is Fine-Grained Recognition Difficult?

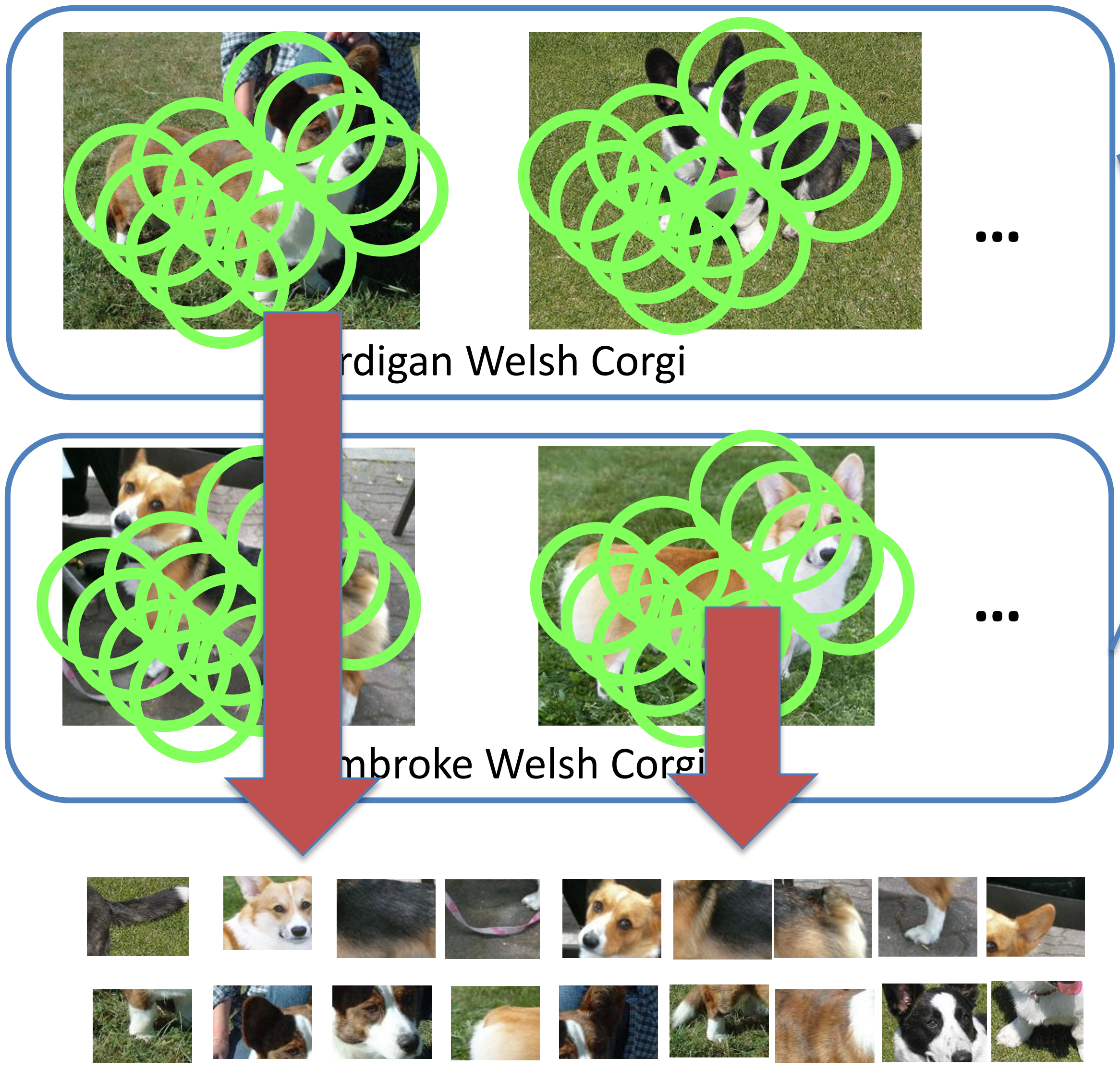
What breed is this dog?

Why is Fine-Grained Recognition Difficult?



Key: Find the right features.

Why is Fine-Grained Recognition Difficult?



Existing Work

[Branson et al. '10]

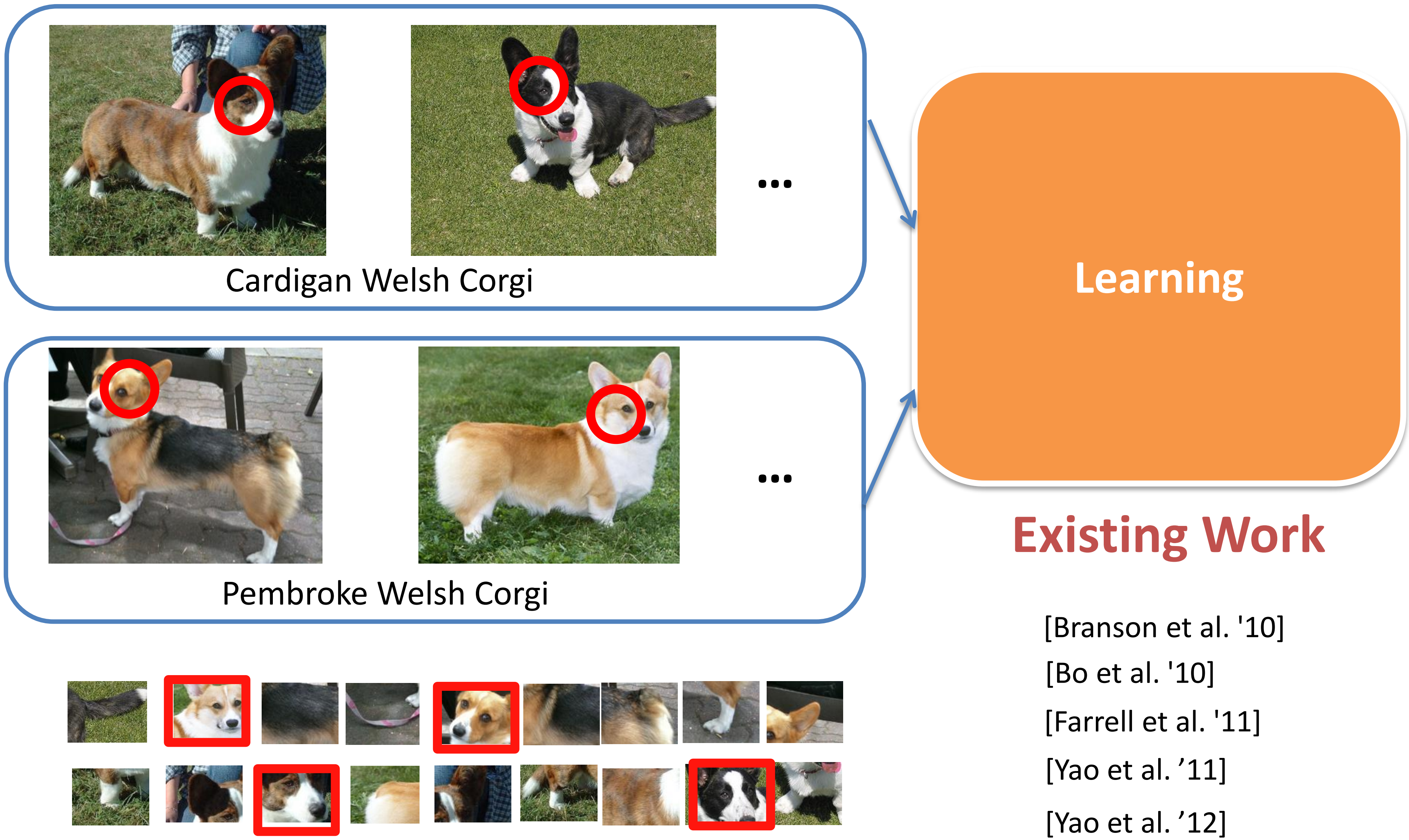
[Bo et al. '10]

[Farrell et al. '11]

[Yao et al. ’11]

[Yao et al. ’12]

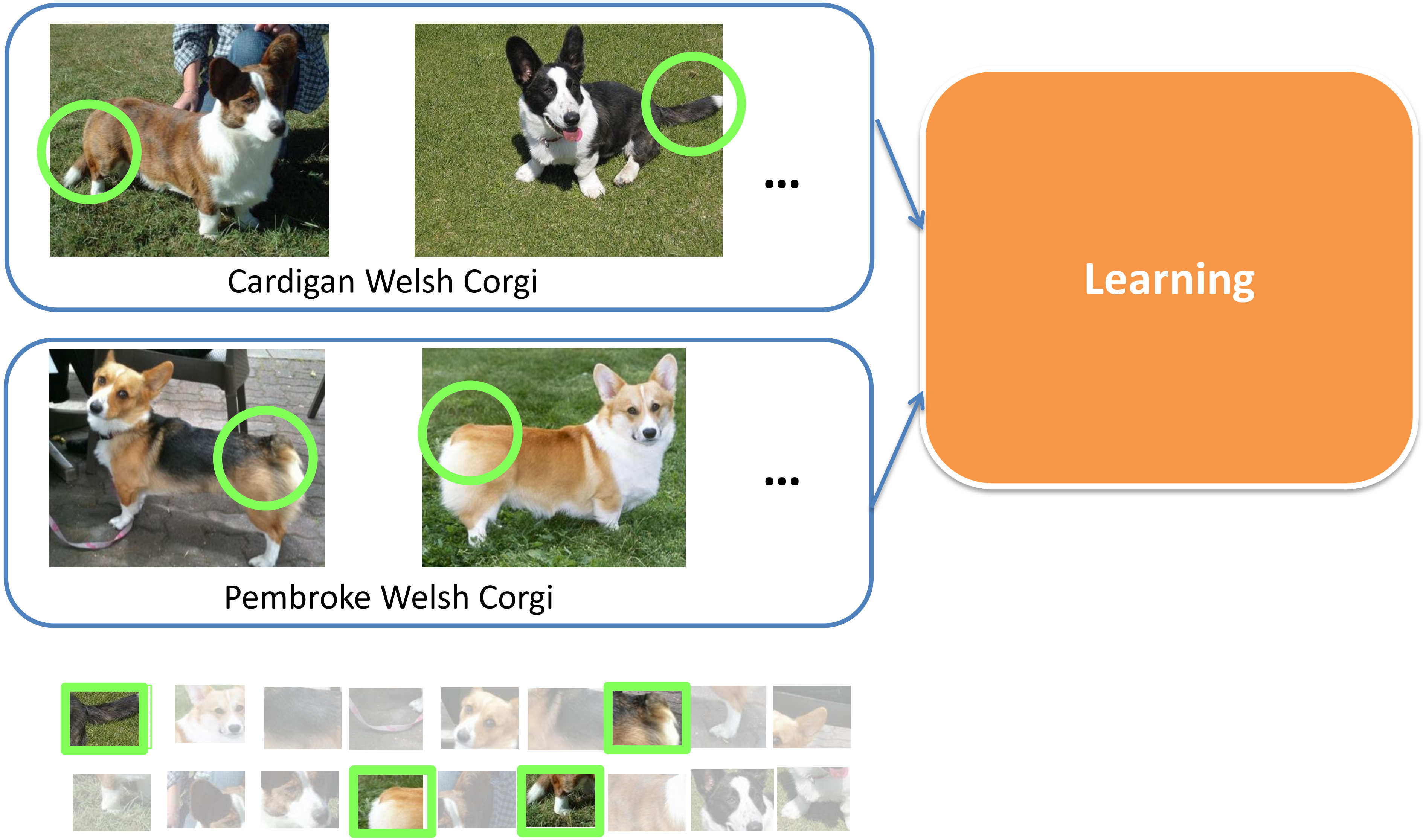
Why is Fine-Grained Recognition Difficult?



Existing Work

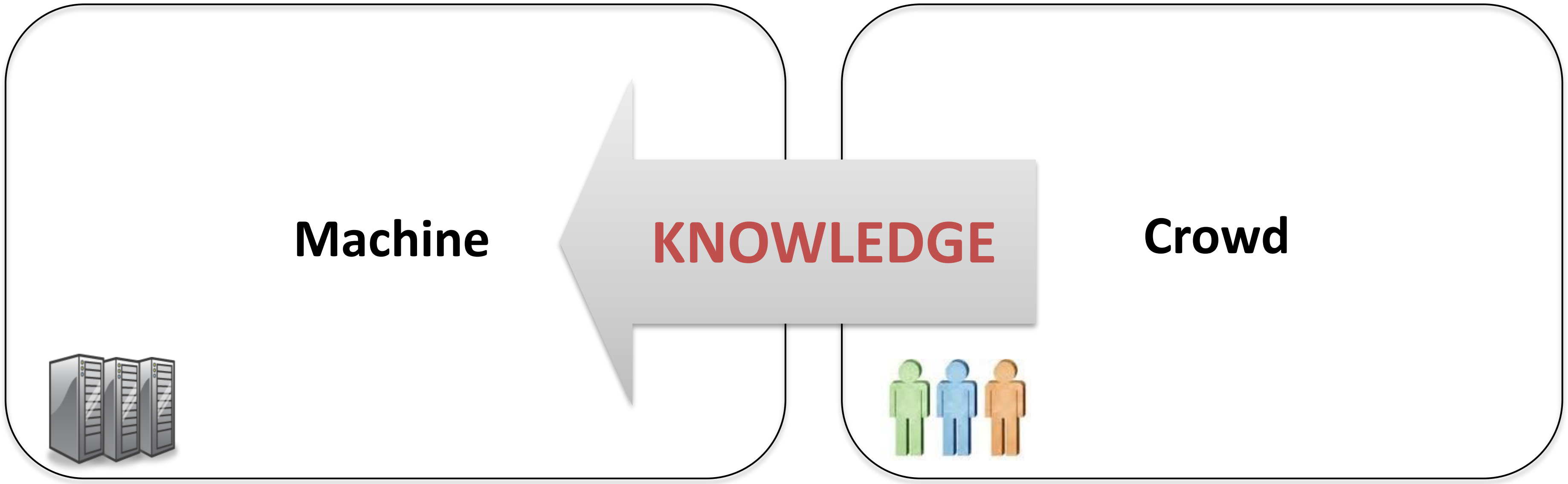
[Branson et al. '10]
[Bo et al. '10]
[Farrell et al. '11]
[Yao et al. ’11]
[Yao et al. ’12]

Why is Fine-Grained Recognition Difficult?

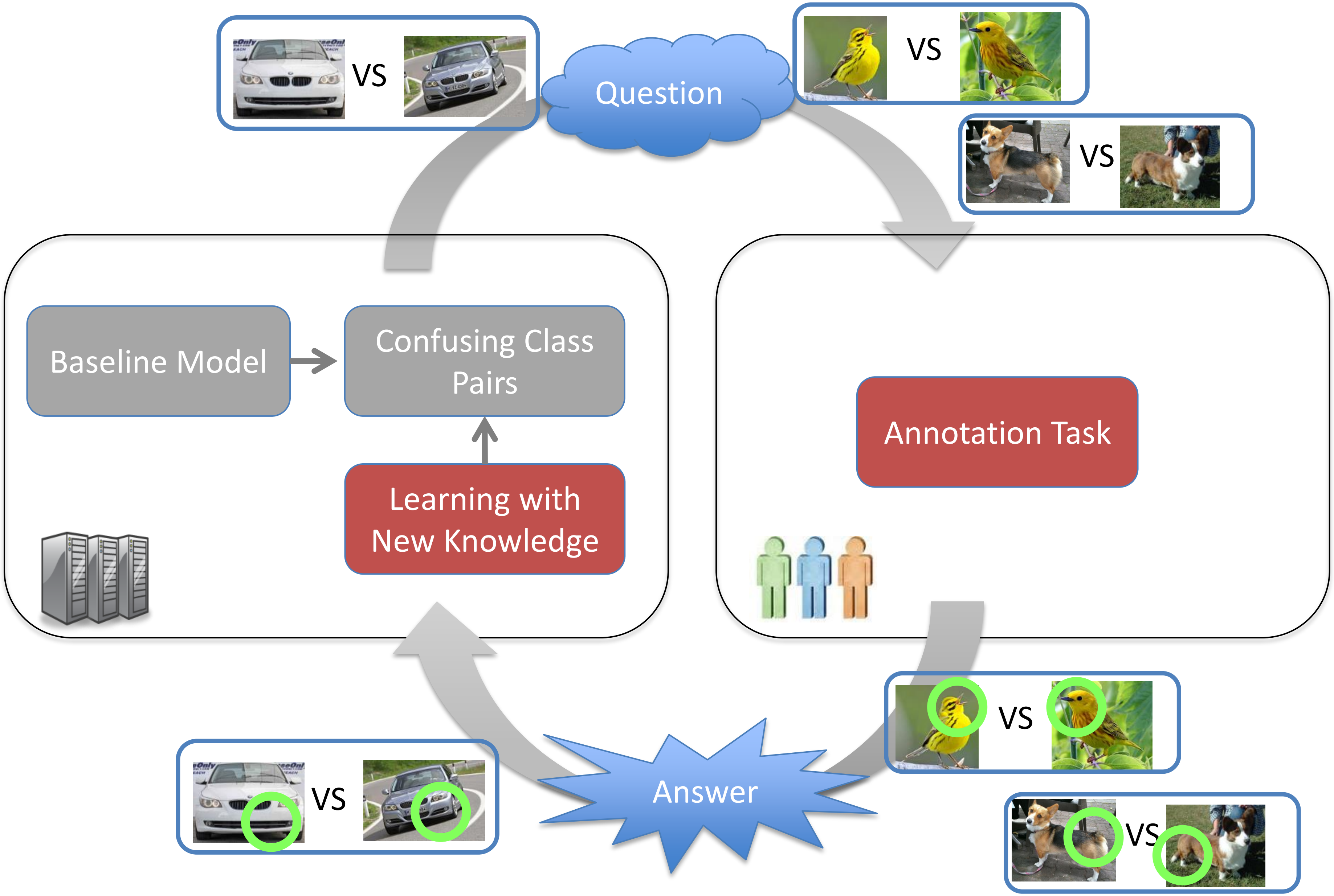


How to help computers select features?

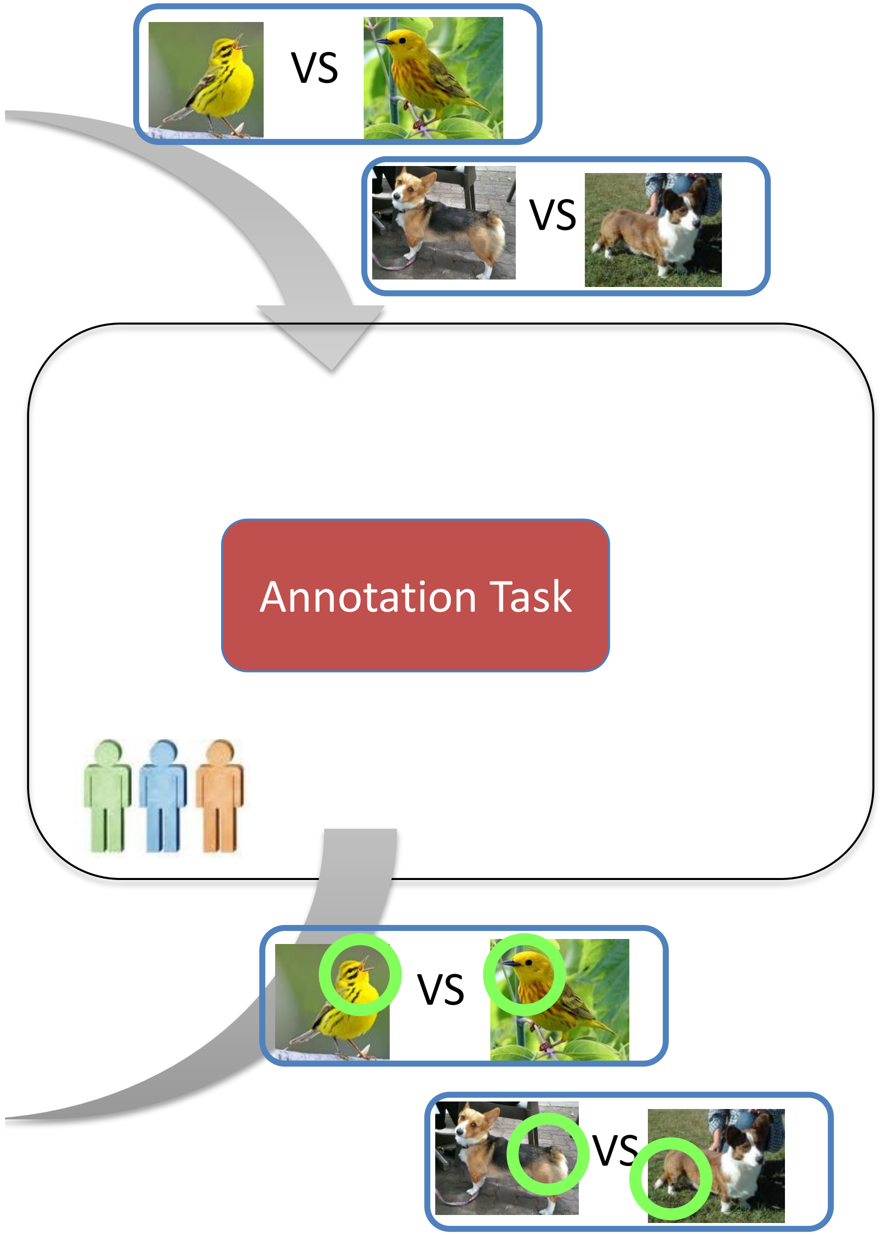
Machine-Crowd Collaboration



Machine-Crowd Collaboration

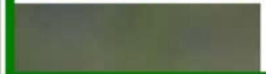


Machine-Crowd Collaboration



Click Me or Press 1

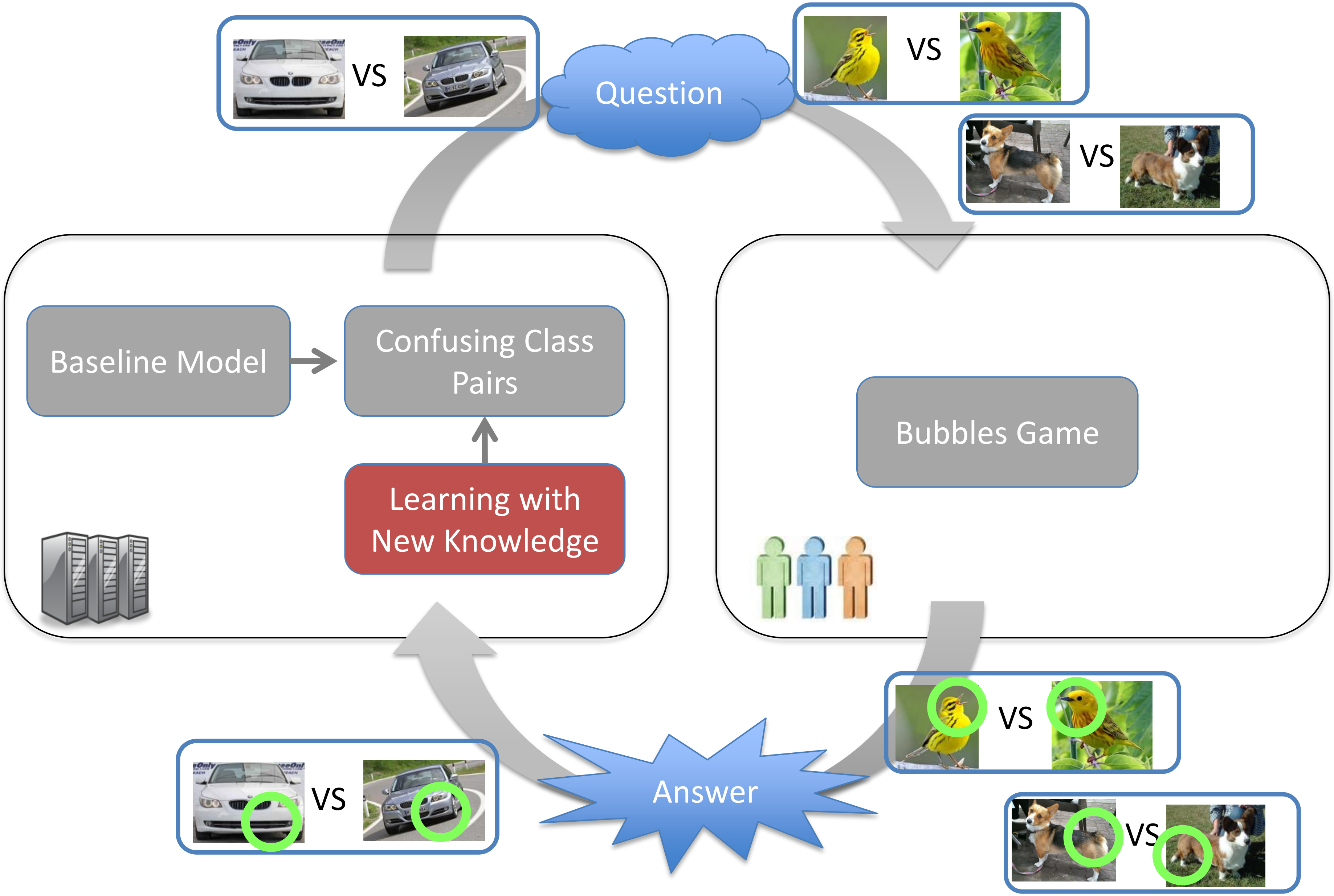
Prairie Warbler (wikipedia)



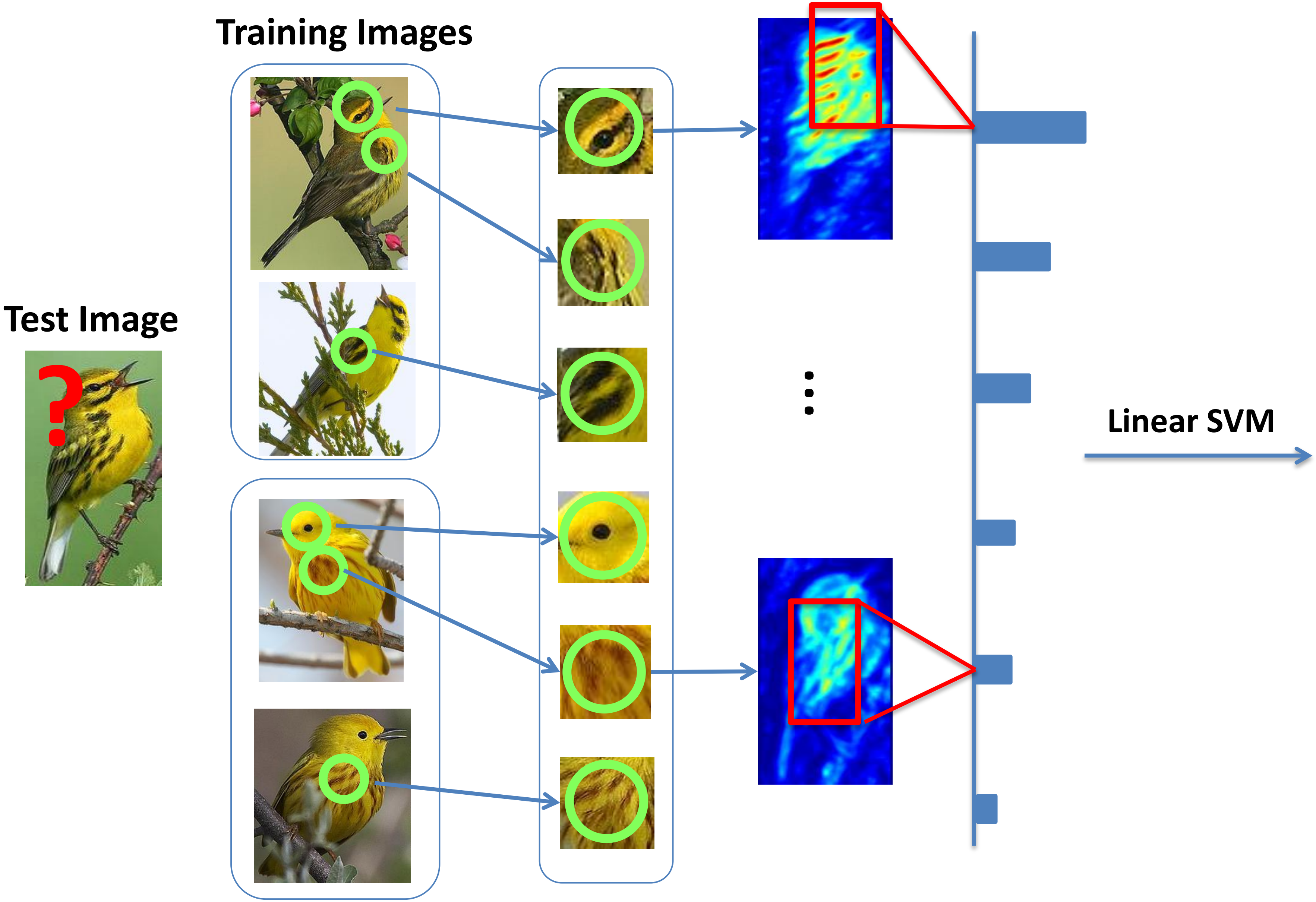
Click Me or Press 2

Yellow Warbler (wikipedia)

Machine-Crowd Collaboration

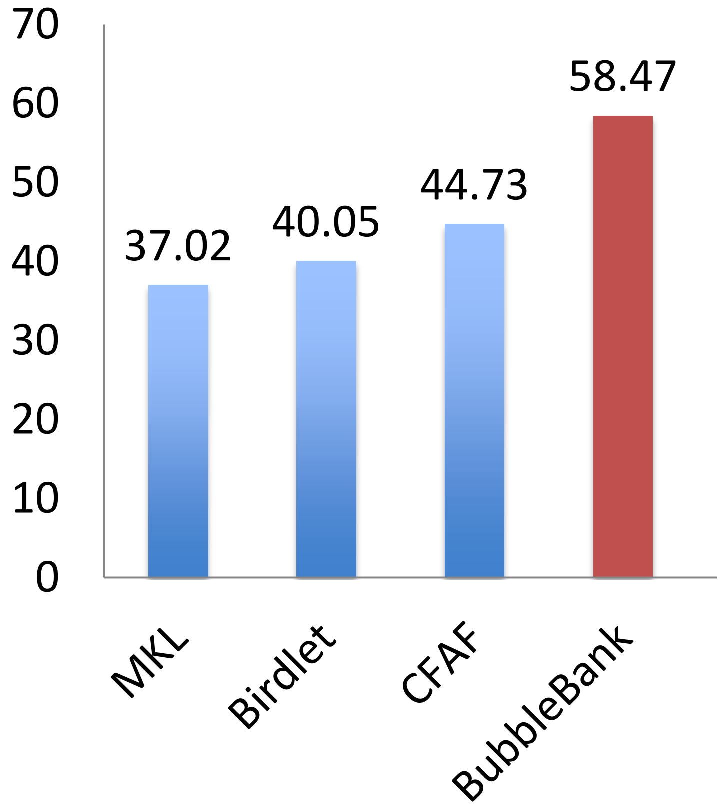


The BubbleBank Representation



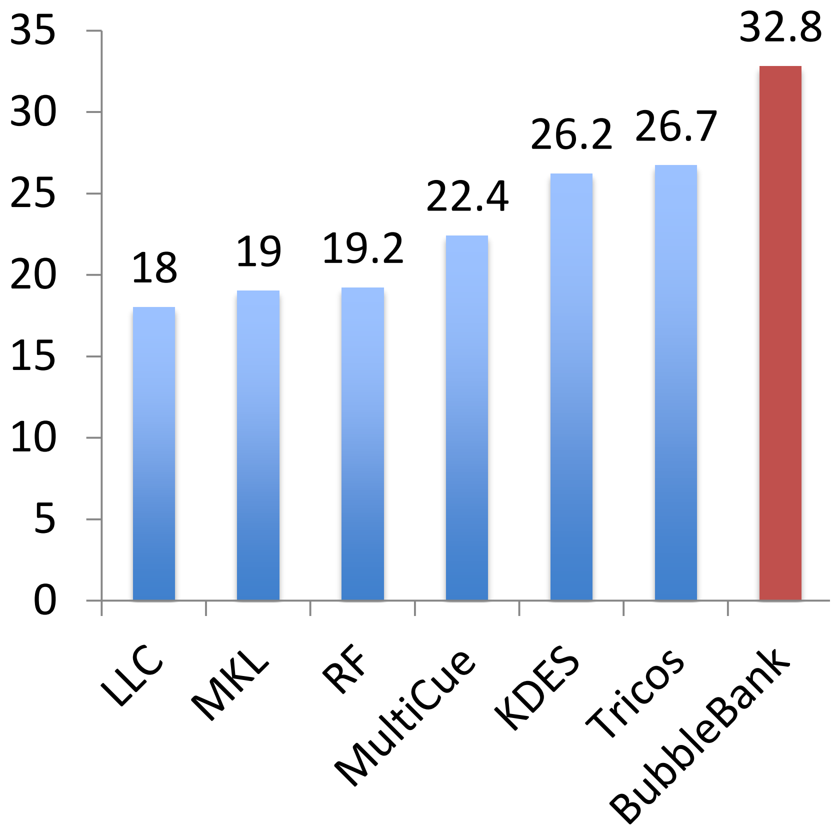
Deng, Krause, & Fei-Fei, *CVPR2013*

mAP on CUB-14 [Welinder et al. 10]



MKL [Branson et al. '10]
Birdlet [Farrell et al. '11]
CFAF [Yao et al.'12]

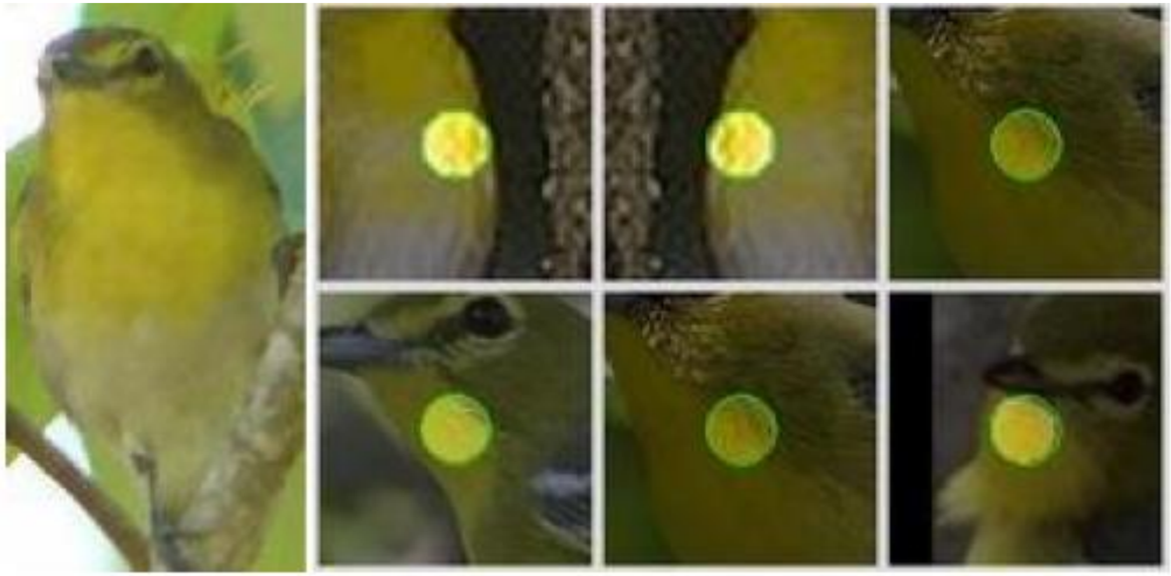
Accuracy on CUB-200 [Welinder et al. 10]



MKL [Branson et al. ‘10]
LLC [Wang et al. ‘09]
RF [Yao et al. '11]
MultiCue [Khan et al.'11]
KDES [Bo et al. ’10]
Tricos [Chai ’12]

Deng, Krause, & Fei-Fei, *CVPR2013*

Top Activated Bubbles (successful predictions)



Deng, Krause, & Fei-Fei, *CVPR2013*

Agenda

How to build a large-scale recognition engine using big data

STEP 1: Build a Large Knowledge Base (ImageNet)

STEP 2: Fine-Grained Recognition (Bubbles)

STEP 3: ?

Agenda

How to build a large-scale recognition engine using big data

STEP 1: Build a Large Knowledge Base (ImageNet)

STEP 2: Fine-Grained Recognition (Bubbles)

STEP 3: Putting a label on "everything"

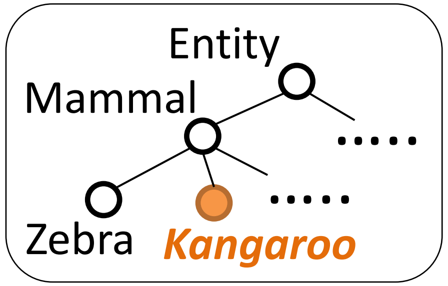
The Current State of the Art

10K classes	32.6%	Krizhevsky et al. NIPS 2012
20K classes	15%	Le et al. NIPS 2012

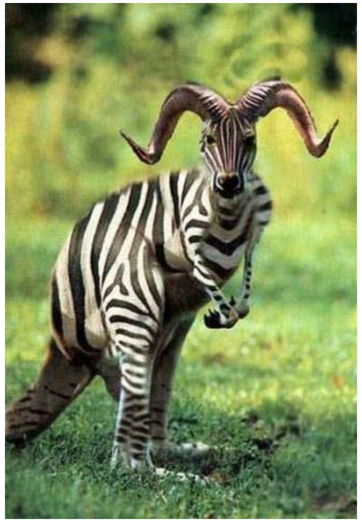
Not quite practical yet…

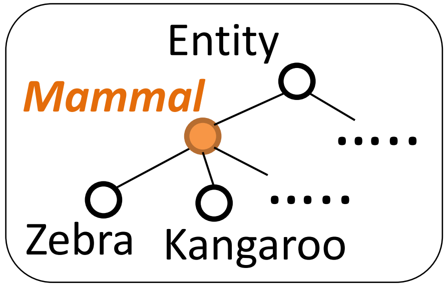
But we are measuring the very fine-grained level

Hedging: Be as informative as possible with few mistakes



Kangaroo ✓

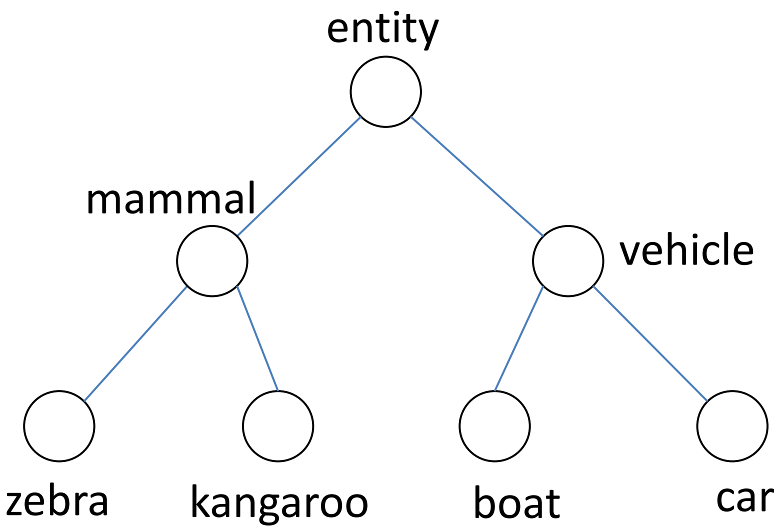




Mammal ✓

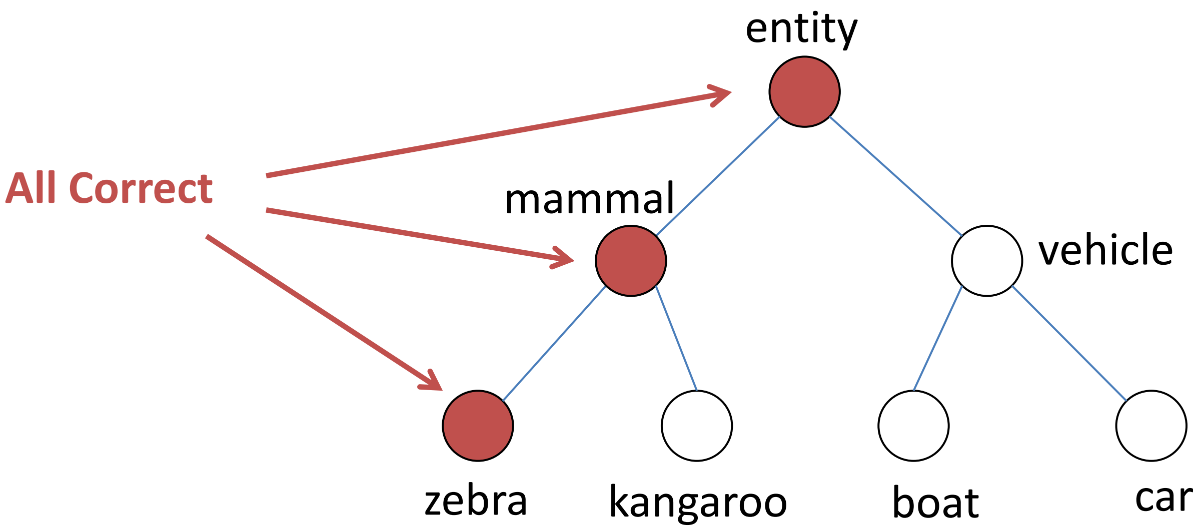
Deng, Krause, Berg, Fei-Fei, CVPR2012

Formal Problem Statement



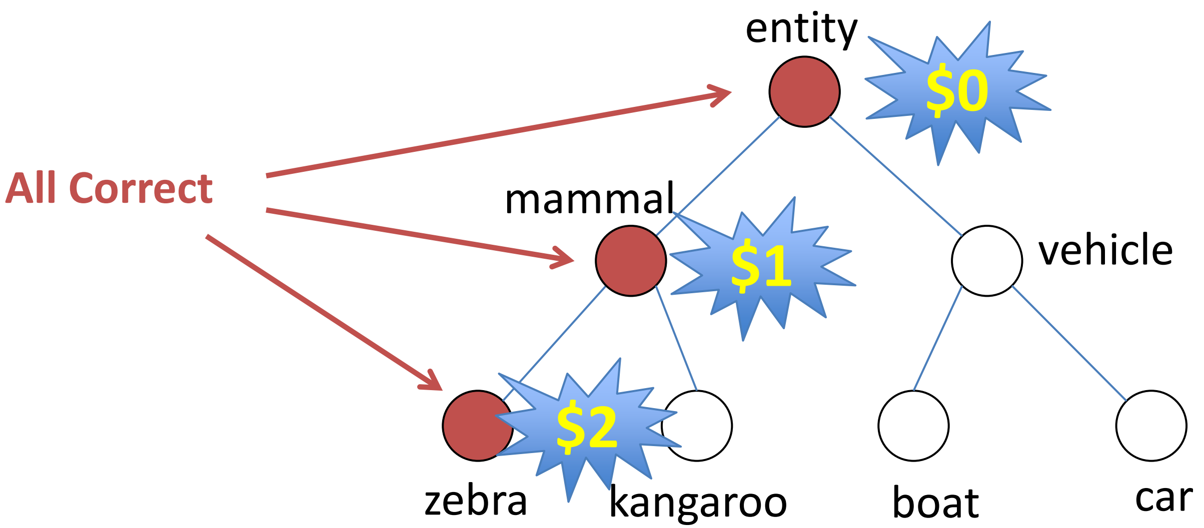
Deng, Krause, Berg, Fei-Fei, CVPR2012

Formal Problem Statement



Deng, Krause, Berg, Fei-Fei, CVPR2012

Formal Problem Statement



Deng, Krause, Berg, Fei-Fei, CVPR2012

Formal Problem Statement

Assumptions

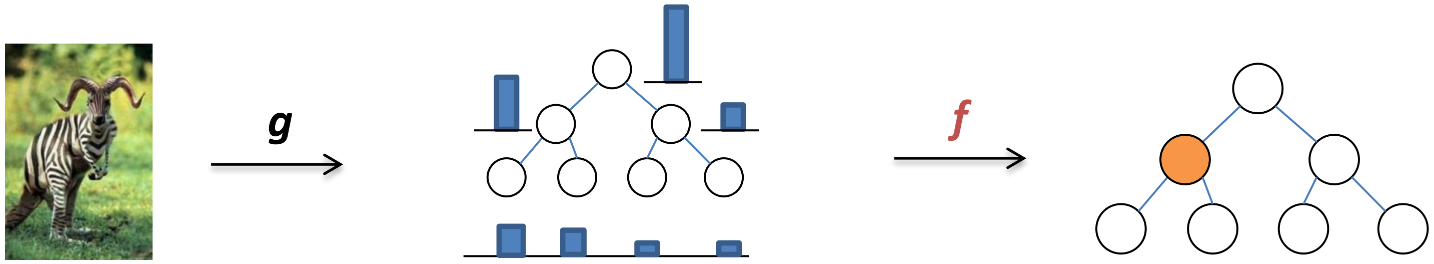
- Same distribution for training and test.
- A base classifier ***g*** that gives posterior probability on the hierarchy.

Goal

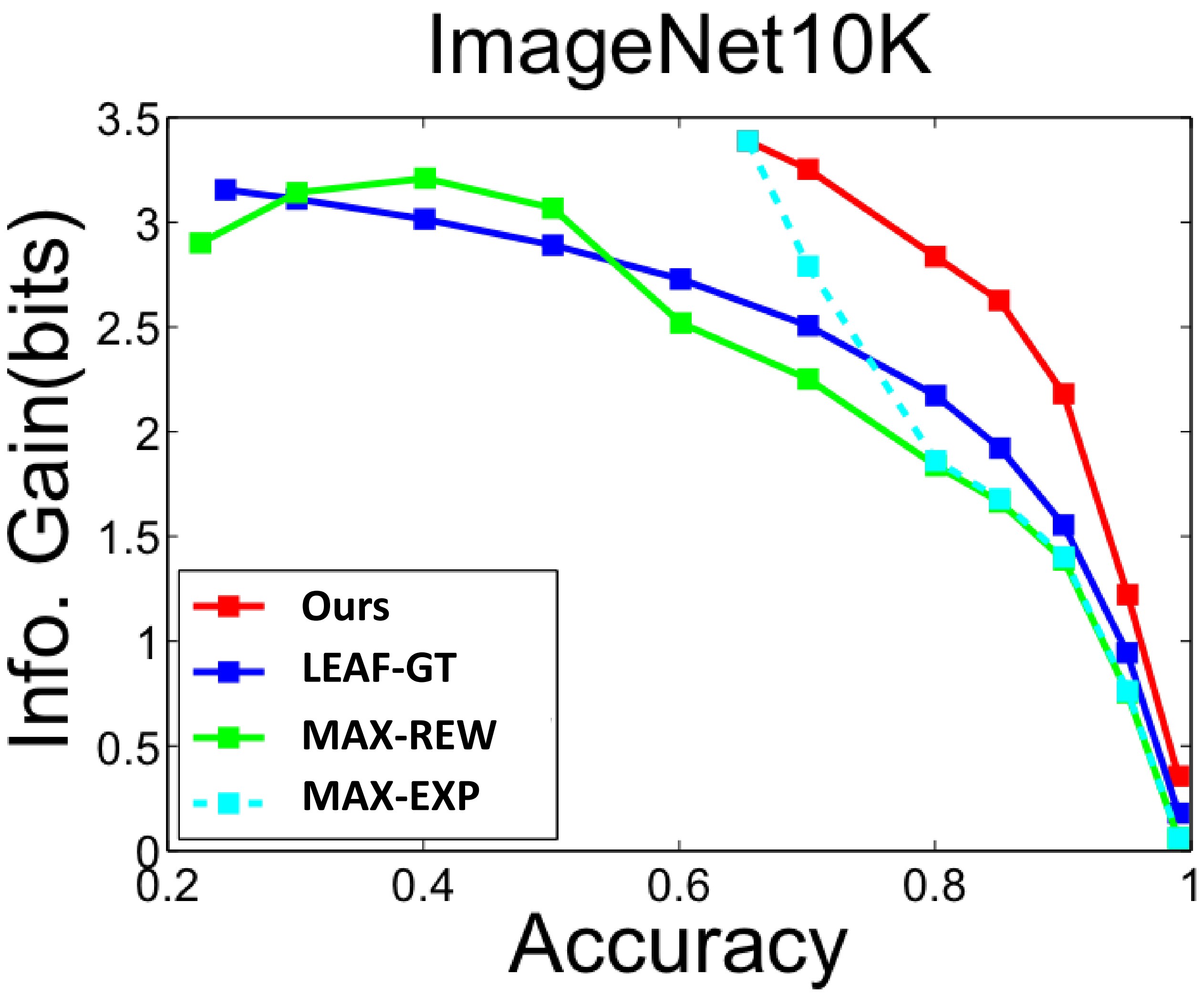
- Find a ***decision rule f***
  - Expected accuracy ***A(f)*** is at least ***1-ε***
  - Maximize expected reward ***R(f)***

$$\underset{f}{\textit{Maximize}} \quad R(f)$$

$$\textit{Subject to} \quad A(f) \geq 1 - \epsilon$$



Deng, Krause, Berg, Fei-Fei, CVPR2012



Deng, Krause, Berg, Fei-Fei, CVPR2012

EVA

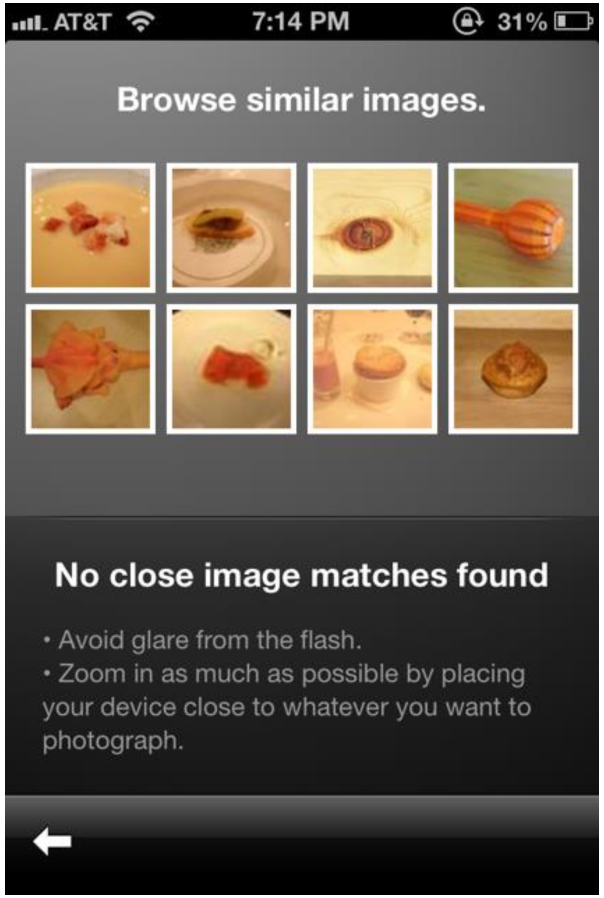
Engine for Visual Annotation

The EVA system, powered by **ImageNet**, can annotate images with guaranteed accuracies. It currently recognizes over **10,000** visual categories. See the project page to find out more.

Paste a URL | Upload an image

ANNOTATE

Google Goggles
Use pictures to search the web.

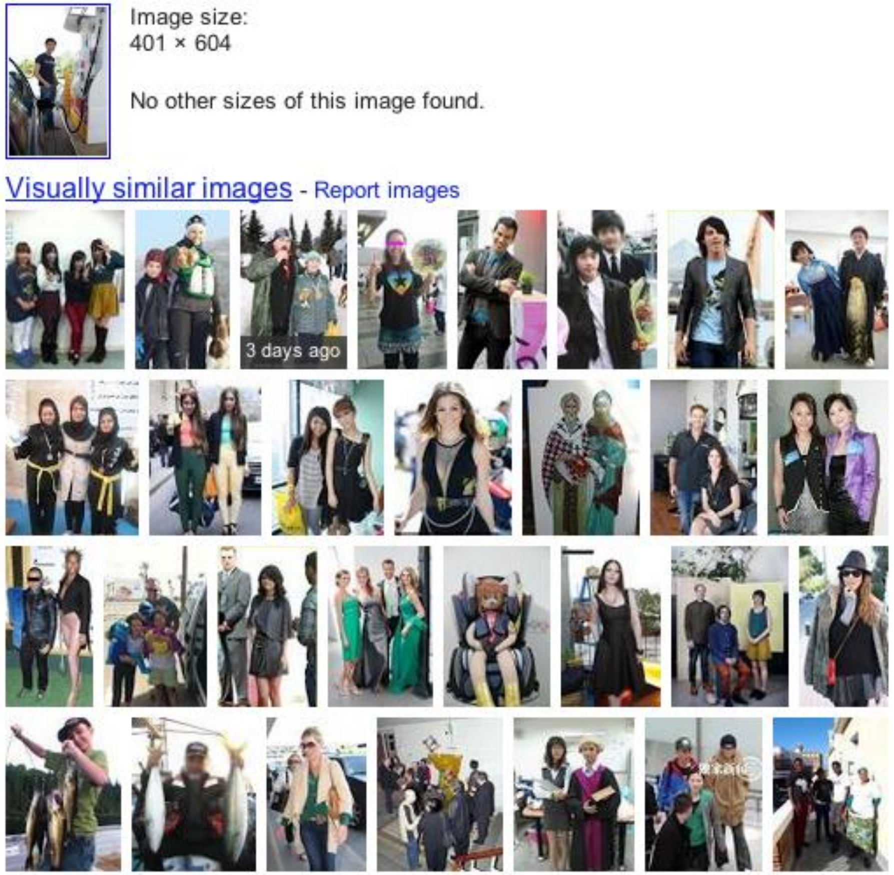


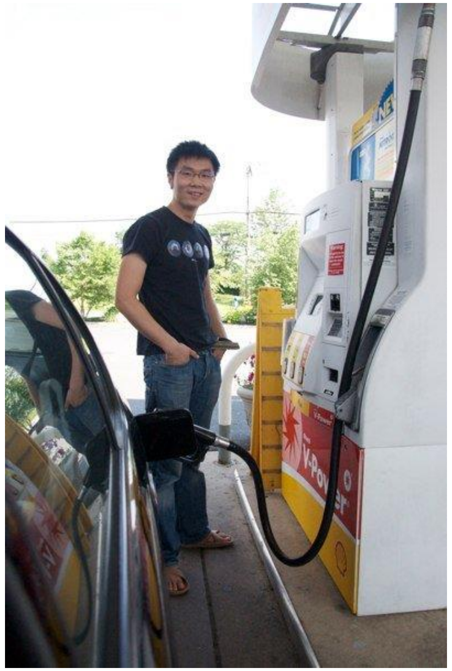
EVA
Engine for Visual Annotation

0.95 coffee mug
0.97 mug
0.99 drinking vessel

Google images

EVA
Engine for Visual Annotation





0.87 face , gas pump, person

0.90 face , gas pump

EVA

Engine for Visual Annotation

0.75 artifact, crater, matter, vertebrate

0.77 crater, matter, vertebrate

0.78 chordate, crater, matter

0.86 animal, matter

0.87 animal

Agenda

How to build a large-scale recognition engine using big data

- **STEP 1:** Build a Large Knowledge Base (ImageNet)
- **STEP 2:** Fine-Grained Recognition (Bubbles)
- **STEP 3:** Putting a label on “everything” (Hedging)

Conclusion & Future Work

- **Harvesting Knowledge**
  - Crowd-Machine Collaboration
  - Visual Representation
  - Active Learning
- **Visual Turing Test**
  - Vision and Language
  - Visual Reasoning
- **Managing Big Visual Data**
  - Large-Scale Learning
  - Indexing and Retrieval
- **Knowledge Transfer**
  - Exploiting Data Biases
  - Domain Adaptation
- **Mining Big Visual Data**
  - Visual Knowledge Graph
  - Social Media

Conclusion & Future Work

- **Harvesting Knowledge**
  - Crowd-Machine Collaboration
  - Visual Representation
  - Active Learning
- **Visual Turing Test**
  - Vision and Language
  - Visual Reasoning
- **Managing Big Visual Data**
  - Large-Scale Learning
  - Indexing and Retrieval
- **Knowledge Transfer**
  - Exploiting Data Biases
  - Domain Adaptation
- **Mining Big Visual Data**
  - Visual Knowledge Graph
  - Social Media

Thank you!

Prof. Kai Li
Princeton U.

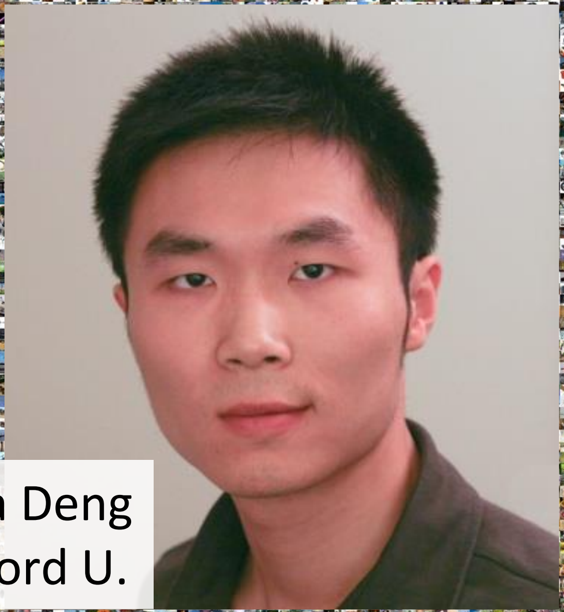
Prof. Alex Berg
Stony Brook U.

Sanjeev Satheesh
Stanford U.

Jonathan Krause
Stanford U.

Zhiheng Huang
Stanford U.

Olga Russakovsky
Stanford U.



Dr. Jia Deng
Stanford U.