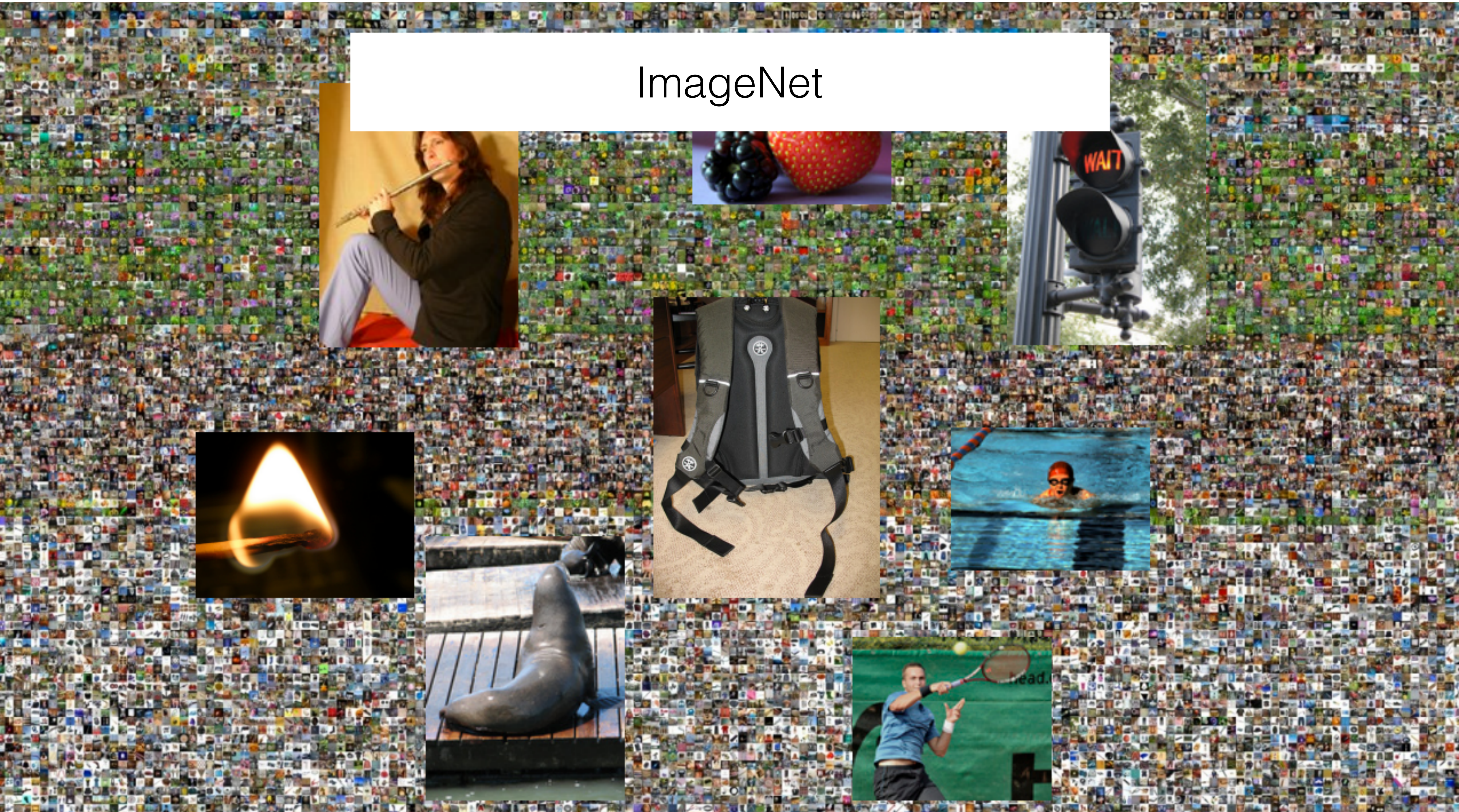
# Exploiting noisy web data for large-scale visual recognition

Lamberto Ballan

*University of Padova, Italy*

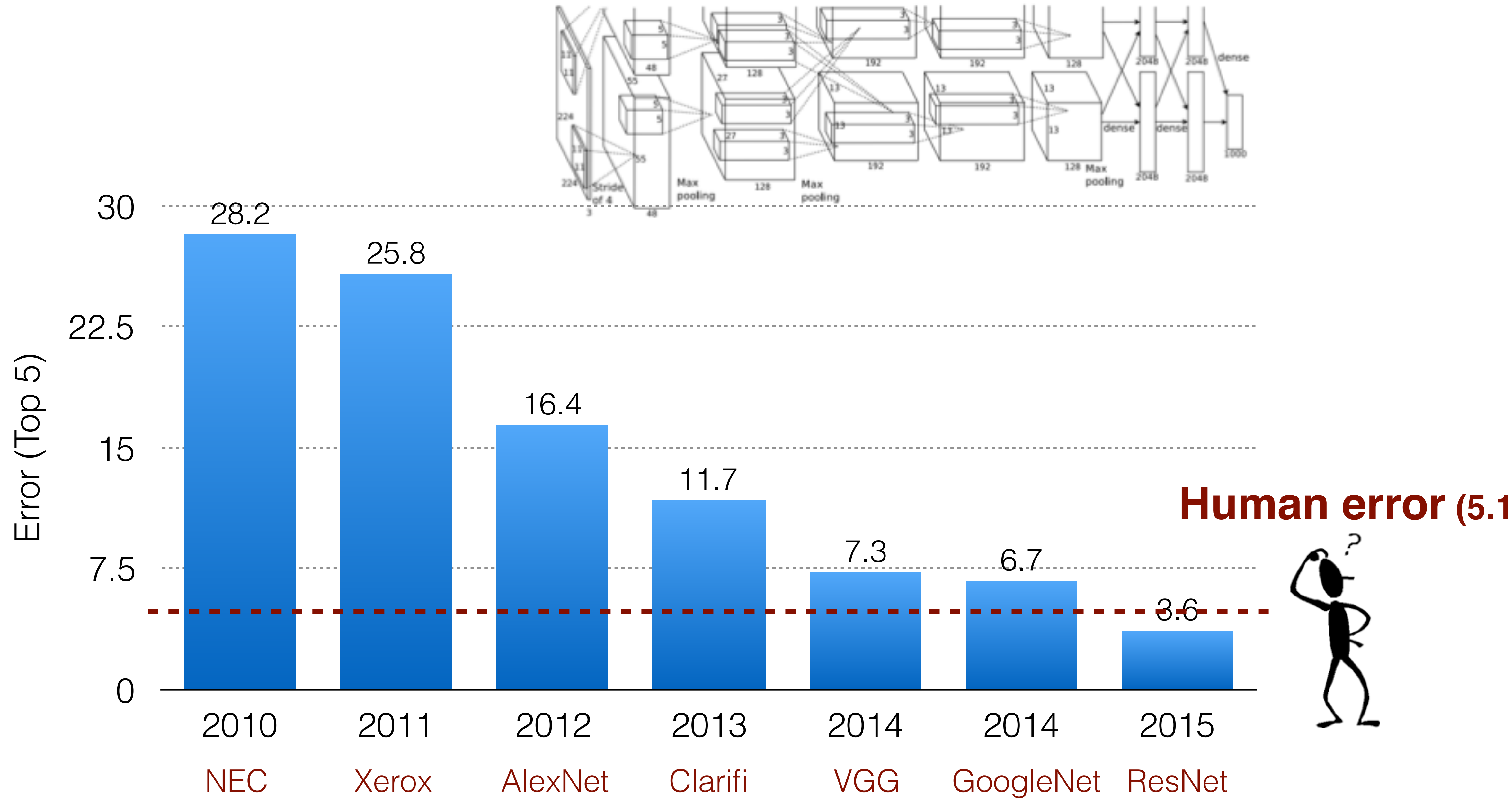
CVPRW WebVision - Jul 26, 2017

# Datasets drive computer vision progress



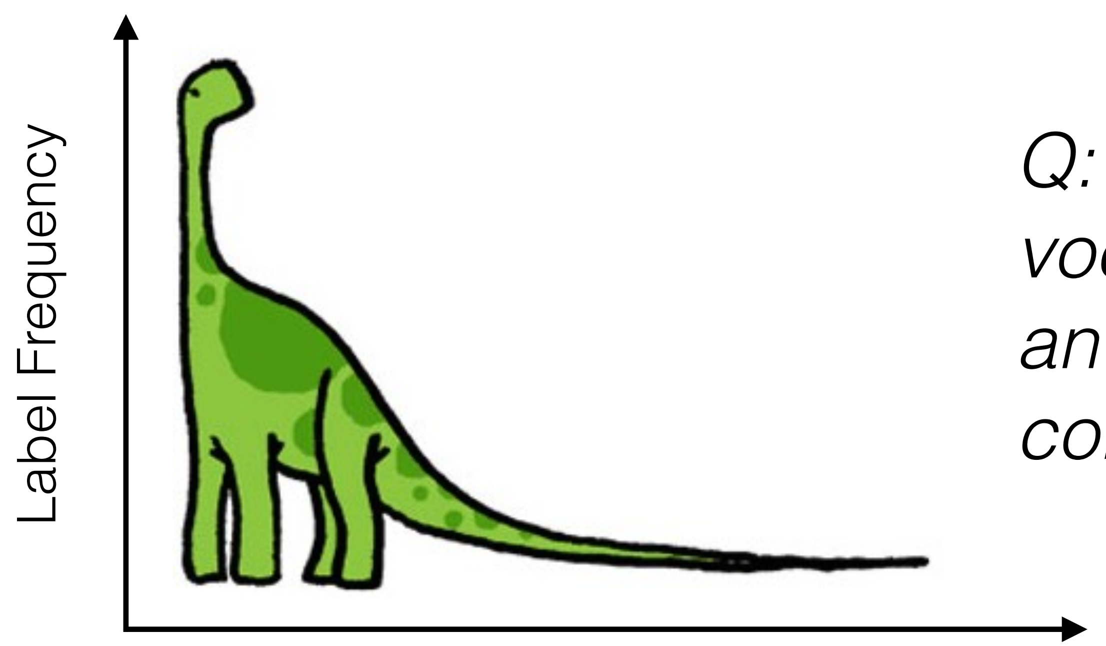
# ImageNet: ILSVRC results

- Result in ILSVRC (classification) over the years



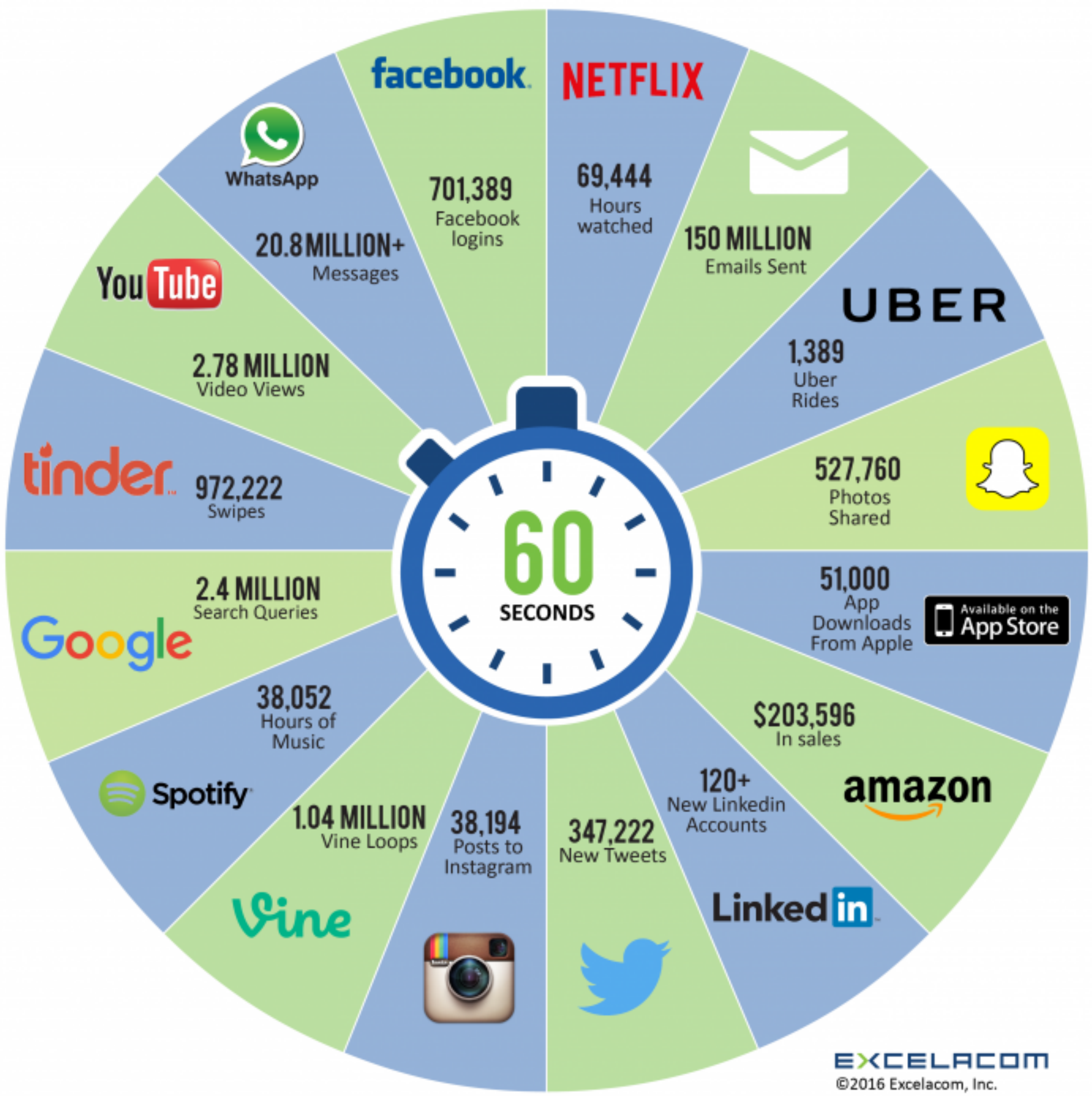
# The long tail

- A small number off generic objects/entities/labels appear very often while most others appear rarely
- There are a few real-world scenarios in which we have access to 1M+ images uniformly belonging to a set of 1000+ classes



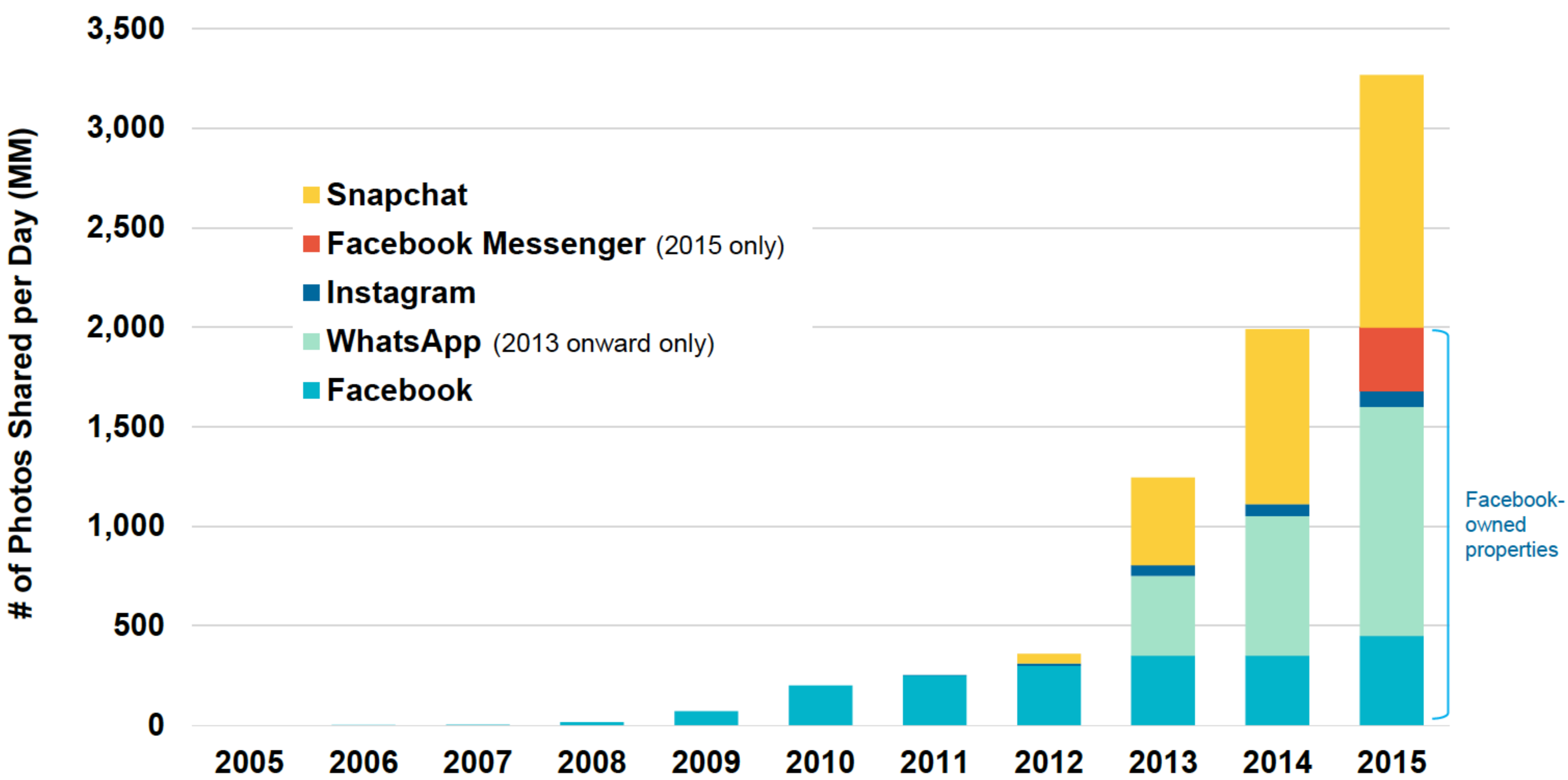

*Q: How to scale up to very large vocabularies (infrequent labels) and a scenario where it is hard to collect ground truth data?*

# Images want to be shared




*Almost all these services allow users to tag, rate, like, and swipe photos*

# Daily number of shared photos



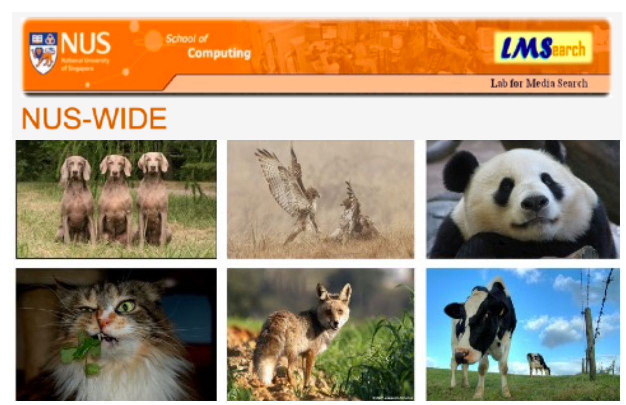



Source: Mary Meeker Internet Trends 2016

# What datasets?

- Ideally the entire Web!
- In practice:


MIRFLICKR-25K
MIRFLICKR-1M




NUS-WIDE
(~260K Flickr images)


YFCC100M
(100M Flickr images)


WebVision
(~2.4M images)

*2008* *2009* *2015* *2017*

# Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement and Retrieval

**Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement, and Retrieval**

XIRONG LI, Renmin University of China
TIBERIO URICCHIO, University of Florence
LAMBERTO BALLAN, University of Florence, Stanford University
MARCO BERTINI, University of Florence
CEES G. M. SNOEK, University of Amsterdam, Qualcomm Research Netherlands
ALBERTO DEL BIMBO, University of Florence

Where previous reviews on content-based image retrieval emphasize what can be seen in an image to bridge the semantic gap, this survey considers what people tag about an image. A comprehensive treatise of three closely linked problems (i.e., image tag assignment, refinement, and tag-based image retrieval) is presented. While existing works vary in terms of their targeted tasks and methodology, they rely on the key functionality of tag relevance, that is, estimating the relevance of a specific tag with respect to the visual content of a given image and its social context. By analyzing what information a specific method exploits to construct its tag relevance function and how such information is exploited, this article introduces a two-dimensional taxonomy to structure the growing literature, understand the ingredients of the main works, clarify their connections and difference, and recognize their merits and limitations. For a head-to-head comparison with the state of the art, a new experimental protocol is presented, with training sets containing 10,000, 100,000, and 1 million images, and an evaluation on three test sets, contributed by various research groups. Eleven

# What has been done so far, our survey

- An open test-bed for benchmarking image tagging, tag refinement and image retrieval approaches:
  - Jingwei: https://github.com/li-xirong/jingwei
  - Implemented 10 methods; train on 10K-100K-1M Flickr images, test on MIRFLICKR-25K and NUS-WIDE
- Taxonomy of previous works for image tagging:
  - learning: *instance-based*, *model-based*, *transduction*
  - media/modality: *tag*, *tag+image*, *tag+image+user*

[X.Li*, T.Uricchio*, **L.Ballan**, M.Bertini, C.Snoek, A.DelBimbo - ACM CSUR 2016 (*equal contrib.)]

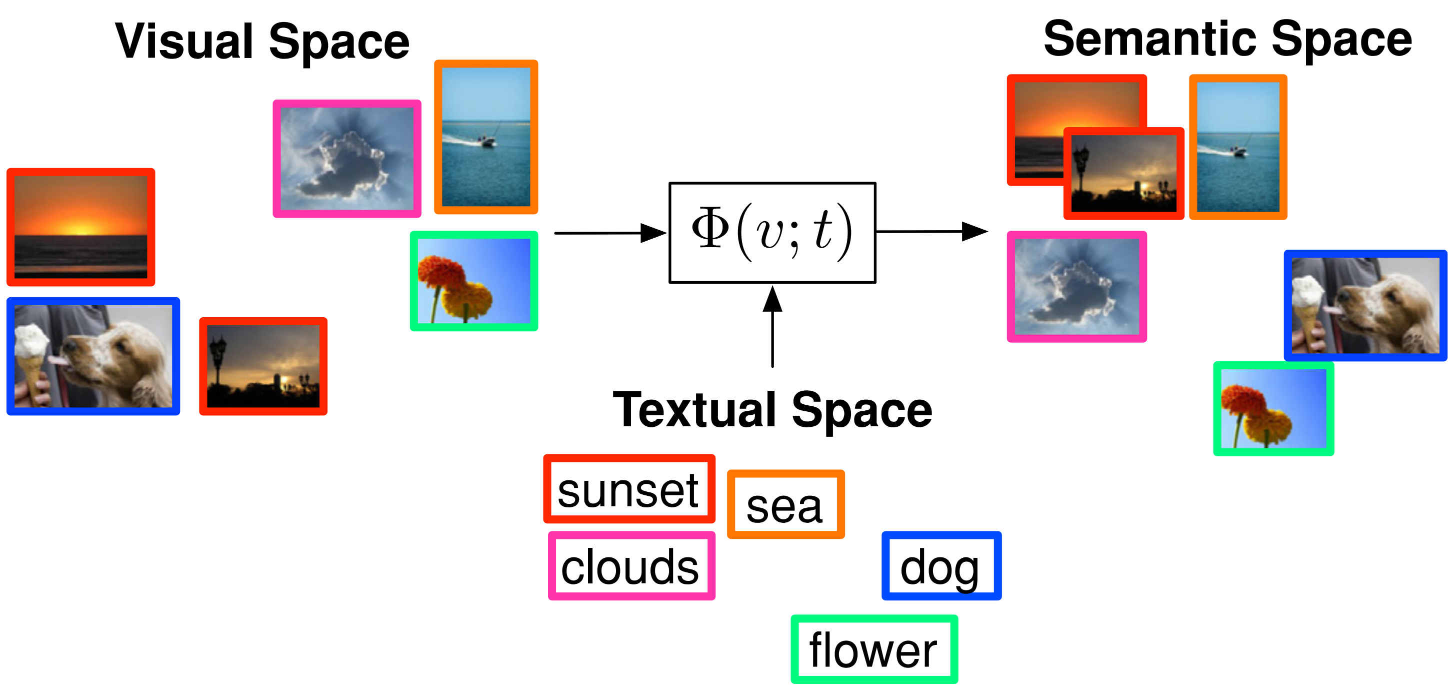
# Instance-based a.k.a. lazy learning

- A popular line of works is based on non-parametric models where labels are transferred to new samples
  - *"Images similar in appearance are likely to share labels"*
  - *e.g.* JEC [IJCV'10], TagProp [CVPR'09], 2PKNN [ECCV'12,IJCV'17]
  - works also in a cross-domain img2video scenario [CVIU'15]
- **pros:** can adapt to new labels and large vocabularies
- **cons:** *i)* it is a memory-based approach so it does not scale well at test time; *ii)* frequent labels dominate

[**L.Ballan**, M.Bertini, G.Serra, A.DelBimbo - CVIU 2015]

# Label transfer in the semantic space

- Labels associated to the training images can be used to re-arrange the original features space




*i) CCA-based embedding (expert labels or tags)*

*ii) Label transfer in the "semantic space"*

*iii) significant improvements in performance at low cost*

[T.Uricchio, **L.Ballan**, L.Seidenari, A.DelBimbo - PR 2017]

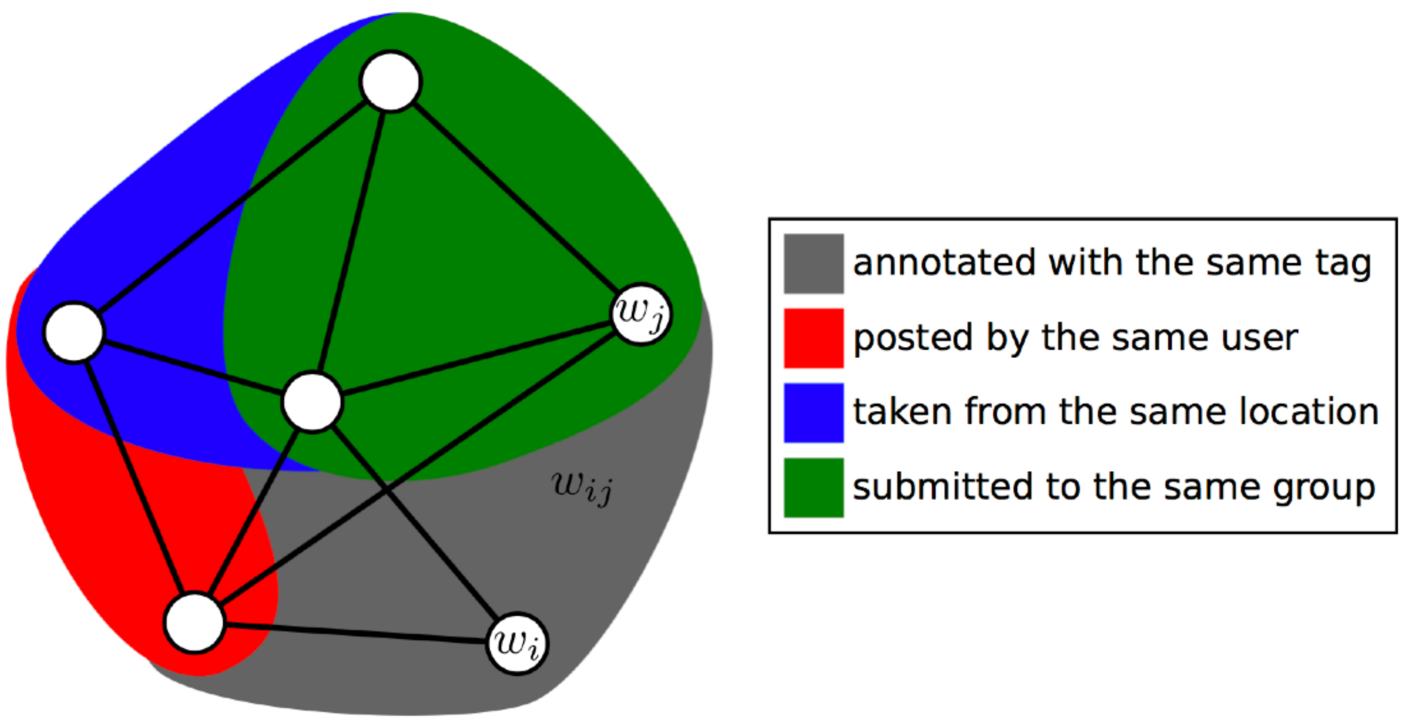
# Model-based methods

- Learn a model for each label following a (usually) fully supervised approach
  - i.e. train a large CNN network
  - e.g. WARP: deep convolutional ranking for multi-label image annotation [ICLR'14]
  - state-of-the-art results on NUS-WIDE using an AlexNet architecture trained on Flickr images and ranking loss

[Y.Gong, Y.Jia, T.Leung, A.Toshev, S.Ioffe - ICLR 2014]

# Web images are not only pixels

- Can we use contextual information such as social-network metadata to improve image classification?
  - Image Labeling on a Network: Using Social-Network Metadata for Image Classification [ECCV'12]




*Relational network model where each node represents an image, with cliques formed from images sharing common properties*

[J.McAuley, J.Leskovec - ECCV 2012]

# Automatic image annotation by exploiting image metadata and weak labels

[J.Johnson*, **L.Ballan***, L.Fei-Fei - ICCV 2015 (*equal contrib.)]

# Motivation

- Can you guess what's in the image?


petal?

fruit?

tentacle?

# Motivation

- Let's try to add more context…


**Tags:**

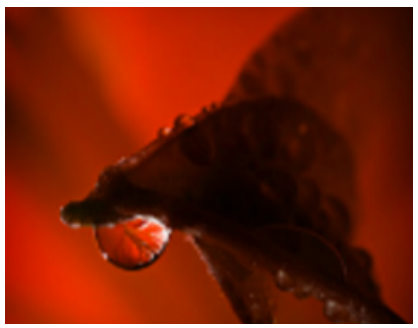
*flower*
*petal*
*closeup*
*water*

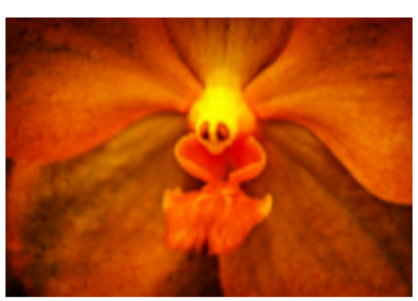
**GPS**
**groups**
**…**

flickr

# Motivation

- In the context of images which share similar metadata it is easier to give the right answer





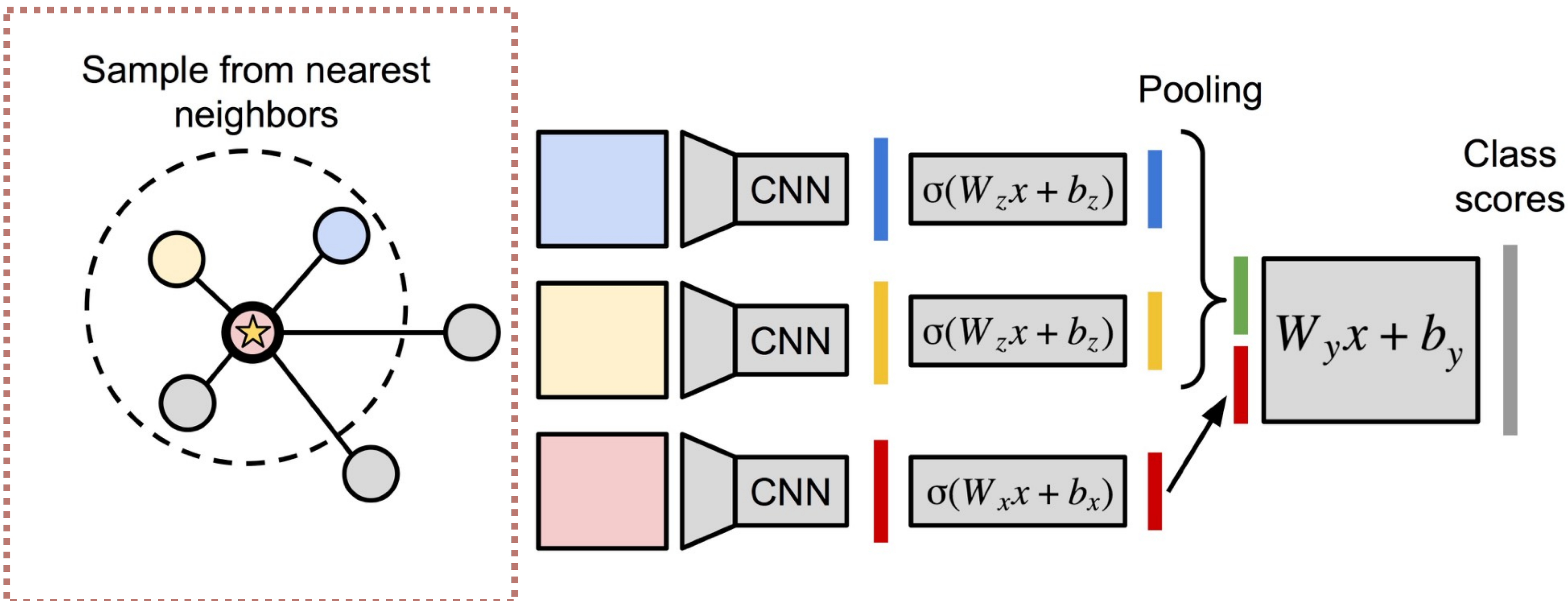


# Approach

- For an image $x \in X$ and neighborhood $z \in Z_x$, we use a function $f$ parameterized by $w$ to predict labels
  - We compute hidden state representations for the image and its neighbors
  - Then we operate on the concatenation of these two representations to compute label scores
- We demonstrate that our model can:
  - handle *different types of image metadata*
  - adapt to *changing vocabularies*

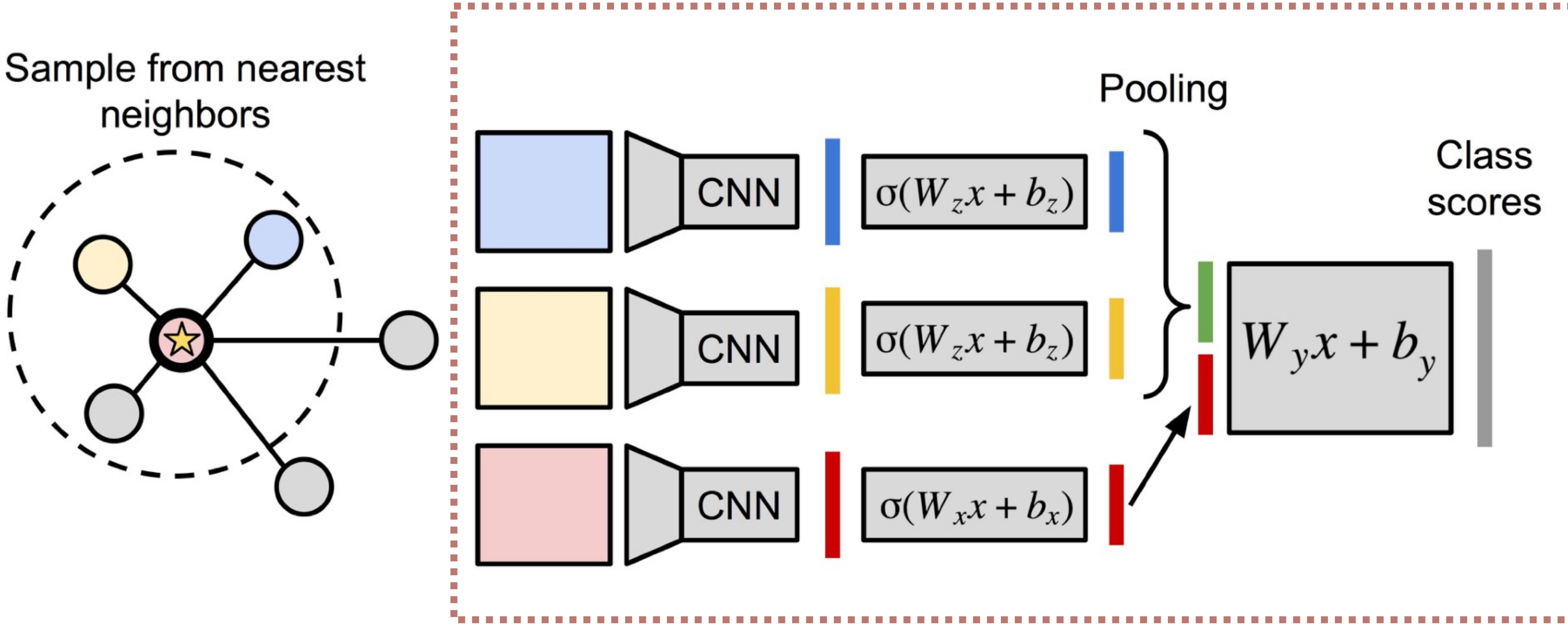
# Approach

- (1) *non-parametric* step to build a neighborhood



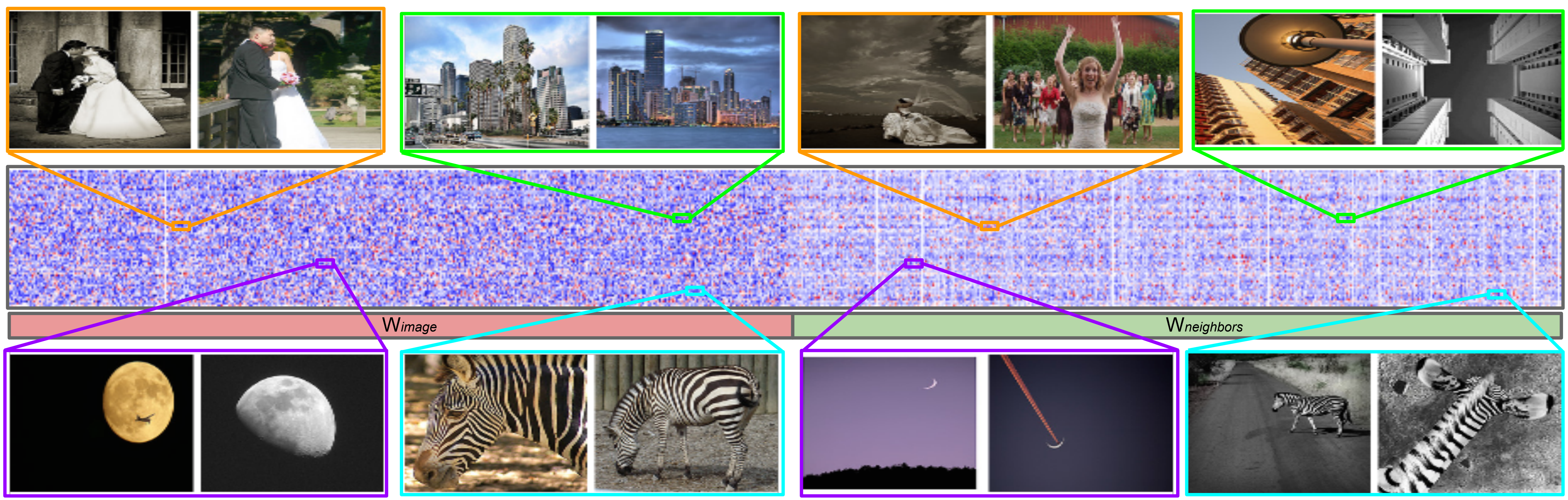
# Approach

- (2) *deep neural network* to blend visual information from the image and its neighbors



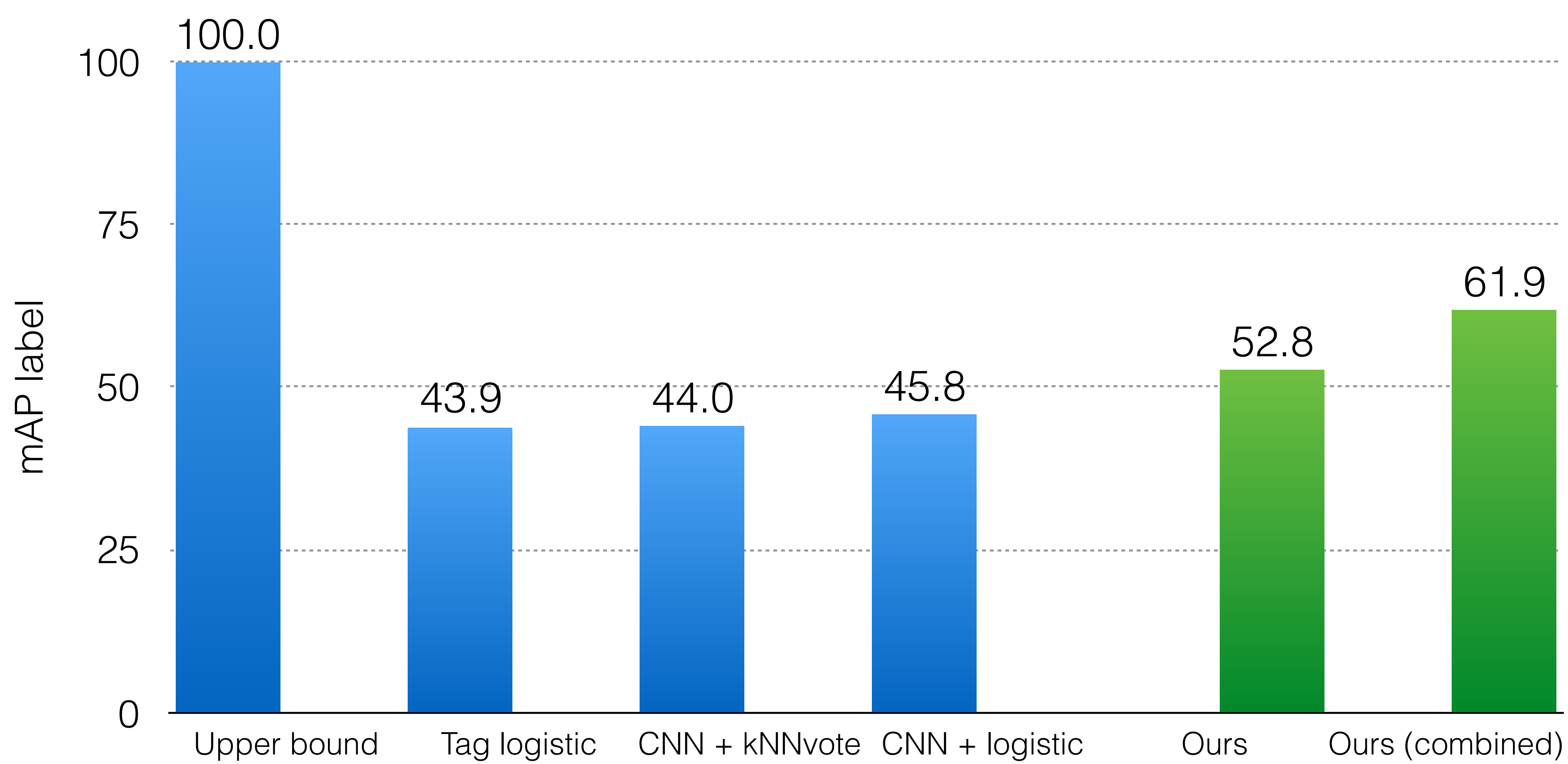
# Approach

- In this way the model uses features from both the image and its neighbors



# Results

- Multi-label image annotation results on the NUS-WIDE dataset (~240K Flickr images)



# Results

- Multi-label image annotation results on the NUS-WIDE dataset (~240K Flickr images)

| Method | $mAP_L$ | $mAP_I$ | $Rec_L$ | $Prec_L$ | $Rec_I$ | $Prec_I$ |
|---|---|---|---|---|---|---|
| Tag-only Model + linear SVM [37] | 46.67 | - | - | - | - | - |
| Graphical Model (all metadata) [37] | 49.00 | - | - | - | - | - |
| CNN + softmax [15] | - | - | 31.22 | 31.68 | 59.52 | 47.82 |
| CNN + ranking [15] | - | - | 26.83 | 31.93 | 58.00 | 46.59 |
| CNN + WARP [15] | - | - | 35.60 | 31.65 | 60.49 | 48.59 |
| Upper bound | 100.00±0.00 | 100.00±0.00 | 68.52±0.35 | 60.68±1.32 | 92.09±0.10 | 66.83±0.12 |
| Tag-only + logistic | 43.88±0.32 | 77.06±0.14 | 47.52±2.59 | 46.83±0.89 | 71.34±0.16 | 51.18±0.16 |
| CNN [27] + kNN-voting [36] | 44.03±0.26 | 73.72±0.10 | 30.83±0.37 | 44.41±1.05 | 68.06±0.15 | 49.49±0.11 |
| CNN [27] + logistic (visual-only) | 45.78±0.18 | 77.15±0.11 | 43.12±0.39 | 40.90±0.39 | 71.60±0.19 | 51.56±0.11 |
| Image neighborhoods + CNN-voting | 50.40±0.23 | 77.86±0.15 | 34.52±0.47 | **56.05**±1.47 | 72.12±0.21 | 51.91±0.20 |
| Our model: tag neighbors | 52.78±0.34 | **80.34**±0.07 | 43.61±0.47 | 46.98±1.01 | 74.72±0.16 | **53.69**±0.13 |
| Our model: tag neighbors + tag vector | **61.88**±0.36 | 80.27±0.08 | **57.30**±0.44 | 54.74±0.63 | **75.10**±0.20 | 53.46±0.09 |

Table 2: Results on NUS-WIDE. Precision and recall are measured using $n = 3$ labels per image. Metrics are reported both per-label ($mAP_L$) and per-image ($mAP_I$). We run on 5 splits of the data and report mean and standard deviation.

# Qualitative results



Neighborhood

# Qualitative results



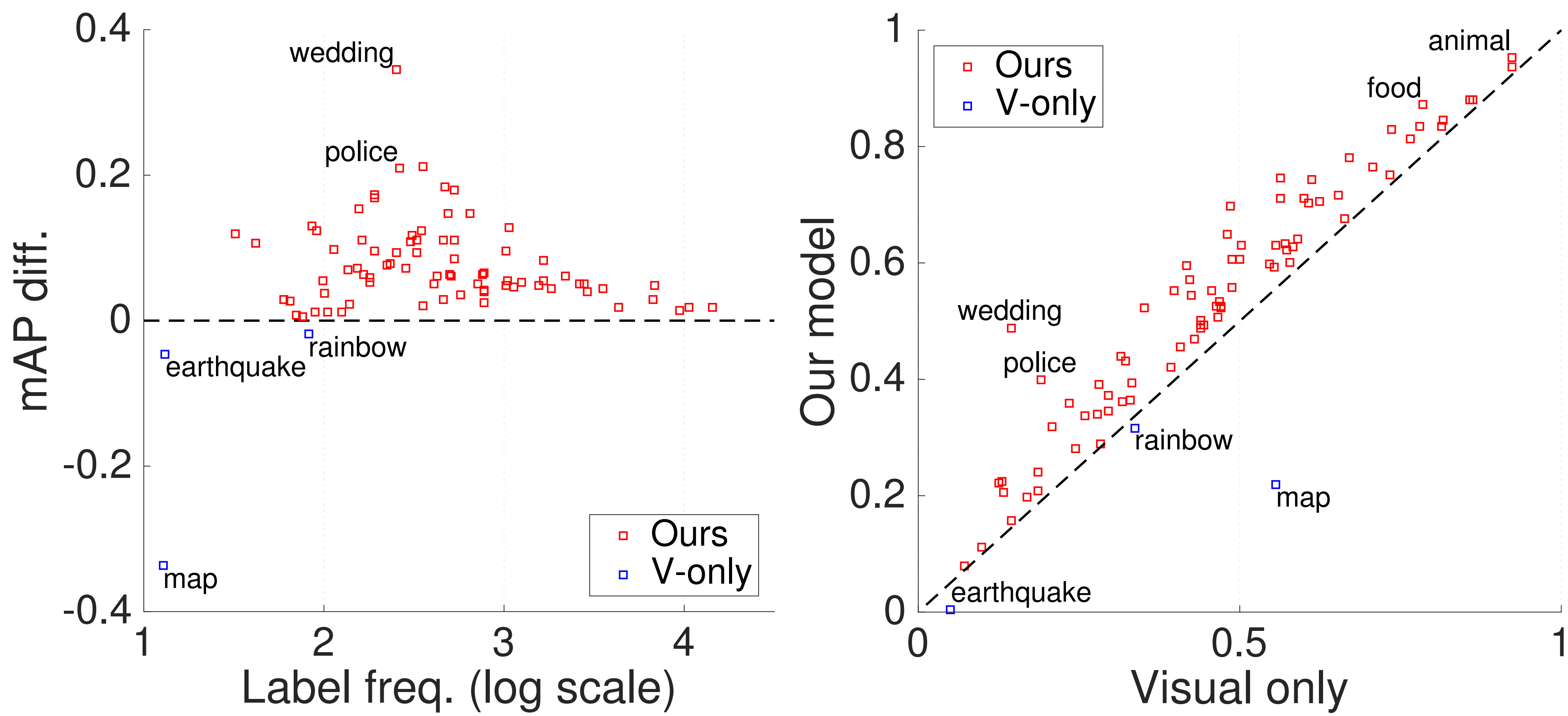
Neighborhood

# Qualitative results


Neighborhood

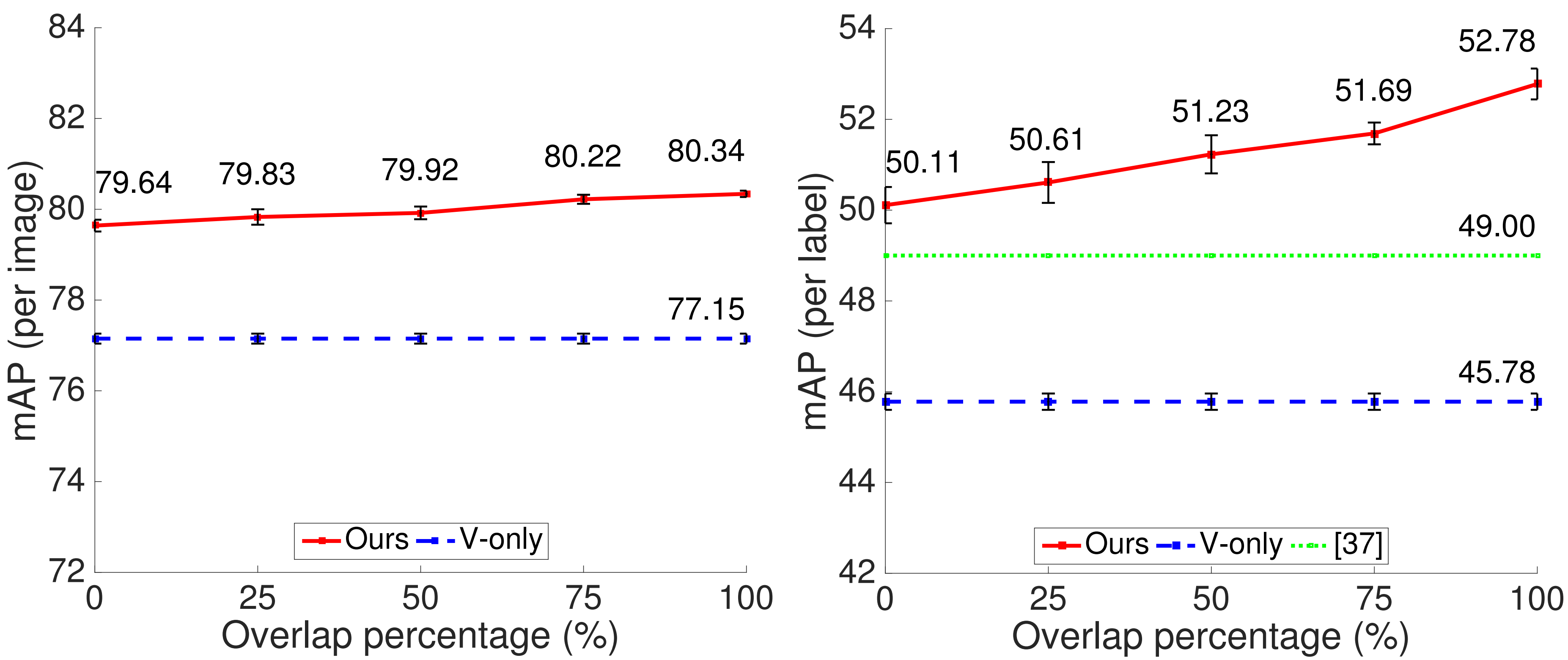
# Results: ours *vs* CNN baseline

- Experiment 1: evaluates AP for each label of our model vs the visual-only CNN baseline



# Results: generalization

- Experiment 2: vocabulary generalization



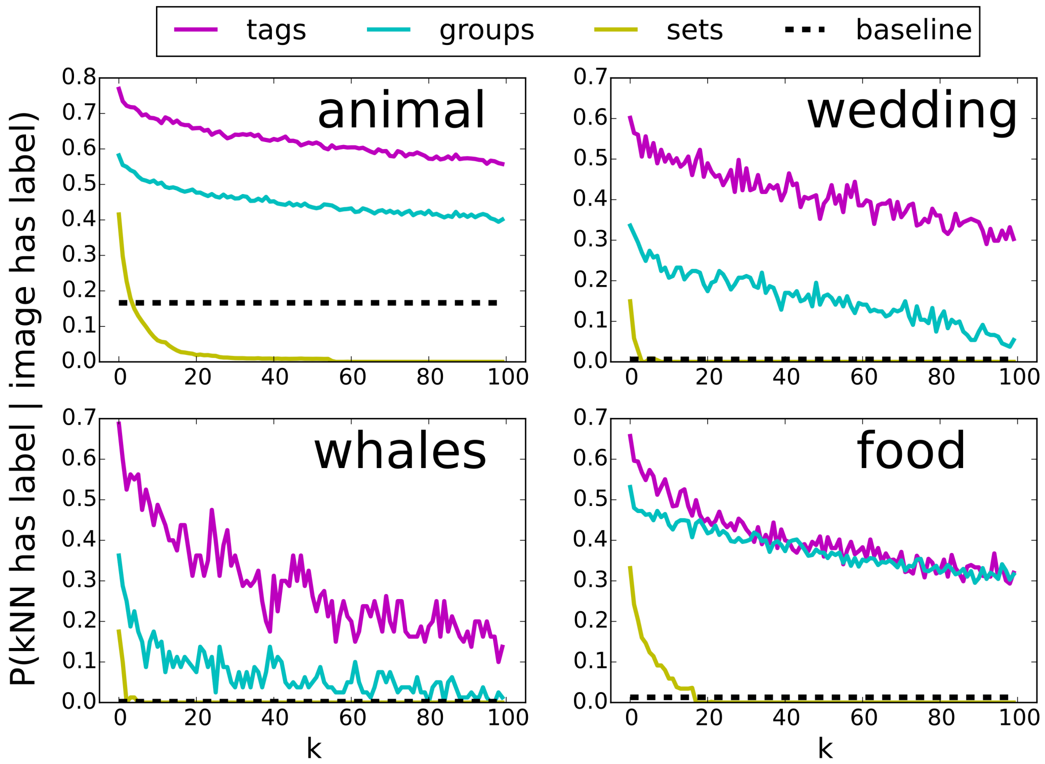

Performance as we vary overlap between tag vocabularies used for training and testing: strong results even in the case of disjoint vocabularies

# Results: generalization

- Experiment 3: metadata generalization

| Train: \ Test: | Tags | Sets | Groups |
|---|---|---|---|
| Tags | 52.78±0.34 | 47.12±0.35 | 48.14±0.33 |
| Sets | 52.21±0.29 | 48.02±0.33 | 48.49±0.16 |
| Groups | 50.32±0.28 | 47.82±0.24 | 48.87±0.22 |

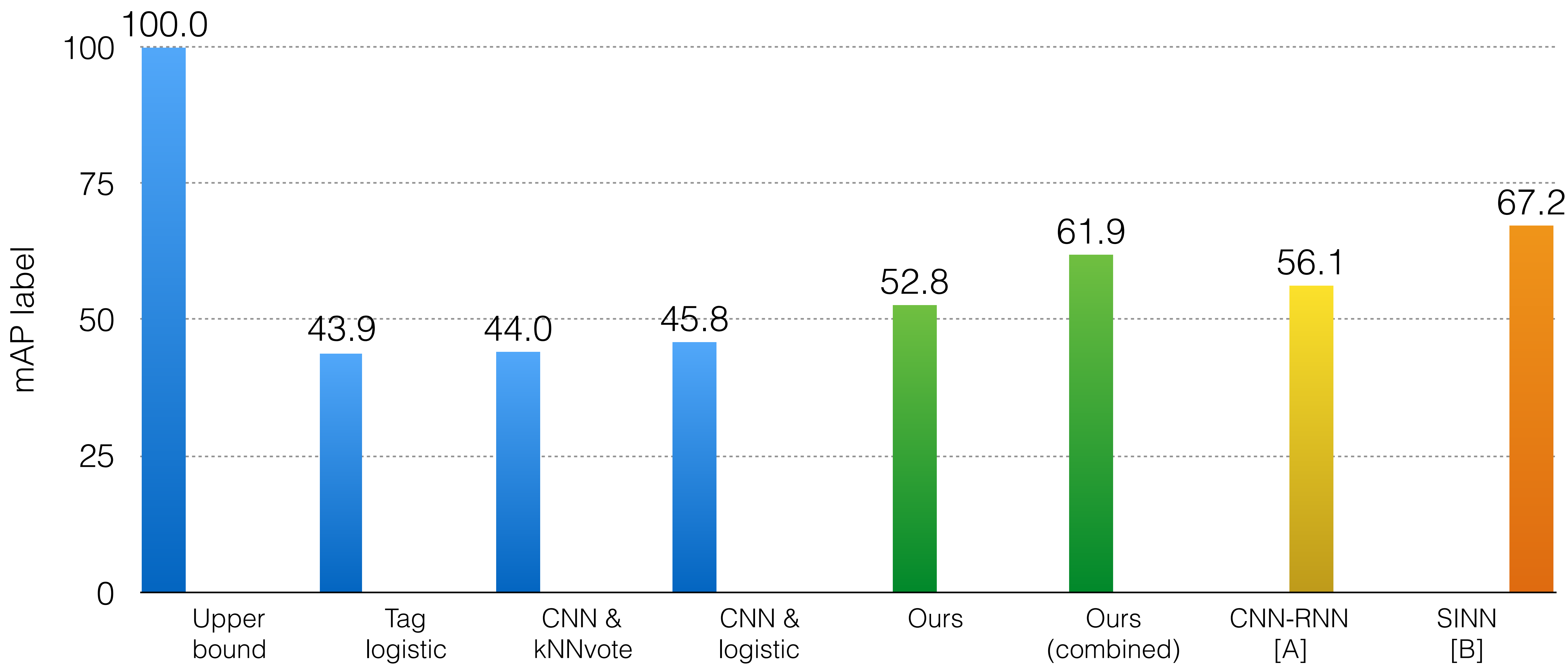
*Results using different types of metadata for training and testing*




*Probability that the k-th neighbor of an image has a label given that the image has the label*

# Results using label relations

- Other recent results on NUS-WIDE by learning label relations




[A] Wang, Yang, Mao, Huang, Huang, Xu - CVPR 2016
[B] Hu, Zhou, Deng, Liao, Mori - CVPR 2016

# Summary

- We really need better datasets and evaluation protocols to evaluate web-vision models
- Visual recognition and learning benefits from:
  - large collections of noisy web data
  - good results even when the model is forced to generalize to new types of metadata at test time

# Next steps

- Use a graph(network)-based representation to find image neighborhoods
- Explore new datasets with a larger label space and noisy annotations
- Visual data mining: share and infer properties based on image similarity on the network

# Acknowledgements

Xirong Li

Tiberio Uricchio

Marco Bertini

Cees Snoek

Alberto Del Bimbo

Lorenzo Seidenari

Justin Johnson

Li Fei-Fei

**Contact Info**

lamberto.ballan@unipd.it

www.lambertoballan.net