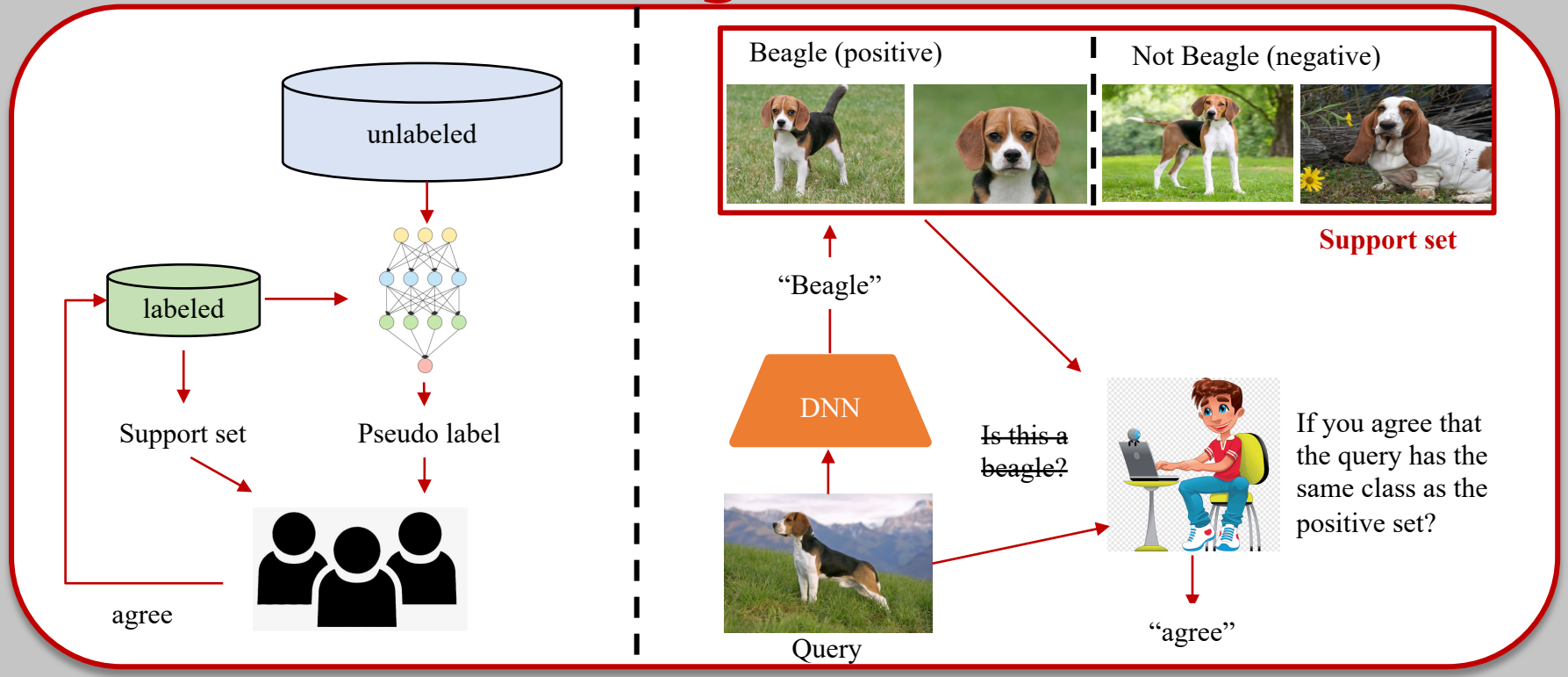
Towards Professional Level Crowd Annotation of Expert Domain Data

Pei Wang and Nuno Vasconcelos

Poster ID: TUE-AM-301

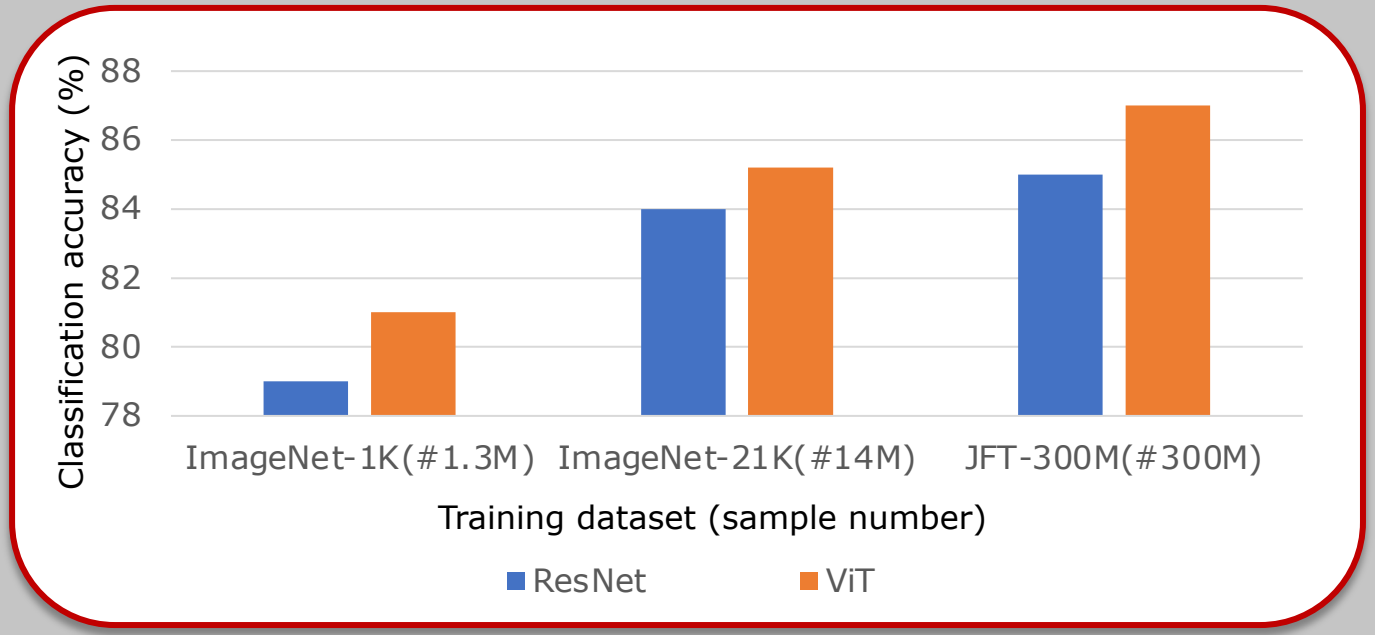
Statistical Visual Computing Lab, UC San Diego

Semi-supervised Learning with Human Filtering



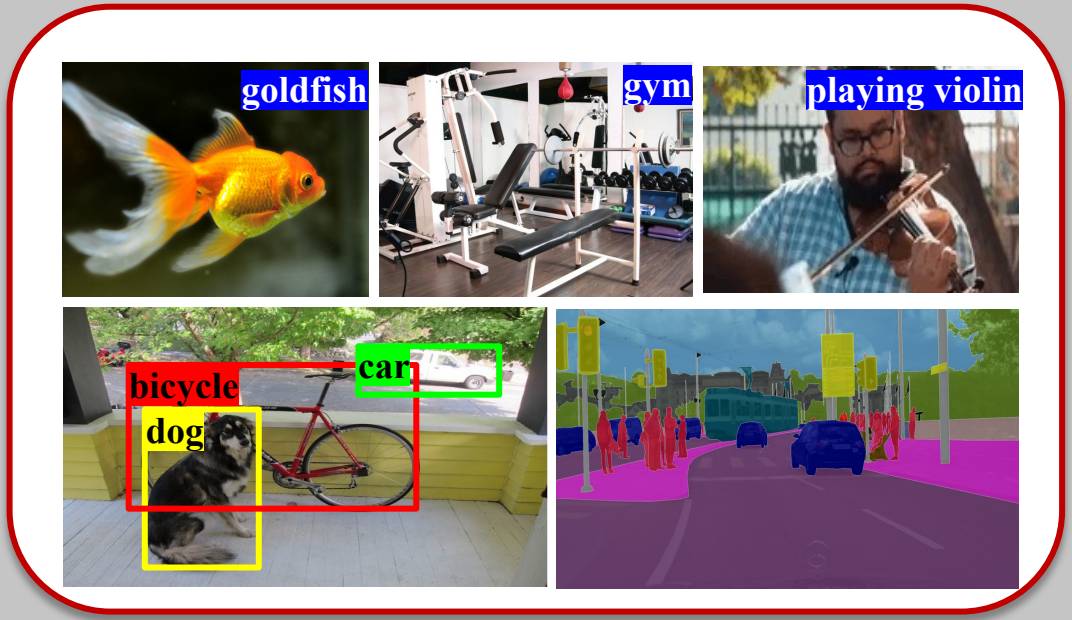
Benefits from Larger Datasets

- Deep classifiers can always benefit from larger datasets.



Common Datasets

- Contain common objects
- Labeled on crowd source platforms
- No prior knowledge needed for the annotation
- Annotation is cheap and easy
- Can be scaled up

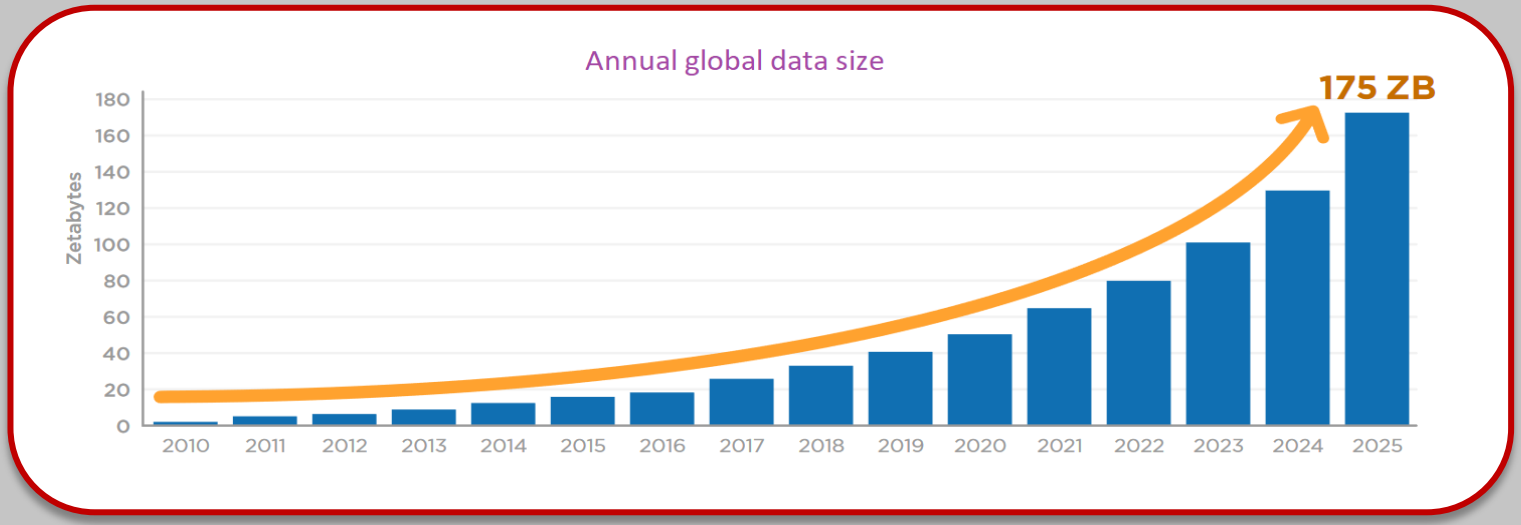


Expert Domain Datasets

- Classes are fine-grained
- Domain knowledge needed for the annotation
- Expert annotation is expensive and even inaccessible
- Difficult to be scaled up

The Era of Big Data

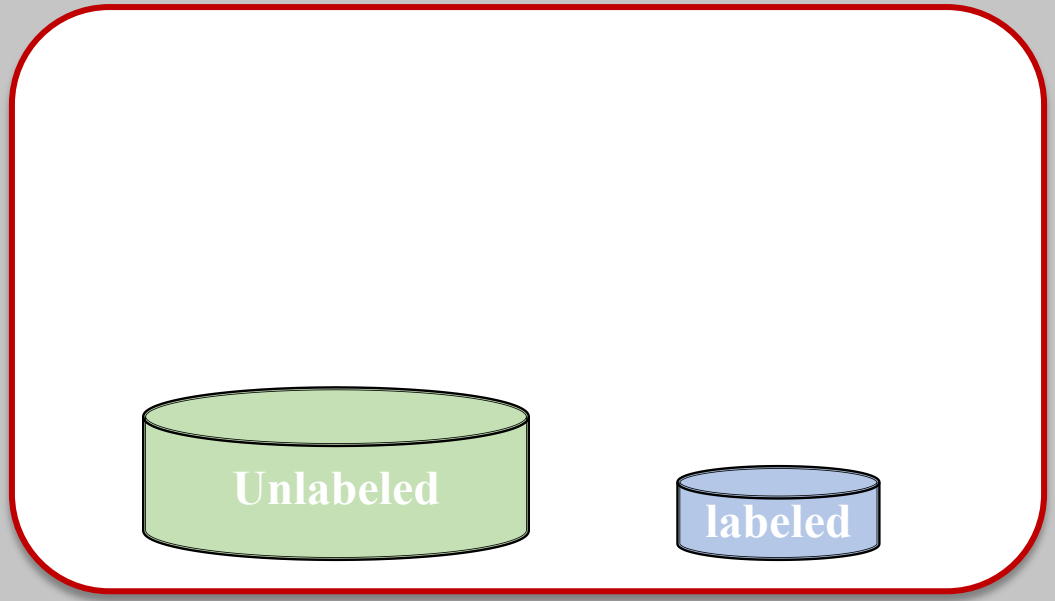
- Data volume grows exponentially.
- The difficulty is not to collect data but to label it.

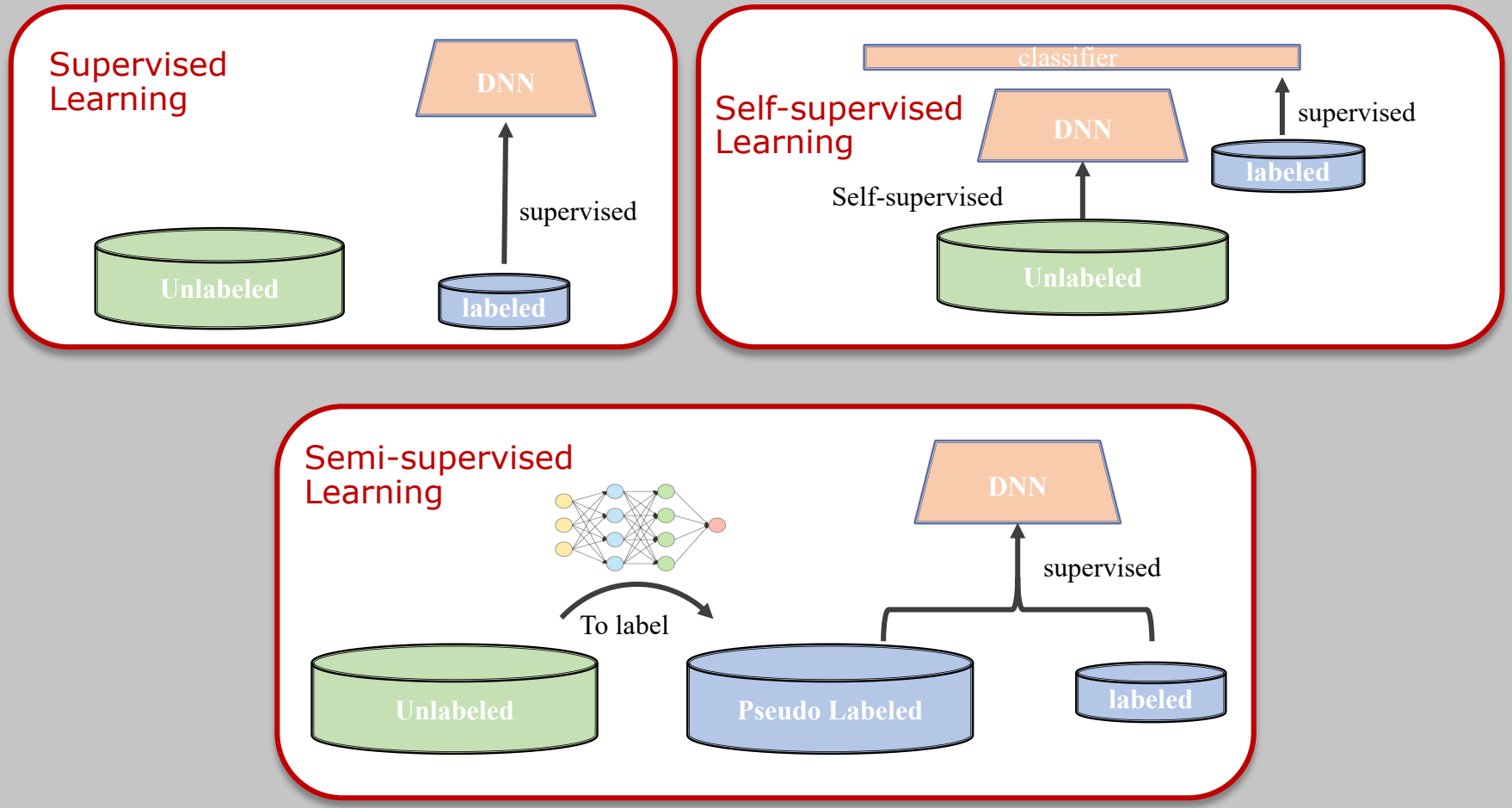


Data Age 2025 report, sponsored by Seagate with data from IDC Global Datasphere, 2018

How to Train a Good Fine-grained Classifier

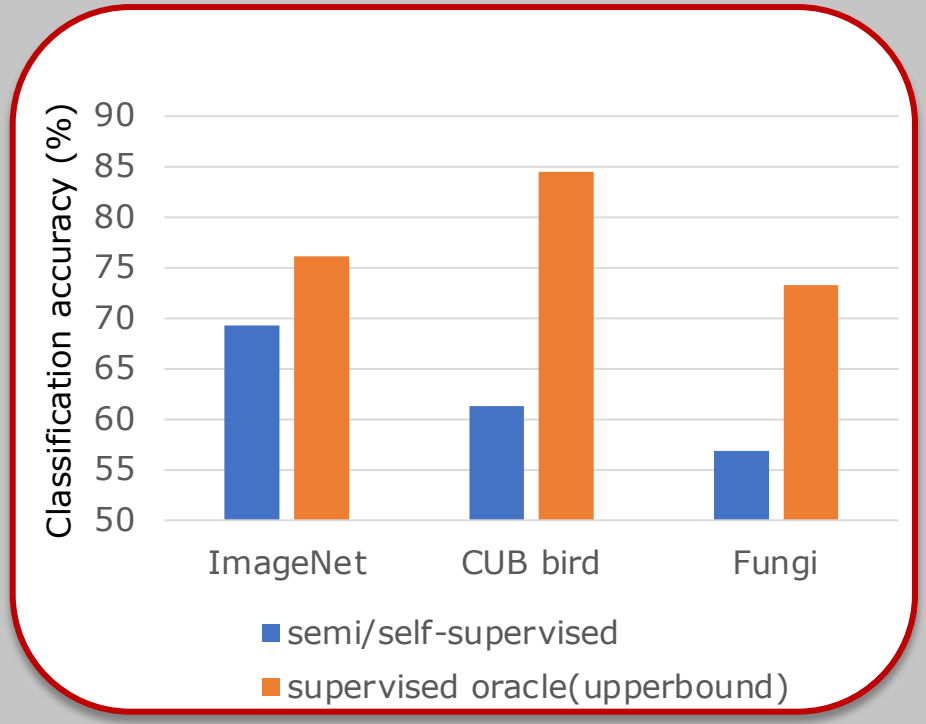
- In expert domains: large unlabeled + small labeled




Supervised Learning
DNN
supervised
Unlabeled
labeled
Self-supervised Learning
classifier
DNN
supervised
labeled
Self-supervised
Unlabeled
Semi-supervised Learning
DNN
supervised
To label
Unlabeled
Pseudo Labeled
labeled

Expert Domains

- Existing solutions
  - Supervised learning
  - Self-supervised learning
  - Semi-supervised learning
- OK on coarse-grained but collapse on fine-grained
- Fine-grained differences are very subtle; not well developed

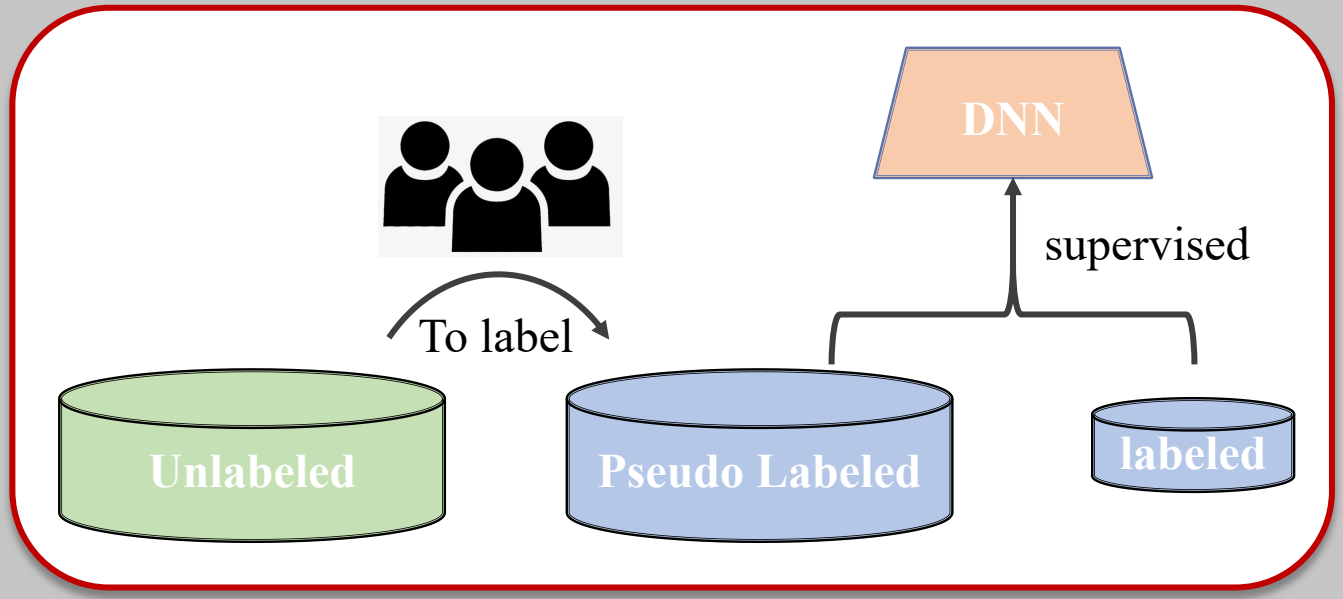


Motivation

- Self/Semi-supervised learning try to learn from the small labeled data.
- Upper bounded by supervised oracle with a big gap in expert domains.
- We have known that learning from large labeled data can work the best.
- An alternative solution - We make small labeled data bigger.
- We aim to scale up the annotation by crowd sourcing platforms

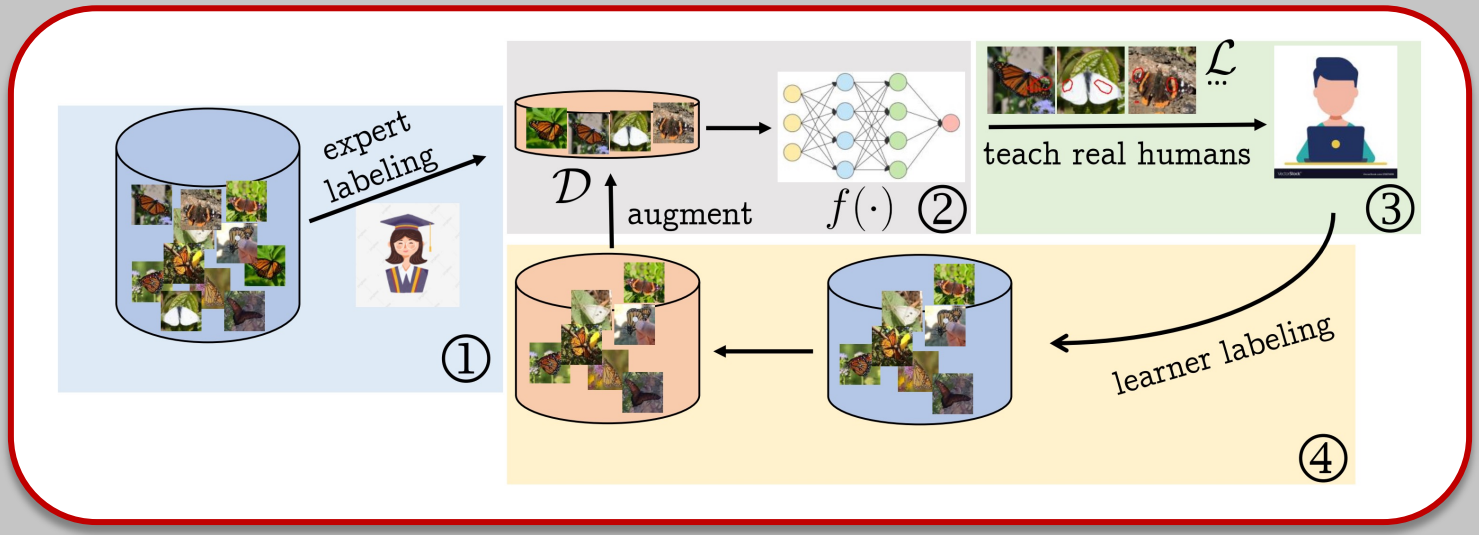
Crowd Source Annotation

- Pursue professional level pseudo labels for the unlabeled from crowdsourcing



A Existing solution: MEMORABLE

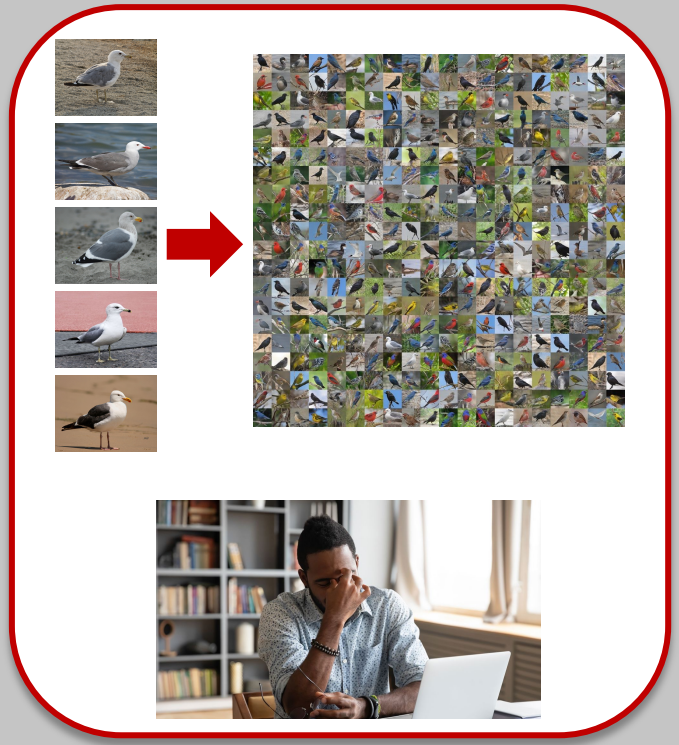
- Machine tEaching fraMewORk for scAlaBLe rEcognition (MEMORABLE)



Wang and Vasconcelos, A new machine teaching framework for scalable recognition, ICCV 2021

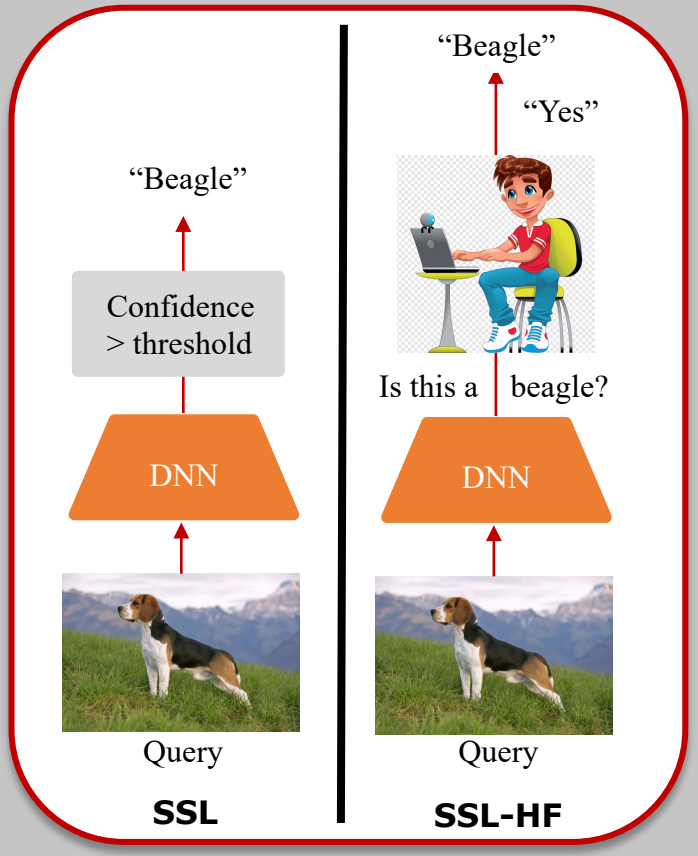
The Problems of MEMORABLE

- The difficulty and challenge of teaching a large amount of classes
  - Short-term memory
- Solution: humans act as filters, not classifiers



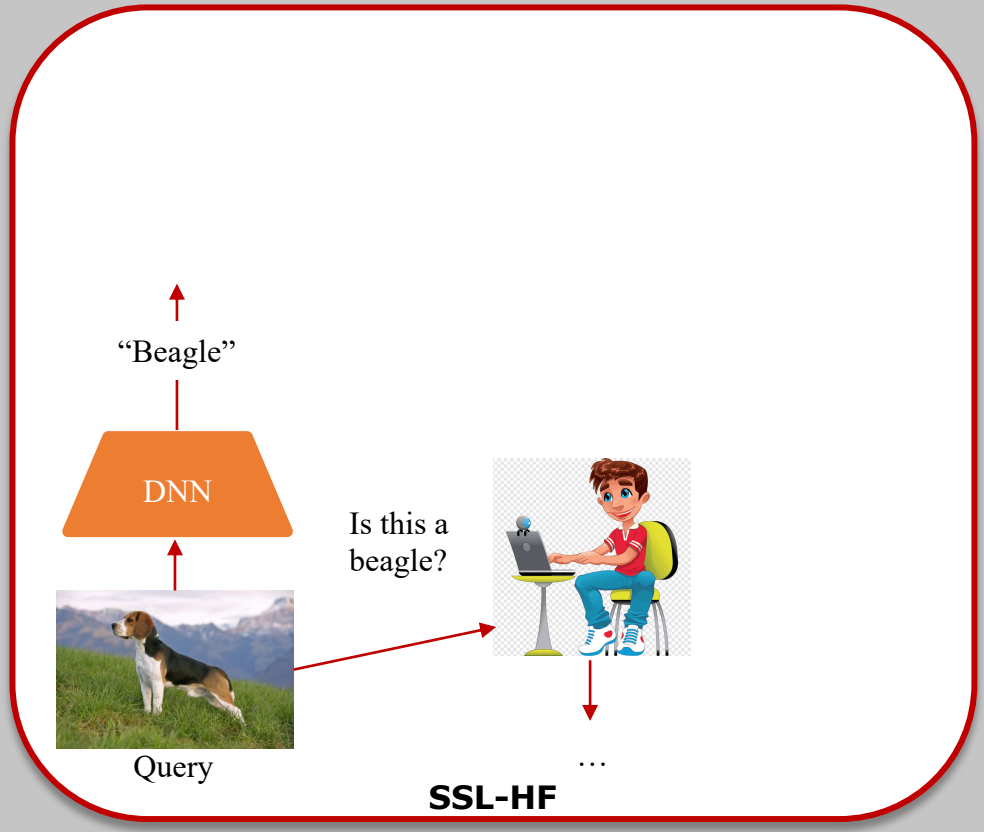
Inspiration

- Pseudo-label filtering is used in SSL by confidence thresholding
- SSL fails in fine-grained
  - bad accuracy when training set is small
  - Poor confidence score estimation
- Replace confidence thresholding with human filtering
  - Strong low-shot learning ability
  - Strong confidence calibration ability



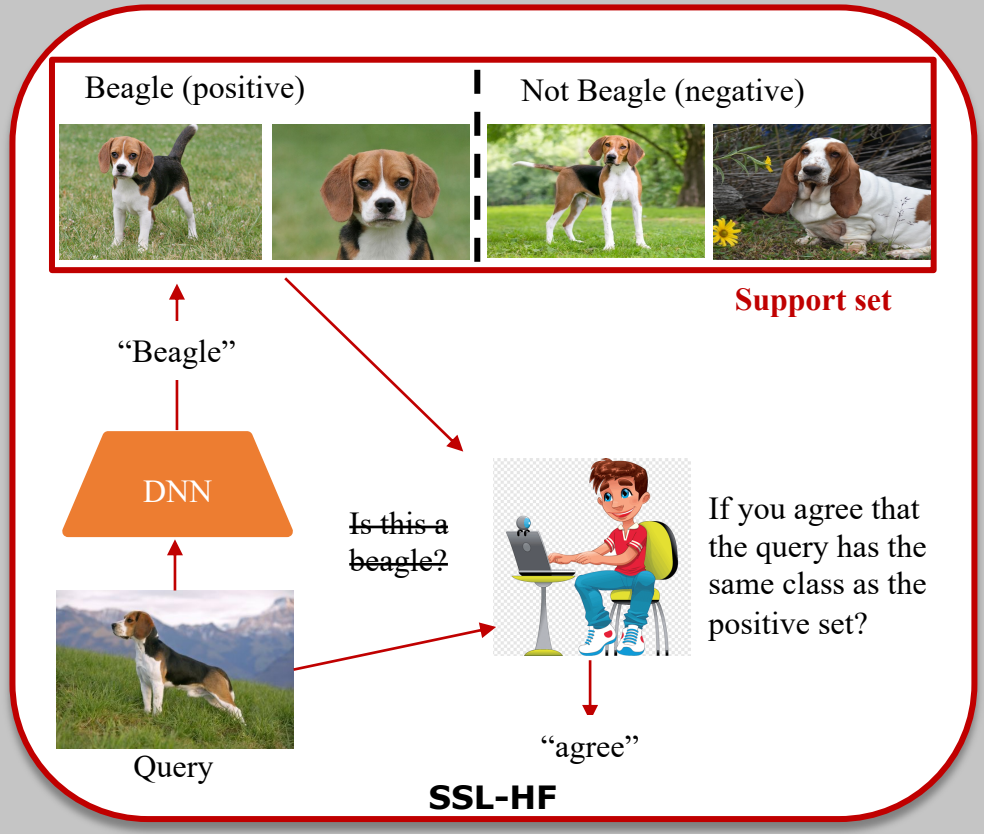
Difficulties

- Annotators may not know the fine-grained categories.



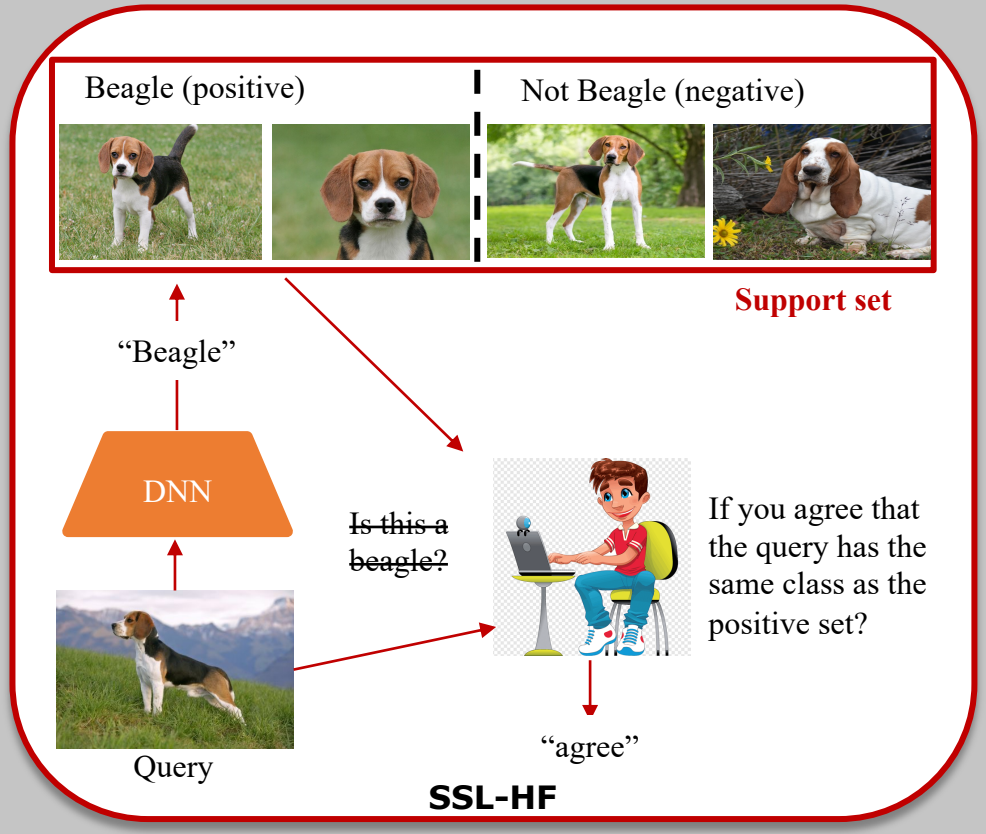
Difficulties

- Annotators may not know the fine-grained categories.
- Solutions
  - Ask question implicitly, by introducing the support set. 'Agree'/'Disagree'/'IDK



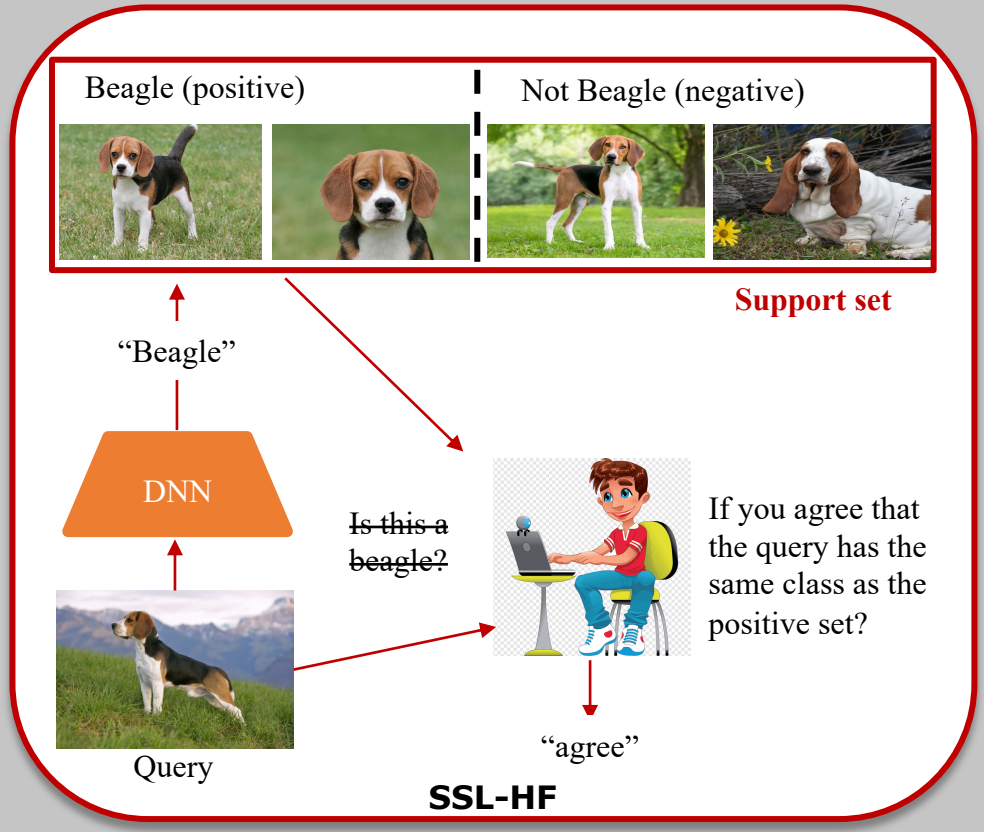
Difficulties

- Do not need to know fine-grained classes
- Just need to compare visual similarity
- No need to teach annotators fine-grained classes
- Eliminate the short-term memory constraints of machine teaching



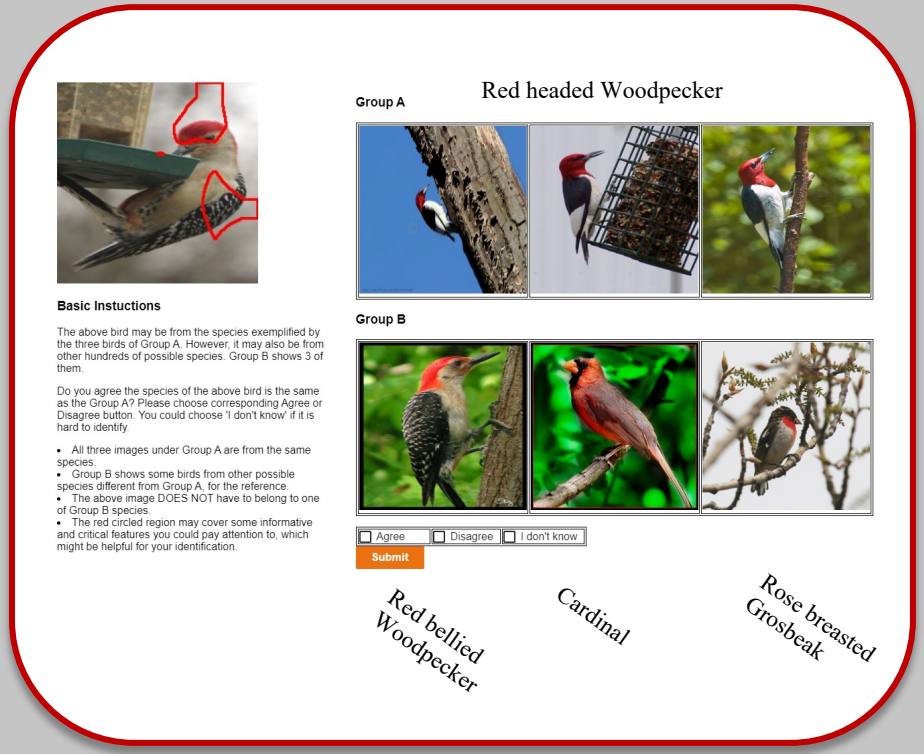
Difficulties

- Solutions
  - Ask question implicitly, by introducing the support set. 'Agree'/'Disagree'/'IDK'
  - Highlight informative image regions



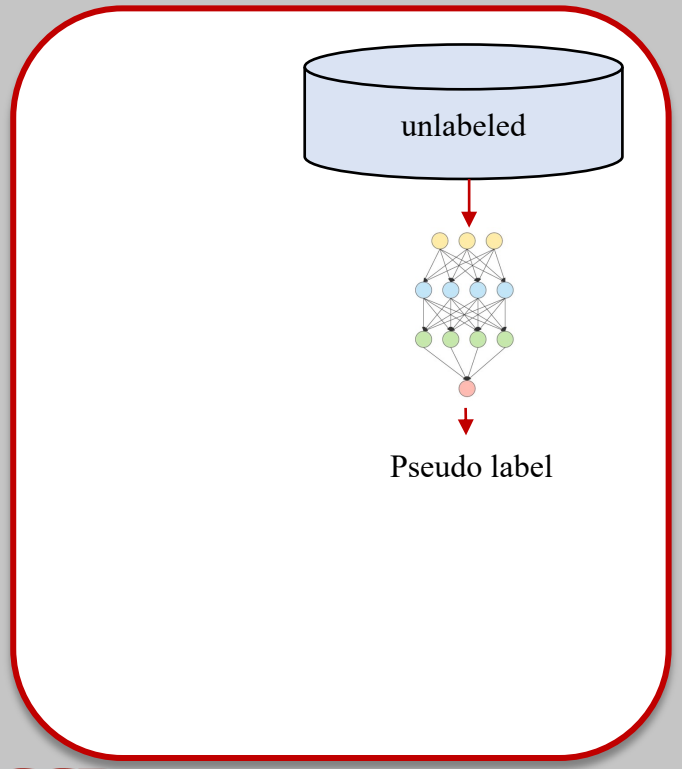
Support Set Generation

- Pseudo label $\hat{y} = f(\boldsymbol{q})$
- Both sets have K images
- Positive set: images sampled from training set of ground truth label $\hat{y}$
- Negative set: K images from K classes other than $\hat{y}$ of largest probabilities in $f(\boldsymbol{q})$ prediction



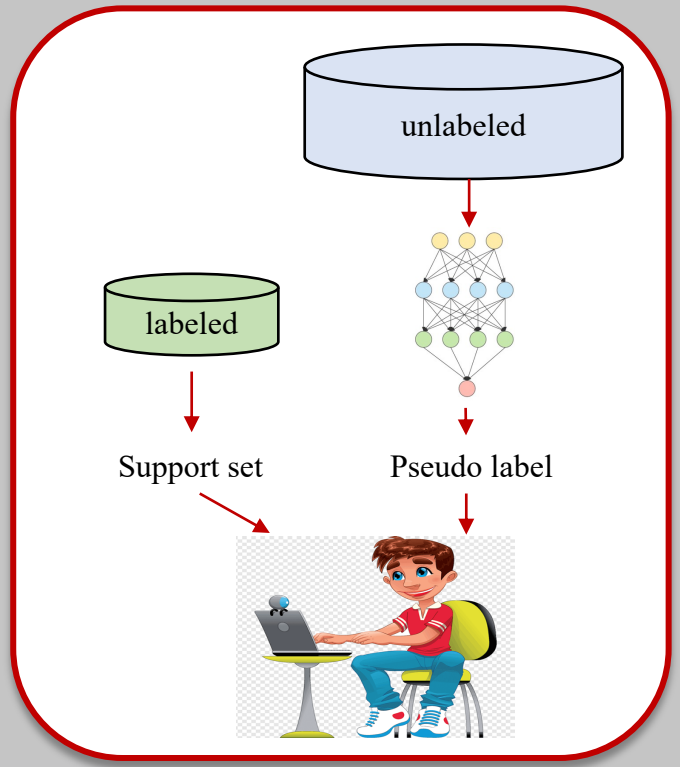
Implementation via an Iterative Progressive Strategy

- Predict labels for each example of unlabeled set



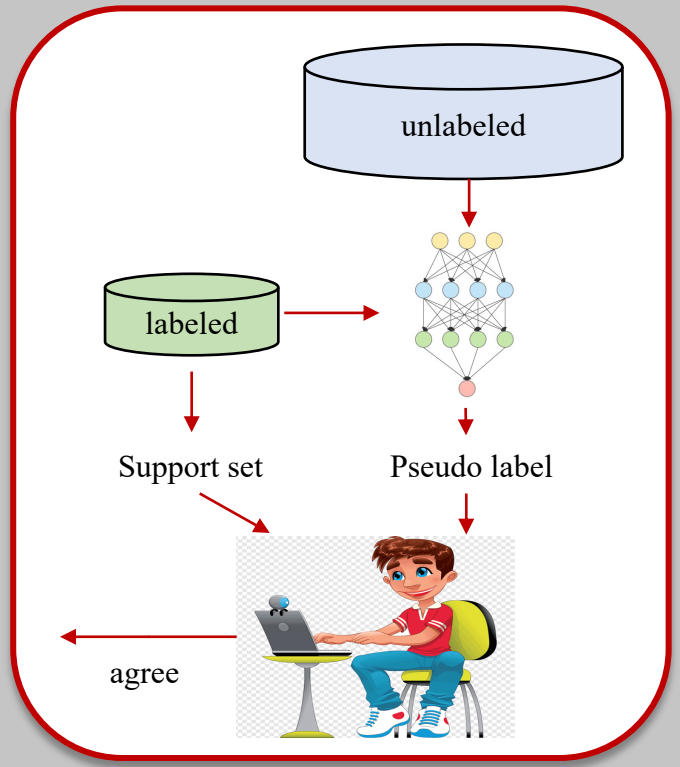
Implementation via an Iterative Progressive Strategy

- Predict labels for each example of unlabeled set
- Generate support sets for high-confidence examples and forward them to annotators



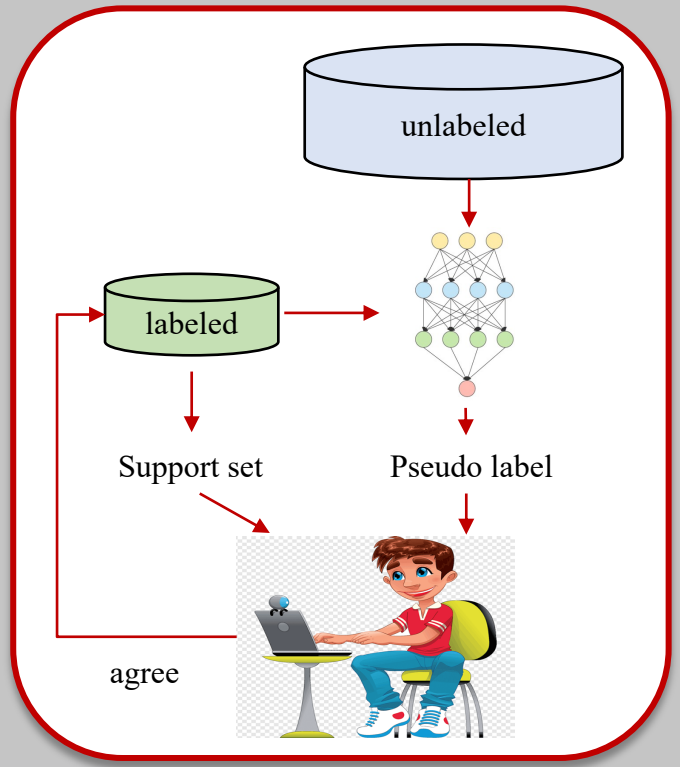
Implementation via an Iterative Progressive Strategy

- Predict labels for each example of unlabeled set
- Generate support sets for high-confidence examples and forward them to annotators
- Annotators produce decisions



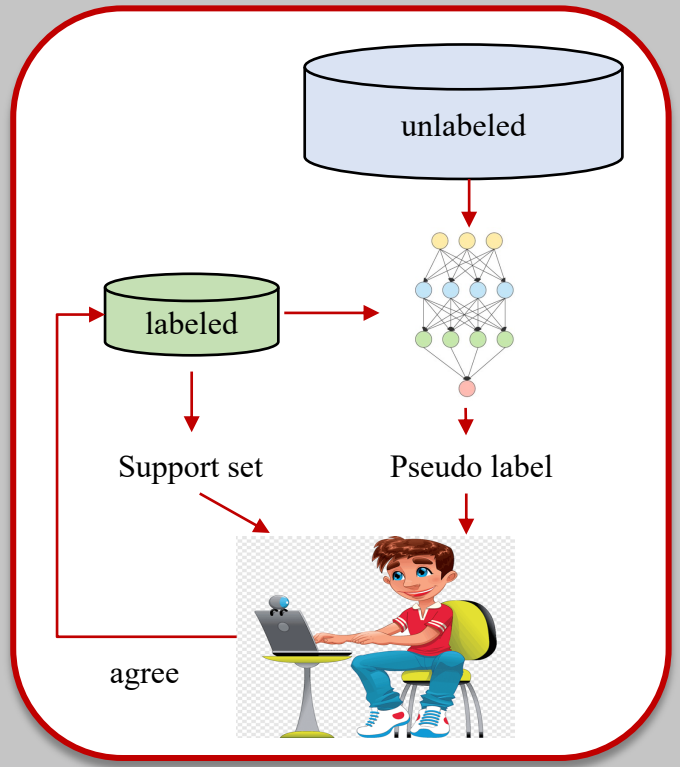
Implementation via an Iterative Progressive Strategy

- Predict labels for each example of unlabeled set
- Generate support sets for high-confidence examples and forward them to annotators
- Annotators produce decisions
- 'Agree' examples are moved from the unlabeled set and added to the labeled set



Implementation via an Iterative Progressive Strategy

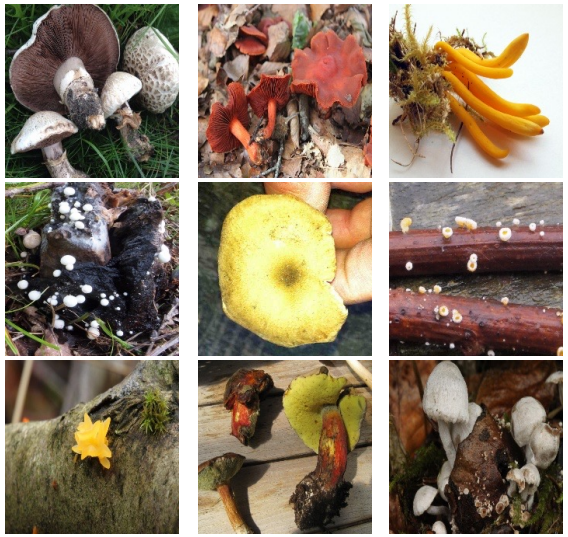
- Predict labels for each example of unlabeled set
- Generate support sets for high-confidence examples and forward them to annotators
- Annotators produce decisions
- 'Agree' examples are moved from the unlabeled set and added to the labeled set
- Update the classifier on the labeled set



Experiments

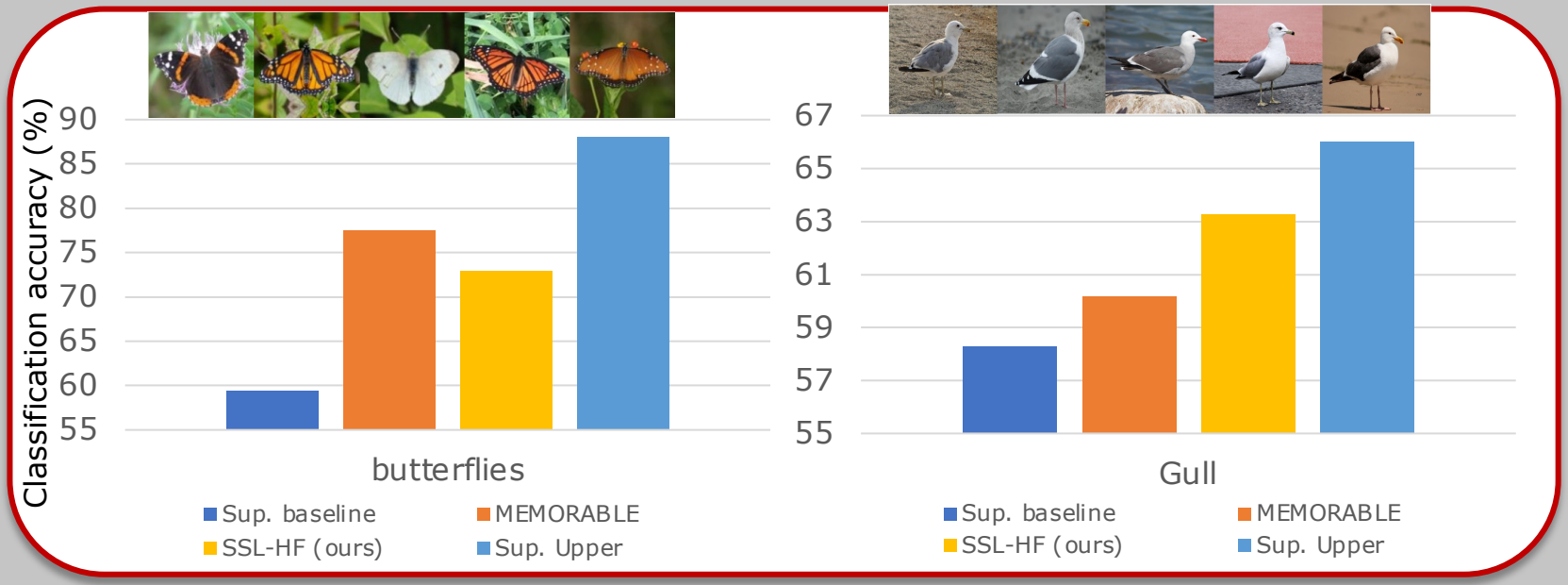
- Datasets

Birds (100 categories)



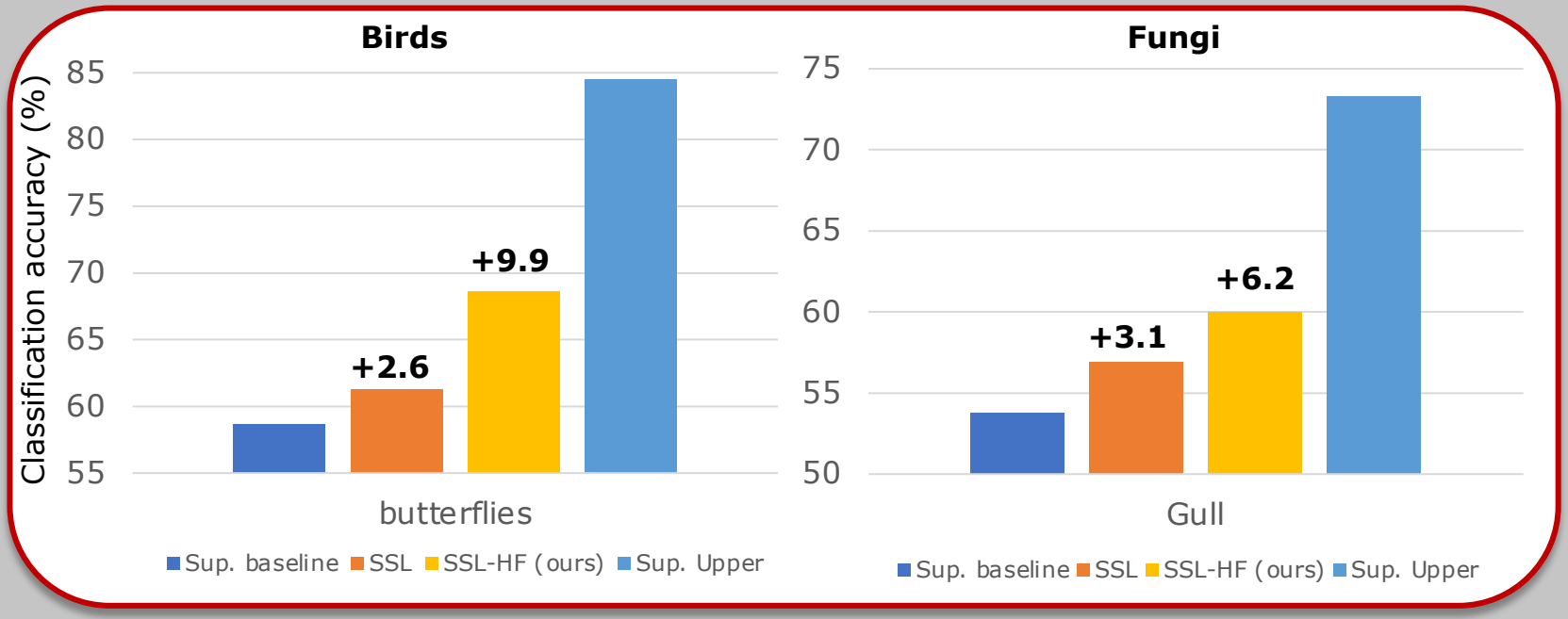
Fungi (200 categories)

Comparison to Machine Teaching



- Our SSL-HF is inferior to MEMORABLE on easier tasks but superior on harder tasks.

Comparison to SSL



- SSL-HF achieves significant gain over SSL.

Conclusion

- Aim to scale up the annotation on expert domains.
- Proposed a crowd source annotation methods.
- Achieved notable improvements.