Embracing Limited and Imperfect Data: A Review on Plant Stress Recognition Using Deep Learning

2023 ASABE lightning talk

Presented by **Mingle Xu**

Core Research Institute of Intelligent Robots

Jeonbuk National University, South Korea

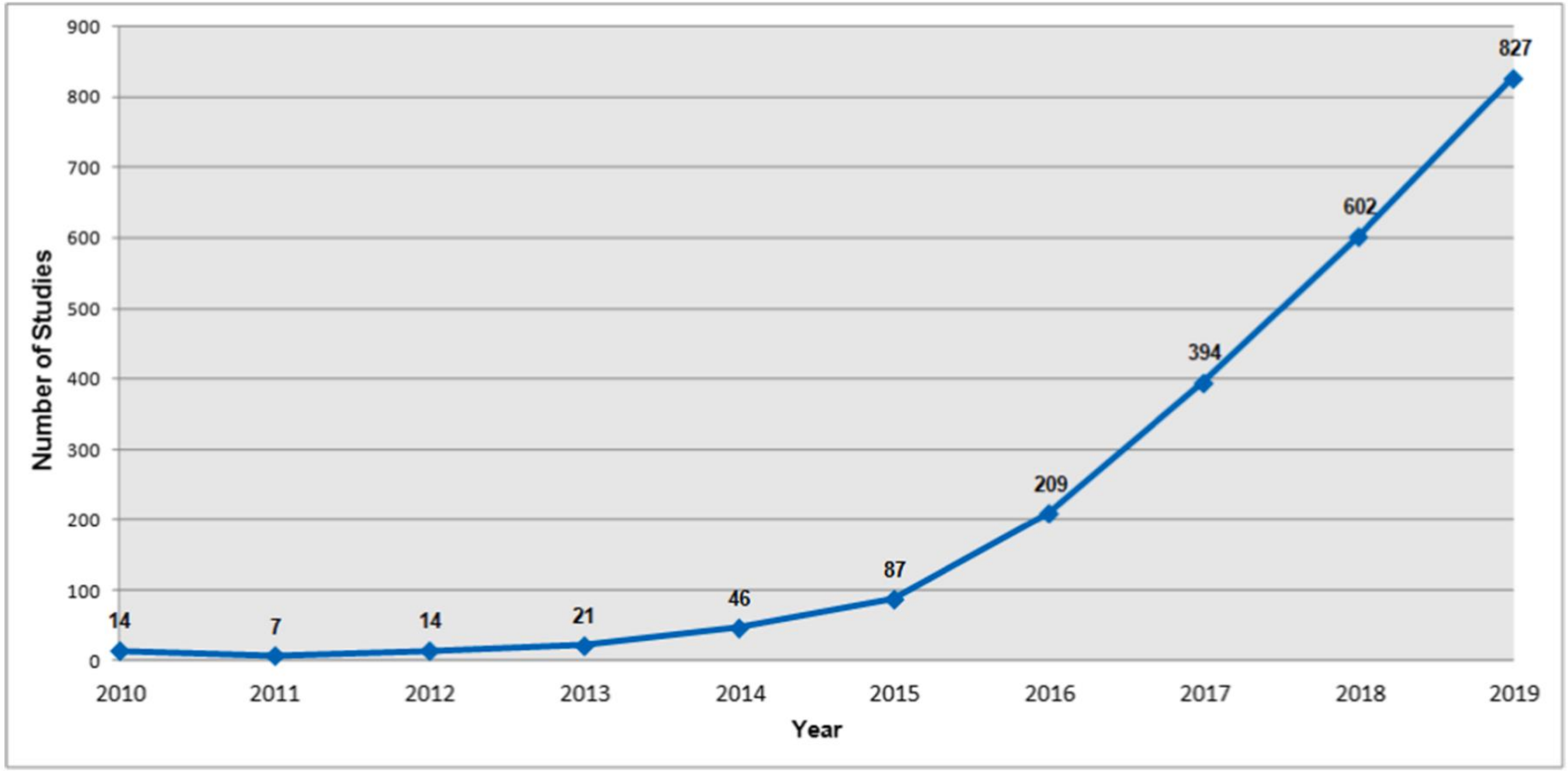
Author: Mingle Xu 1, Hyongsuk Kim, Jucheng Yang 2, Alvaro Fuentes, Yao Meng, Sook Yoon, Taehyun Kim, Dong Sun Park

Content

- A brief history of plant stress recognition using deep learning
- Challenge: a good training dataset is often required but its collecting is not easy.
  - Limited data
    - Reasons: model and task
    - Class-level: few-shot and class imbalance
    - Dataset-level: domain shift and unknown class
  - Imperfect data
    - Basic information: desired annotation strategy EEP
    - Incomplete annotation
    - Inexact annotation
    - Inaccurate annotation
- Concluding remarks

A quick history

The number of publications related to plant stress recognition using deep learning is increasing.



"Plant diseases recognition on images using convolutional neural networks: A systematic review." *Computers and Electronics in Agriculture* 185 (2021).

A quick history

The performance is decent in most of the publications.

Table 8 (*continued*).

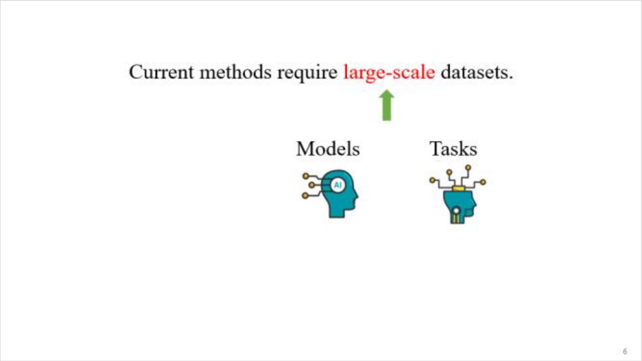
Author	Technique	Dataset	Capturing Condition	Acc %
Abbas, Jain, Gour, and Vankudothu (2021)	Conditional GAN, DenseNet121	20,012 tomato leaf images in 10 classes from PlantVillage dataset	Laboratory	97.11
Zhao et al. (2021)	DoubleGA, WGAN, VGG16, ResNet50, and DenseNet121	31,361 leaf images from PlantVillage dataset	Laboratory	–
	DenseNet121	PlantVillage dataset	–	99.53
Ji and Wu (2022)	DeepLabV3+, ResNet50	500 grape leaf image	In-field	97.75

"Trends in vision-based machine learning techniques for plant disease identification: A systematic review." *Expert Systems with Applications* (2022).

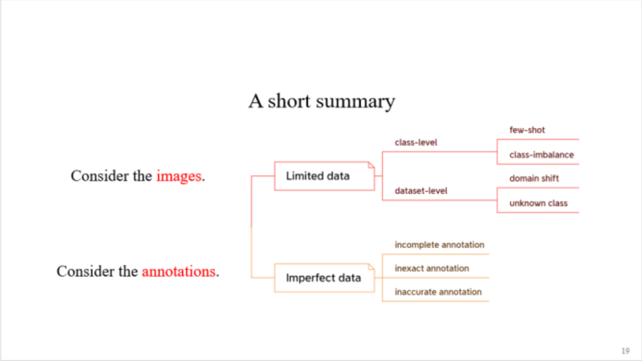
However, current deep learning-based models may suffer in real-world applications.

Because a good training dataset is often required but its collecting is **difficult and expensive**, and even impractical.

Large-scale dataset.

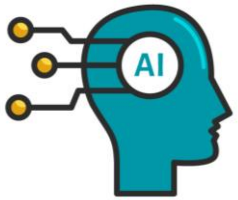


Annotated properly.

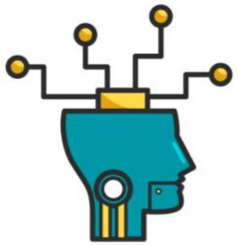


Current methods require large-scale datasets.

Models

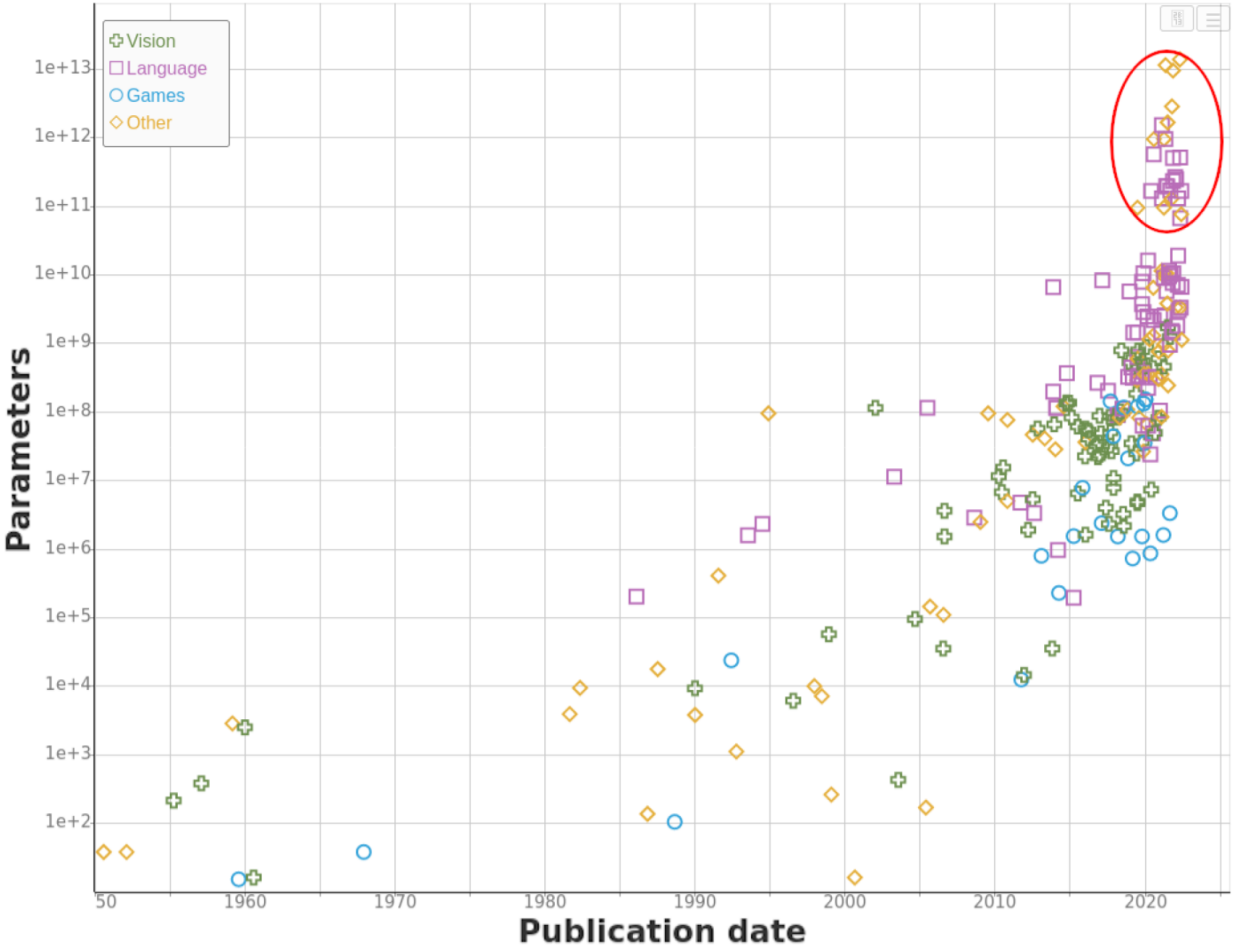


Tasks



Models

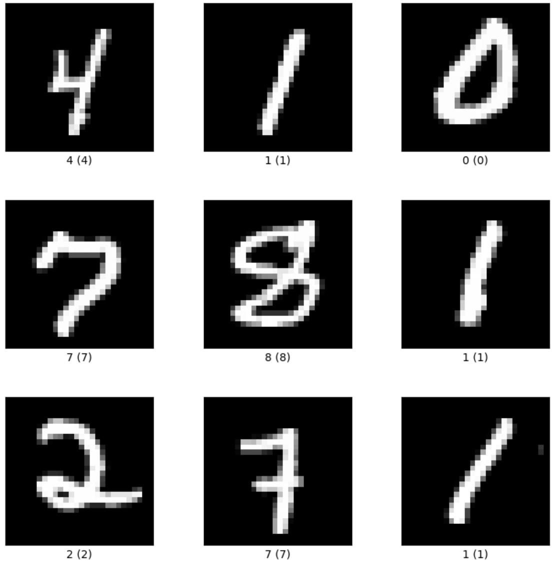
- Bigger models tend to have more data to be trained well, otherwise may be overfitting.



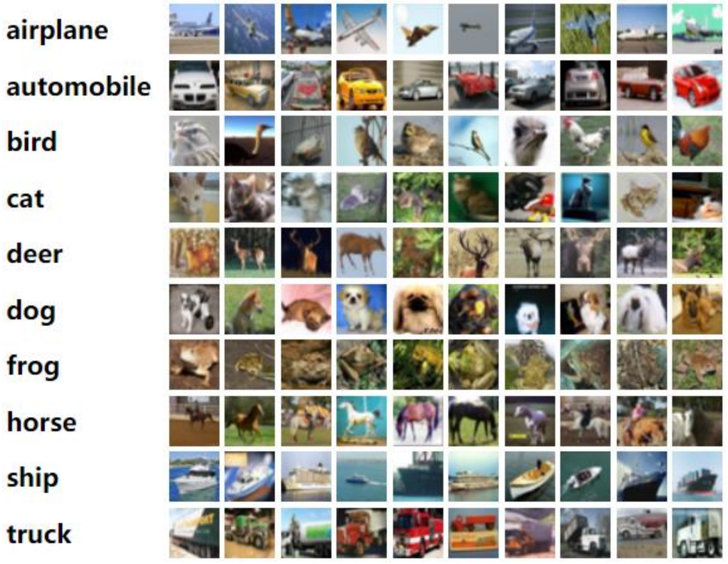
"Machine Learning Model Sizes and the Parameter Gap." *arXiv* (2022).

Task: Plant Stress Recognition

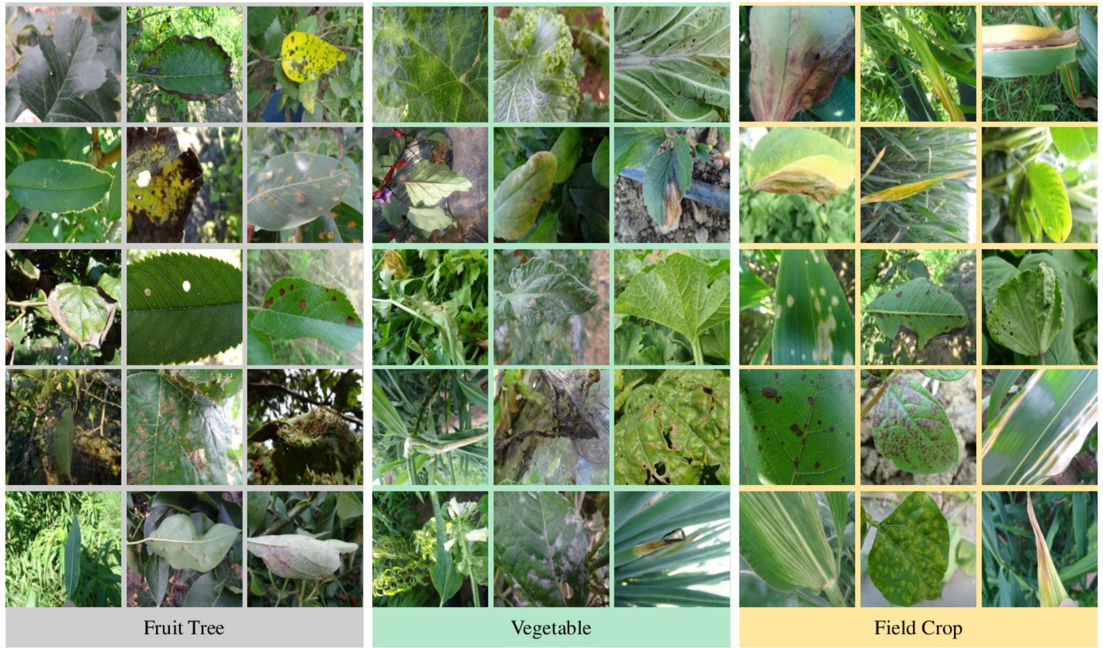
- Plant stress recognition in real-world applications is complex.



MNIST



CIFAR10



Plant stress images from PDD271

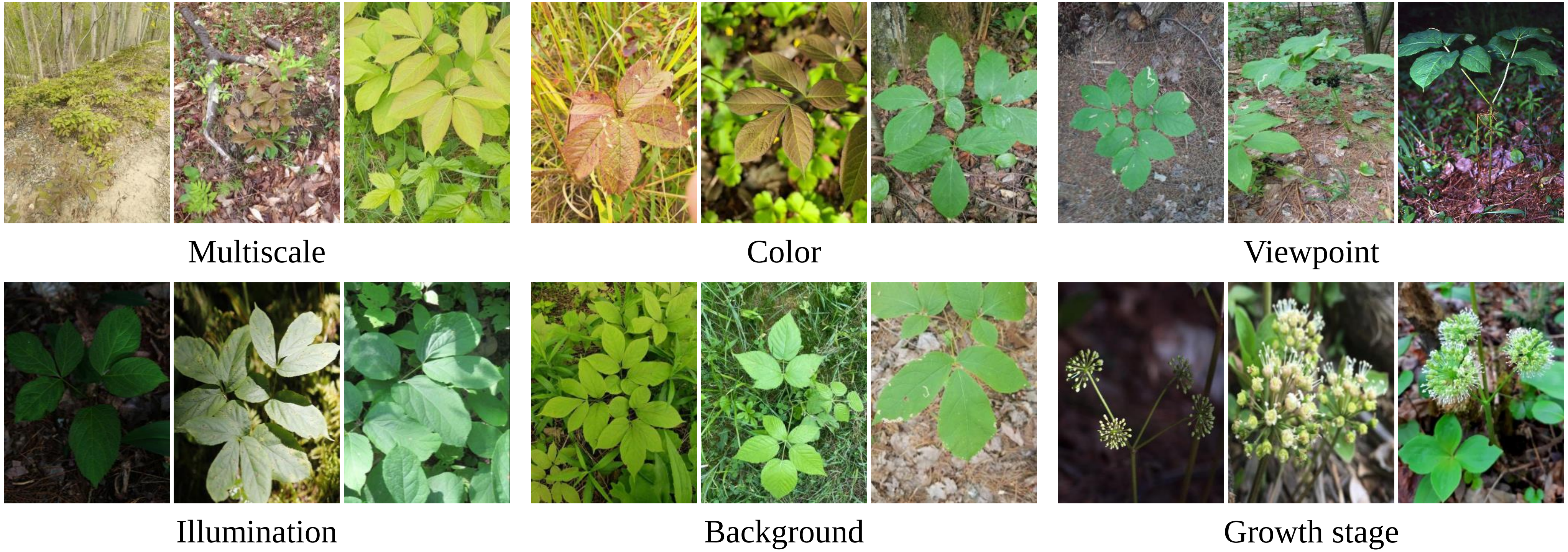
"Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998)

"Learning multiple layers of features from tiny images." (2009)

“Plant disease recognition: A large-scale benchmark dataset and a visual region and loss reweighting approach”, TIP 2021.

Task: Plant Stress Recognition

- Images in plant stress recognition include huge intra-class variations and similar inter-class variations. → need more data to train models with decent performance.
- **A good dataset should cover these intra-class variations**.



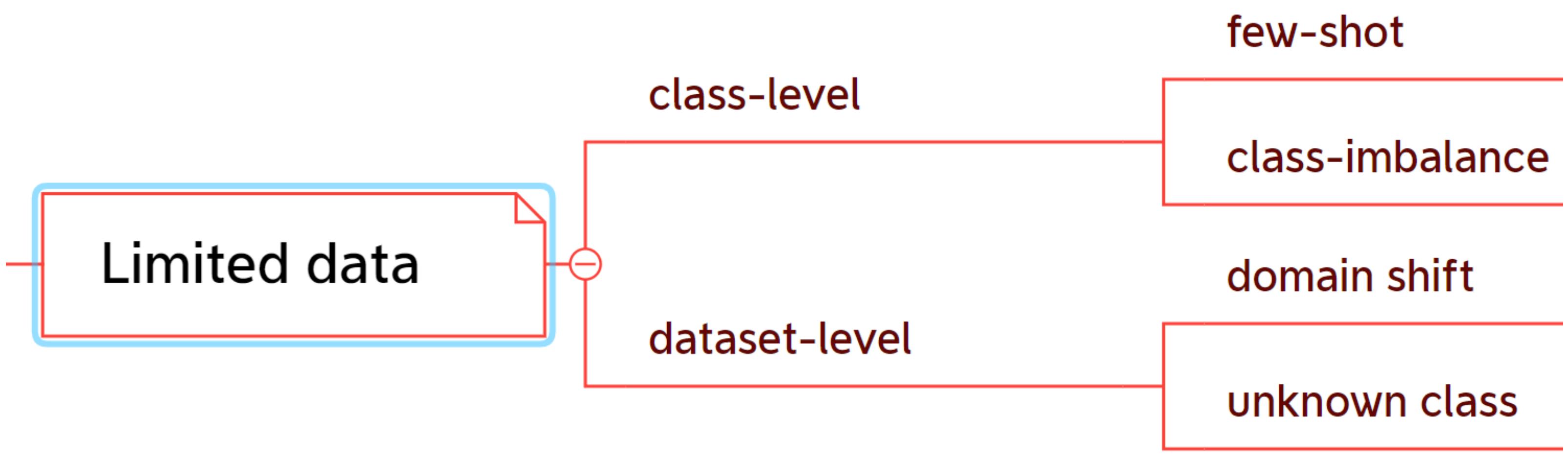
Images from the same class (PlantCLEF2022)

- More examples of intra-class variation.

Different types of tomato leaves

Different stages of tomato stress

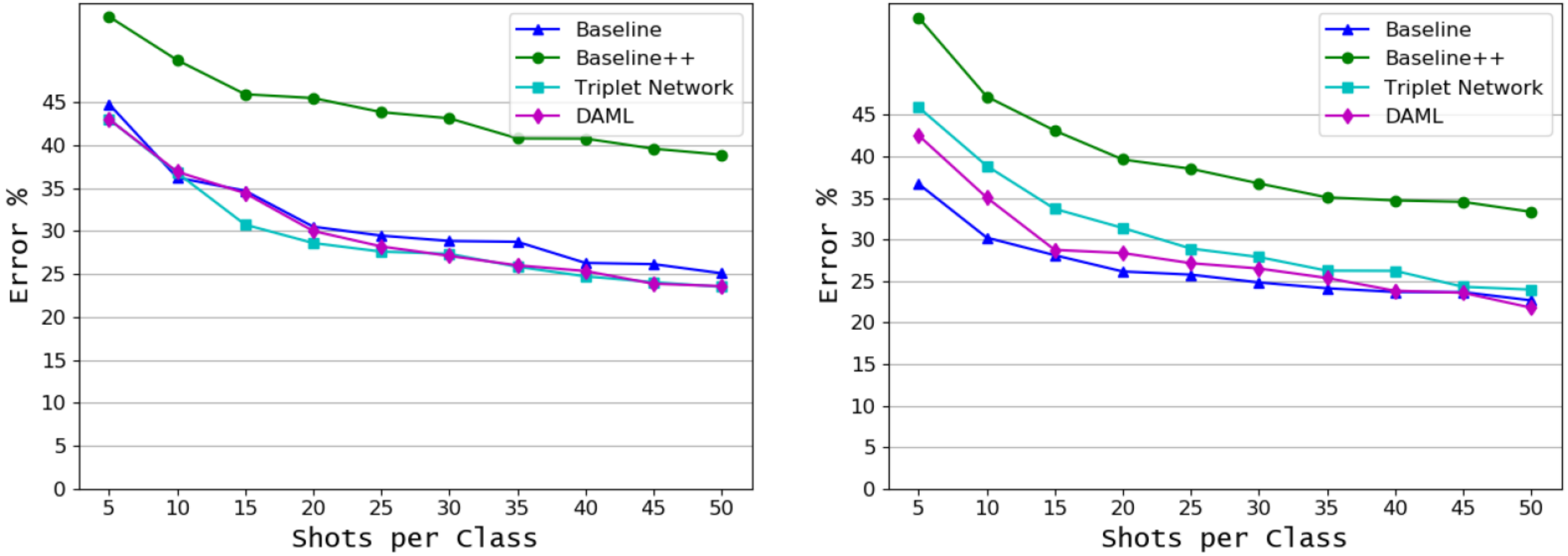
If our training dataset is **not** in large-scale, …



Few-shot

- Every class has the same number and few images, such as 10, 20, and 100.
- In this scenario, holistic performance is not good, either for every class.

Experiment in PlantVillage.



(a) ResNet18 (b) ResNet34

"Convolutional neural network for automatic identification of plant diseases with limited data." *Plants* (2020).

Class imbalance

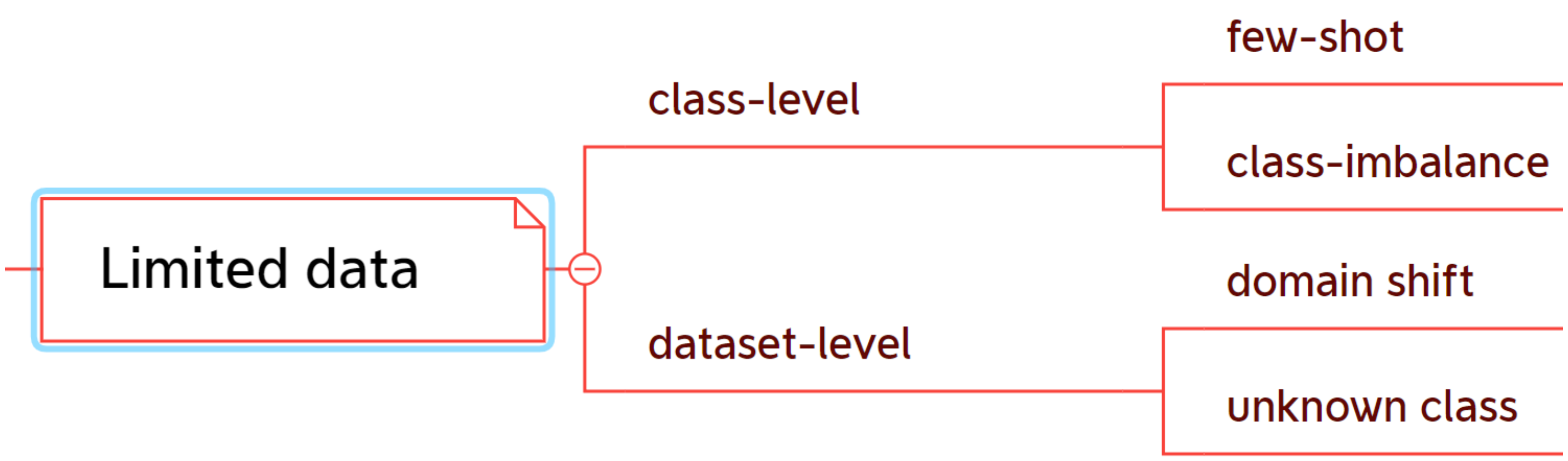
- One class has much more images than another class, $n_{X_i} \gg n_{X_j}$.
- In this scenario, the former class may have much better performance than the latter.

Class	Data set A		Data set B
	Training	Validation	Testing
Healthy (H)	4,000	717	1,046
MYSV (H)	4,000	745	2,034
Brown Spot (B)	2,000	784	1,220
Powdery Mildew (P)	2,000	796	89
Total	12,000	3,042	4,389

Class	# of test images	Baseline (%)
Healthy (H)	1,046	**85.1**
MYSV (M)	2,034	75.4
Brown Spot (B)	1,220	62.8
Powdery Mildew (P)	89	61.8
Average		71.3

"Leafgan: An effective data augmentation method for practical plant disease diagnosis." *IEEE Transactions on Automation Science and Engineering (2020).*

A short summary



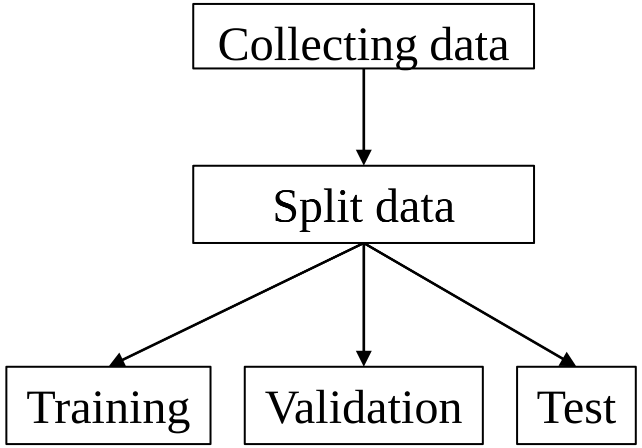
Consider the situation within the training datasets.

Consider the difference between the test and training datasets.

A basic assumption of deep learning

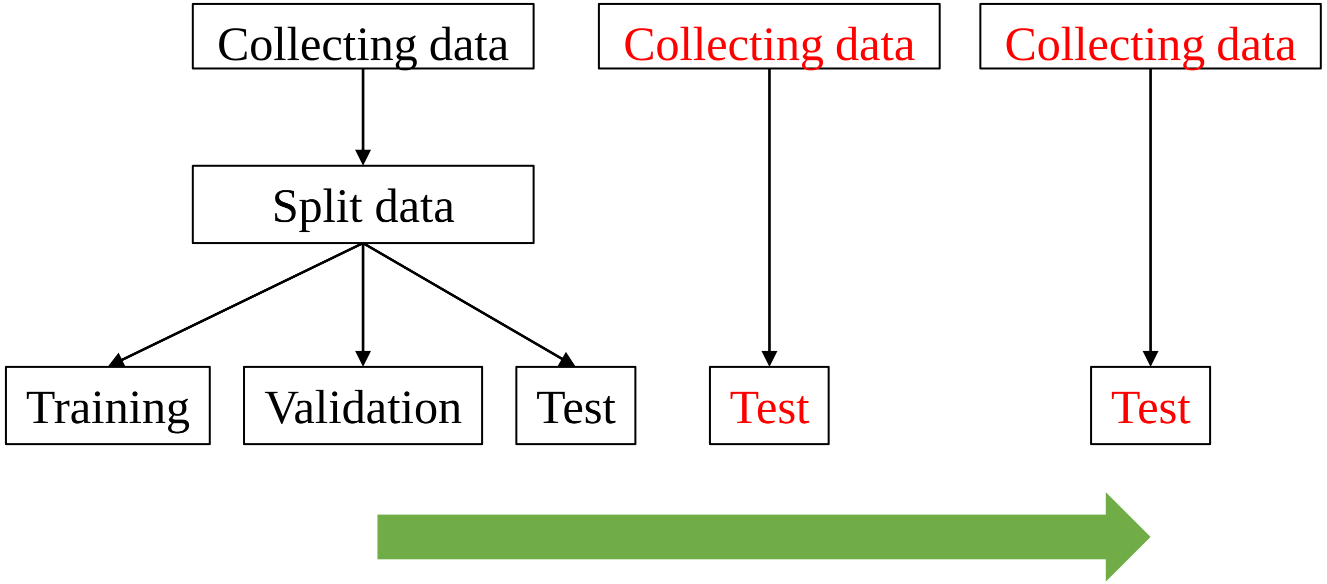
The training and test datasets are in same or similar distribution: $P_{train} \approx P_{test}$.

A widely adopted strategy to evaluate models.



The split **almost** guarantees they are in a similar distribution.

The assumption is strong. Scenarios in real-world applications.



Spatial and temporary changing

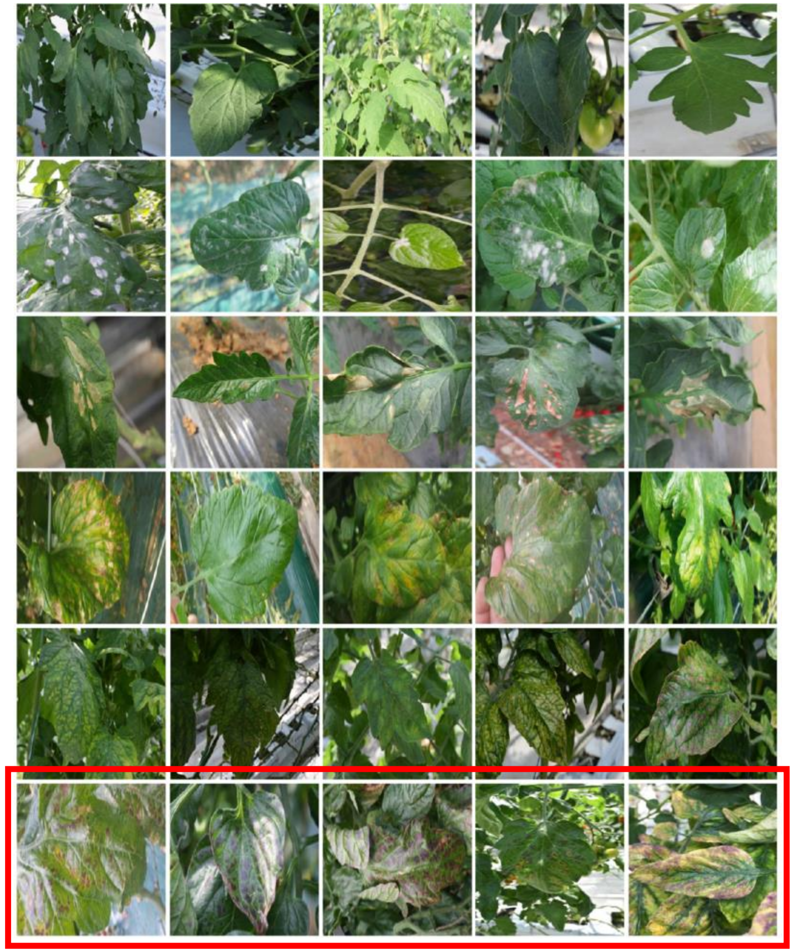
If you can collect a dataset large enough, no problem.

Dataset-level: unknown class

What will happen if we can not collect a large-scale dataset to train models?

Unknown classes may happen in test process.

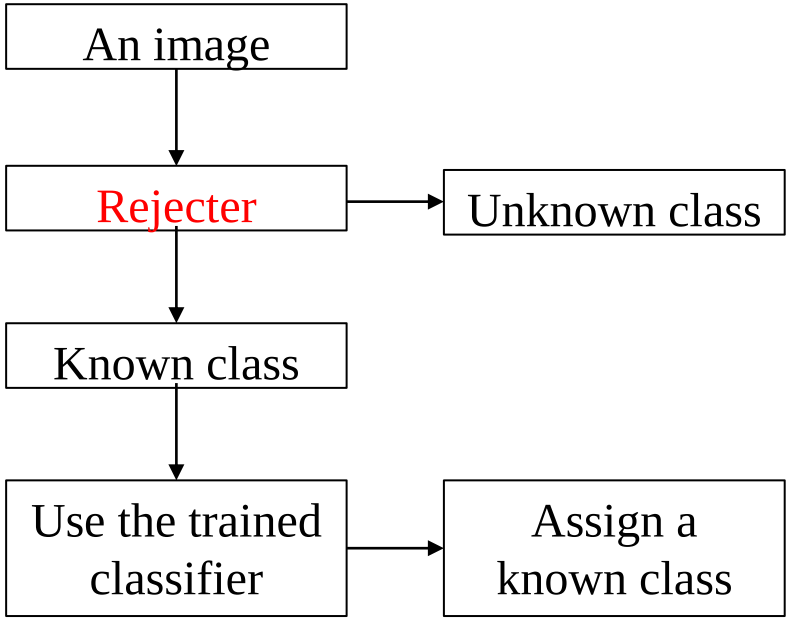
A new class in
the test dataset.



Dataset-level: unknown class

Unknown class challenge: most of the current models just assign an image from a new class as a known class existing in the training dataset.

But the new class may result in big trouble.



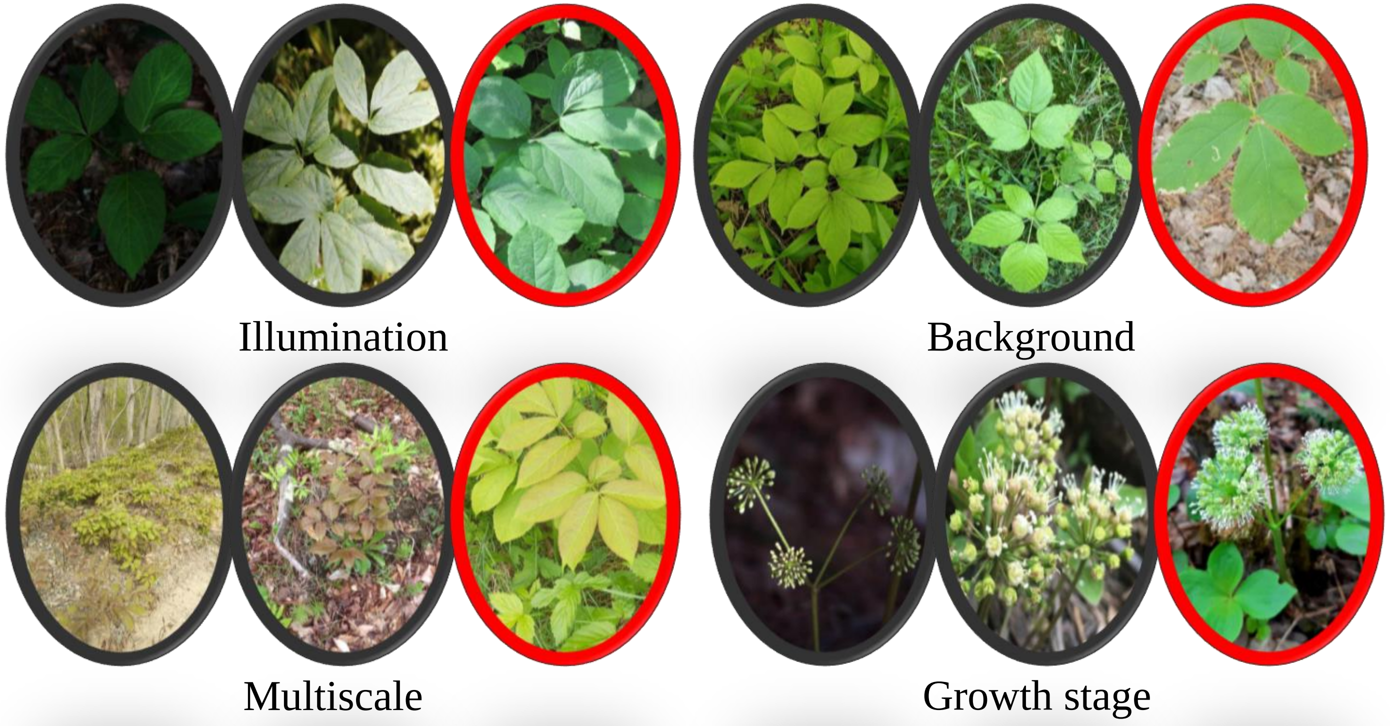
What is desired? Open set recognition (OSR).

“Known and Unknown Class Recognition on Plant Species and Diseases”, submitted (2023).

Dataset-level: domain shift

Domain shift: the test and training datasets are in different distributions (variations).

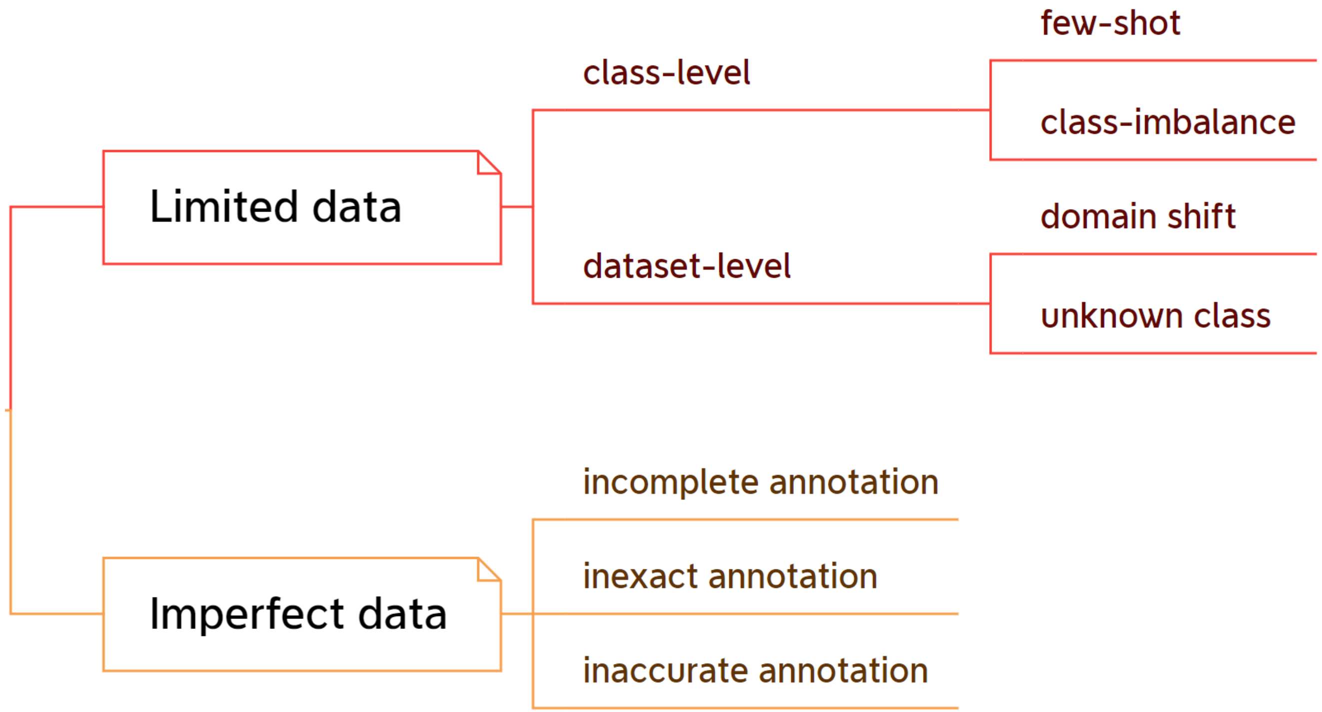
Challenge: low performance to recognize those images from the different distributions.



A short summary

Consider the images.

Consider the annotations.



Annotation and its strategies

Different computer vision objectives have different annotations.

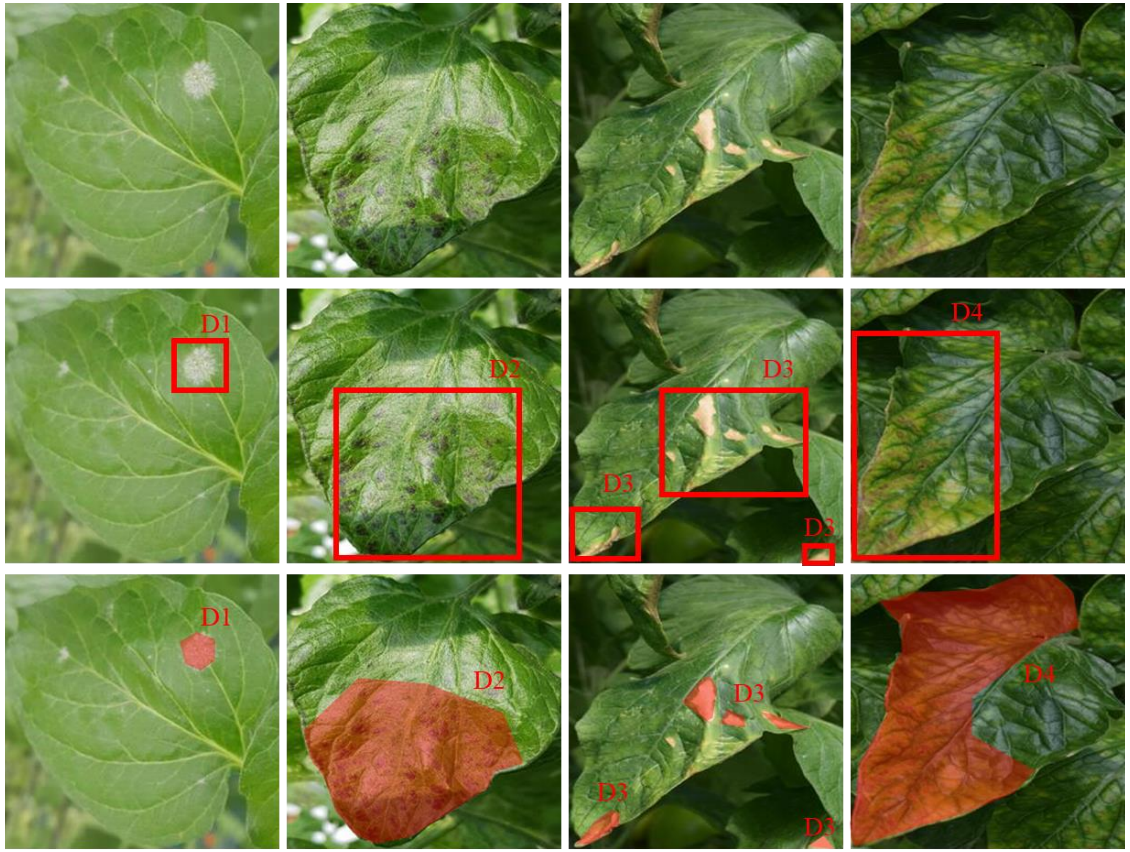


Image classification

Assume that one image only covers one class and every image should be annotated with one label.

Object detection

Allow that an image has multi-class and every class is annotated with a pair of label and bounding box.

Segmentation

Every pixel should be annotated with labels.

Desired annotation strategy: EEP

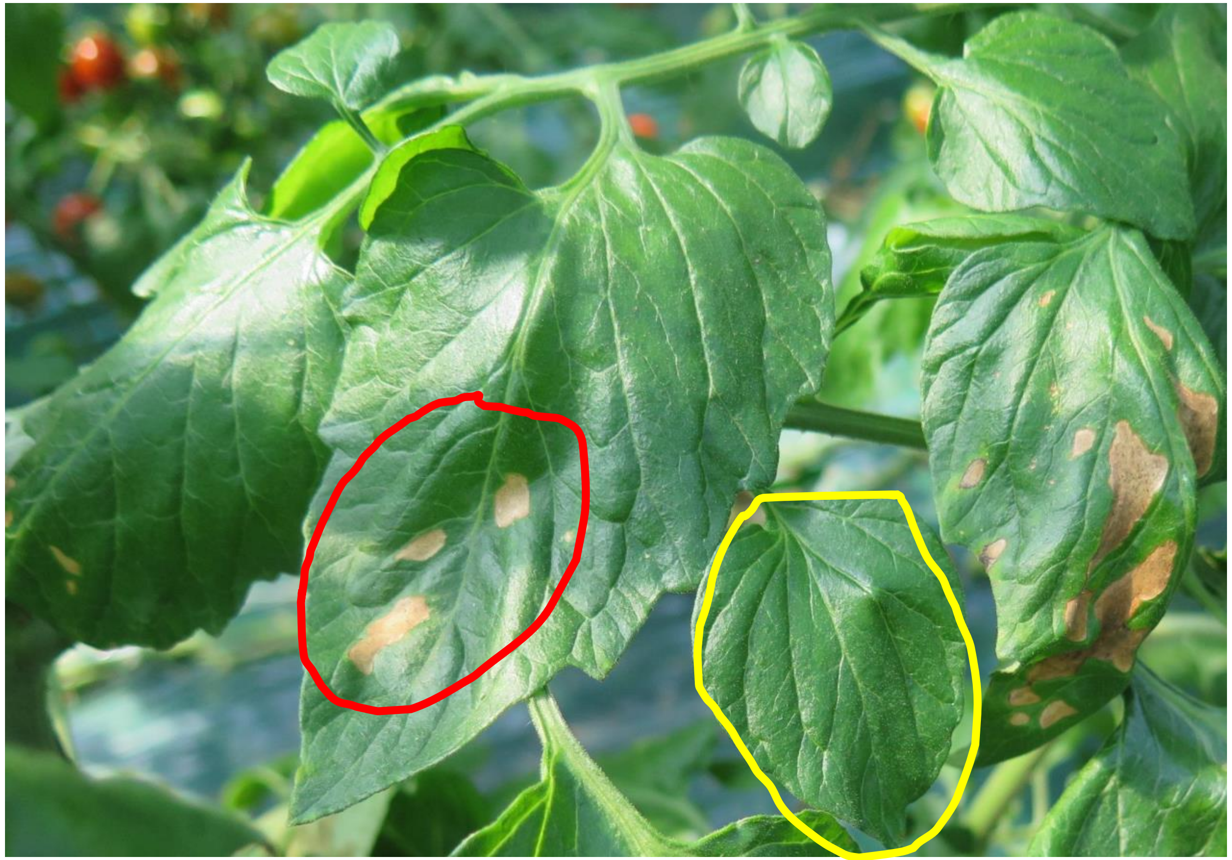
Exclusive: one annotation has only one class.

Extensive: every class in one image should be annotated.

Precise: annotations should be precise.

Image classification

Labeled as canker?
But it also has healthy leaves.



Desired annotation strategy: EEP

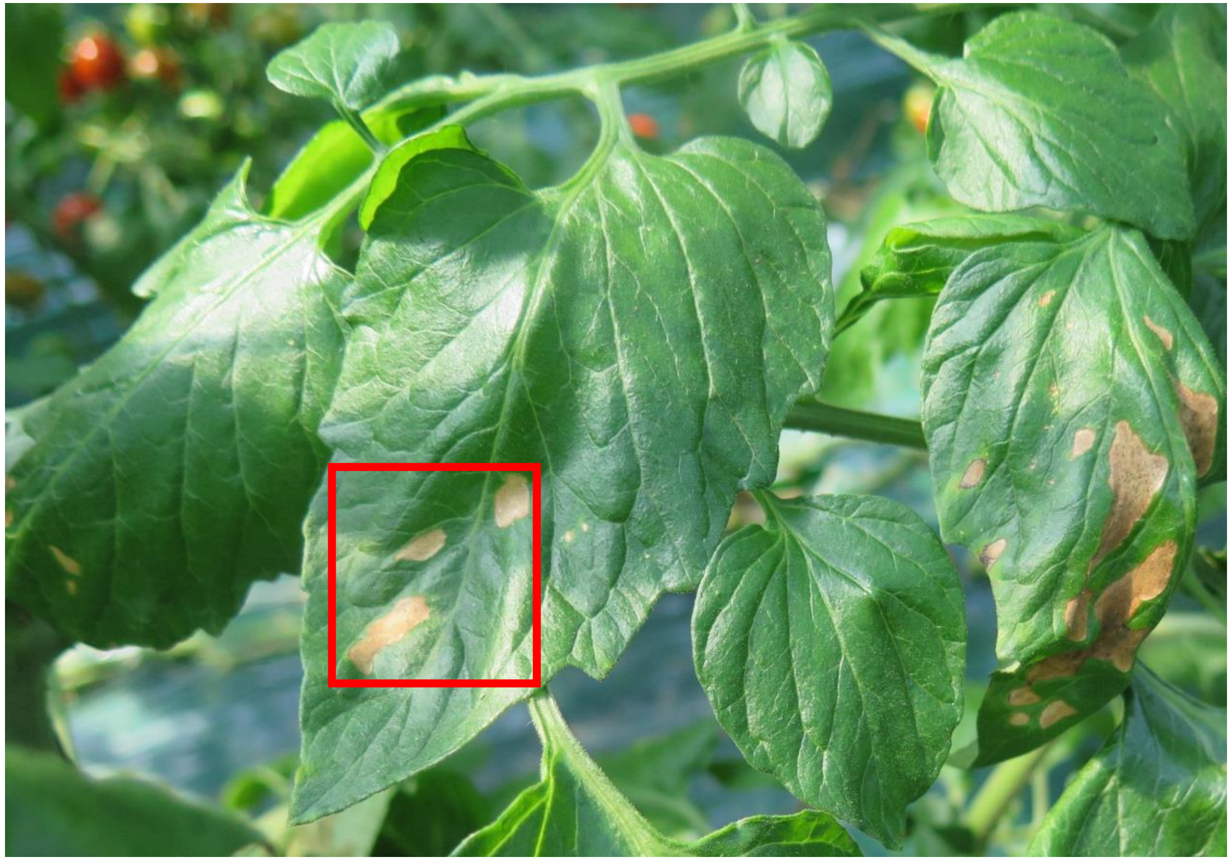
Exclusive: one annotation has only one class.

Extensive: every class in one image should be annotated.

Precise: annotations should be precise.

Object detection

Just annotate one canker?
But it also has one more.



Violation towards the EEP annotation strategy

Incomplete annotation: some images or symptoms are not annotated.

Unlabeled images are much cheaper and may be useful.

Challenge: how to use the unannotated images or symptoms.

Classification

Labeled Unlabeled

Object detection

Labeled Unlabeled

Desired annotation strategy: EEP

Exclusive: one annotation has only one class.

Extensive: every class in one image should be annotated.

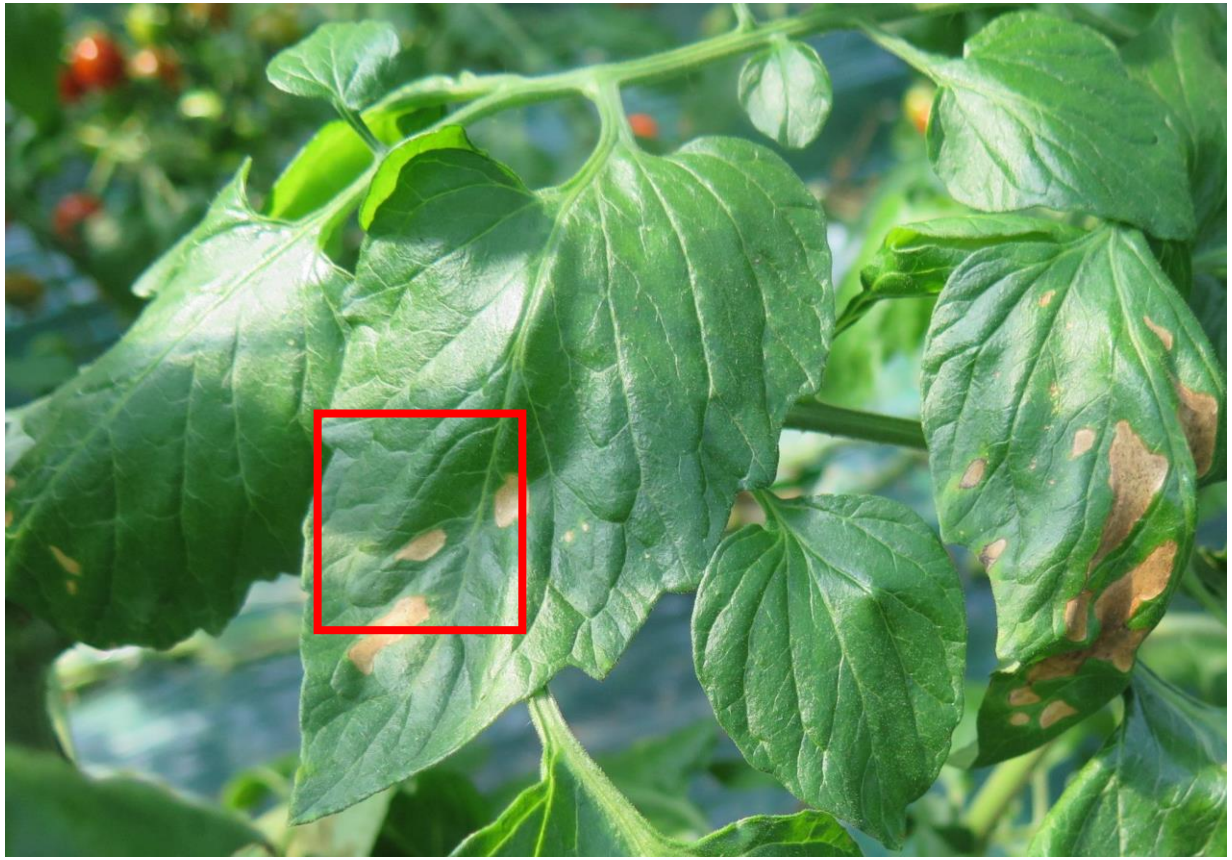
Precise: annotations should be precise.

Object detection

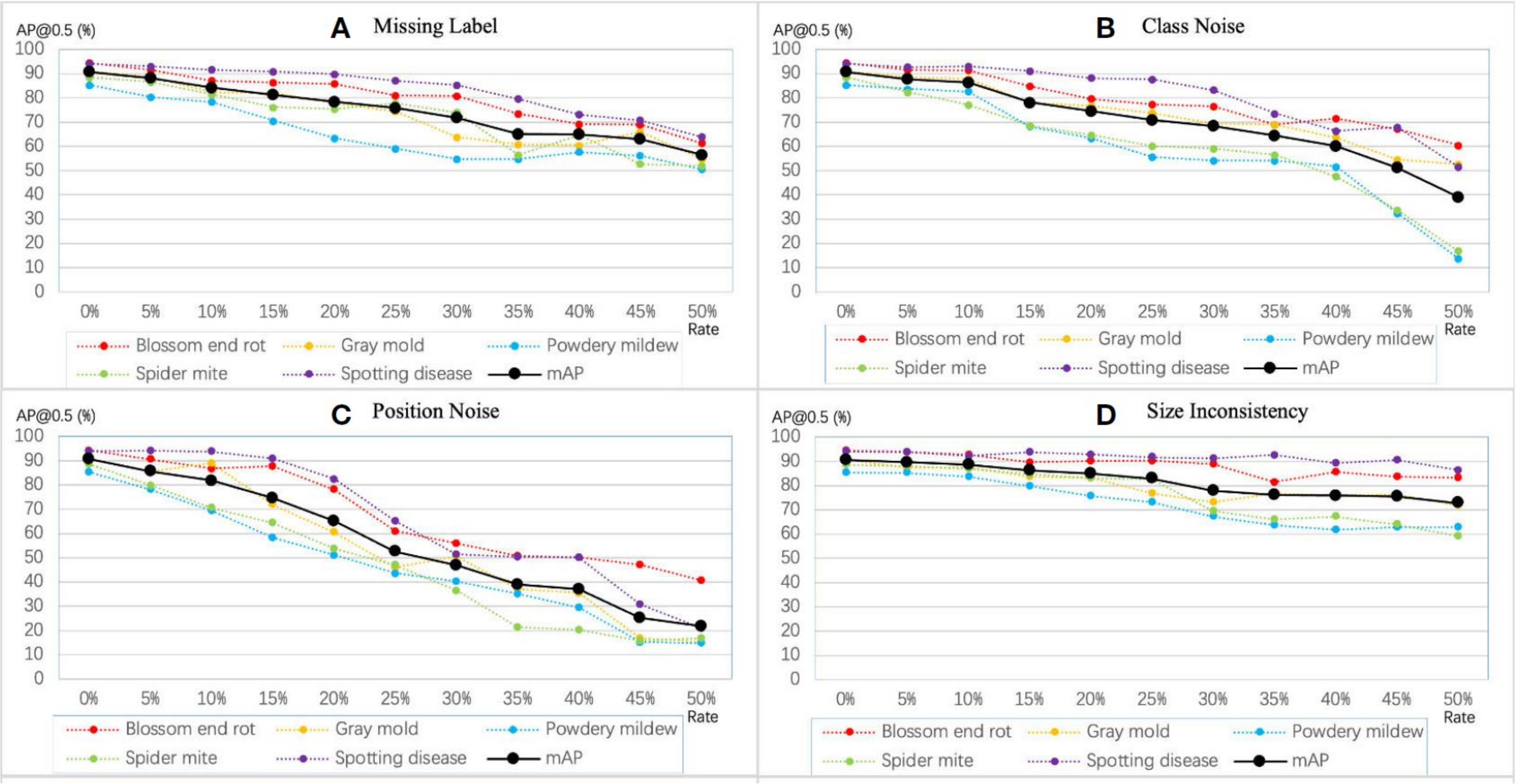
The bounding box is not precise → bounding box is imprecise.

Annotate it as Powdery Mildew →label is imprecise, actually wrong.

Inaccurate annotation: annotations are not always correct, such as wrong labels, and not precise bounding boxes.



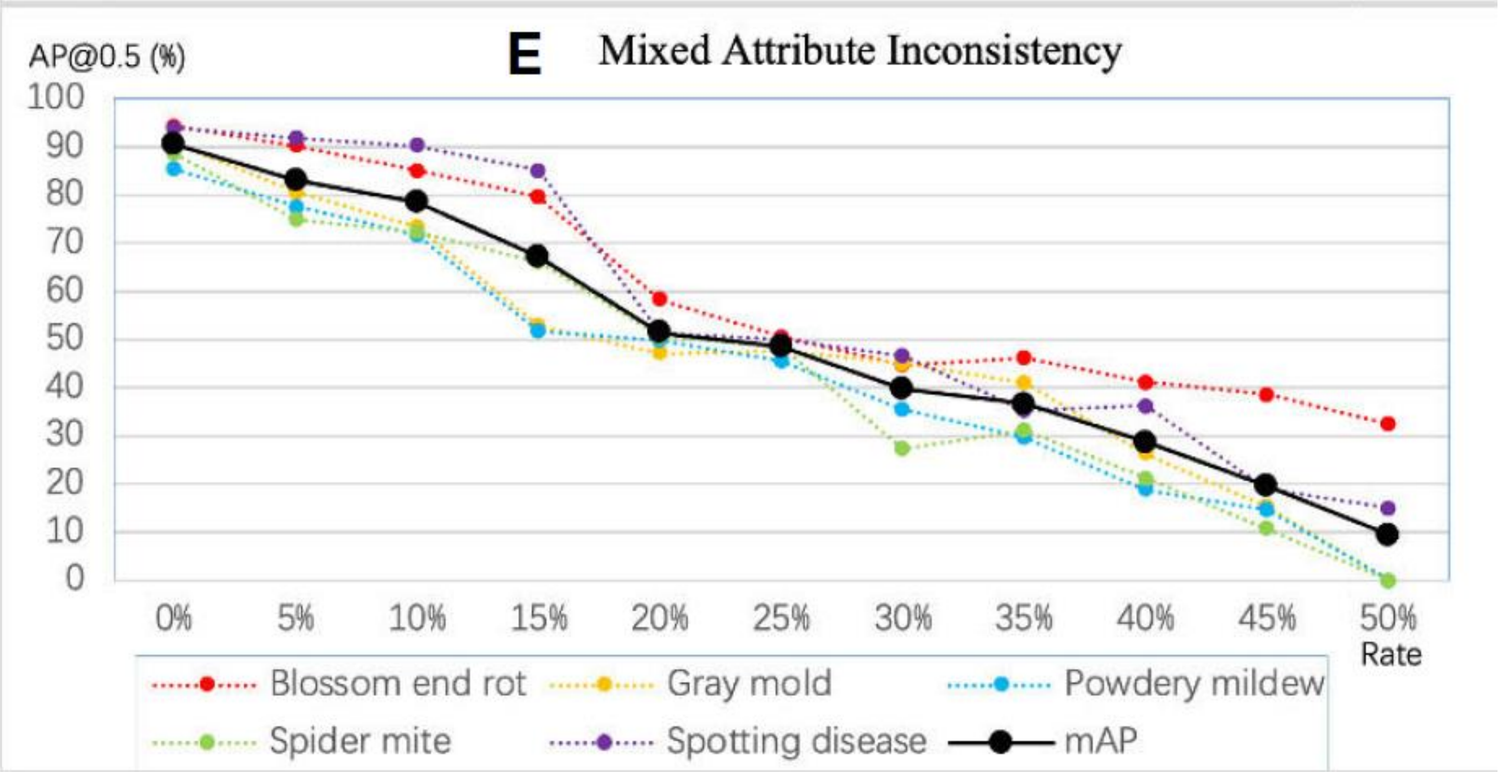
Violation towards the EEP annotation strategy



"Data-centric annotation analysis for plant disease detection: Strategy, consistency, and performance." *Frontiers in Plant Science* 13 (2022).

Violation towards the EEP annotation strategy

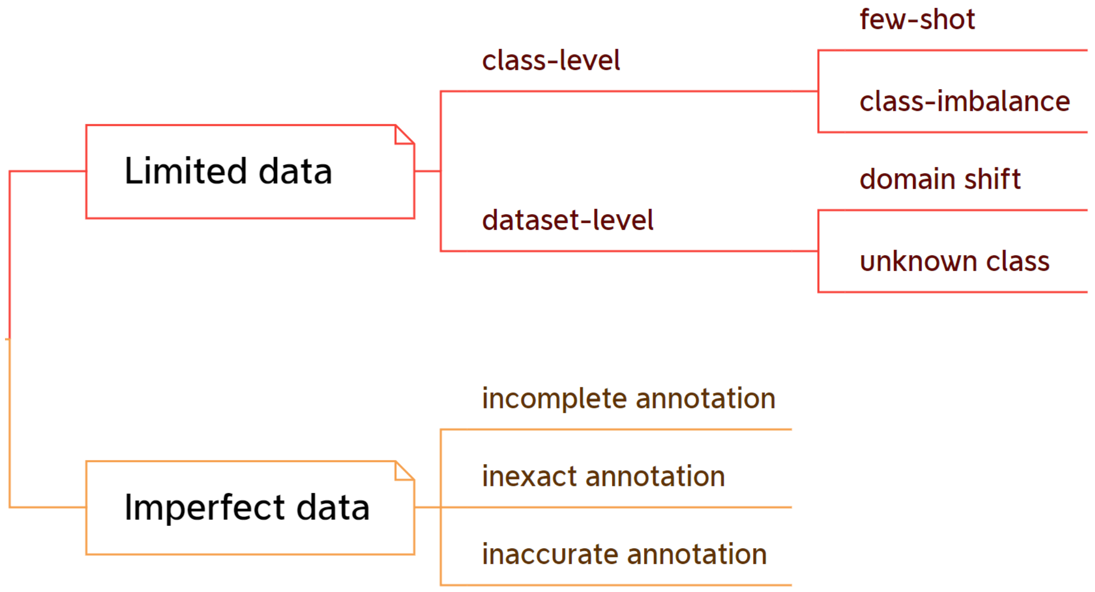
The mixed impact.



"Data-centric annotation analysis for plant disease detection: Strategy, consistency, and performance." *Frontiers in Plant Science* 13 (2022).

Conclusion

Current deep learning models may suffer in real-world because collecting a desired dataset is difficult, expensive, and even impractical. Embracing limited and imperfect data is a way to address the challenge.



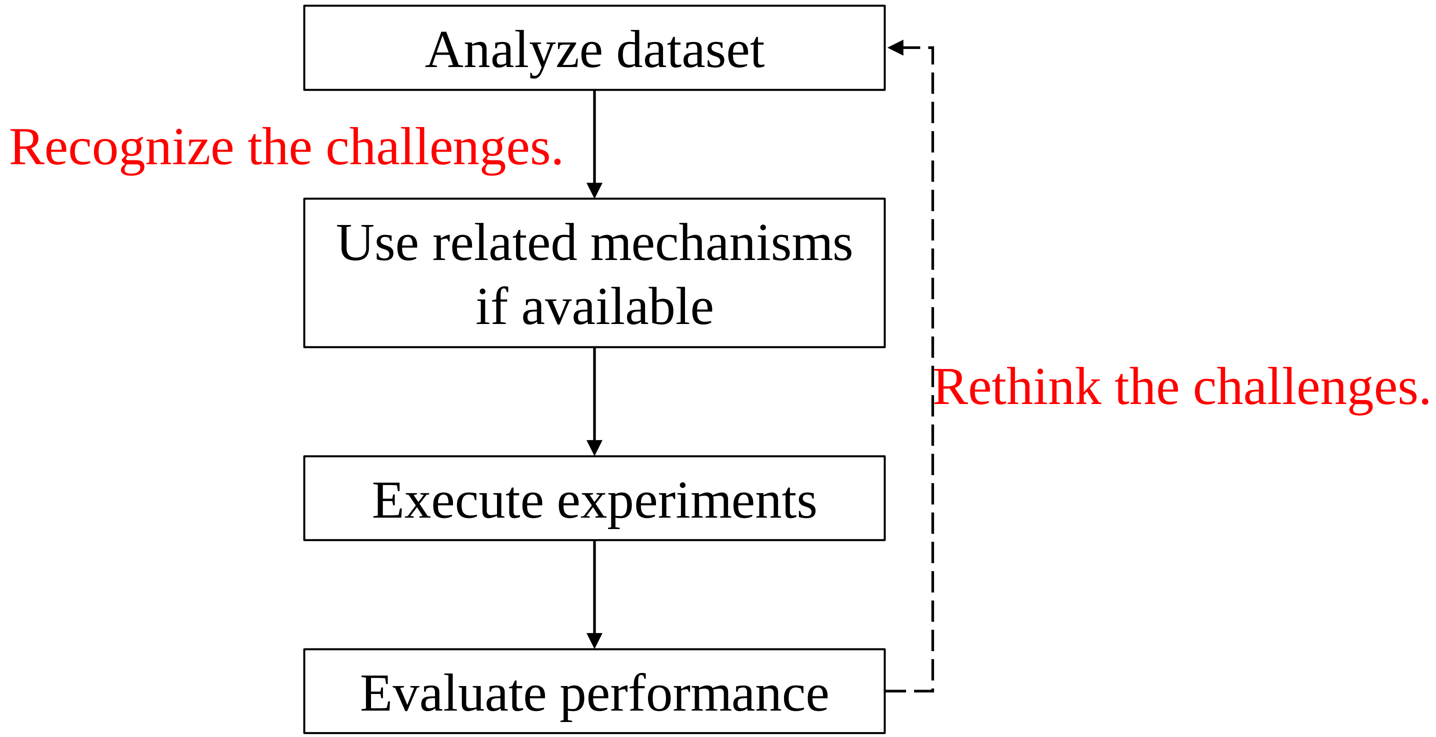
Some mechanisms are proposed to address the challenges but not enough, referring to **our preprint paper**.

Xu, Mingle, et al. "Embracing Limited and Imperfect Data: A Review on Plant Stress Recognition Using Deep Learning." *arXiv* (2023).

Concluding remark

Analyzing datasets is essential for practical applications.
Multiple issues may exist simultaneously and are more difficult.

Process for a real-world project.



Xu, Mingle, et al. "Embracing Limited and Imperfect Data: A Review on Plant Stress Recognition Using Deep Learning." *arXiv* (2023).

Concluding remark

Sometimes we need to rethink our objective and dataset.

High-quality datasets are essential for the community.

Set an objective in a systematic perspective.

Dataset

1. Design strategies to collect and annotate data.
2. Collect data.
3. Annotate data.
4. Analyze the challenges in the annotated dataset.

Algorithm

1. Design a deep learning-based model.
2. Design the training strategies.
3. Train and evaluate the model.
4. Test the model.

Evaluate the objective.

One disease may be very similar to another one where models can not classify them.

Rethink the dataset.

The training dataset is poor and we may need to collect another one.

Xu, Mingle, et al. "Embracing Limited and Imperfect Data: A Review on Plant Stress Recognition Using Deep Learning." *arXiv* (2023).

Thank You

Questions and Comments

Email: xml@jbnu.ac.kr to Mingle Xu

Public slides: https://xml94.github.io/presentations.html

The complete paper is available at https://arxiv.org/abs/2305.11533

Examples of real scenarios.



Images vary in different farms such as the background.

"Open Set Self and Across Domain Adaptation for Tomato Disease Recognition With Deep Learning Techniques." *Frontiers in Plant Science* (2021)

Desired annotation strategy: EEP

Exclusive: one annotation has only one class.

Extensive: every class in one image should be annotated.

Precise: annotations should be precise.

Image classification

Labeled as powdery mildew?
But it also has ToCV.

Preprocessing: crop.

Incomplete annotation: some images or symptoms are not annotated.

Unlabeled images are much cheaper and may be useful.

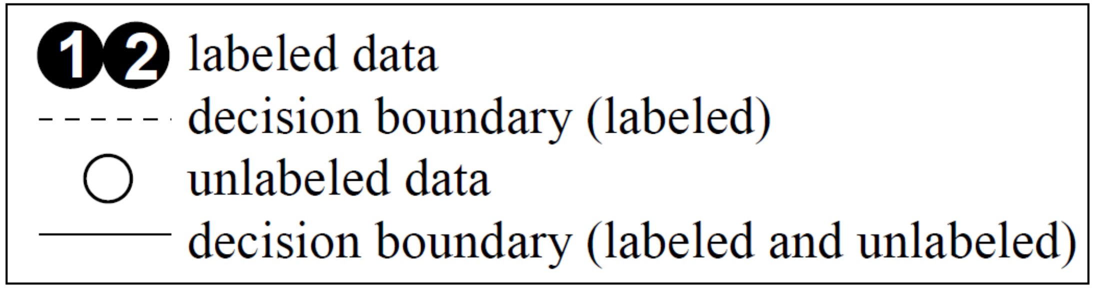
Challenge: how to use the annotated images or symptoms.

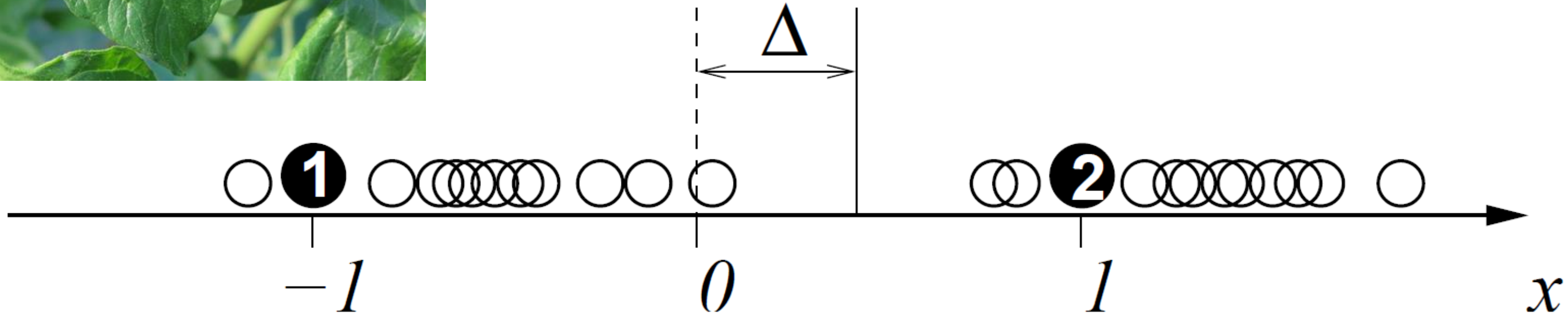
Classification

Labeled Unlabeled

Object detection

Labeled Unlabeled





Introduction to Semi-Supervised Learning, ICML tutorial 2007.

When the EEP annotation strategy is violated, imperfect data appear.

Imperfect data
- incomplete annotation
- inexact annotation
- inaccurate annotation