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

2023 July

Presented by **Mingle Xu**

Core Research Institute of Intelligent Robots

Jeonbuk National University, South Korea

Author list: Mingle Xu 1,†, Hyongsuk Kim 1,†, Jucheng Yang 2,∗, Alvaro Fuentes 1, Yao Meng 1, Sook Yoon 3,∗, Taehyun Kim 4 and Dong Sun Park 1

Content

- A brief history
- The dataset challenge to deploying deep learning in real applications for plant stress recognition
  - Model: deep learning models need more data
  - Task: plant stress recognition has image variations
- Challenges
  - Limited data
    - Class-level: few-shot and class imbalance
    - Dataset-level: domain shift and unknown class
  - Imperfect data
    - Basic information: annotation strategies
    - Incomplete annotation
    - Inexact annotation
    - Inaccurate annotation
    - Incompatible annotation

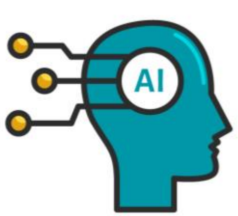
- Limited data
  - class-level
    - few-shot
    - class-imbalance
  - dataset-level
    - domain shift
    - unknown class
- Imperfect data
  - incomplete annotation
  - inexact annotation
  - inaccurate annotation

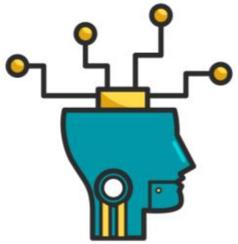
It may **not** be reliable to deploy current deep learning-based models in real-world applications.

The training dataset is one of the main reasons.

Current methods require large-scale datasets.

Models Tasks

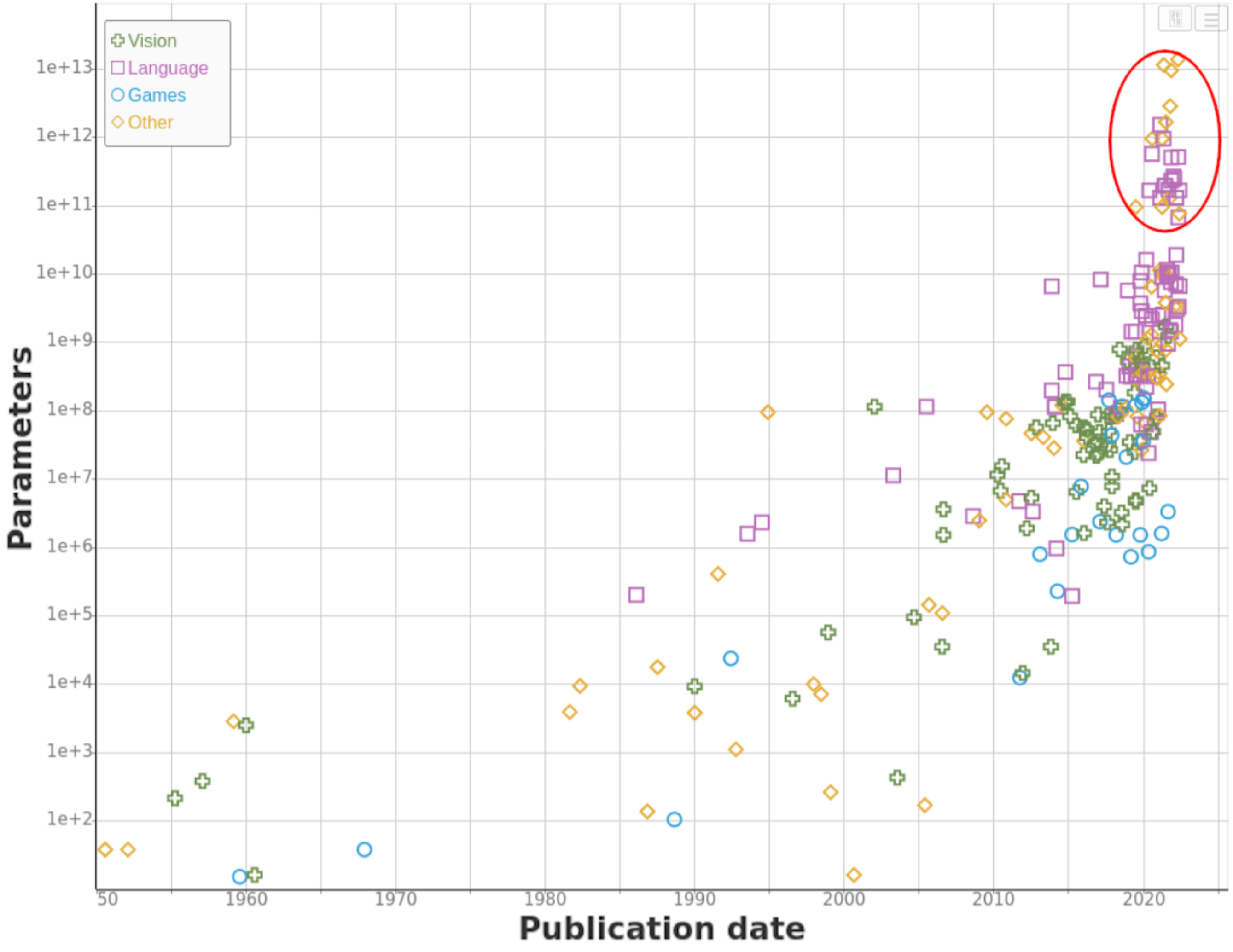


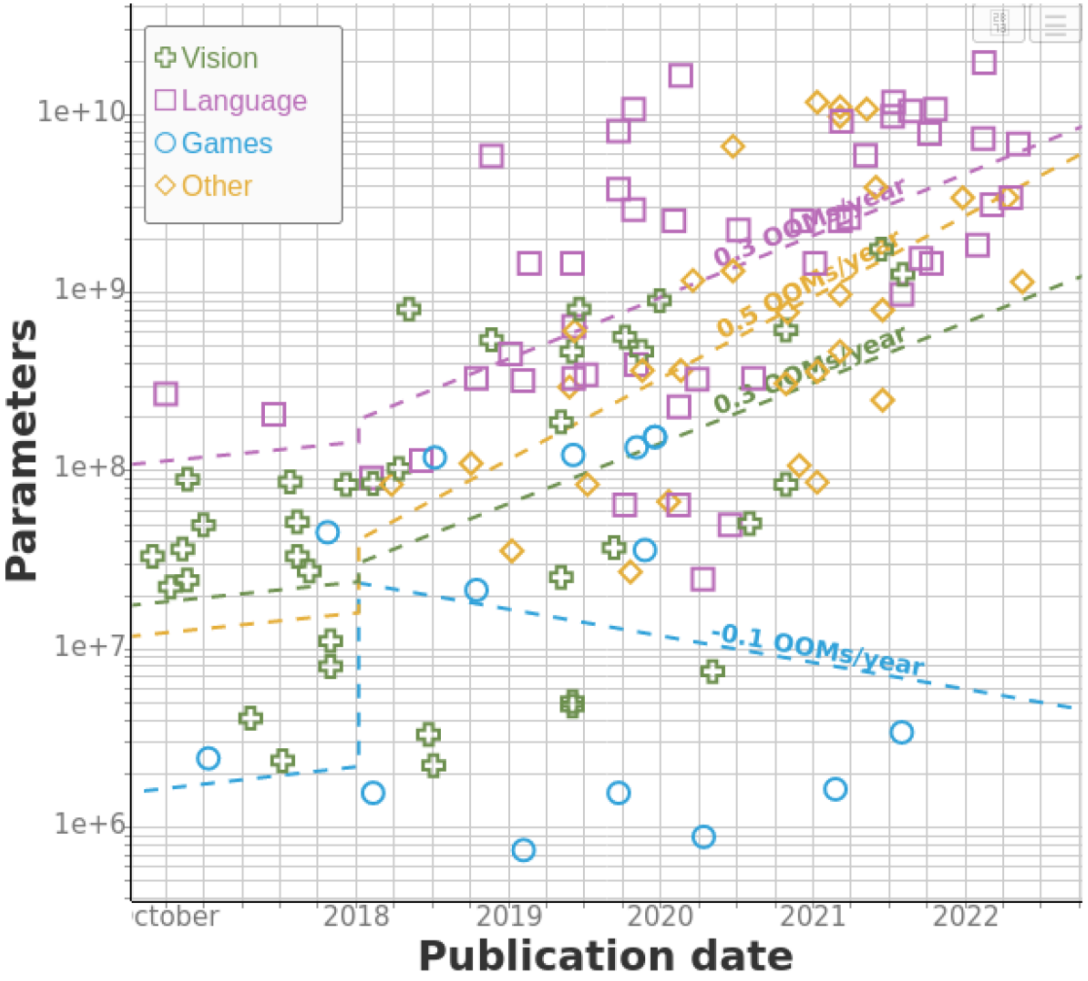


The objective of this presentation:

Models

- Bigger models tend to have more data to be trained well.

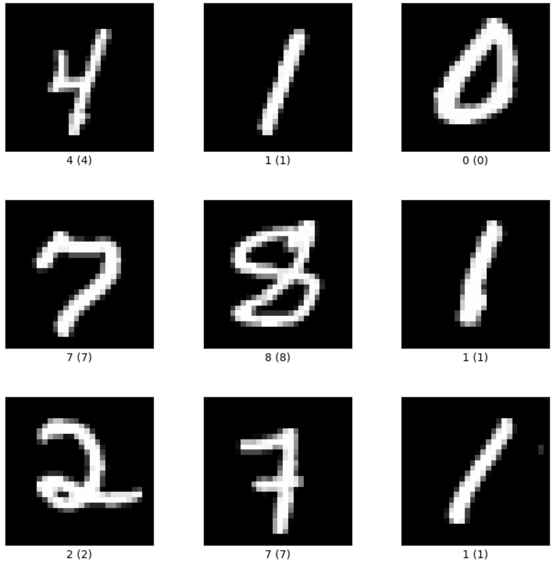




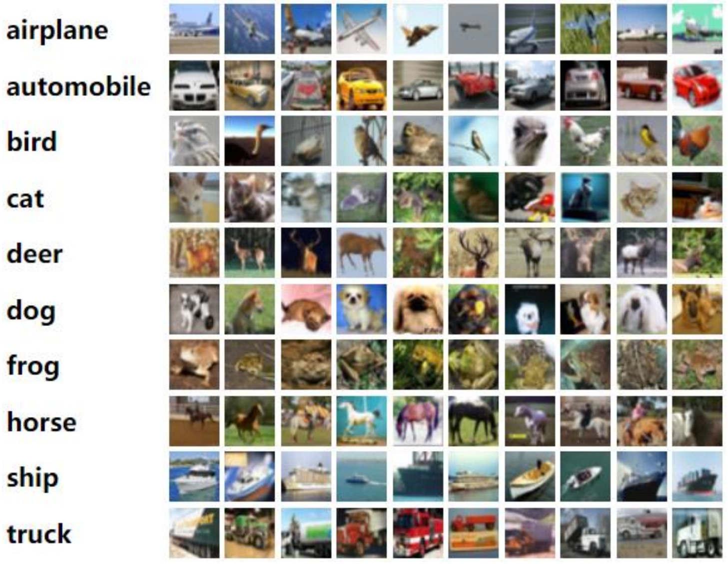
"Machine Learning Model Sizes and the Parameter Gap." *arXiv* (2022).

Task: Plant Stress Recognition

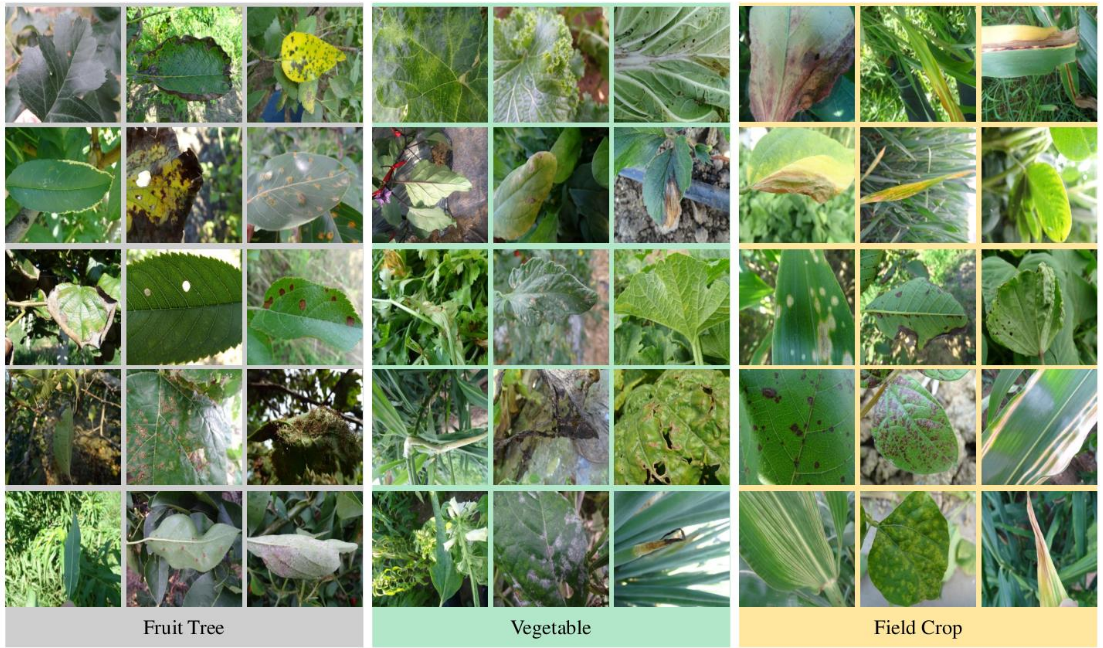
- Plant stress recognition in <u>real-world applications</u> is complex.



MNIST



CIFAR10



Plant stress images from PDD271

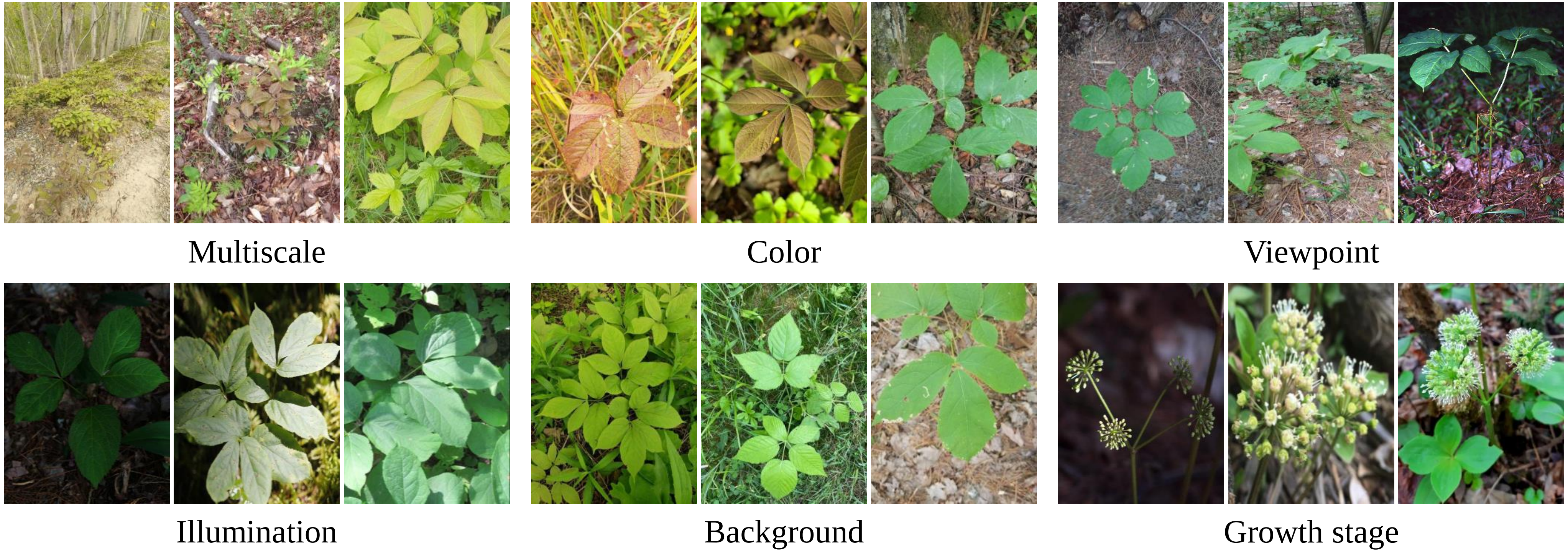
"Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998)

"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.



Images from the same class (PlantCLEF2022)

- More examples of intra-class variation.

Different types of tomato leaves

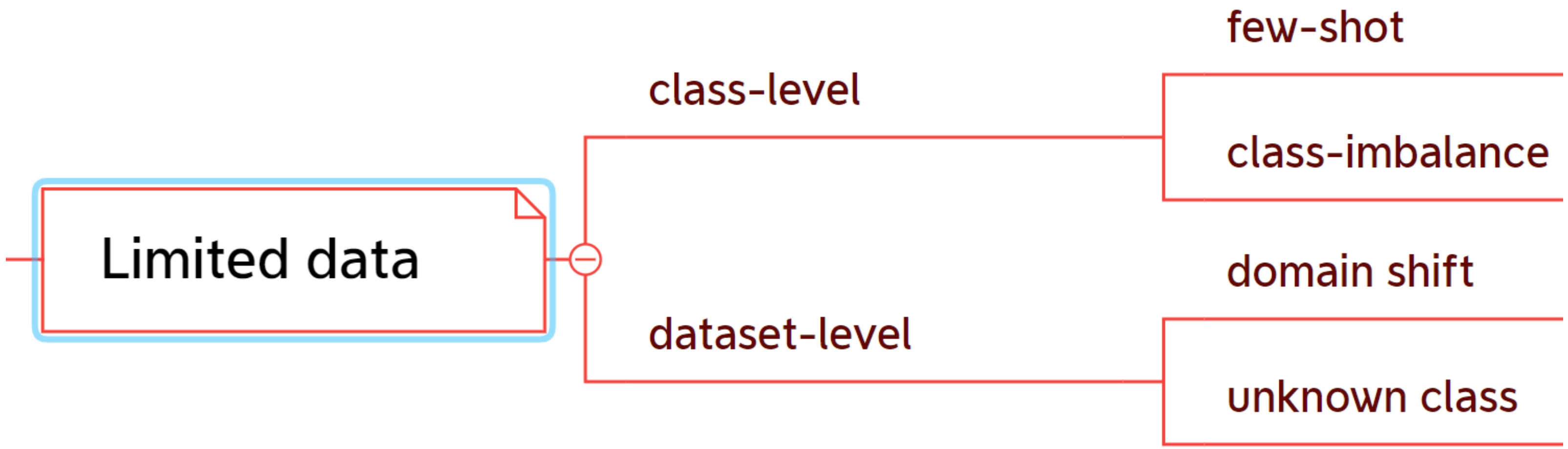
Different stages of tomato stress

- Inter-class variation, a relative issue.
- (b) (d) is more similar to (a) (c).

(a) Tomato leaf mold (b) Tomato magnesium deficiency

(c) Tomato canker (d) Tomato chlorosis virus (ToCV)

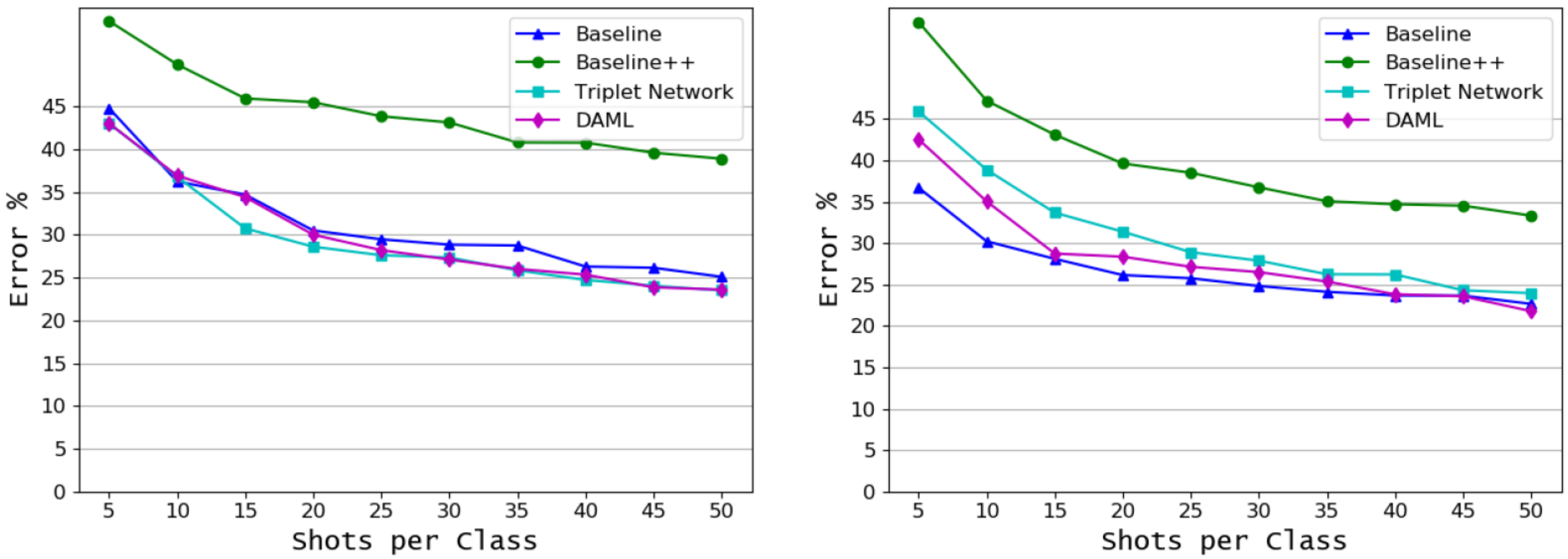
If our training dataset is **not** in large-scale, …



Few-shot

- Strictly, every class has the same number and few images, such as M =10, 20, and 100.
- $n_{X_i} = M$. n_{X_i} is the number of images for class i in a training dataset, M is the exact number.
- In this scenario, holistic performance is not good.
- The performance for every class is not good either.

Experiment in PlantVillage.



(a) ResNet18 (b) ResNet34

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

Few-shot

- We define generalized few-shot, where classes have similar lower performance.
- $n_{X_i} \approx M$ and $\boldsymbol{M}$ **may be large**, such as 200 (depending on task complexities).
  - If a task is more complex, larger M may still result in a few-shot issue → collecting more data.

- Advantage:
  - Recognizing the situation of applications
  - Use corresponding methods.

"Embracing Limited and Imperfect Data: A Review on Plant Stress Recognition Using Deep Learning." arXiv (2023).

Class imbalance

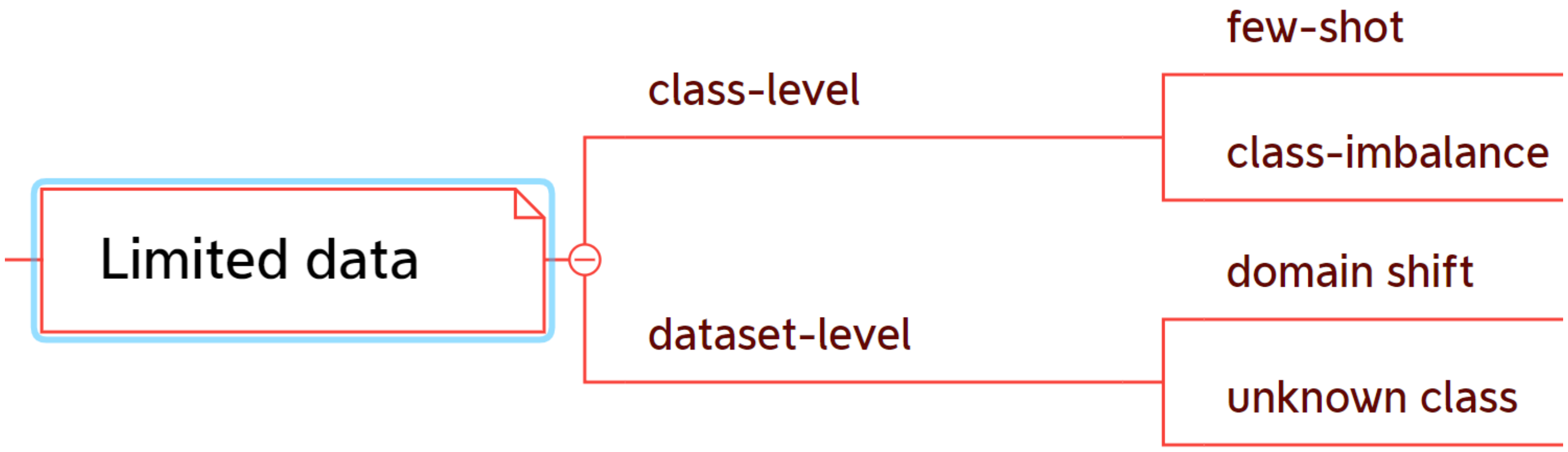
- One class has much more images than another class, $n_{X_i} \gg n_{X_j}$.
- In this scenario, the former class may have much better performance than the latter.

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



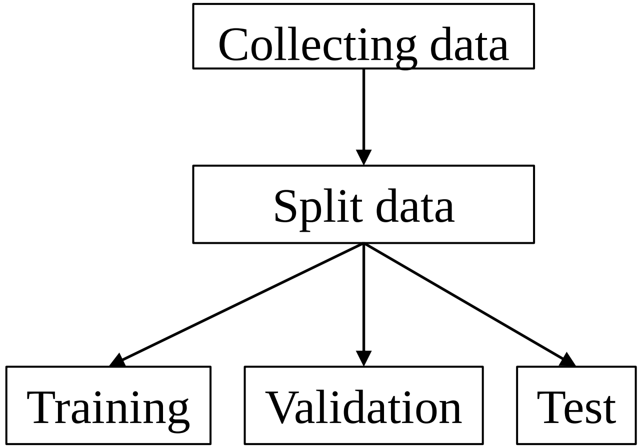
Just describes the situation within the training datasets.

Considers the difference between the test and training datasets.

A basic assumption of deep learning

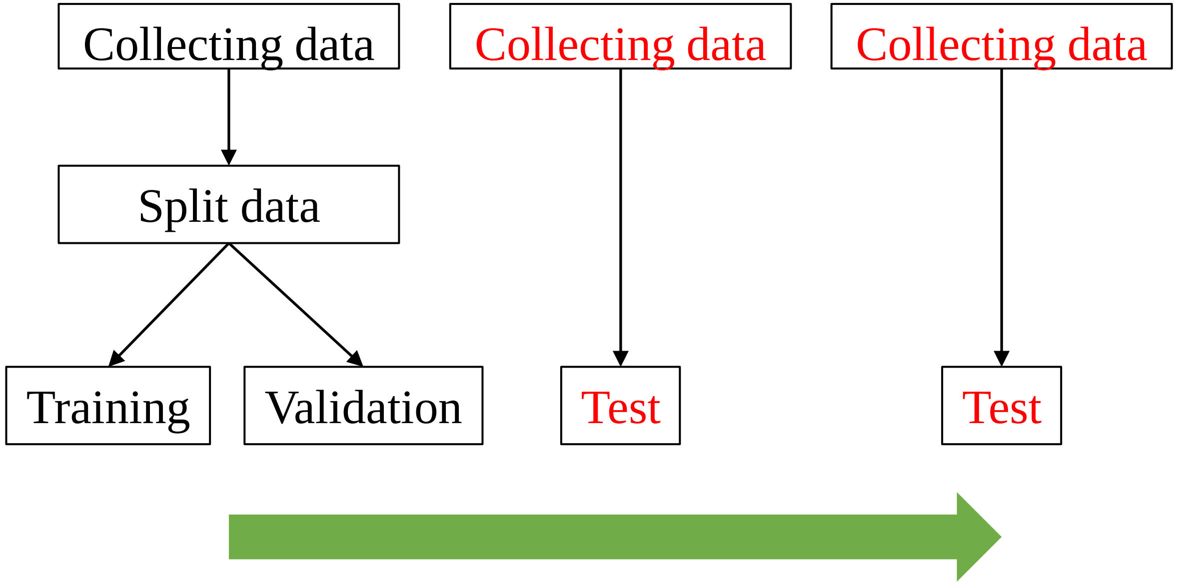
The training and test datasets are in same or similar distribution: $P_{train} \approx P_{test}$.

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.

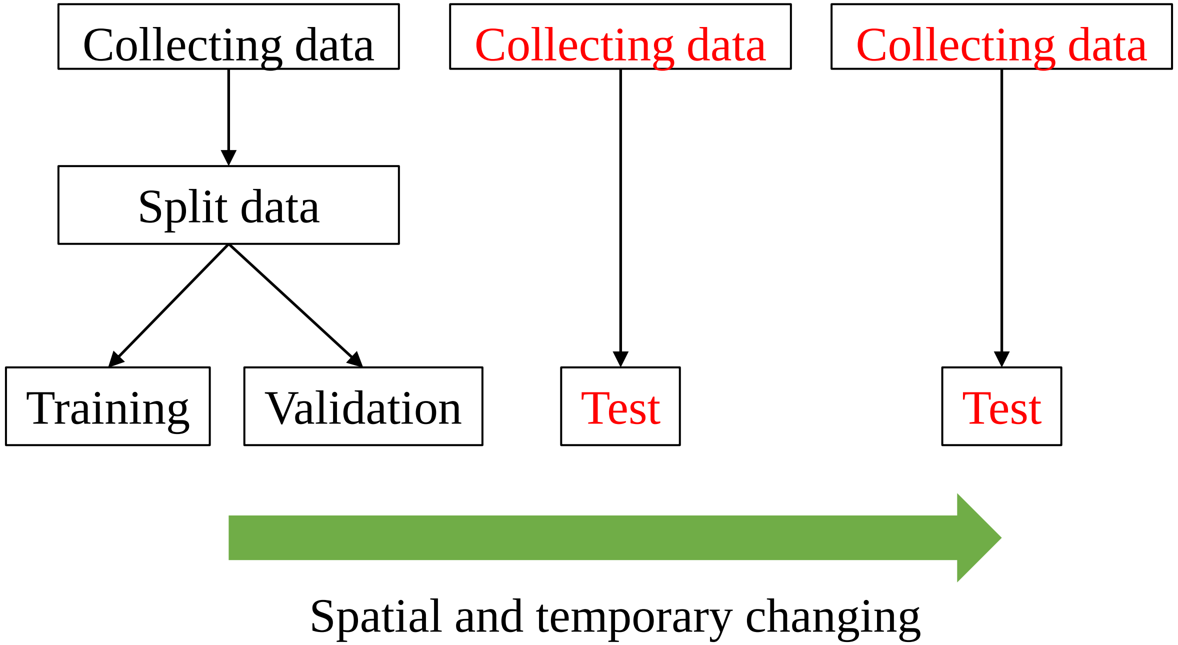


Spatial and temporary changing

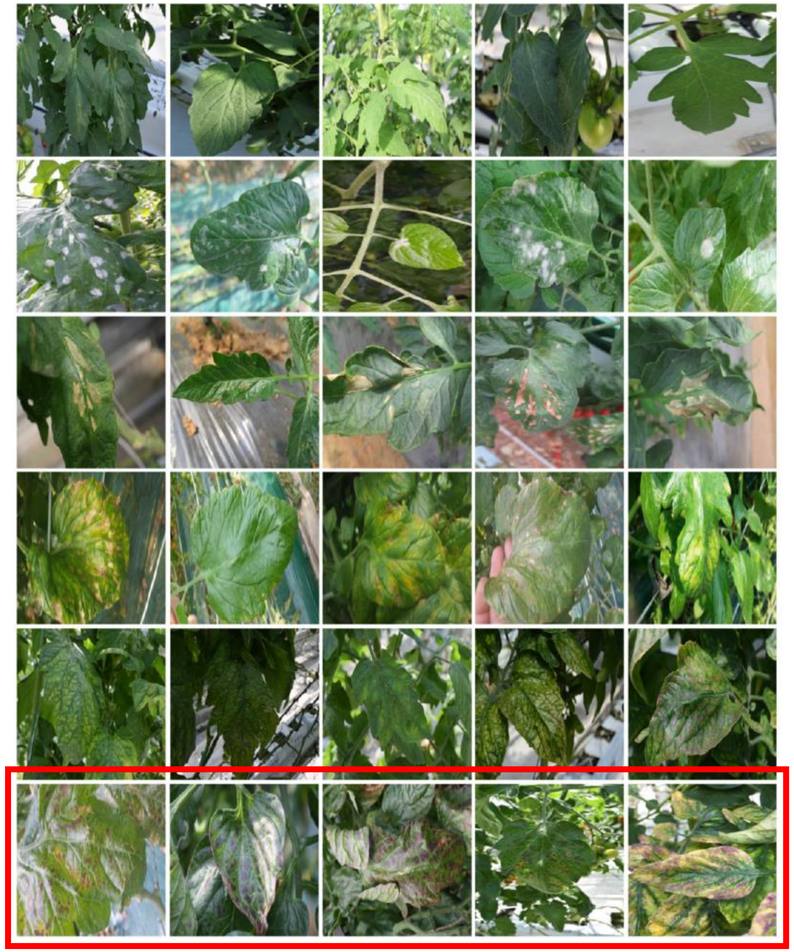
If you can collect a dataset large enough, no problem.

Dataset-level: unknown class

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



New types of plant stress may appear.

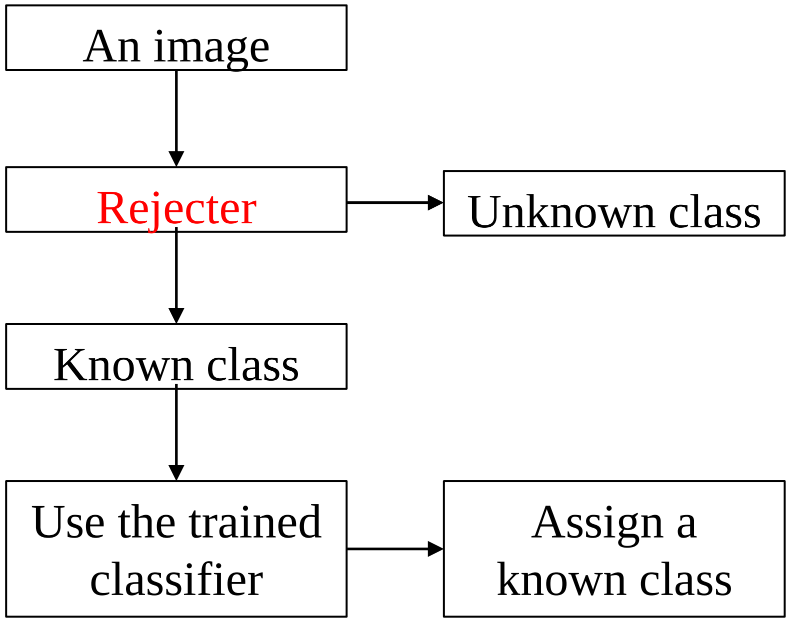


A new class in the test dataset.

Dataset-level: 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.

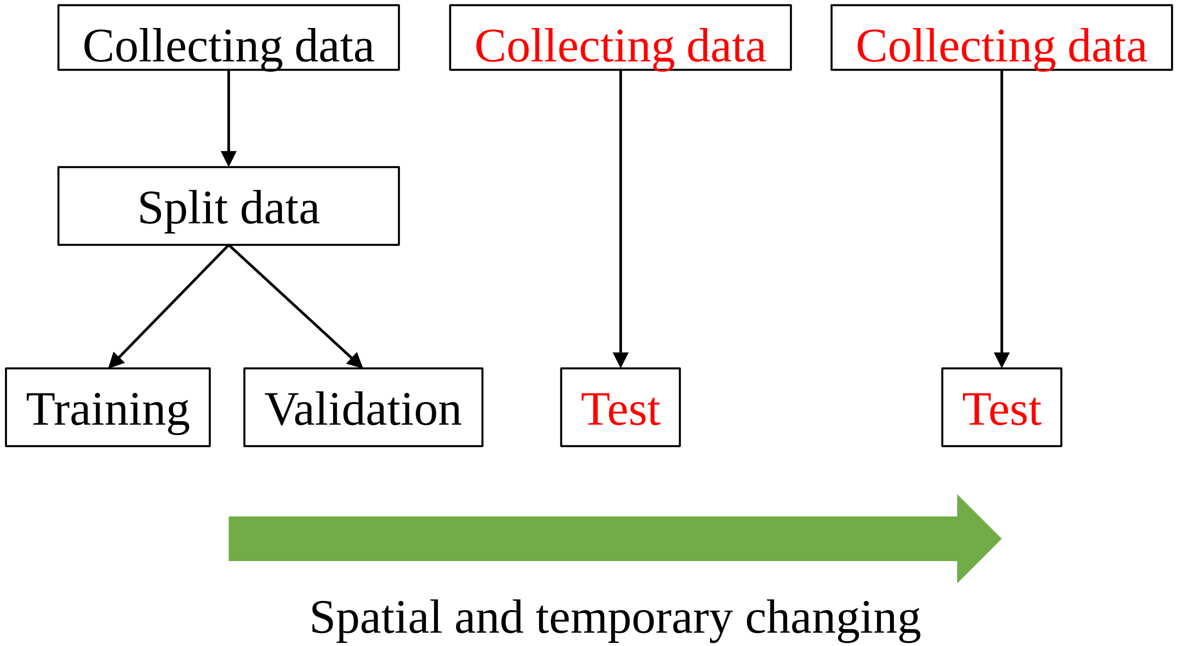
What is desired? Open set recognition (OSR).



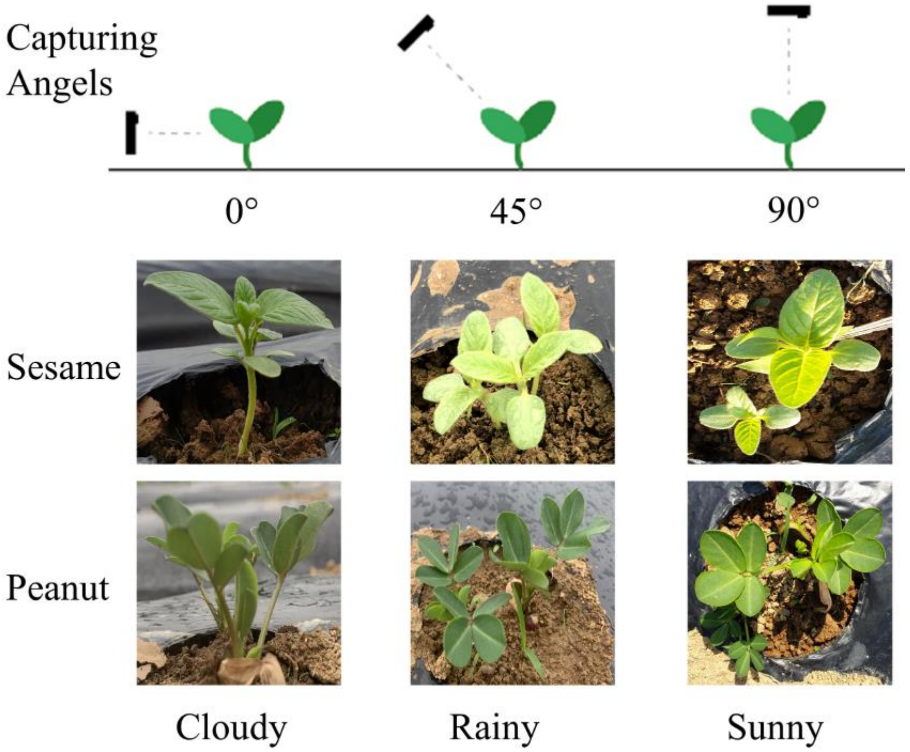
“Known and Unknown Class Recognition on Plant Species and Diseases”, submitted (2023).

Dataset-level: domain shift

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



Plant stresses have a new intra-class variation.



"CWD30: A Comprehensive and Holistic Dataset for Crop Weed Recognition in Precision Agriculture." *arXiv* (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.

Examples of real scenarios.

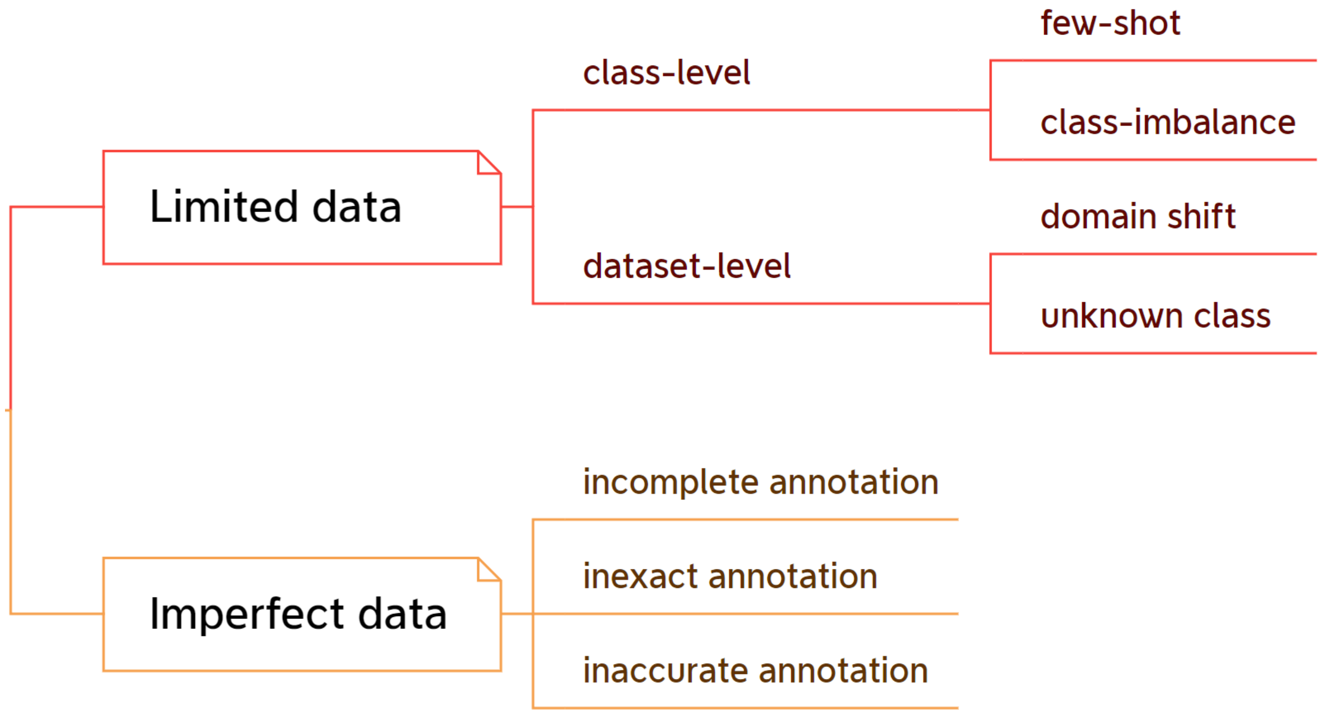
Images vary in different farms such as the background.

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

A short summary

Consider the images.

Consider the annotations.



Annotation and its strategies

Different computer vision objectives have different annotations.

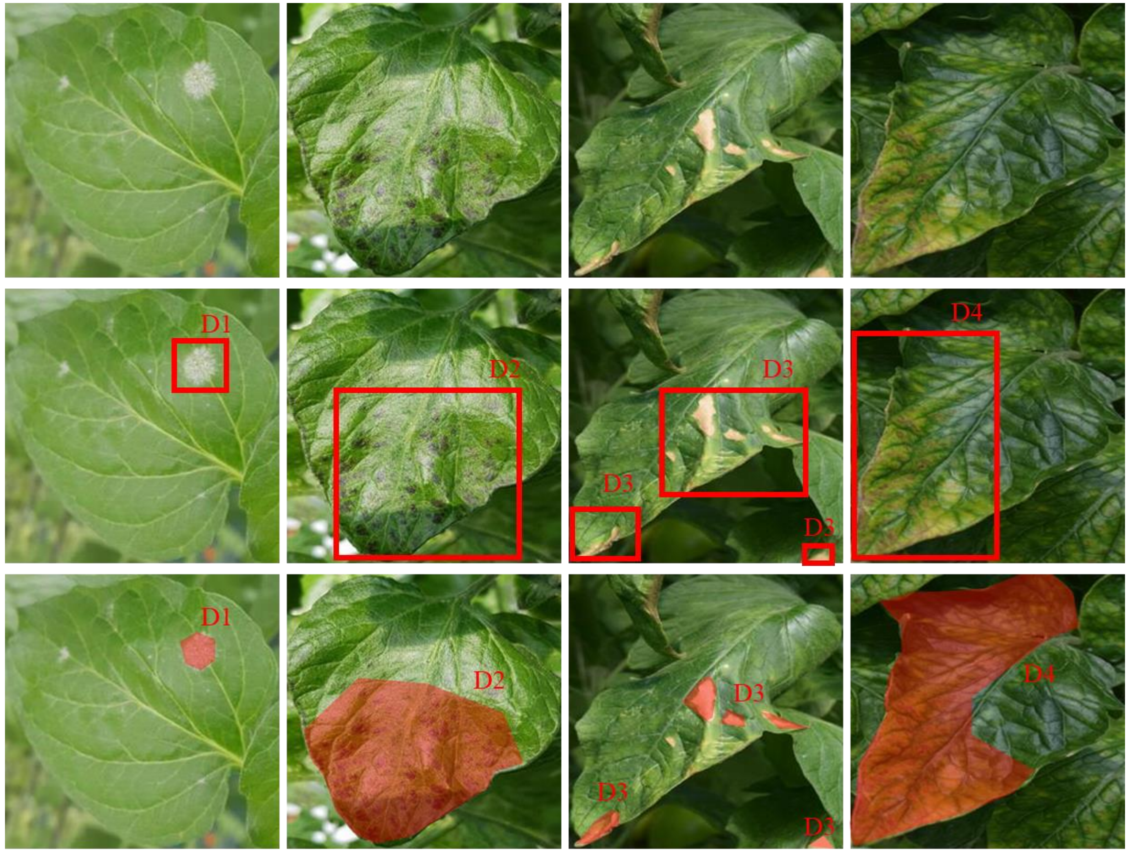


Image classification

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

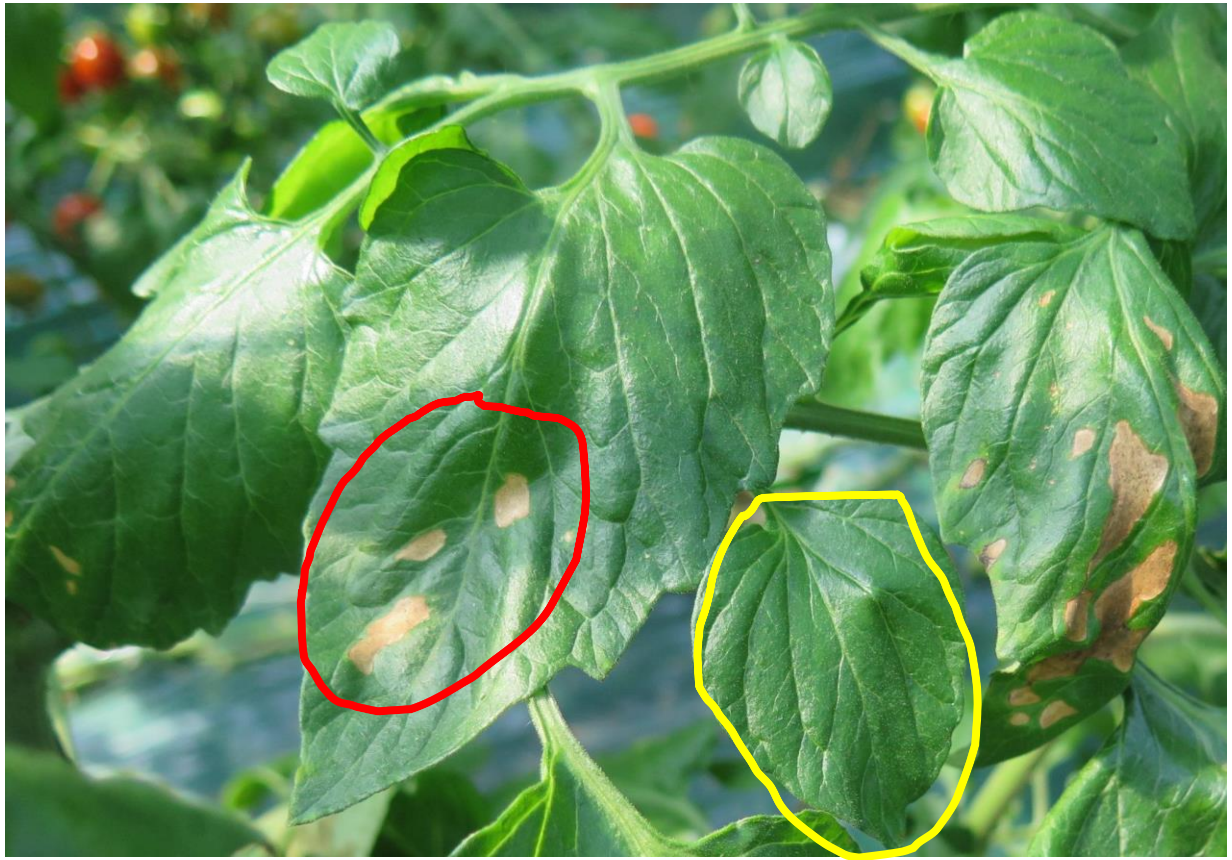
Exclusive: one annotation has only one class.

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.

Image classification

Labeled as powdery mildew?
But it also has ToCV.

Preprocessing: crop.

Desired annotation strategy: EEP

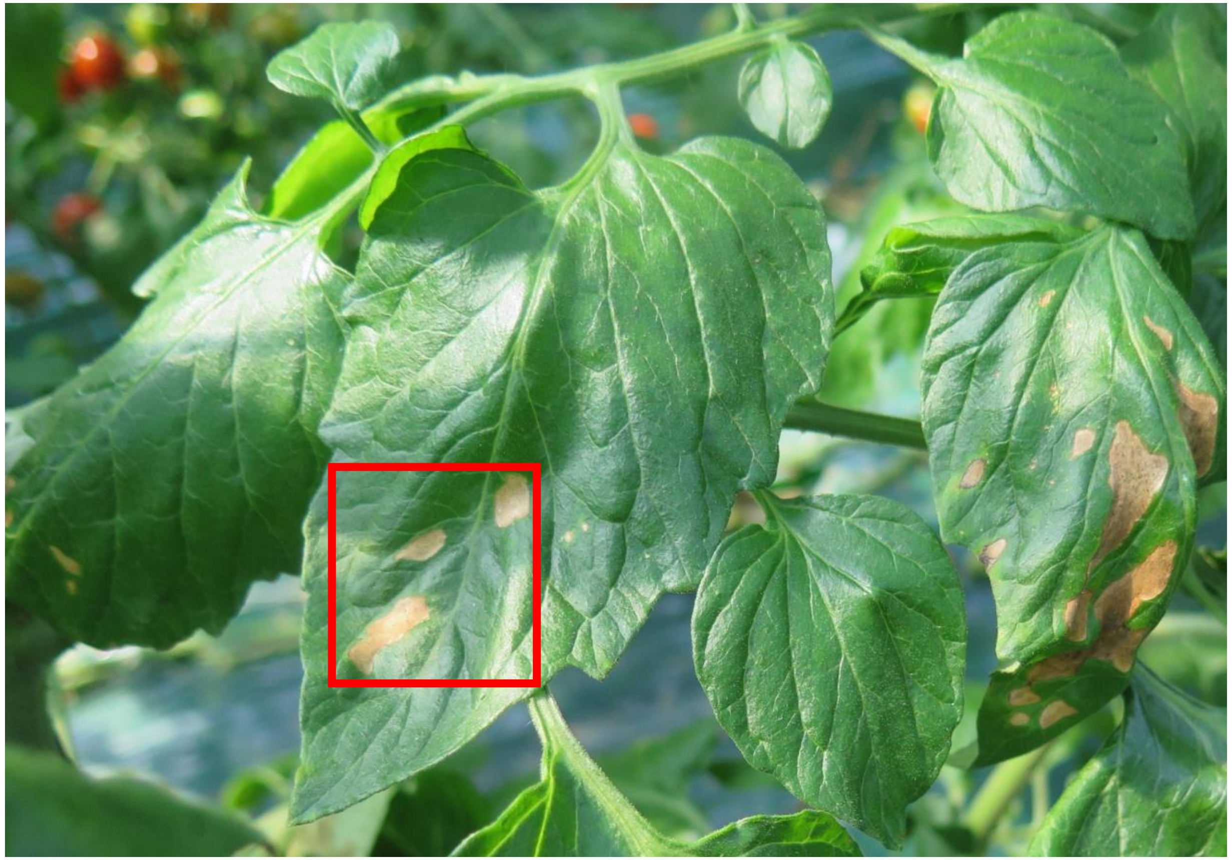
Exclusive: one annotation has only one class.

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.



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.

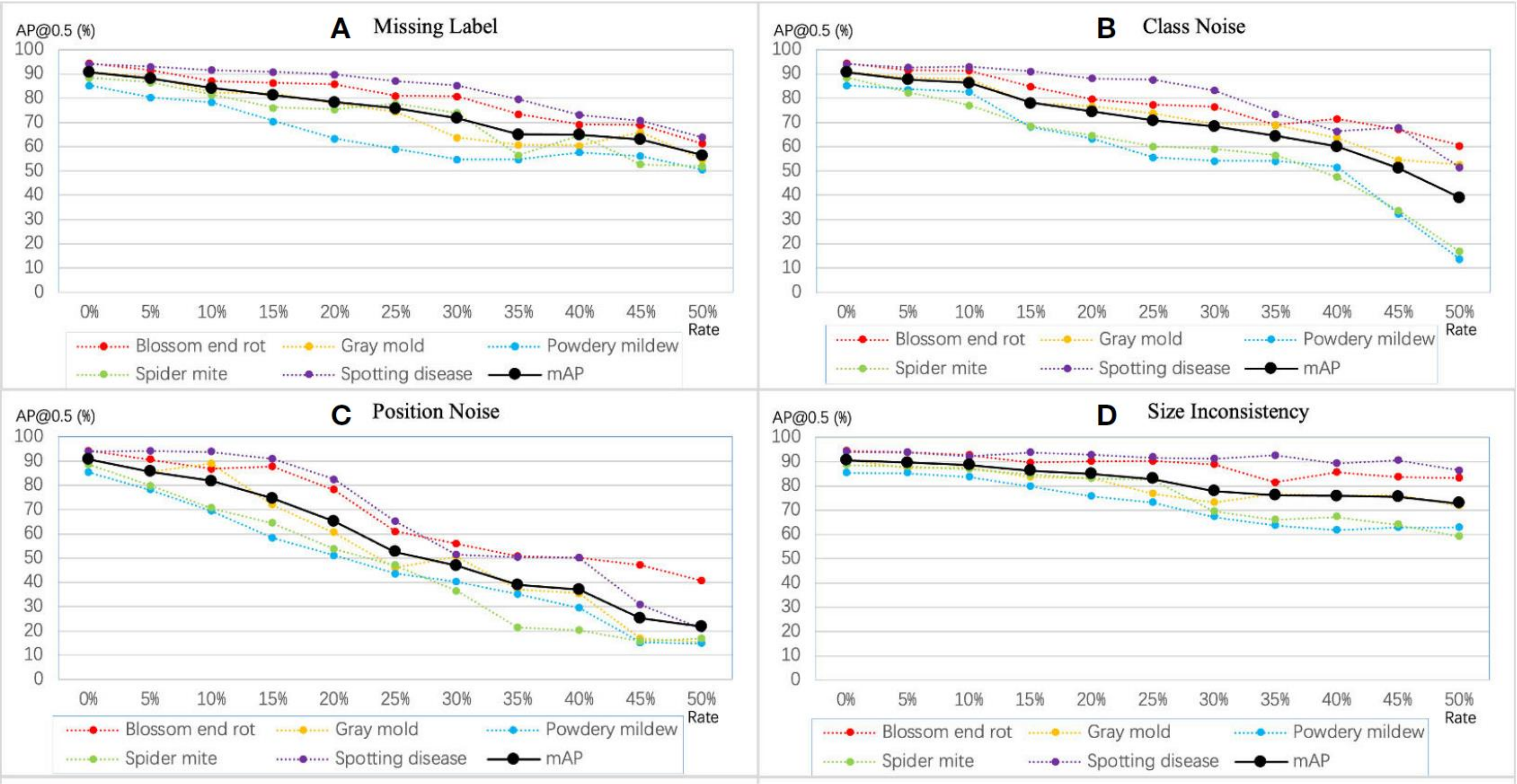
Object detection

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

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

Violating the EEP strategy results in challenges.

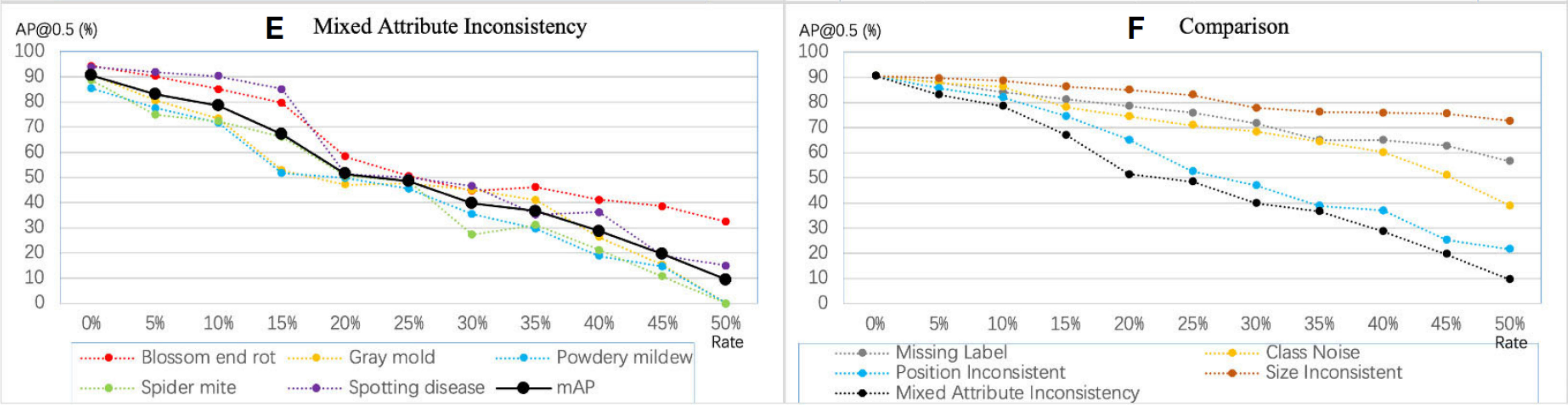
Individual noise



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

Violating the EEP strategy results in challenges.

The mixed impact.



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

Thank You

Questions and Comments

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

Public slides:

Thank you

Photo by istock/fotokostic

Photo by Mingle Xu in RDA in South Korea