# Semi-Supervised Learning in Deep Neural Networks

## MixMatch

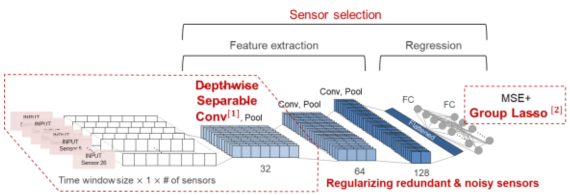
2020. 12. 4.

Data Mining & Quality Analytics Lab.
이민정
Minjung Lee
leemj2520@korea.ac.kr

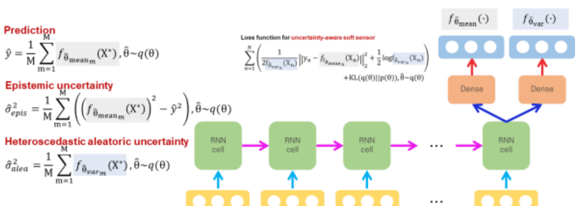
# Presenter

- 이민정 (Minjung Lee)
  - 고려대학교 산업경영공학부 재학 중
  - 석·박사 통합과정(2017.03~) Ph.D. Candidate
  - Data Mining & Quality Analytics Lab(김성범 교수님)
- Research Interest
  - Deep learning for multivariate sensor data
  - Incomplete multivariate sensor data

✓ Regularized-CNN for sensor data




✓ Uncertainty-aware soft sensor




✓ Capsule networks

✓ Deep generative models and inference

✓ Understanding uncertainty and Bayesian convolutional neural networks

✓ Graph convolution networks

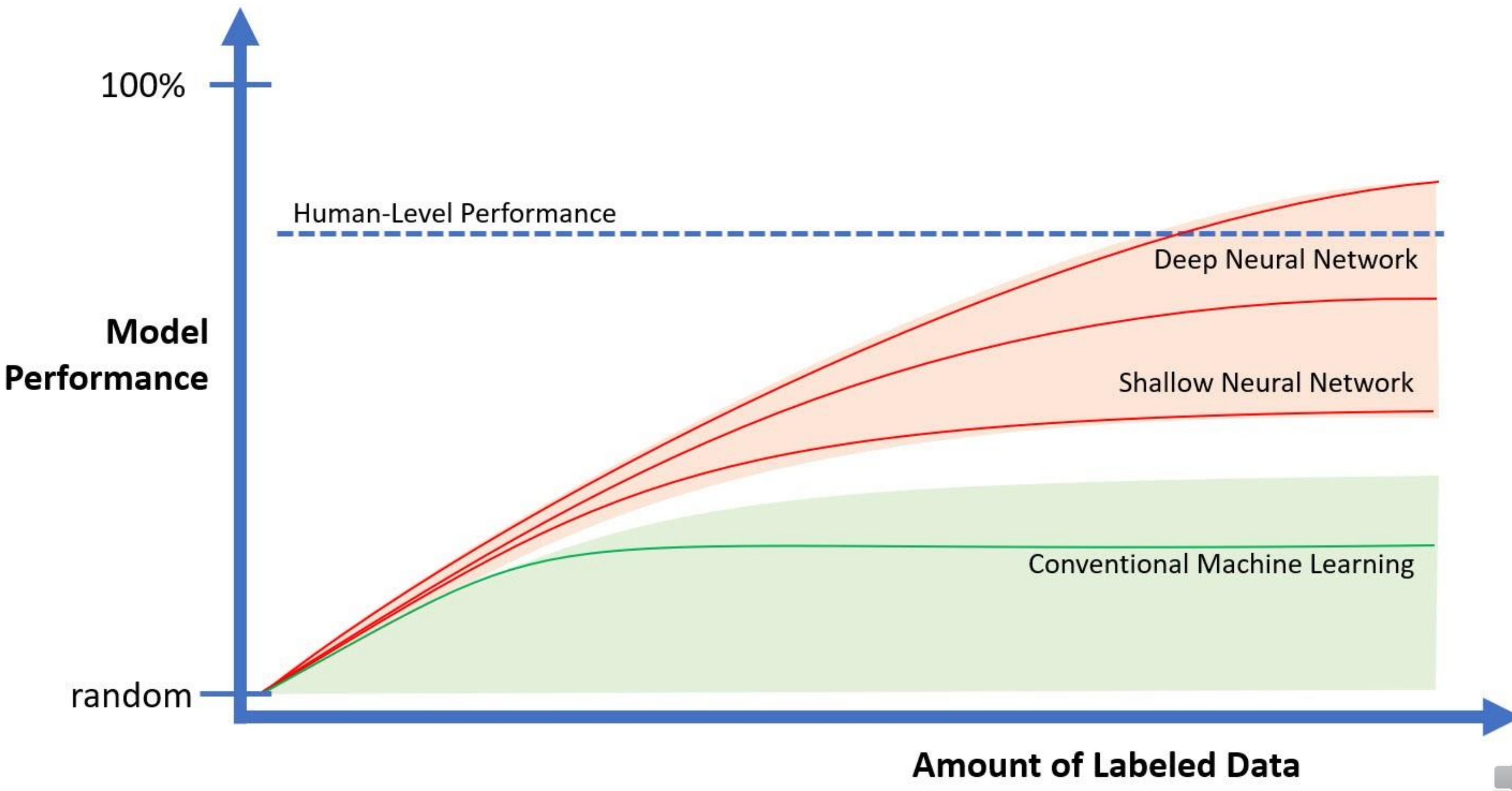
✓ Bayesian optimization

# Contents

# Introduction

딥러닝 모델 우수한 성능을 위해서는

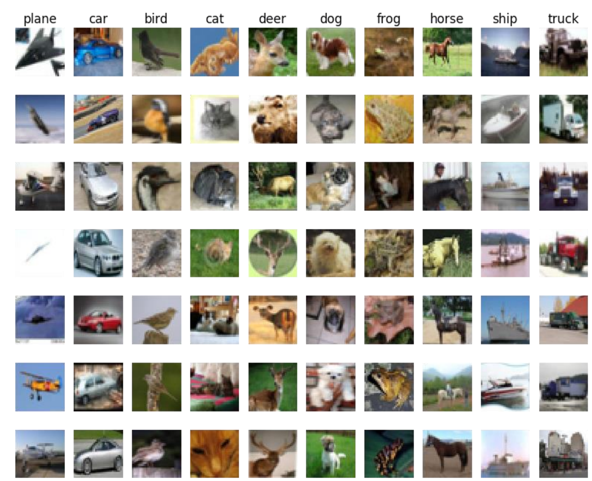
많은 양의 레이블드 데이터(labeled data)가 필요!



# Introduction

## 컴퓨터 비전 문제 해결을 위해 많은 양의 레이블드 데이터를 사용한 다양한 지도 학습 기반 알고리즘 개발이 이루어짐

MNIST




CIFAR-10

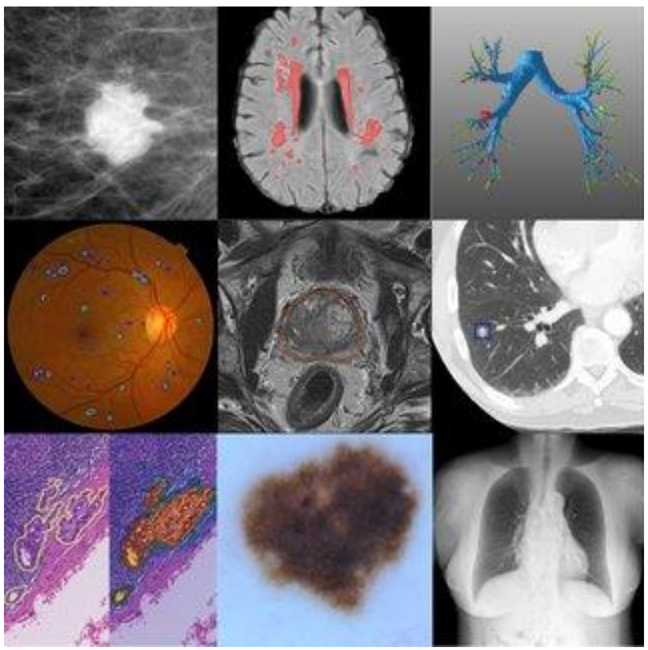

MSCOCO

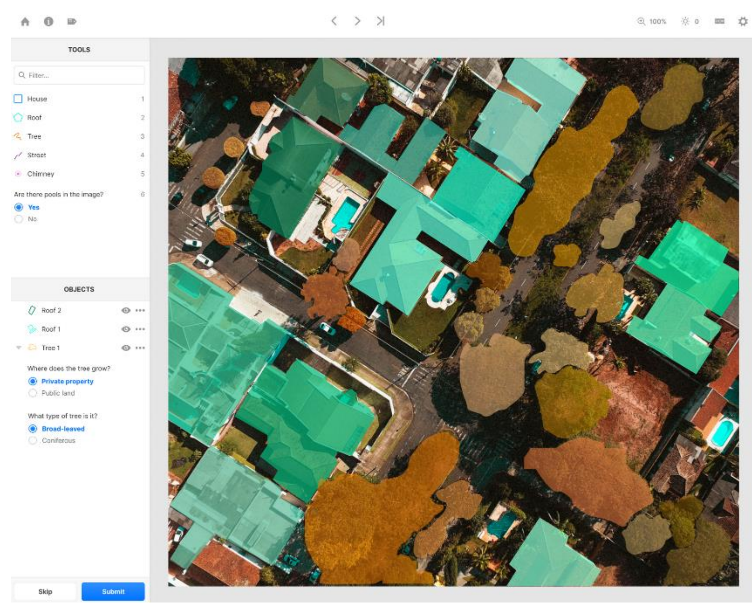
ImageNet

# Introduction

하지만 실제 현실 문제에서는 많은 양의 레이블드 데이터를

확보하는 것은 현실적으로 제약이 따름





# Introduction

이러한 문제 상황을 해결하기 위한

다양한 방식의 연구들이 시도

↓

Transfer learning

Meta learning

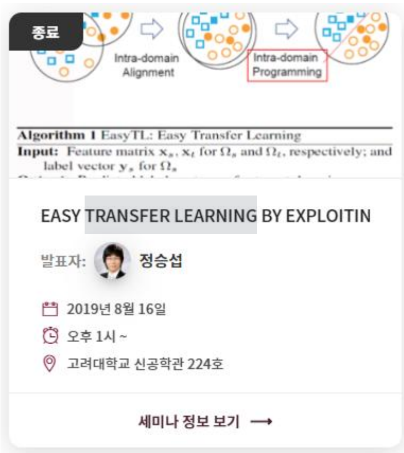
Weakly-supervised learning

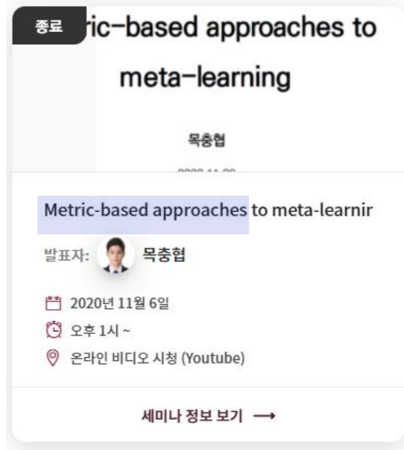
Self-supervised learning

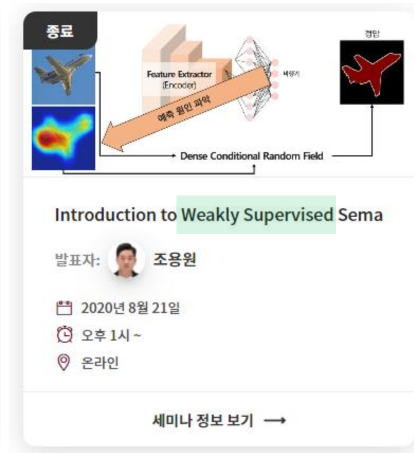
Semi-supervised learning

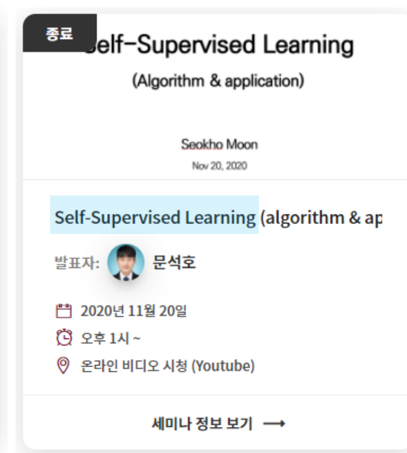
...

# Introduction




Transfer learning

Meta learning

Weakly-supervised learning

Self-supervised learning

Semi-supervised learning

…

# Introduction

오늘의 주제 : Semi-supervised learning

in deep neural networks

Transfer learning

Meta learning

Weakly-supervised learning

Self-supervised learning

Semi-supervised learning

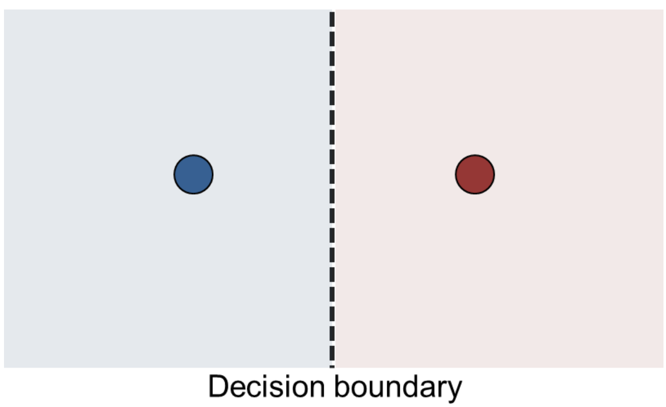
...

# Semi-supervised learning

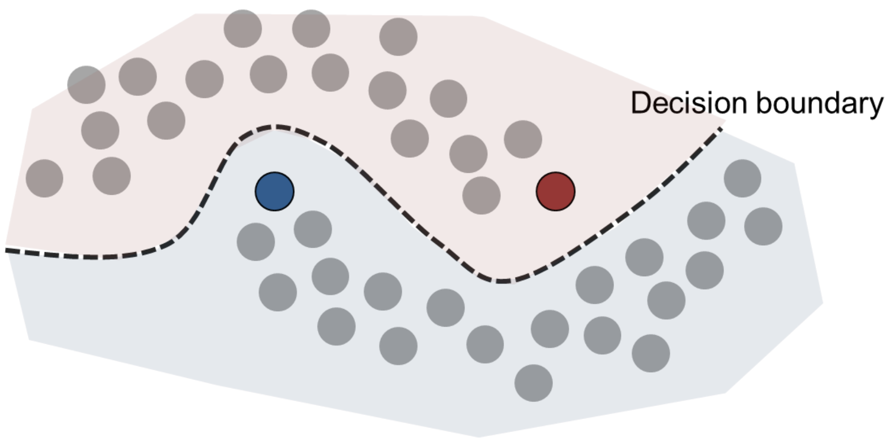
## Semi-supervised learning

Unlabeled 데이터를 사용하여
일반화 성능을 높이는 모델 만들자

Supervised

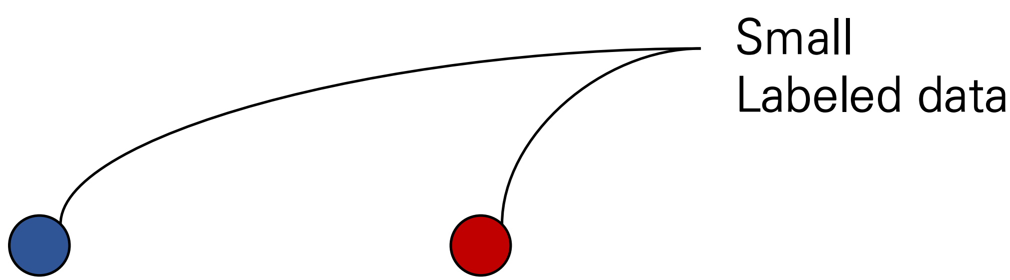



Semi-supervised



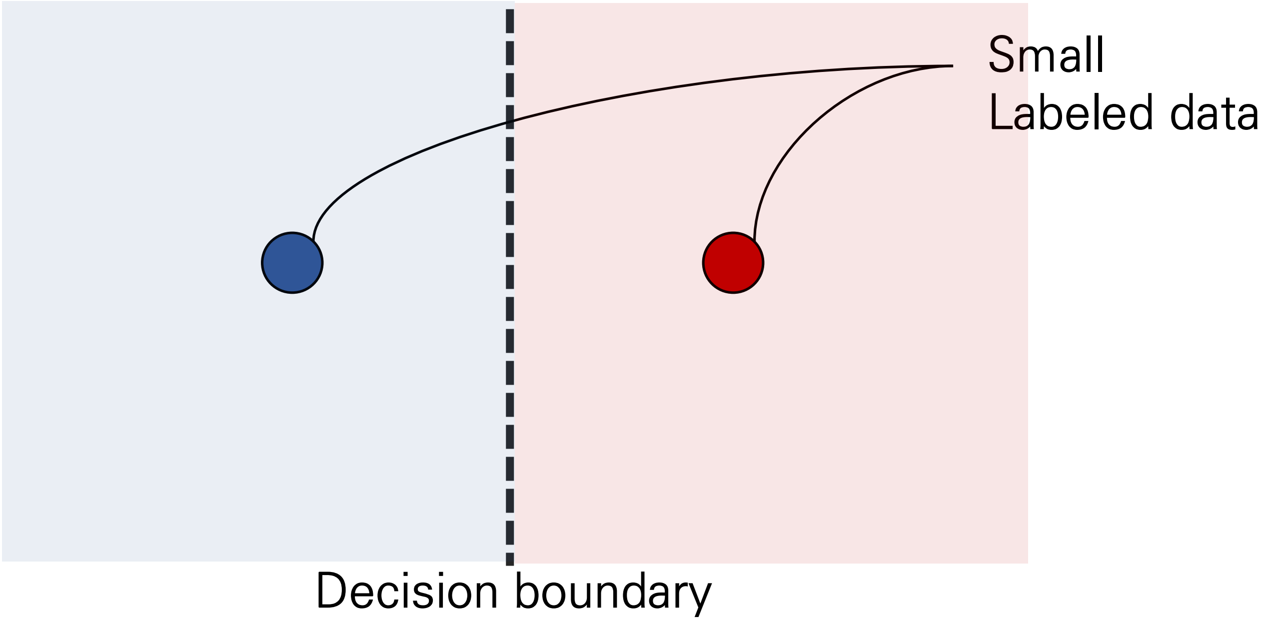
# Semi-supervised learning

## Supervised learning



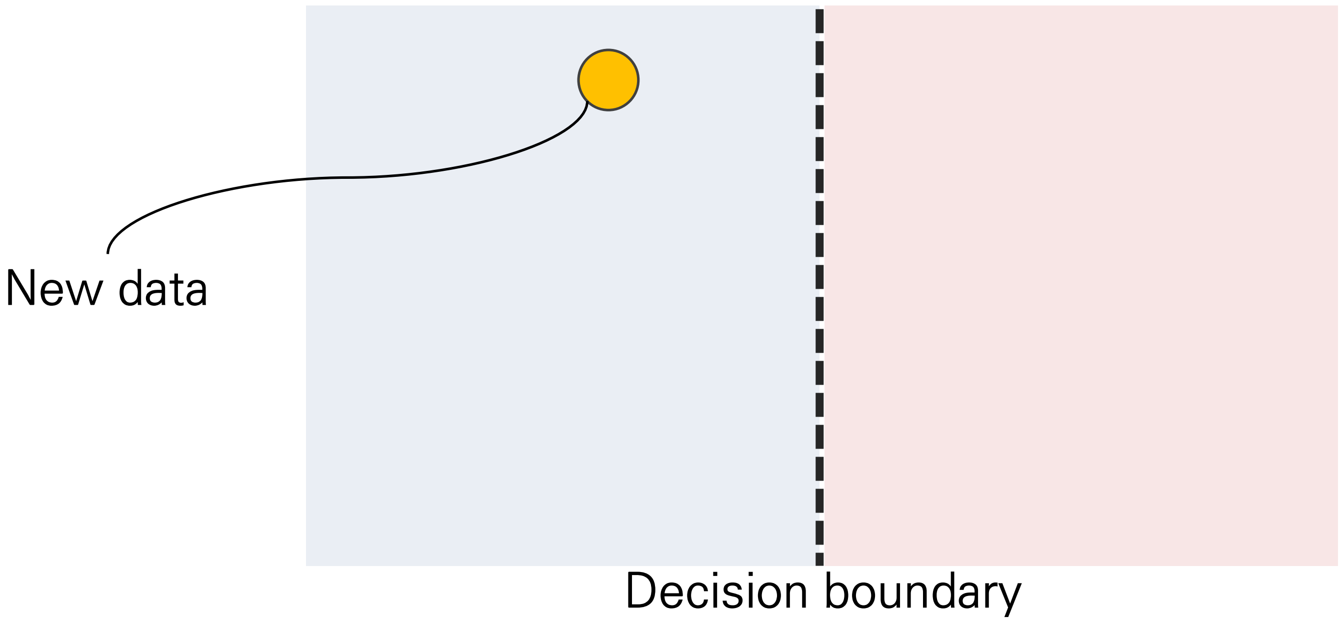
# Semi-supervised learning

## Supervised learning



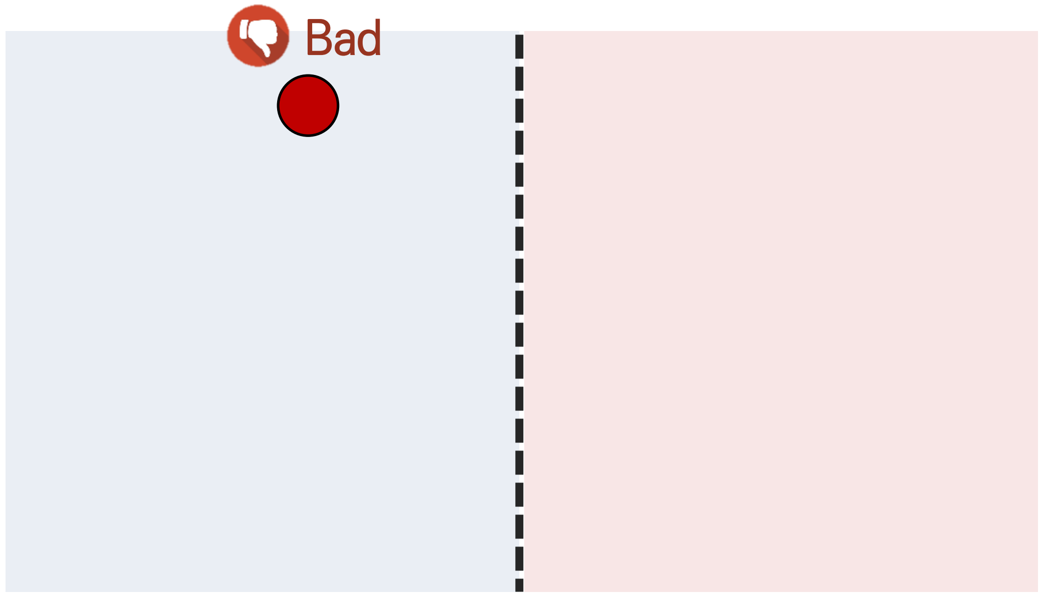
# Semi-supervised learning

## Supervised learning

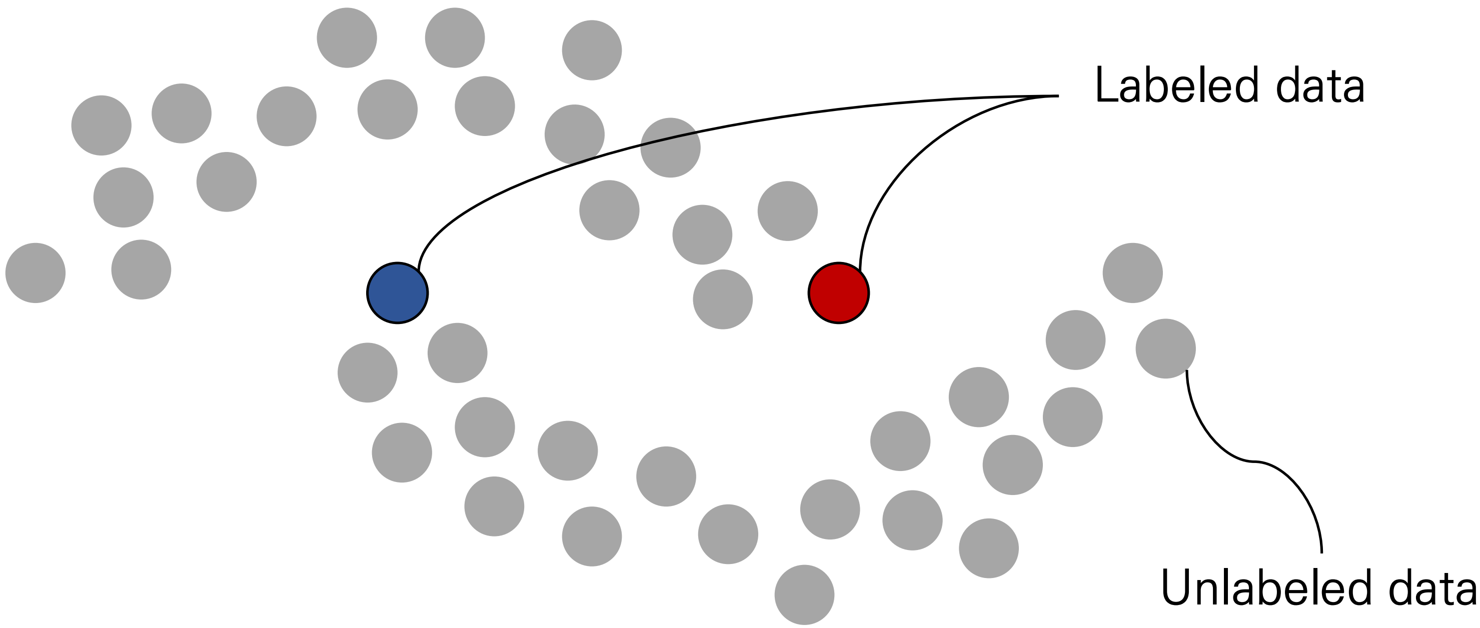


# Semi-supervised learning

## Supervised learning

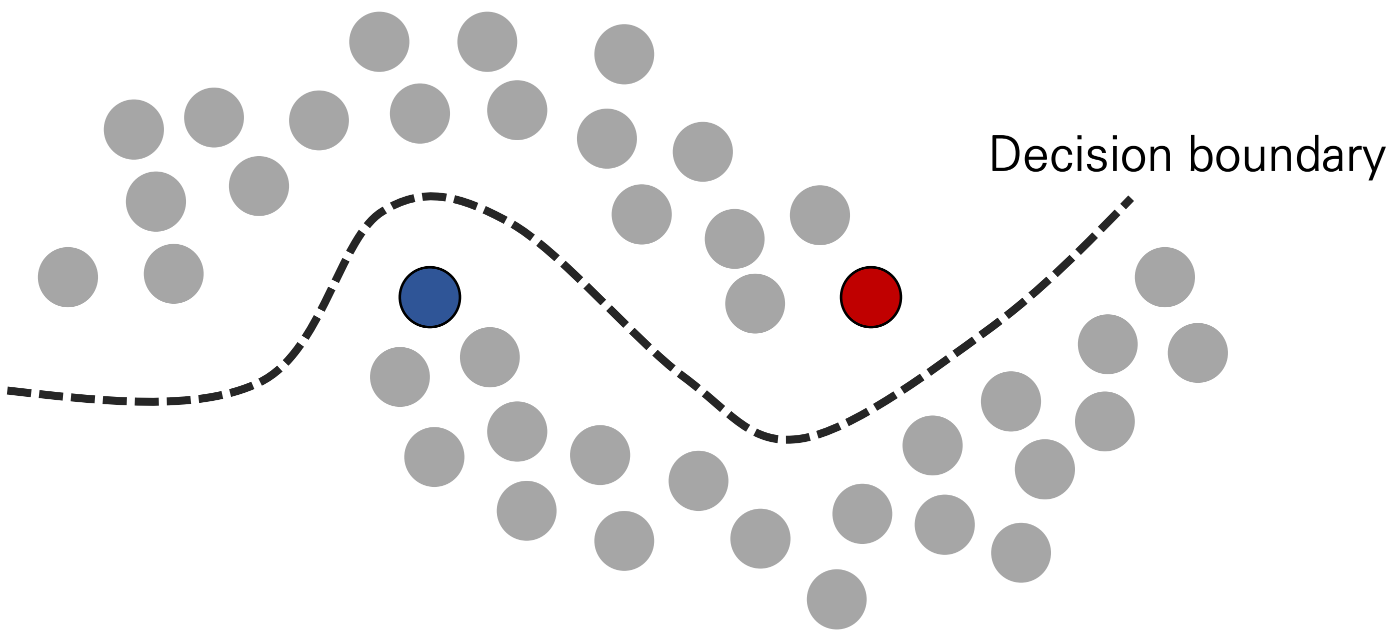


## Semi-supervised learning



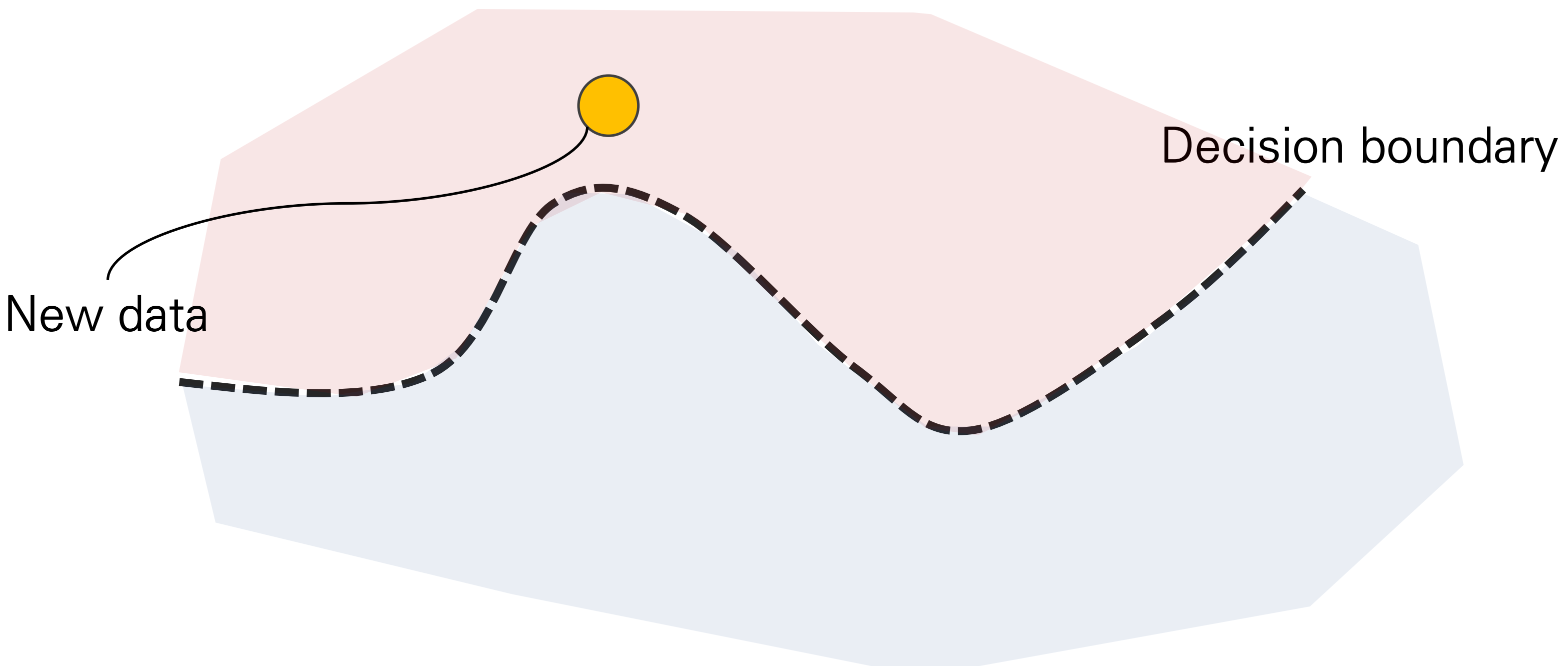
# Semi-supervised learning

## Semi-supervised learning



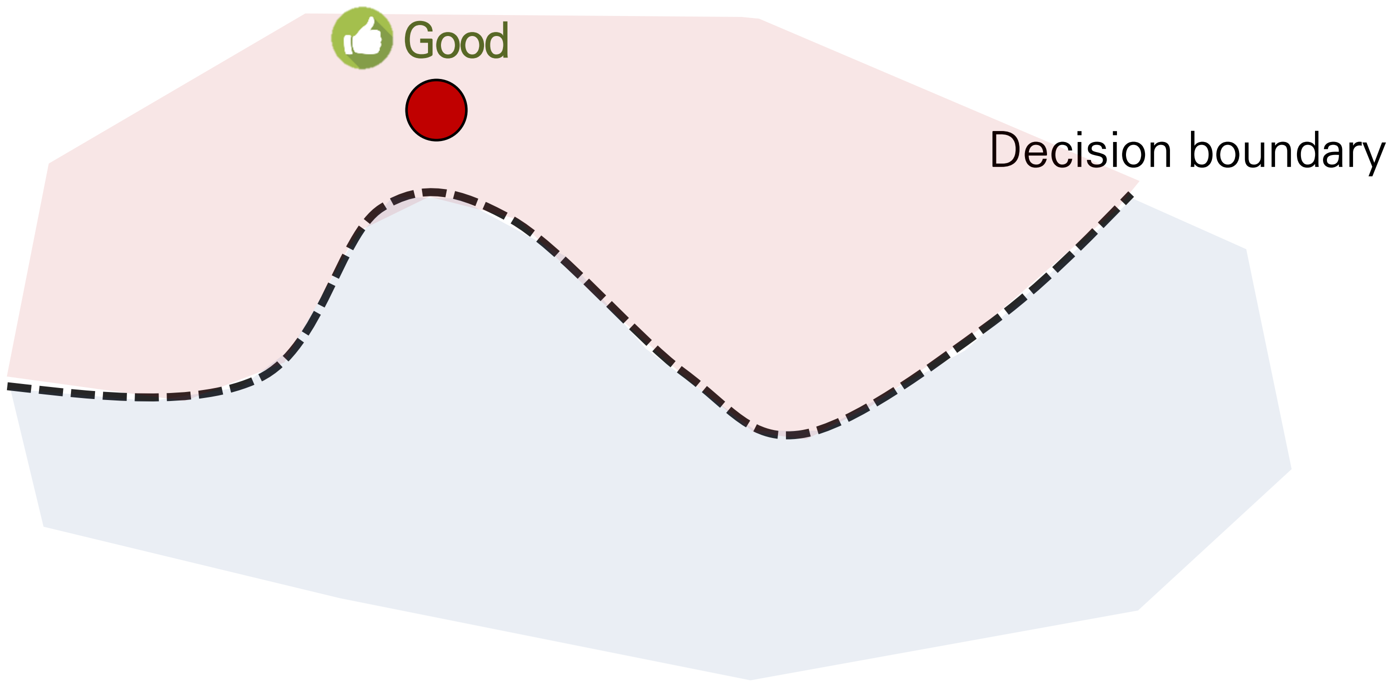
# Semi-supervised learning

## Semi-supervised learning



# Semi-supervised learning

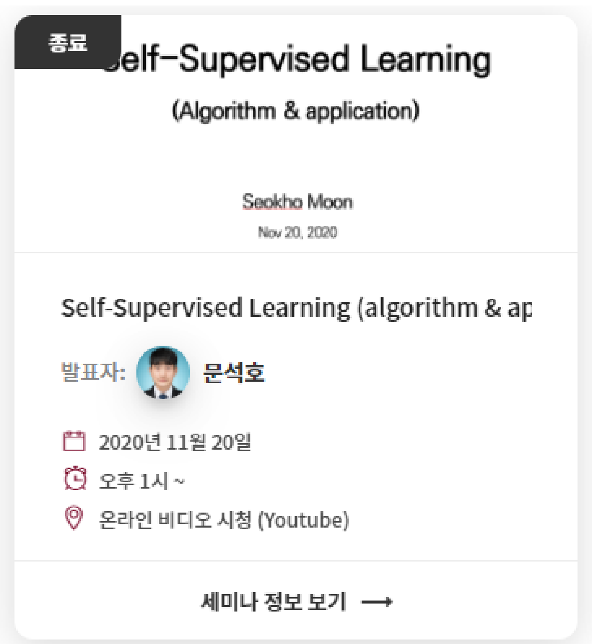
## Semi-supervised learning

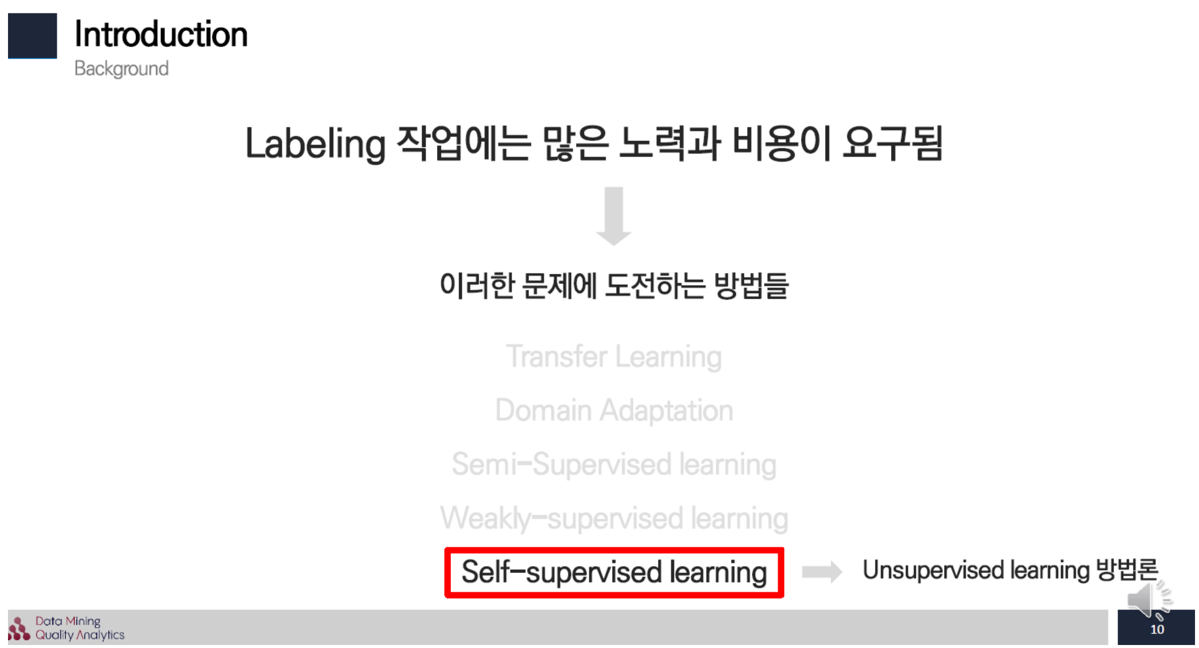


# Semi-supervised learning

## Unlabeled data 사용?

## Self-supervised learning과 차이점은?




2020. 11. 20. 문석호 연구원 Self-Supervised Learning (algorithm & application) 세미나

# Semi-supervised learning

## Unlabeled data 사용?

## Self-supervised learning과 차이점은?

| Self-supervised | Semi-supervised |
| --- | --- |
| Unlabeled data 사용<br>+<br>Two-stage | Unlabeled data 사용<br>+<br>One-stage |

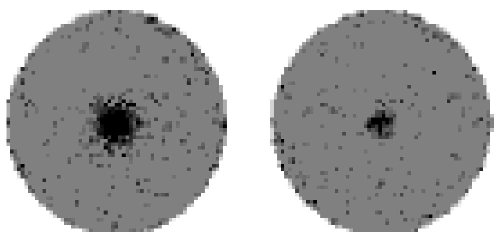
Schmarje, L., Santarossa, M., Schröder, S. M., & Koch, R. (2020). A survey on Semi-, Self-and Unsupervised Learning for Image Classification. arXiv preprint arXiv:2002.08721.

# Semi-supervised learning

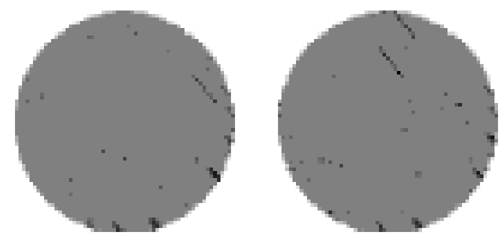
## Self-supervised learning

### Small labeled data

Center



Scratch

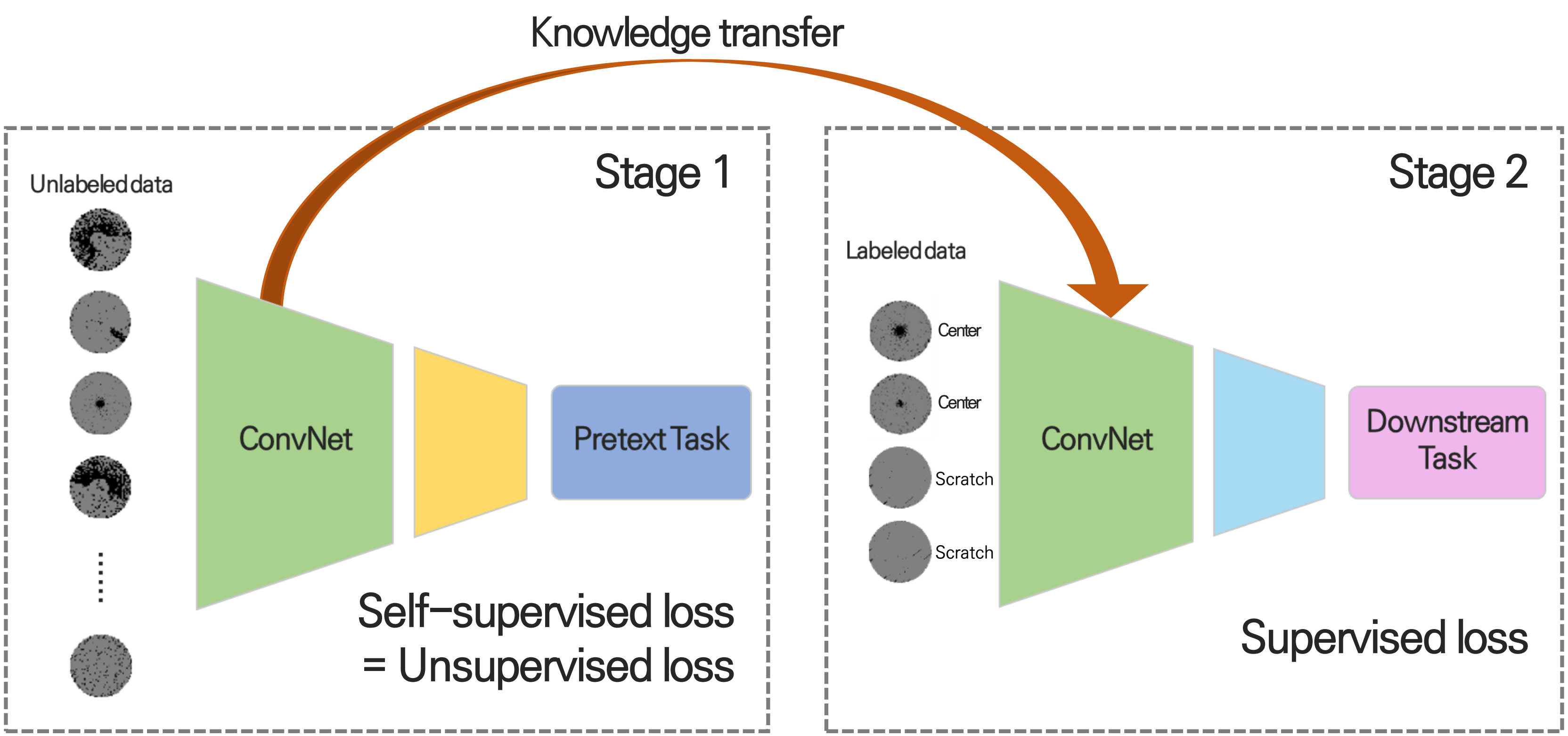


### Large unlabeled data

Self-Supervised Representation Learning for Wafer Bin Map Defect Pattern Classification, IEEE T SEMICONDUCT M 2020 (preprint)

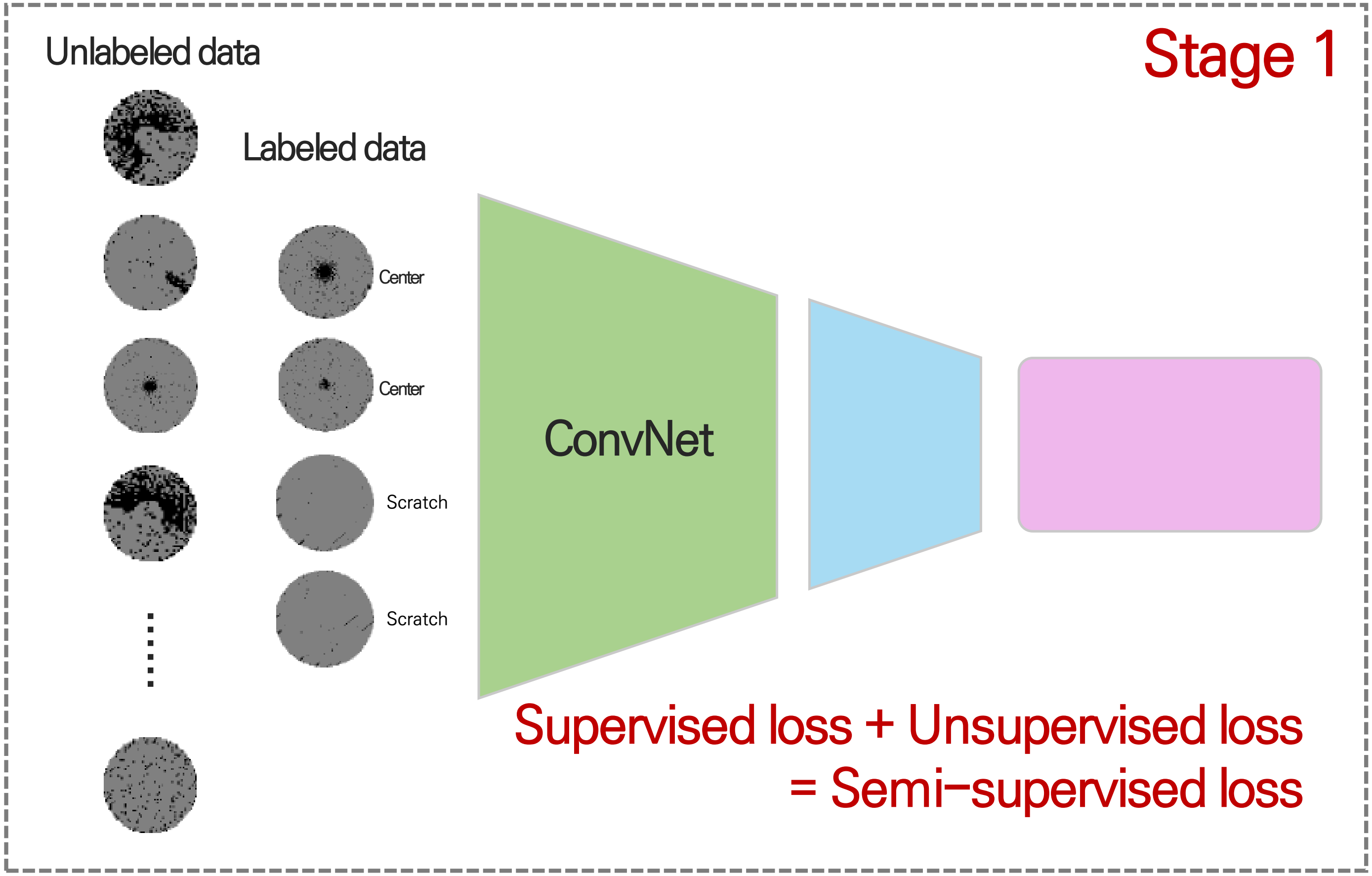
# Semi-supervised learning

## Self-supervised learning



# Semi-supervised learning

## Semi-supervised learning



# MixMatch

❖ MixMatch, NeurIPS 2019

- 2020년 11월 25일 기준 365회 인용

## MixMatch: A Holistic Approach to Semi-Supervised Learning

**David Berthelot**
Google Research
dberth@google.com

**Nicholas Carlini**
Google Research
ncarlini@google.com

**Ian Goodfellow**
Work done at Google
ian-academic@mailfence.com

**Avital Oliver**
Google Research
avitalo@google.com

**Nicolas Papernot**
Google Research
papernot@google.com

**Colin Raffel**
Google Research
craffel@google.com

**Abstract**

Semi-supervised learning has proven to be a powerful paradigm for leveraging unlabeled data to mitigate the reliance on large labeled datasets. In this work, we unify the current dominant approaches for semi-supervised learning to produce a new algorithm, MixMatch, that guesses low-entropy labels for data-augmented unlabeled examples and mixes labeled and unlabeled data using MixUp. MixMatch obtains state-of-the-art results by a large margin across many datasets and labeled data amounts. For example, on CIFAR-10 with 250 labels, we reduce error rate by a factor of 4 (from 38% to 11%) and by a factor of 2 on STL-10. We also demonstrate how MixMatch can help achieve a dramatically better accuracy-privacy trade-off for differential privacy. Finally, we perform an ablation study to tease apart which components of MixMatch are most important for its success. We release all code used in our experiments.[1]

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems* (pp. 5049-5059).

# MixMatch

## MixMatch

### A Holistic Approach to Semi-Supervised Learning

# MixMatch

## A Holistic Approach to Semi-Supervised Learning

| Consistency Regularization | Entropy Minimization | Traditional Regularization (MixUp) |
| --- | --- | --- |

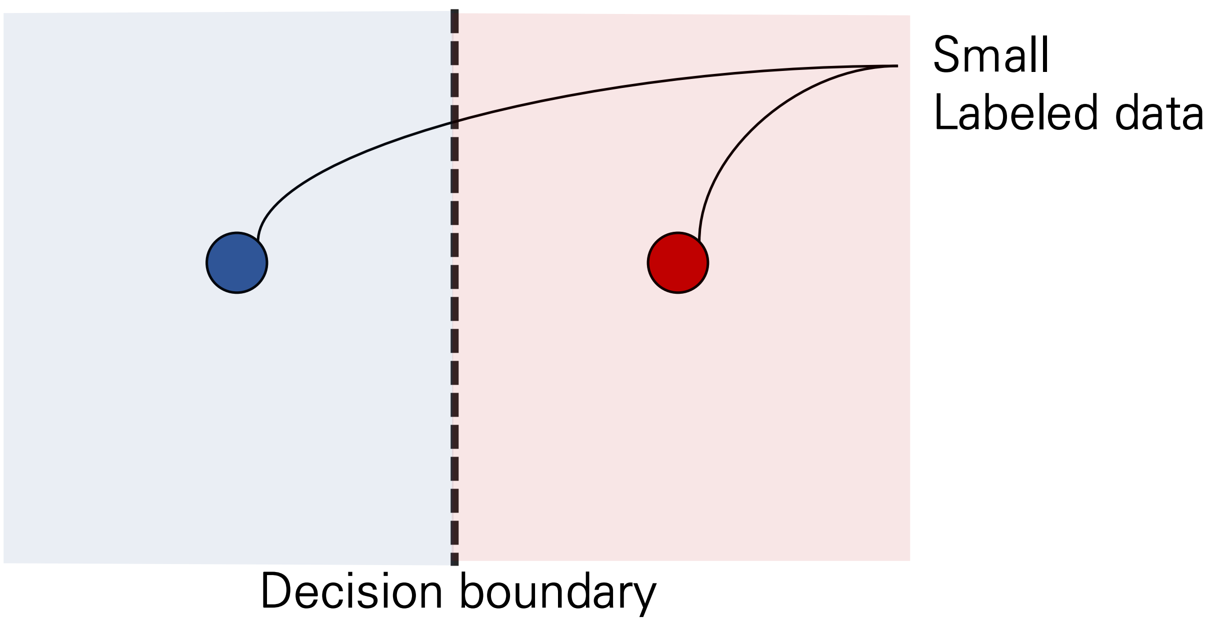
Supervised loss + Unsupervised loss
= Semi-supervised loss

일반화 성능 향상

# MixMatch

## Consistency Regularization

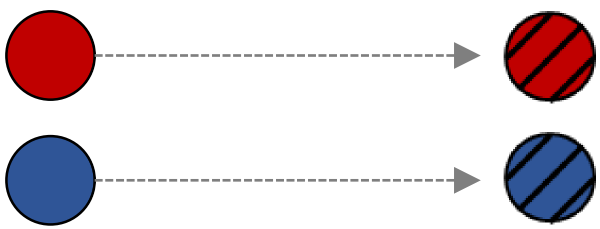
Supervised learning + data augmentation



# MixMatch

## Consistency Regularization

Supervised learning + data augmentation

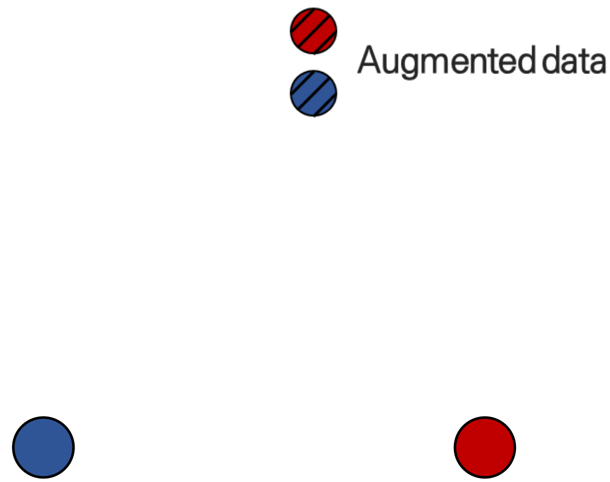


Augmented data

# MixMatch

## Consistency Regularization

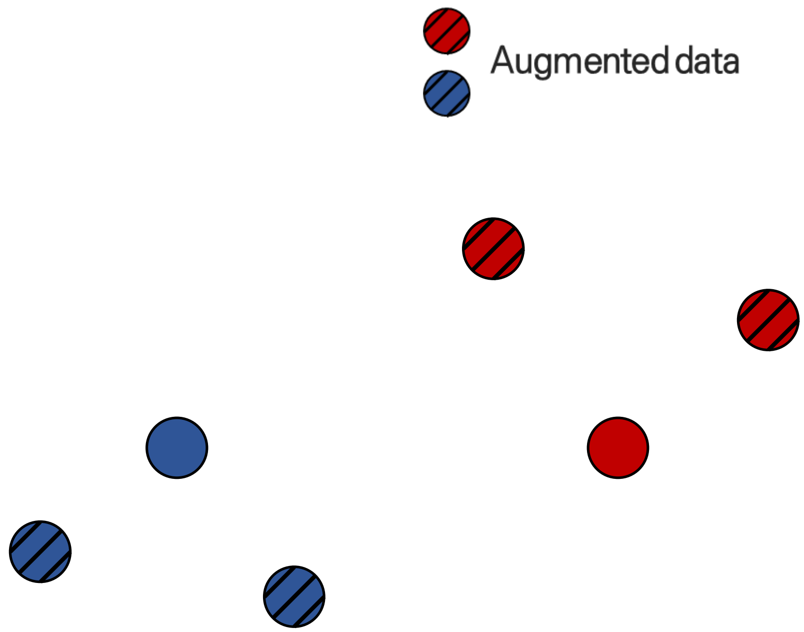
Supervised learning + data augmentation



# MixMatch

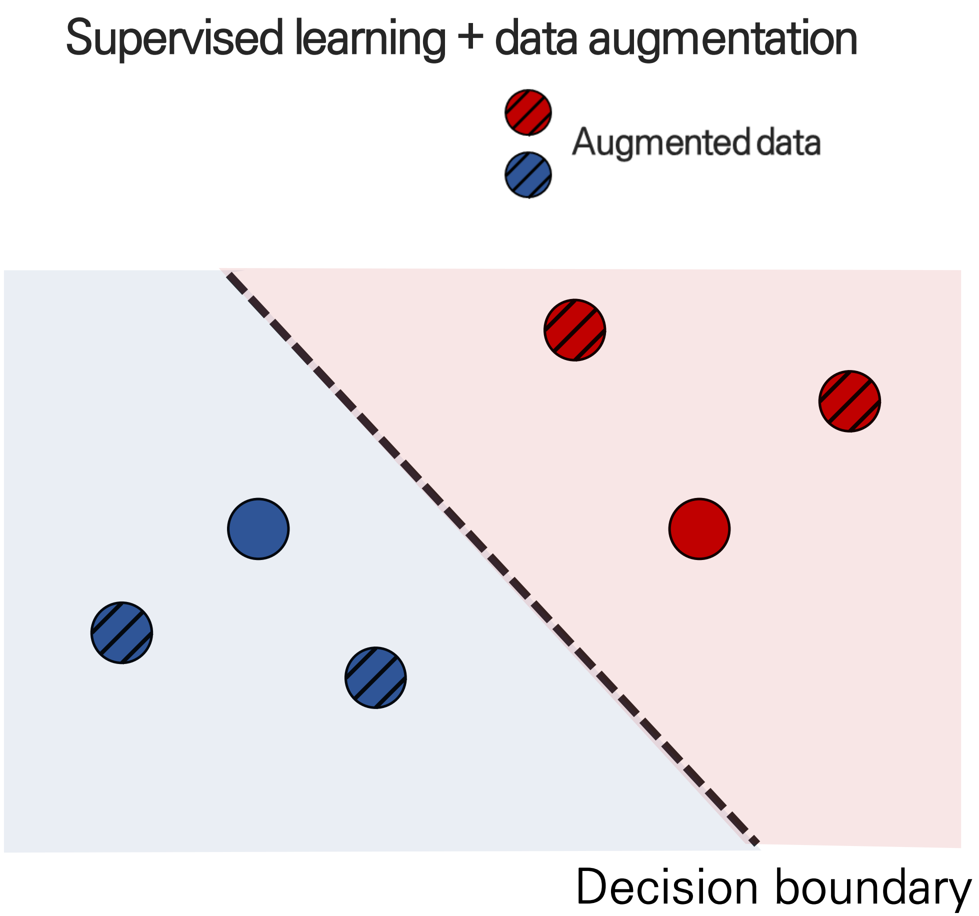
## Consistency Regularization

Supervised learning + data augmentation



# MixMatch

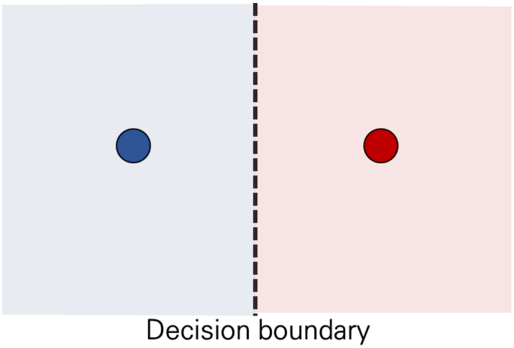
## Consistency Regularization

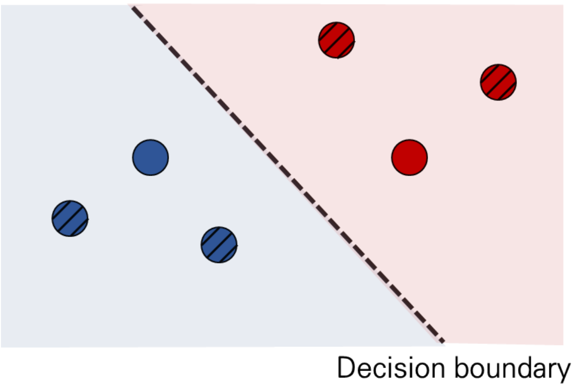


# MixMatch

## Consistency Regularization

Supervised learning + data augmentation

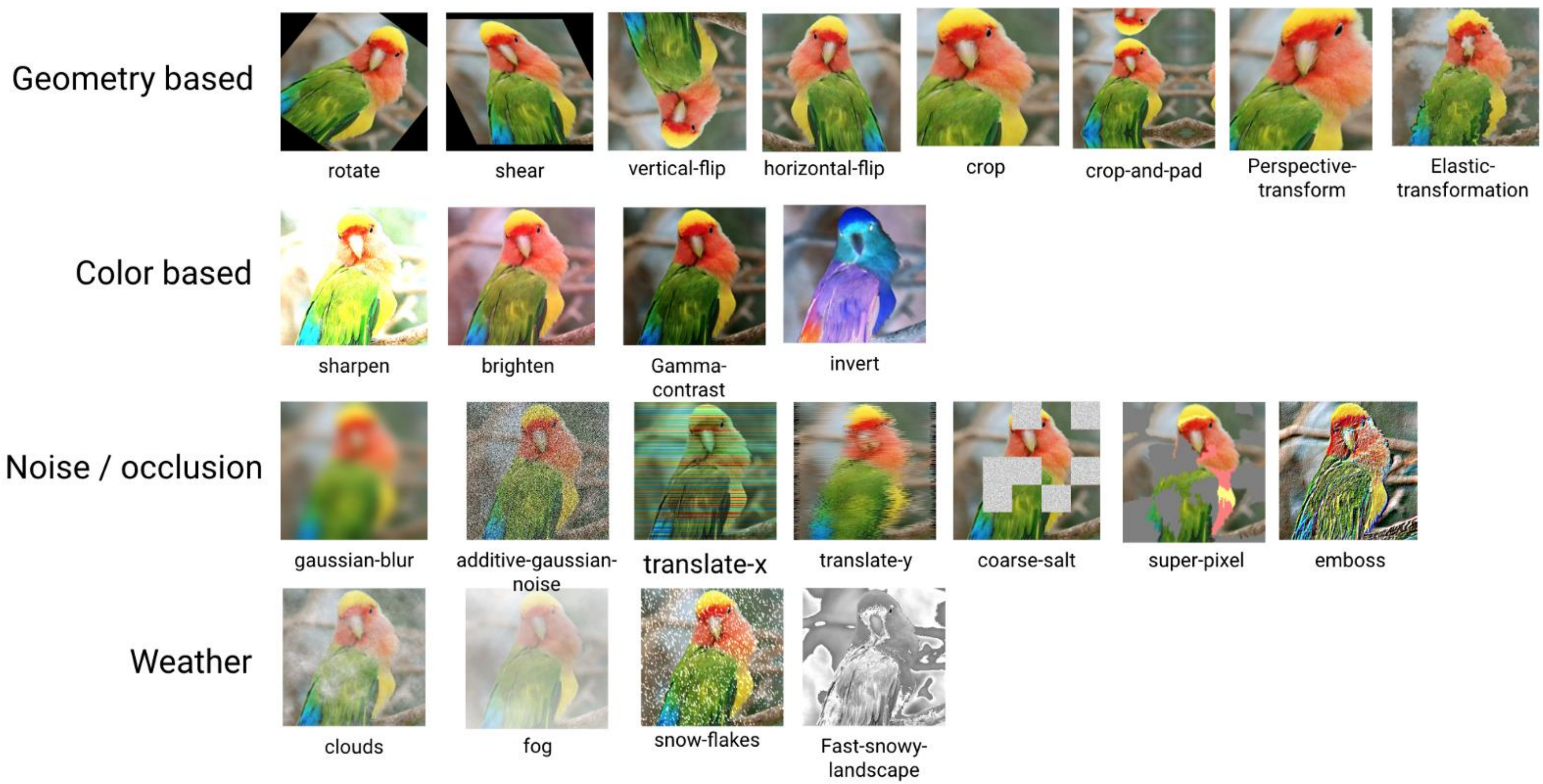



# MixMatch

## Consistency Regularization

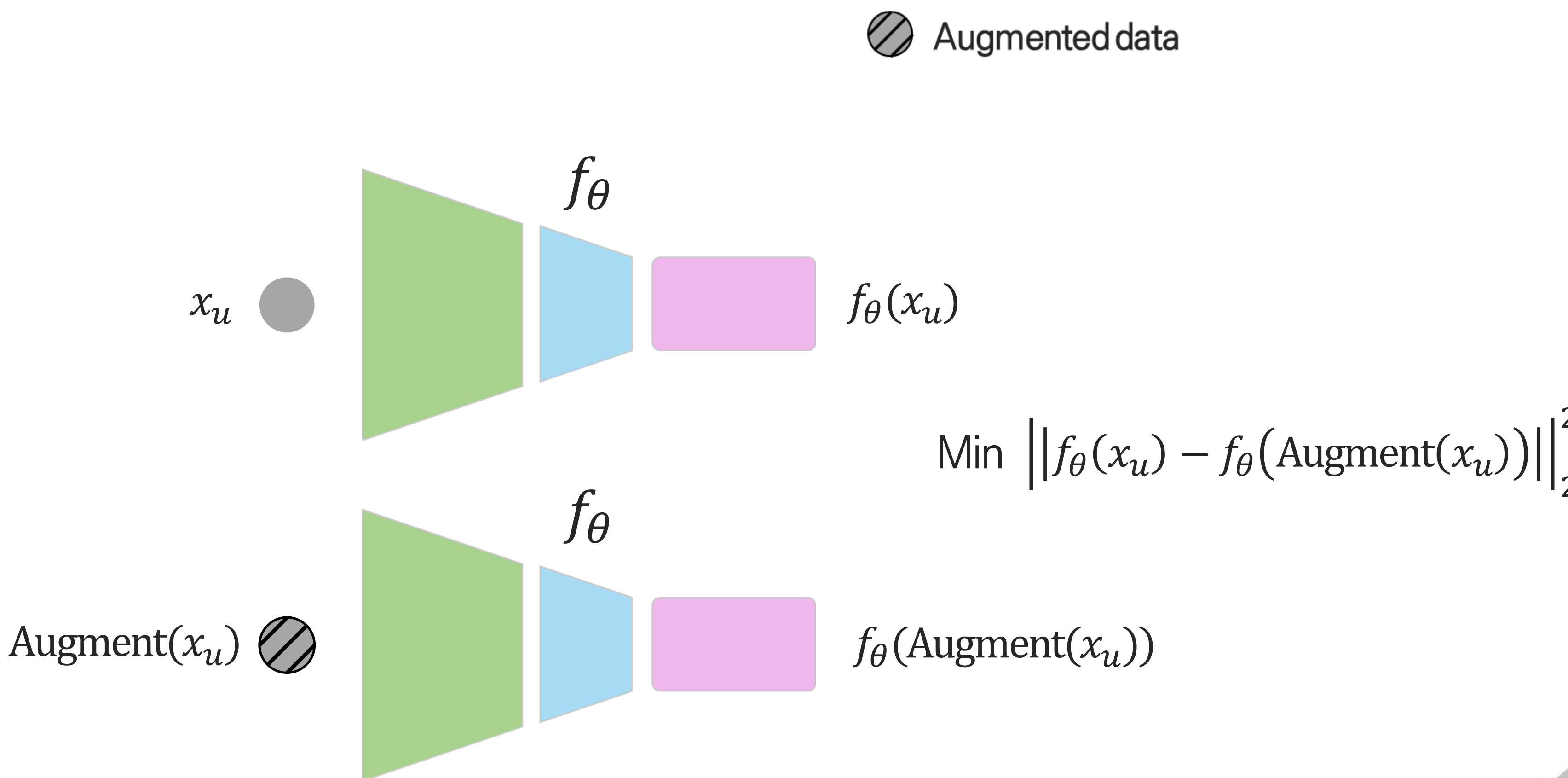
### Supervised learning + data augmentation



# MixMatch

## Consistency Regularization

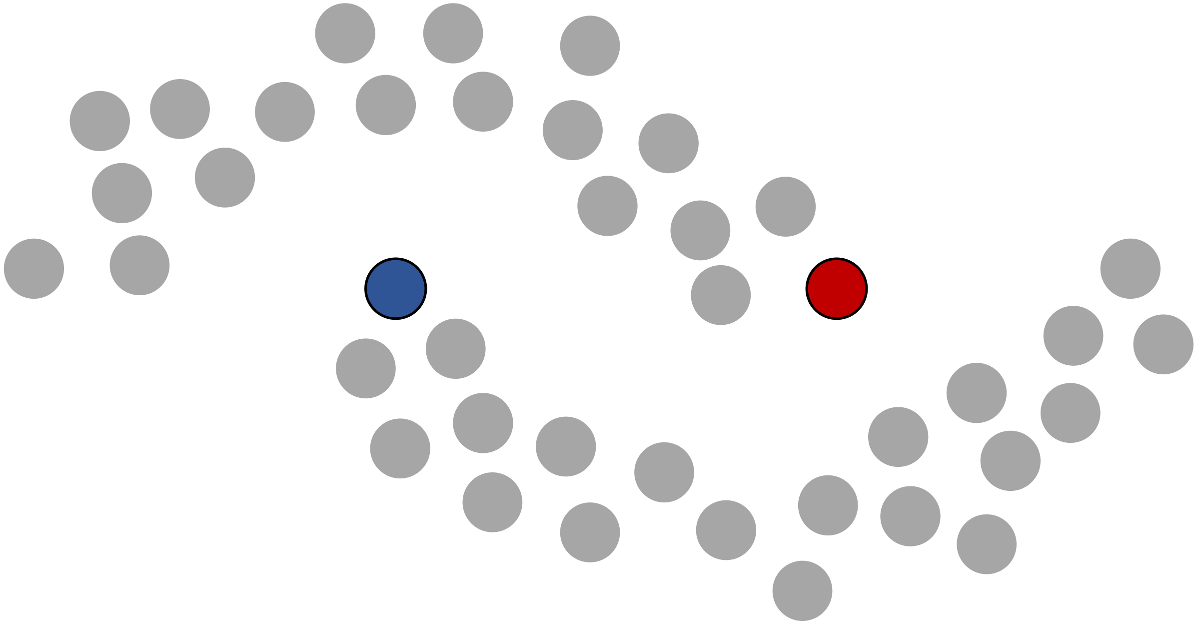
Semi-supervised learning + data augmentation



# MixMatch

## Entropy Minimization

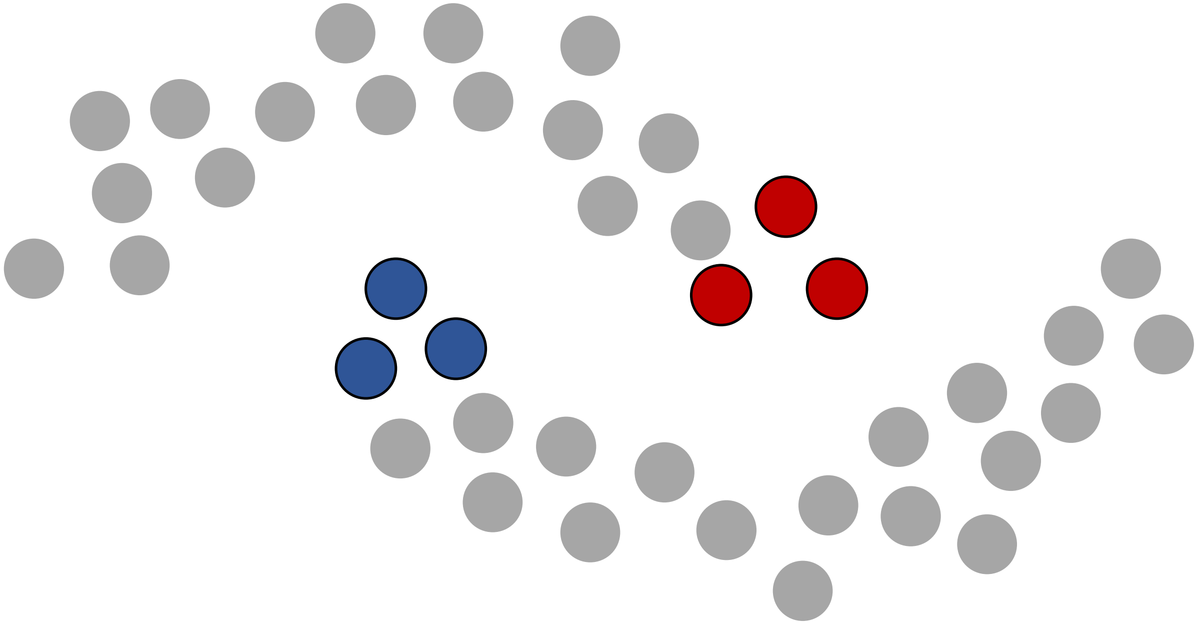
Continuity assumption in semi-supervised learning



# MixMatch

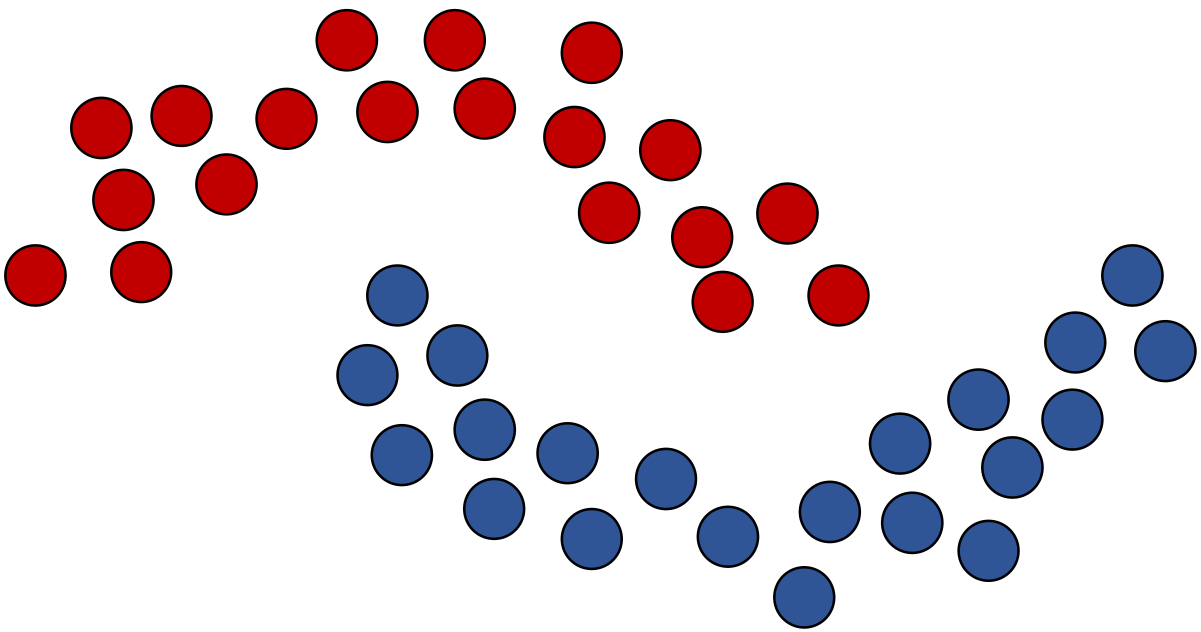
## Entropy Minimization

Continuity assumption in semi-supervised learning



## Entropy Minimization

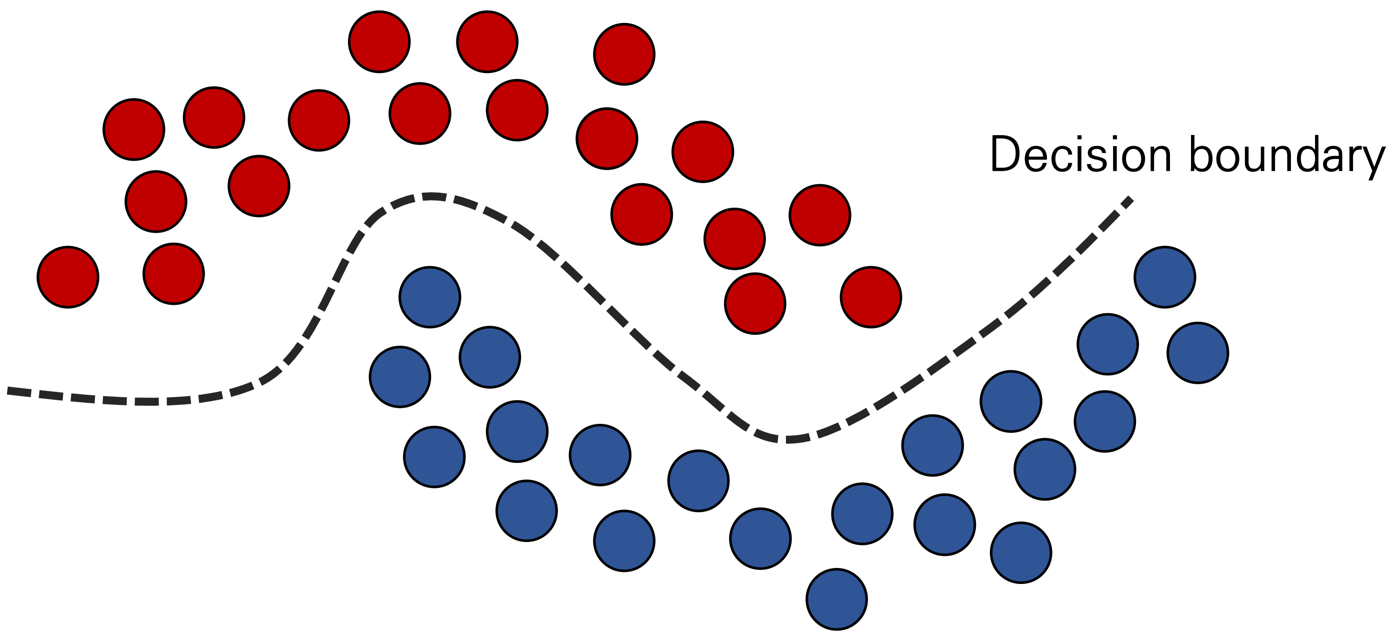
Continuity assumption in semi-supervised learning



# MixMatch

## Entropy Minimization

Continuity assumption in semi-supervised learning

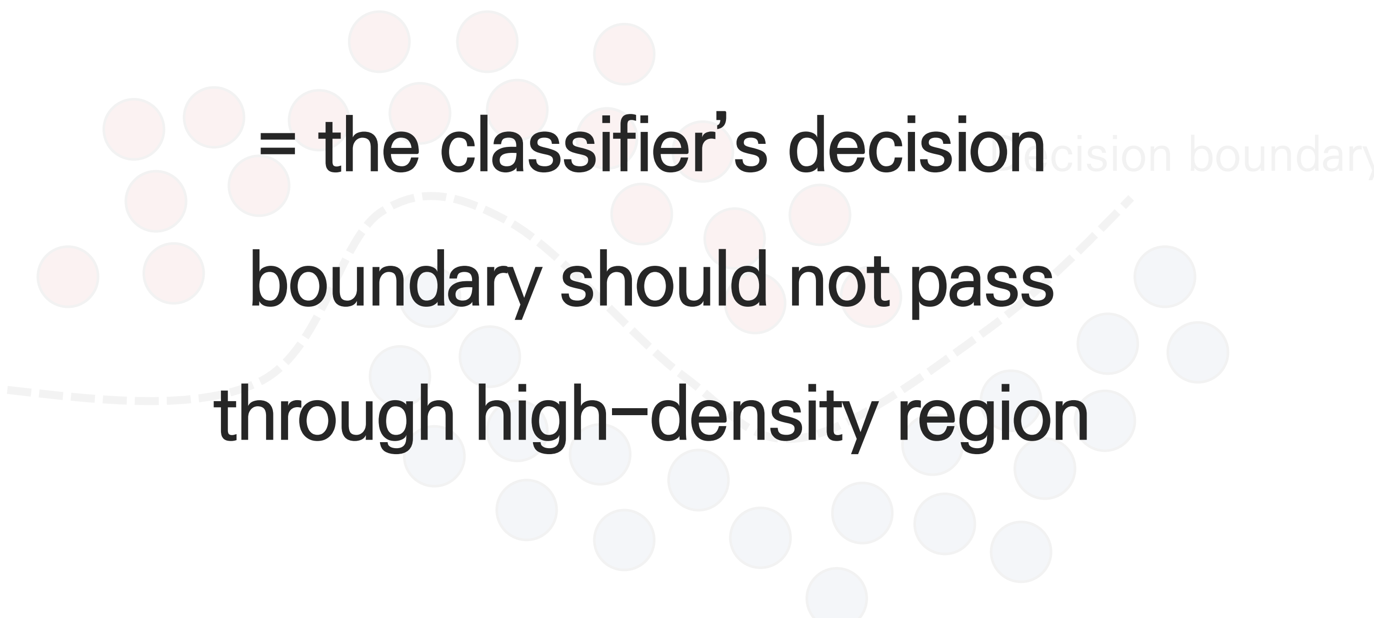


# MixMatch

## Entropy Minimization

Continuity assumption in semi-supervised learning

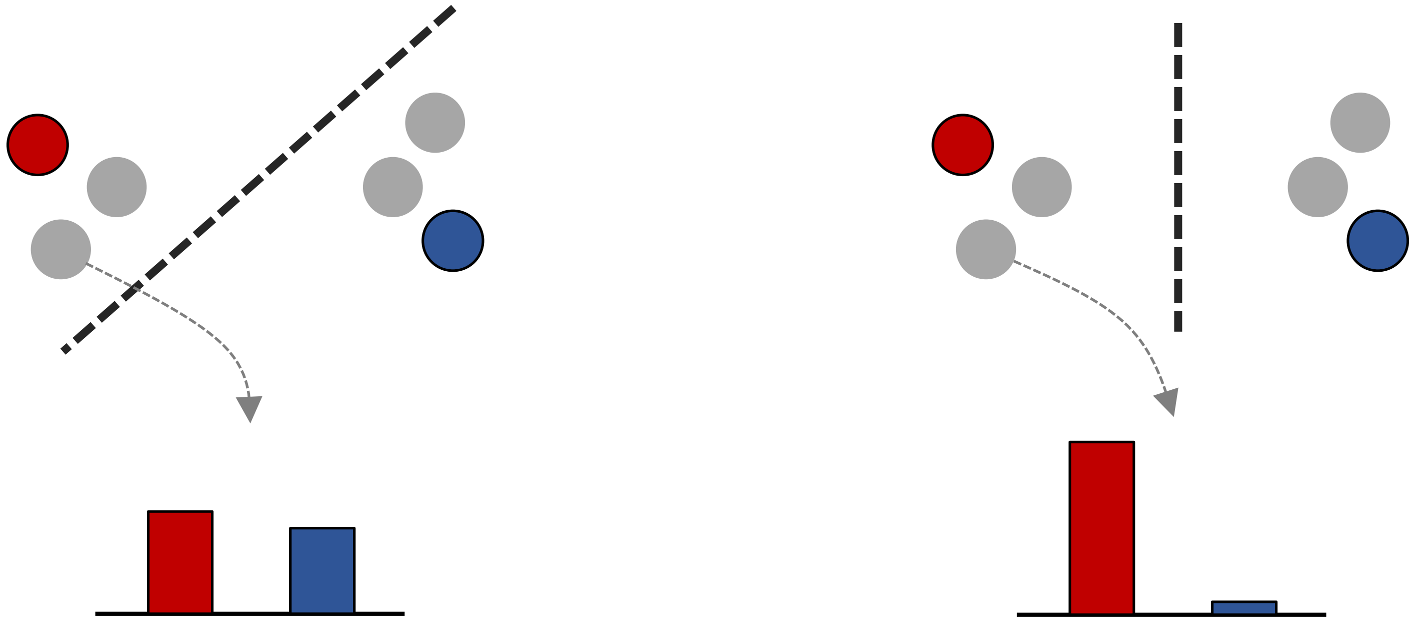
= the classifier's decision boundary should not pass through high-density region



# Entropy Minimization

Continuity assumption in semi-supervised learning

= the classifier's decision boundary should not pass through high-density region

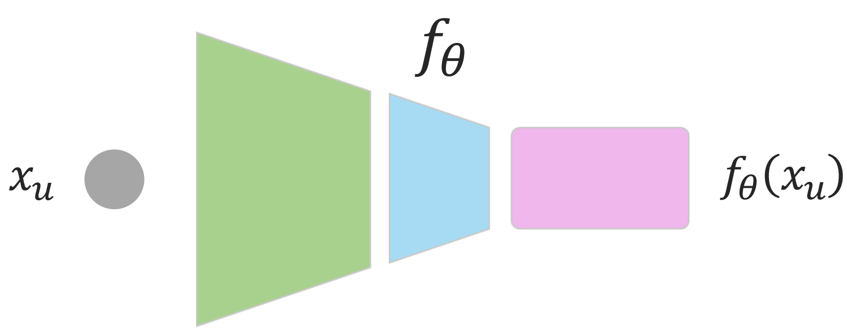


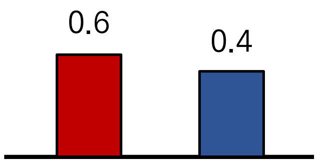
# MixMatch

## Entropy Minimization

Continuity assumption in semi-supervised learning

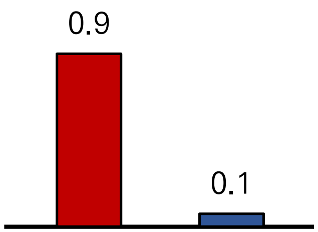
= the classifier's decision boundary should not pass through high-density region




$$-0.6log0.6 - 0.4log0.4 = 0.8563$$

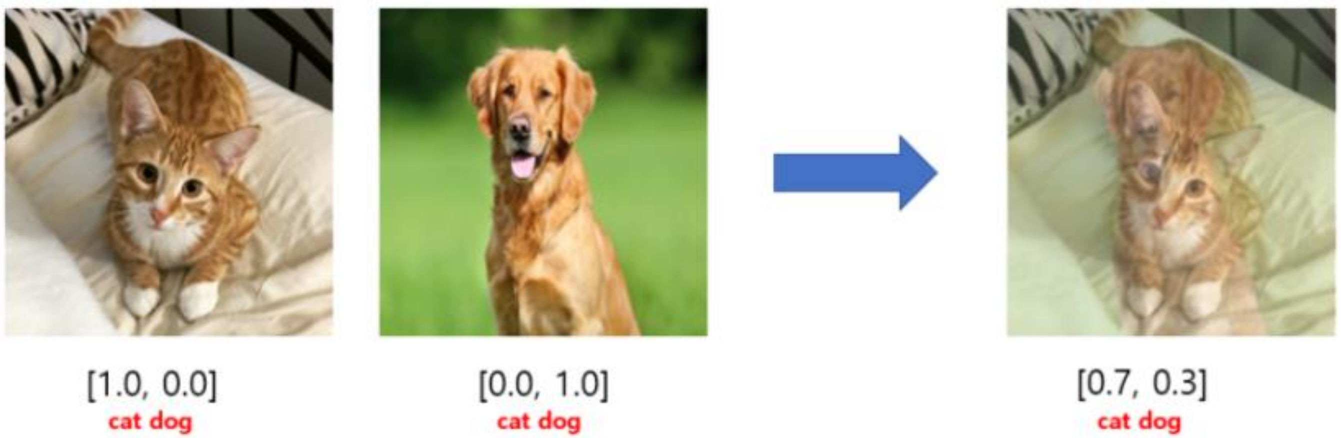
$$H\left(f_\theta(x_u)\right) = -\sum_{i=1}^{\#\ of\ class} P_i log P_i$$




$$-0.9log0.9 - 0.1log0.1 = 0.3251$$

## Traditional Regularization (MixUp)

### MixUp in supervised learning



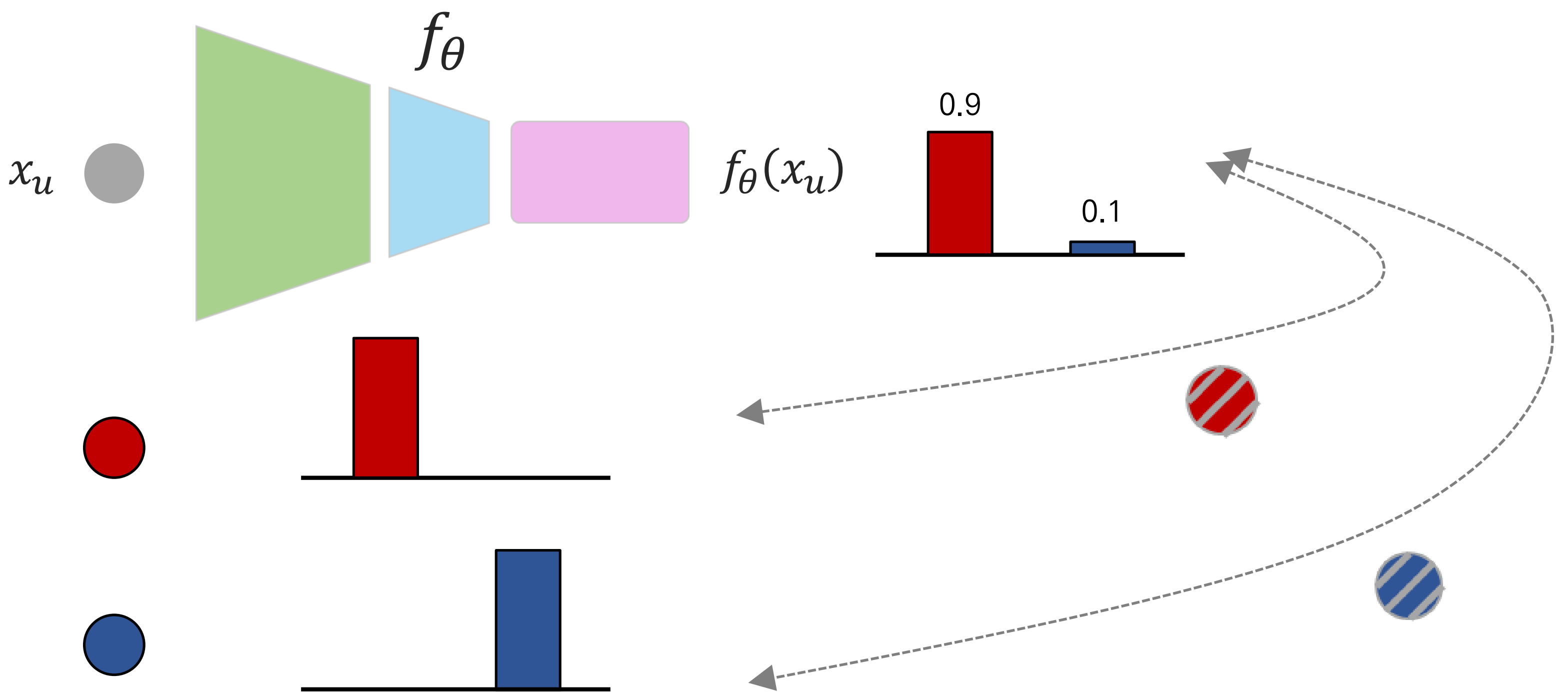

[figure_08] Mixup_pytorch_image

https://blog.airlab.re.kr/2019/11/mixup

# MixMatch

## Traditional Regularization (MixUp)

MixUp in semi-supervised learning

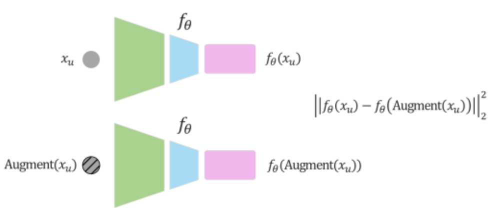


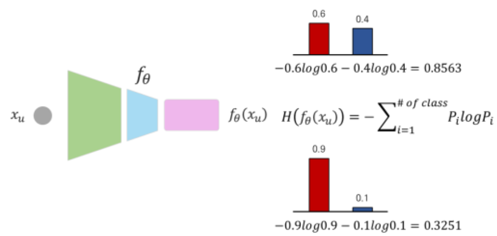
# MixMatch

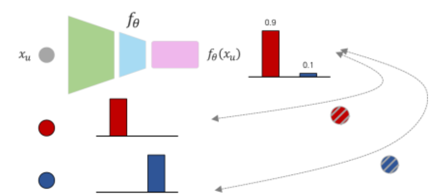
## MixMatch

### A Holistic Approach to Semi-Supervised Learning

| Consistency Regularization | Entropy Minimization | Traditional Regularization (MixUp) |
| --- | --- | --- |




일반화 성능 향상

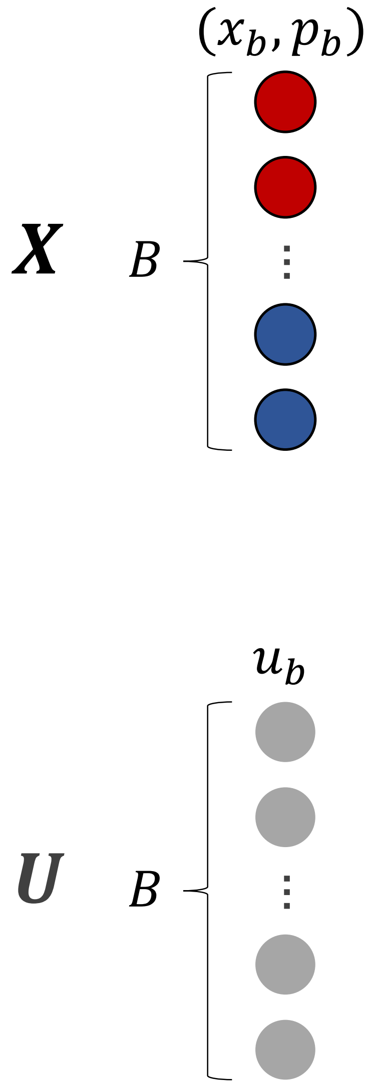
# MixMatch

## A Holistic Approach to Semi-Supervised Learning

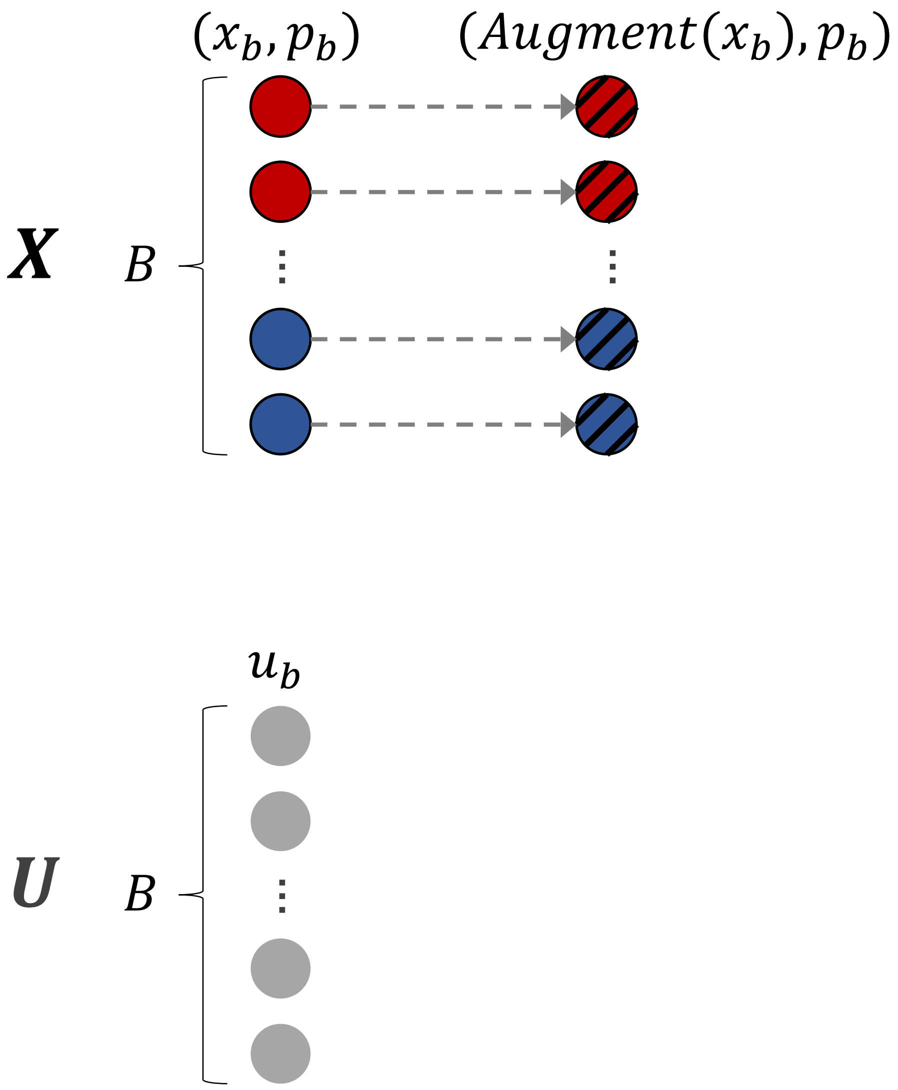
**Algorithm 1** MixMatch takes a batch of labeled data $\mathcal{X}$ and a batch of unlabeled data $\mathcal{U}$ and produces a collection $\mathcal{X}'$ (resp. $\mathcal{U}'$) of processed labeled examples (resp. unlabeled with guessed labels).

**Input:** Batch of labeled examples and their one-hot labels $\mathcal{X} = \big((x_b, p_b); b \in (1, \ldots, B)\big)$, batch of unlabeled examples $\mathcal{U} = \big(u_b; b \in (1, \ldots, B)\big)$, sharpening temperature $T$, number of augmentations $K$, Beta distribution parameter $\alpha$ for MixUp.
**for** $b = 1$ **to** $B$ **do**
  $\hat{x}_b = \text{Augment}(x_b)$ *// Apply data augmentation to $x_b$*
  **for** $k = 1$ **to** $K$ **do**
    $\hat{u}_{b,k} = \text{Augment}(u_b)$ *// Apply $k^{th}$ round of data augmentation to $u_b$*
  **end for**
  $\bar{q}_b = \frac{1}{K} \sum_k \text{p}_{\text{model}}(y \mid \hat{u}_{b,k}; \theta)$ *// Compute average predictions across all augmentations of $u_b$*
  $q_b = \text{Sharpen}(\bar{q}_b, T)$ *// Apply temperature sharpening to the average prediction (see eq. (7))*
**end for**
$\hat{\mathcal{X}} = \big((\hat{x}_b, p_b); b \in (1, \ldots, B)\big)$ *// Augmented labeled examples and their labels*
$\hat{\mathcal{U}} = \big((\hat{u}_{b,k}, q_b); b \in (1, \ldots, B), k \in (1, \ldots, K)\big)$ *// Augmented unlabeled examples, guessed labels*
$\mathcal{W} = \text{Shuffle}\big(\text{Concat}(\hat{\mathcal{X}}, \hat{\mathcal{U}})\big)$ *// Combine and shuffle labeled and unlabeled data*
$\mathcal{X}' = \big(\text{MixUp}(\hat{\mathcal{X}}_i, \mathcal{W}_i); i \in (1, \ldots, |\hat{\mathcal{X}}|)\big)$ *// Apply MixUp to labeled data and entries from $\mathcal{W}$*
$\mathcal{U}' = \big(\text{MixUp}(\hat{\mathcal{U}}_i, \mathcal{W}_{i+|\hat{\mathcal{X}}|}); i \in (1, \ldots, |\hat{\mathcal{U}}|)\big)$ *// Apply MixUp to unlabeled data and the rest of $\mathcal{W}$*
**return** $\mathcal{X}', \mathcal{U}'$

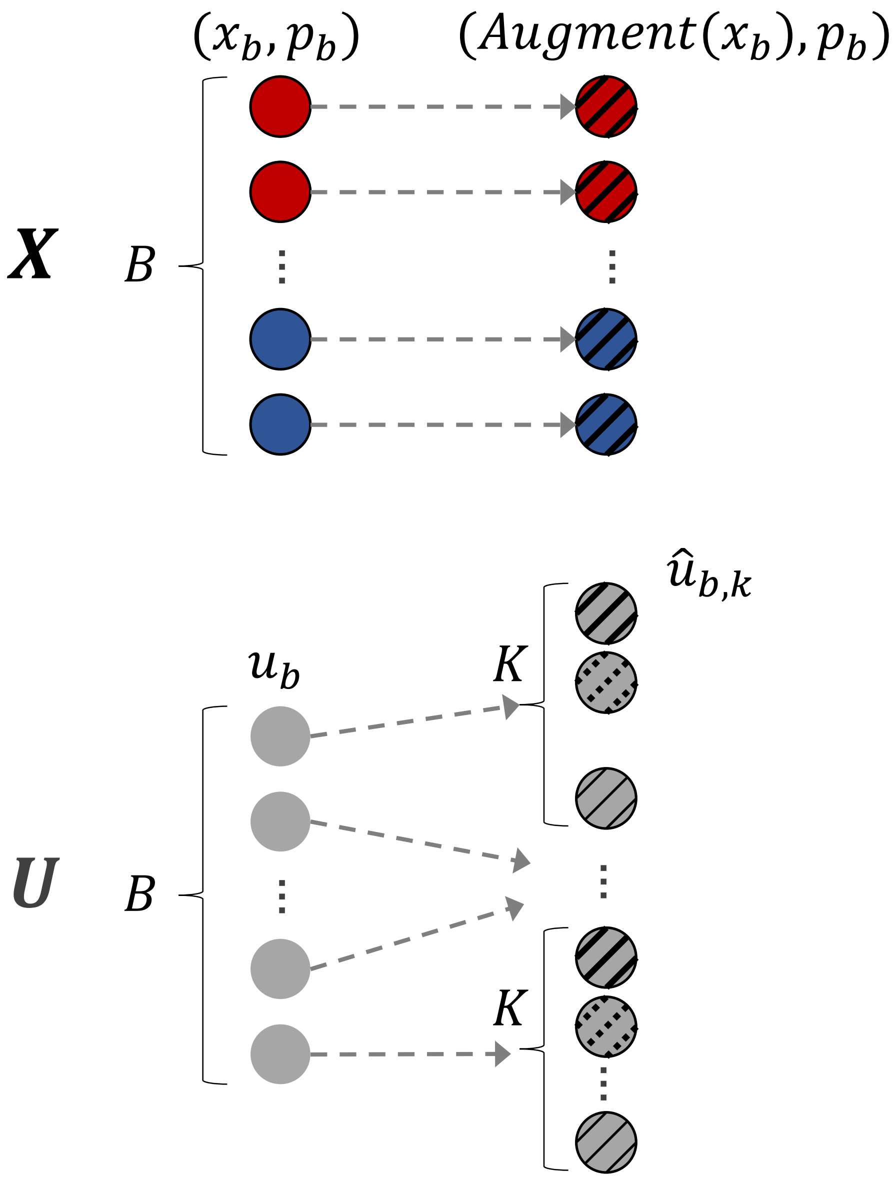
# MixMatch



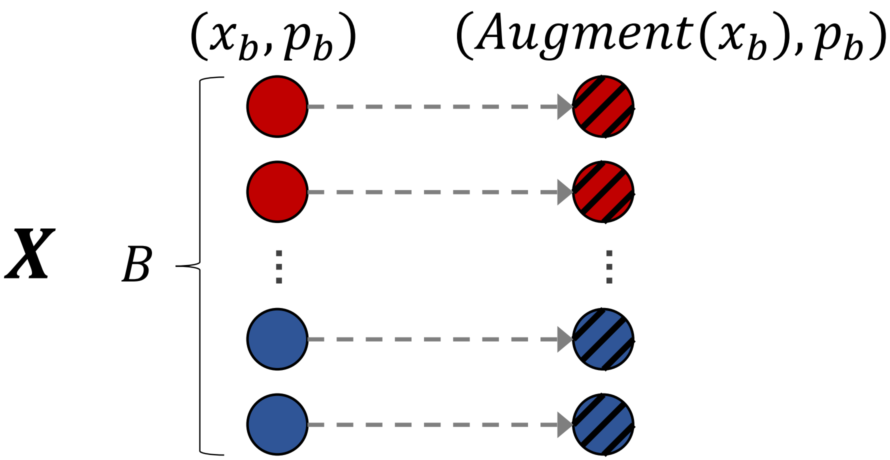
# MixMatch



# MixMatch

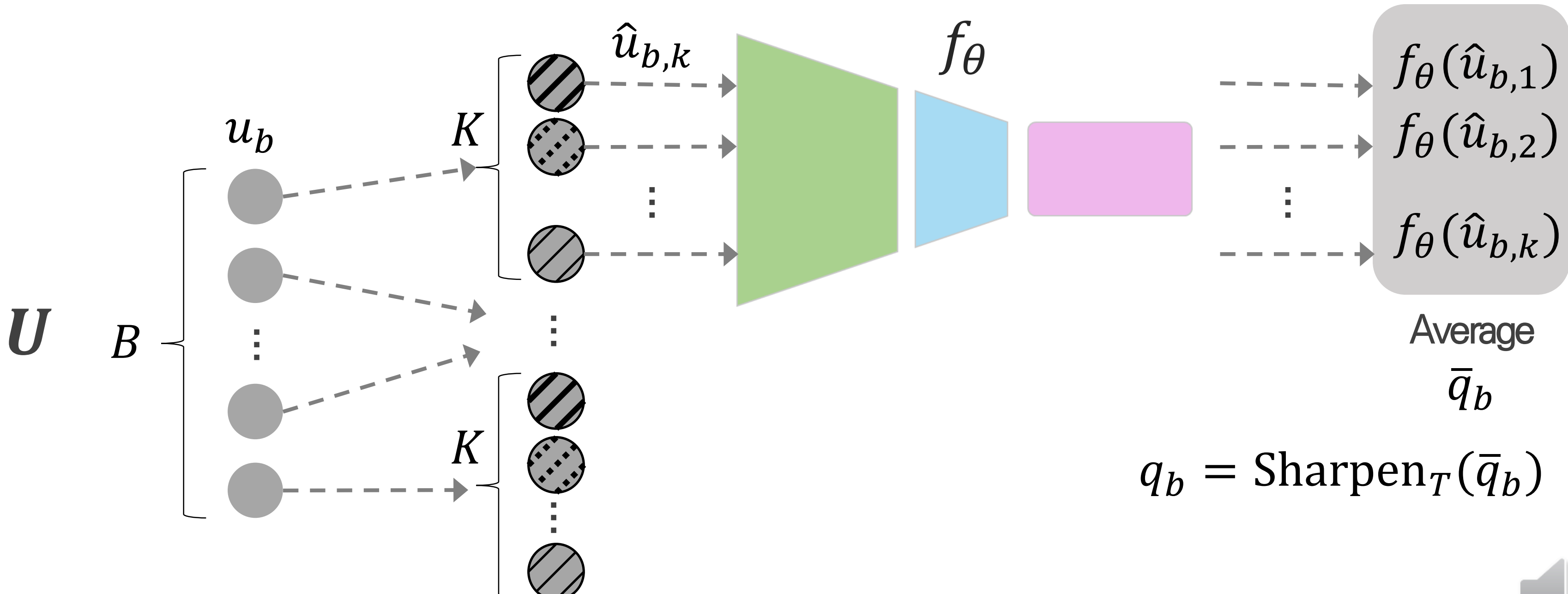


# MixMatch

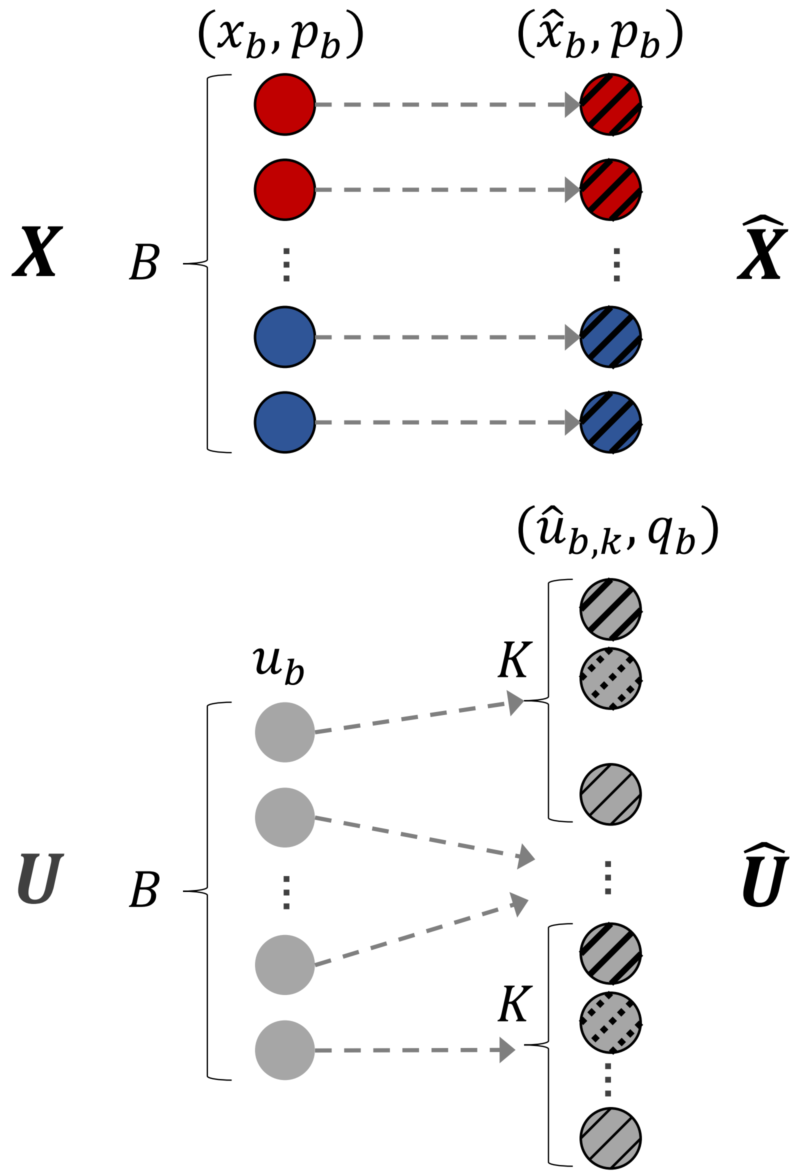



$$Sharpen(p, T)_i = \frac{p_i^{\frac{1}{T}}}{\sum_{j=1}^{\#\ of\ class} p_j^{\frac{1}{T}}},$$

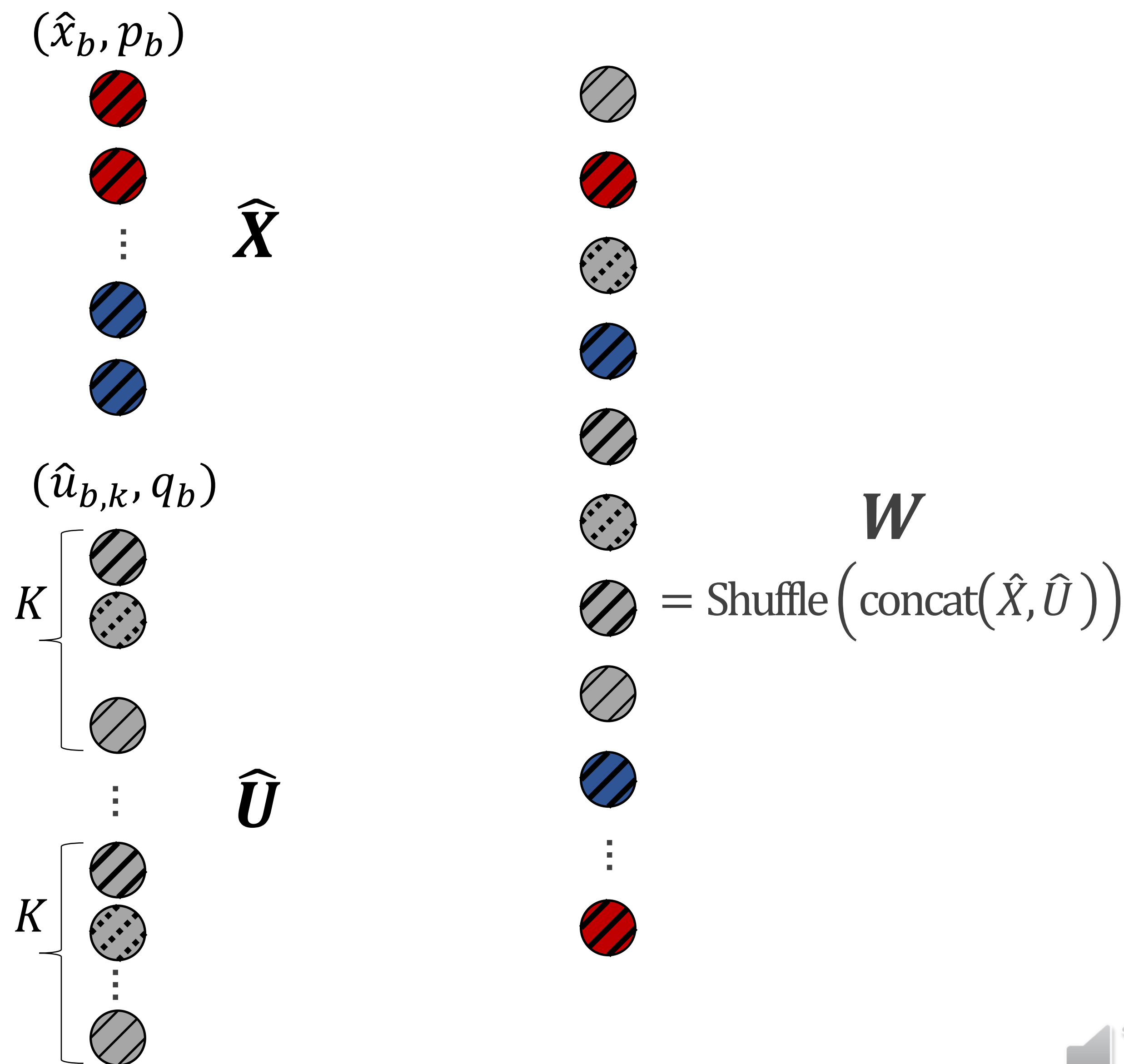
$$T \to 0, Sharpen_T \to \text{one hot}$$



# MixMatch


$(x_b, p_b)$
$(\hat{x}_b, p_b)$
$\boldsymbol{X}$
$B$
$\hat{\boldsymbol{X}}$
$(\hat{u}_{b,k}, q_b)$
$u_b$
$K$
$\boldsymbol{U}$
$B$
$\hat{\boldsymbol{U}}$
$K$

# MixMatch

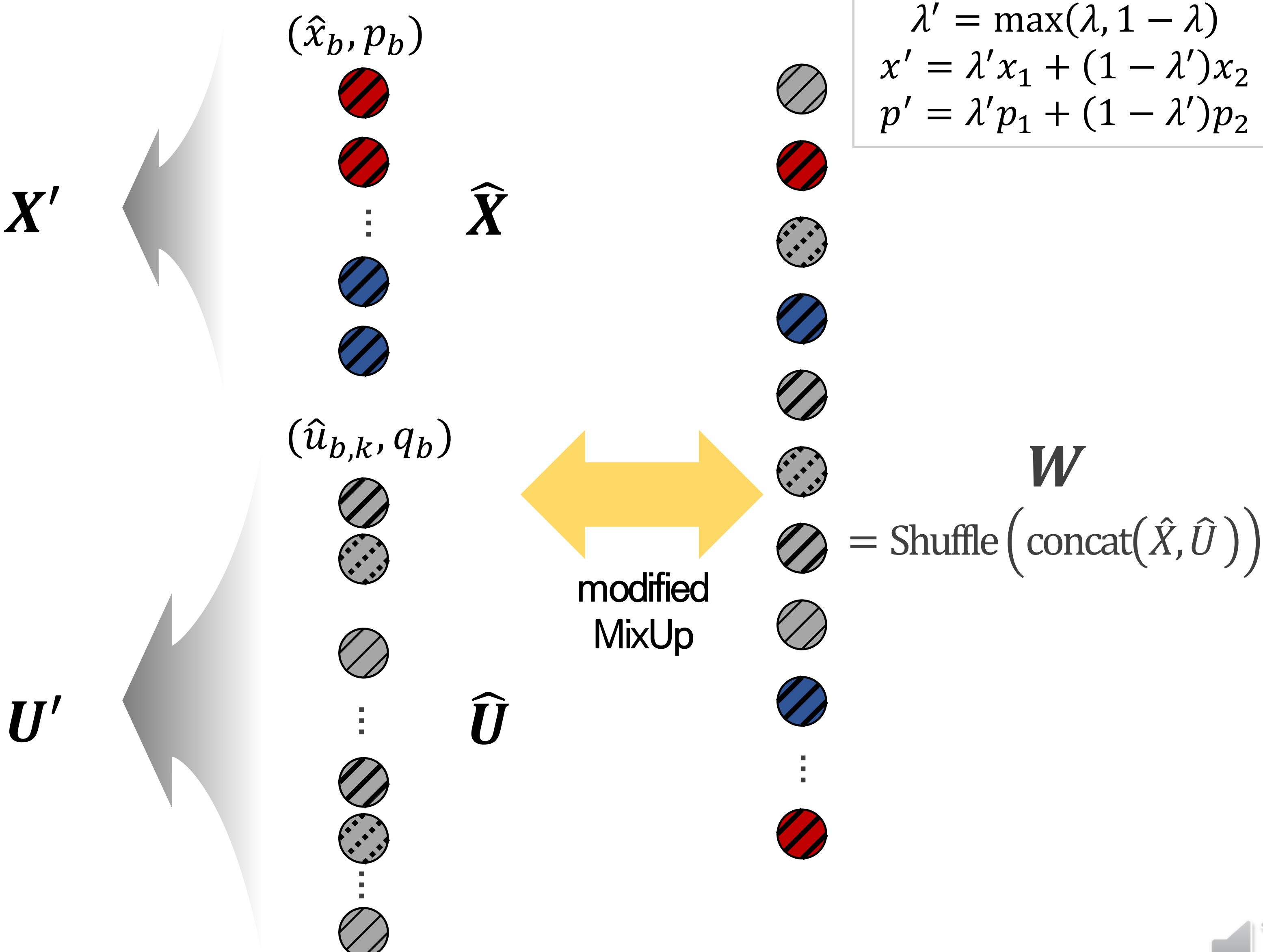


$(\hat{x}_b, p_b)$
$\hat{X}$
$(\hat{u}_{b,k}, q_b)$
$K$
$\hat{U}$
$K$
$W$
$= \text{Shuffle}\left(\text{concat}(\hat{X}, \hat{U})\right)$

# MixMatch

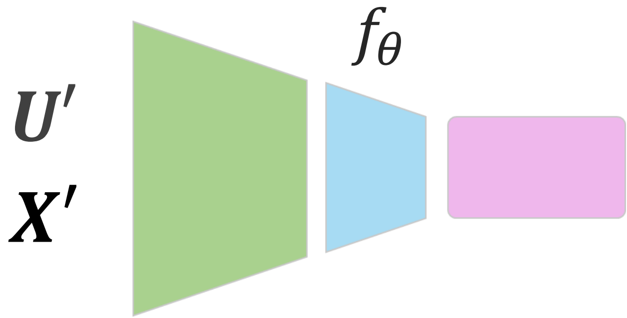
**Modified MixUp**

$$\lambda \sim Beta(\alpha, \alpha)$$
$$\lambda' = \max(\lambda, 1-\lambda)$$
$$x' = \lambda' x_1 + (1-\lambda') x_2$$
$$p' = \lambda' p_1 + (1-\lambda') p_2$$



# MixMatch

A Holistic Approach to Semi-Supervised Learning




Supervised loss($\boldsymbol{L_x}$) + $\lambda_u$·unsupervised loss ($\boldsymbol{L_u}$)
= cross entropy + $\lambda_u$ · consistency regularization

# MixMatch

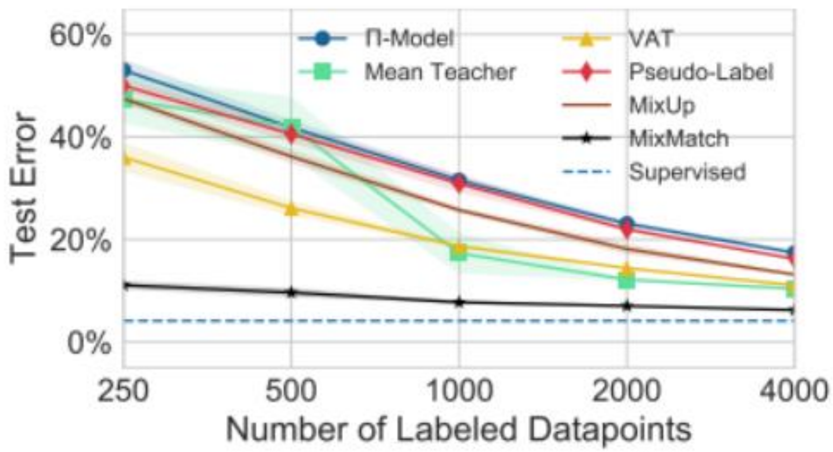



Figure 2: Error rate comparison of MixMatch to baseline methods on CIFAR-10 for a varying number of labels. Exact numbers are provided in table 5 (appendix). "Supervised" refers to training with all 50000 training examples and no unlabeled data. With 250 labels MixMatch reaches an error rate comparable to next-best method's performance with 4000 labels.

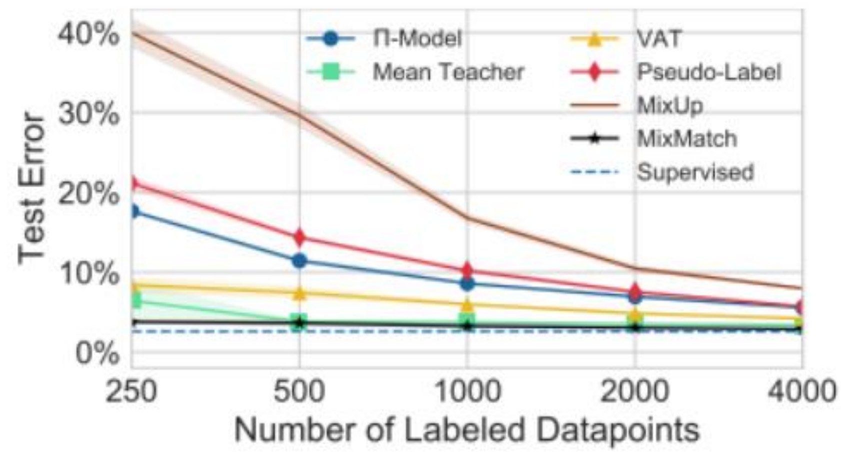



Figure 3: Error rate comparison of MixMatch to baseline methods on SVHN for a varying number of labels. Exact numbers are provided in table 6 (appendix). "Supervised" refers to training with all 73257 training examples and no unlabeled data. With 250 examples MixMatch nearly reaches the accuracy of supervised training for this model.

# MixMatch

| Ablation | 250 labels | 4000 labels |
|---|---|---|
| MixMatch | 11.80 | 6.00 |
| MixMatch without distribution averaging ($K = 1$) | 17.09 | 8.06 |
| MixMatch with $K = 3$ | 11.55 | 6.23 |
| MixMatch with $K = 4$ | 12.45 | 5.88 |
| MixMatch without temperature sharpening ($T = 1$) | 27.83 | 10.59 |
| MixMatch with parameter EMA | 11.86 | 6.47 |
| MixMatch without MixUp | 39.11 | 10.97 |
| MixMatch with MixUp on labeled only | 32.16 | 9.22 |
| MixMatch with MixUp on unlabeled only | 12.35 | 6.83 |
| MixMatch with MixUp on separate labeled and unlabeled | 12.26 | 6.50 |
| Interpolation Consistency Training [45] | 38.60 | 6.81 |

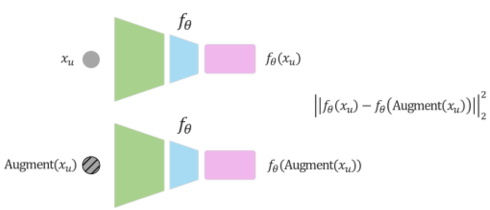
Table 4: Ablation study results. All values are error rates on CIFAR-10 with 250 or 4000 labels.

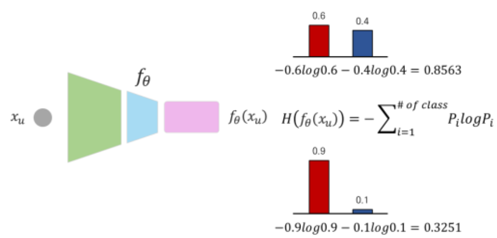
# MixMatch

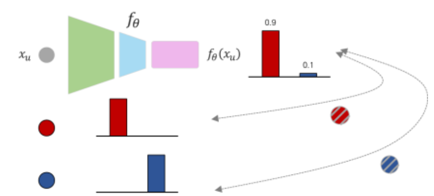
## MixMatch

### A Holistic Approach to Semi-Supervised Learning

| Consistency Regularization | Entropy Minimization | Traditional Regularization (MixUp) |
| --- | --- | --- |




**일반화 성능 향상**

# Conclusions

❖ Realistic evaluation of deep semi-supervised learning algorithms, NeurIPS 2018

- 2020년 11월 25일 기준 322회 인용

## Realistic Evaluation of Deep Semi-Supervised Learning Algorithms

**Avital Oliver*, Augustus Odena*, Colin Raffel*, Ekin D. Cubuk & Ian J. Goodfellow**
Google Brain
{avitalo,augustusodena,craffel,cubuk,goodfellow}@google.com

**Abstract**

Semi-supervised learning (SSL) provides a powerful framework for leveraging unlabeled data when labels are limited or expensive to obtain. SSL algorithms based on deep neural networks have recently proven successful on standard benchmark tasks. However, we argue that these benchmarks fail to address many issues that SSL algorithms would face in real-world applications. After creating a unified reimplementation of various widely-used SSL techniques, we test them in a suite of experiments designed to address these issues. We find that the performance of simple baselines which do not use unlabeled data is often underreported, SSL methods differ in sensitivity to the amount of labeled and unlabeled data, and performance can degrade substantially when the unlabeled dataset contains out-of-distribution examples. To help guide SSL research towards real-world applicability, we make our unified reimplemention and evaluation platform publicly available.[2]

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., & Goodfellow, I. (2018). Realistic evaluation of deep semi-supervised learning algorithms. In Advances in neural information processing systems (pp. 3235-3246).

# Conclusions

❖ Hyperparameter

• ReMixMatch[1], FixMatch[2]

❖ Transfer learning

❖ Unlabeled data contains a different distribution

❖ Regression Task

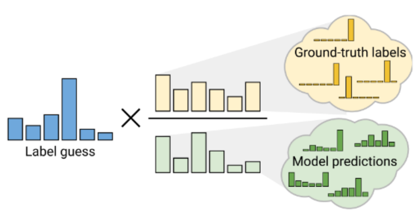



Figure 1: Distribution alignment. Guessed label distributions are adjusted according to the ratio of the empirical ground-truth class distribution divided by the average model predictions on unlabeled data.

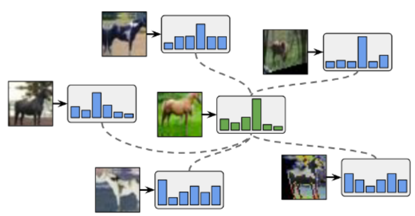


Figure 2: Augmentation anchoring. We use the prediction for a weakly augmented image (green, middle) as the target for predictions on strong augmentations of the same image (blue).

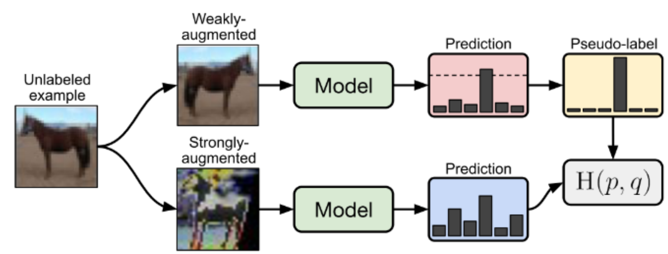



Figure 1: Diagram of FixMatch. A weakly-augmented image (top) is fed into the model to obtain predictions (red box). When the model assigns a probability to any class which is above a threshold (dotted line), the prediction is converted to a one-hot pseudo-label. Then, we compute the model's prediction for a strong augmentation of the same image (bottom). The model is trained to make its prediction on the strongly-augmented version match the pseudo-label via a cross-entropy loss.


[1] Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., & Raffel, C. (2019, September). Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In International Conference on Learning Representations.

[2] Sohn, K., Berthelot, D., Li, C. L., Zhang, Z., Carlini, N., Cubuk, E. D., ... & Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In Advances in neural information processing systems.

# 감사합니다

2020. 12. 4.

Data Mining & Quality Analytics Lab.
이민정
Minjung Lee
leemj2520@korea.ac.kr