The Second Korea-Japan Machine Learning Workshop, Jeju, Korea Feb. 23, 2019

Weakly Supervised Classification, Robust Learning and More:

Overview of Our Recent Advances

Masashi Sugiyama

Imperfect Information Learning Team

RIKEN Center for Advanced Intelligence Project

Machine Learning and Statistical Data Analysis Lab

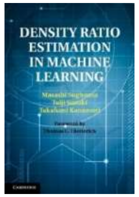
The University of Tokyo

About Myself

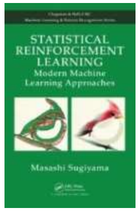
- **Affiliations:**
  - Director: RIKEN AIP
  - Professor: University of Tokyo
  - Consultant: several local startups
- **Research interests:**
  - Theory and algorithms of ML
  - Real-world applications with partners
- **Goal:**
  - Develop practically useful algorithms that have theoretical support

Sugiyama & Kawanabe, Machine Learning in Non-Stationary Environments, MIT Press, 2012

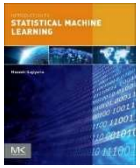
Sugiyama, Suzuki & Kanamori, Density Ratio Estimation in Machine Learning, Cambridge University Press, 2012



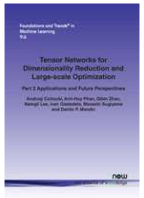
Sugiyama, Statistical Reinforcement Learning, Chapman and Hall/CRC, 2015



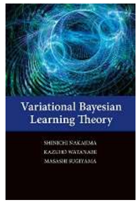
Sugiyama, Introduction to Statistical Machine Learning, Morgan Kaufmann, 2015



Cichocki, Phan, Zhao, Lee, Oseledets, Sugiyama & Mandic, Tensor Networks for Dimensionality Reduction and Large-Scale Optimizations, Now, 2017



Nakajima, Watanabe & Sugiyama, Variational Bayesian Learning Theory, Cambridge University Press, 2019



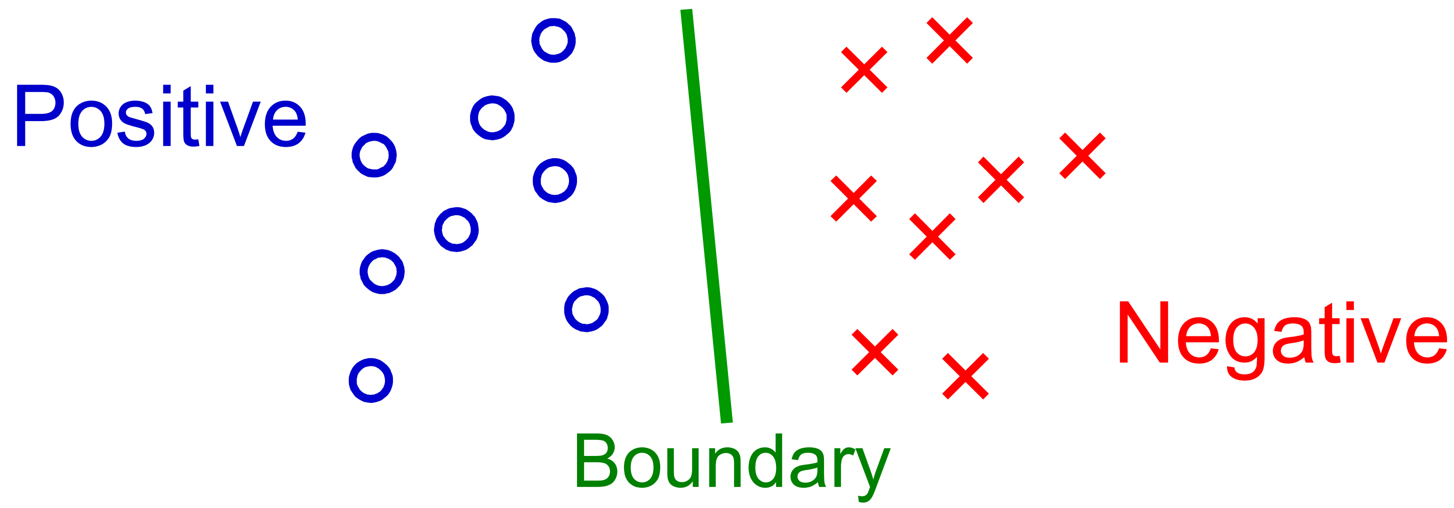
My Talk

1. Weakly supervised classification
2. Robust learning
3. More

What Is This Tutorial about?

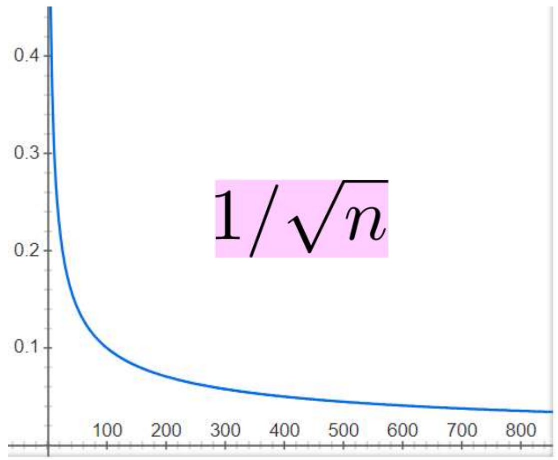
- Machine learning from big labeled data is highly successful.
  - Speech recognition, image understanding, natural language translation, recommendation…
- However, there are various applications where massive labeled data is not available.
  - Medicine, disaster, infrastructure, robotics, …
- Learning from limited information is promising.
  - Not learning from small samples.
  - We need many data, but they can be “weak”.

Our Target Problem: Binary Supervised Classification



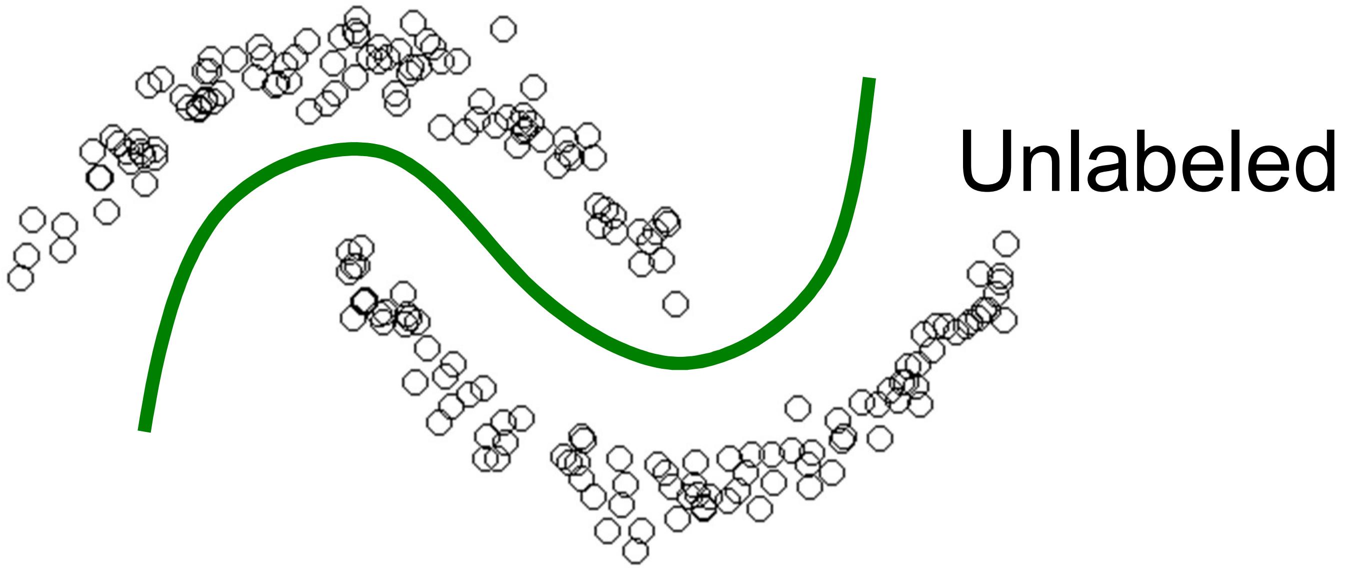
- Larger amount of labeled data yields better classification accuracy.
- Estimation error of the boundary decreases in order $1/\sqrt{n}$.

n : Number of labeled samples



Unsupervised Classification

- Gathering labeled data is costly. Let's use unlabeled data that are often cheap to collect:

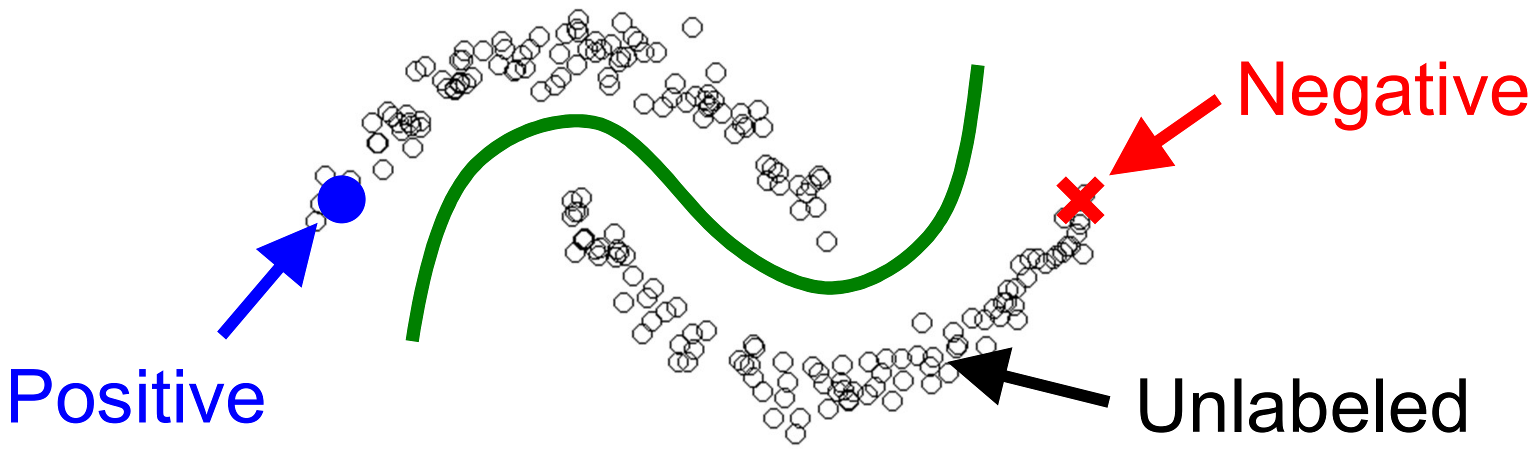


  - Unsupervised classification is typically clustering.
  - This works well only when each cluster corresponds to a class.

Semi-Supervised Classification

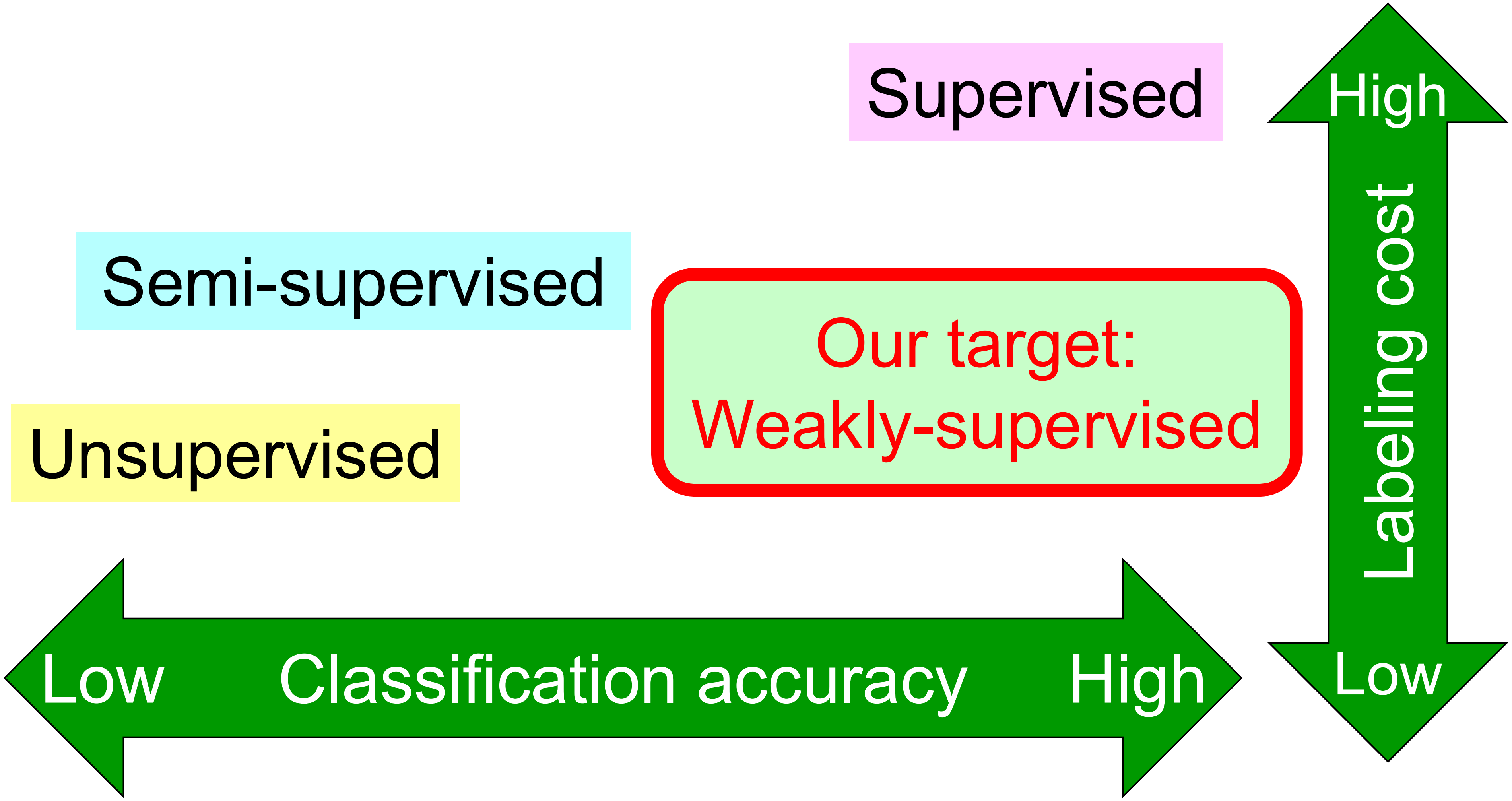
Chapelle, Schölkopf & Zien (MIT Press 2006) and many

- Use a large number of unlabeled samples and a small number of labeled samples.
- Find a boundary along the cluster structure induced by unlabeled samples:
  - Sometimes very useful.
  - But not that different from unsupervised classification.



Weakly-Supervised Learning

- High-accuracy and low-cost classification by empirical risk minimization.



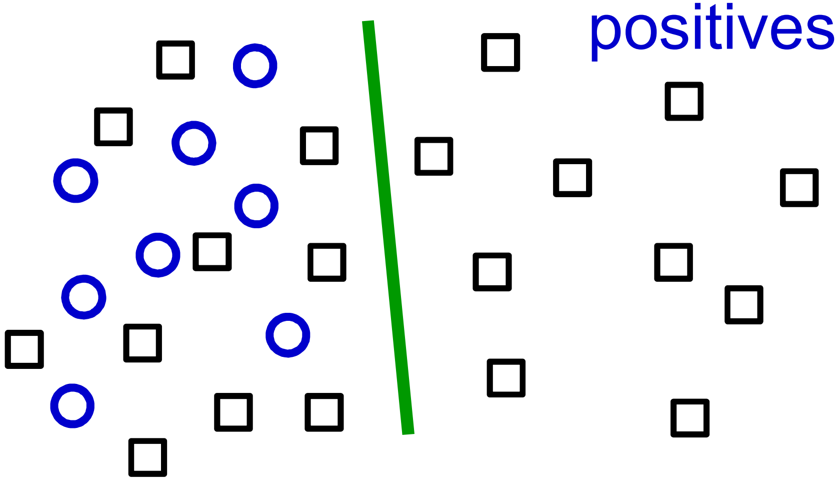
Method 1: PU Classification

du Plessis, Niu & Sugiyama (NIPS2014, ICML2015)
Niu, du Plessis, Sakai, Ma & Sugiyama (NIPS2016), Kiryo, Niu, du Plessis & Sugiyama (NIPS2017)
Hsieh, Niu & Sugiyama (arXiv2018), Kato, Xu, Niu & Sugiyama (arXiv2018)
Kwon, Kim, Sugiyama & Paik (arXiv2019), Xu, Li, Niu, Han & Sugiyama (arXiv2019)

- Only PU data is available; N data is missing:
  - Click vs. non-click
  - Friend vs. non-friend

Positive

Unlabeled (mixture of positives +1 and negatives)



$1/\sqrt{n}$

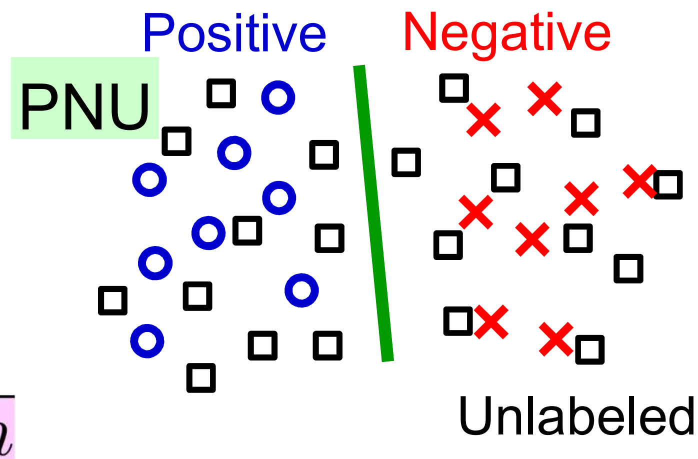
- From PU data, PN classifiers are trainable!

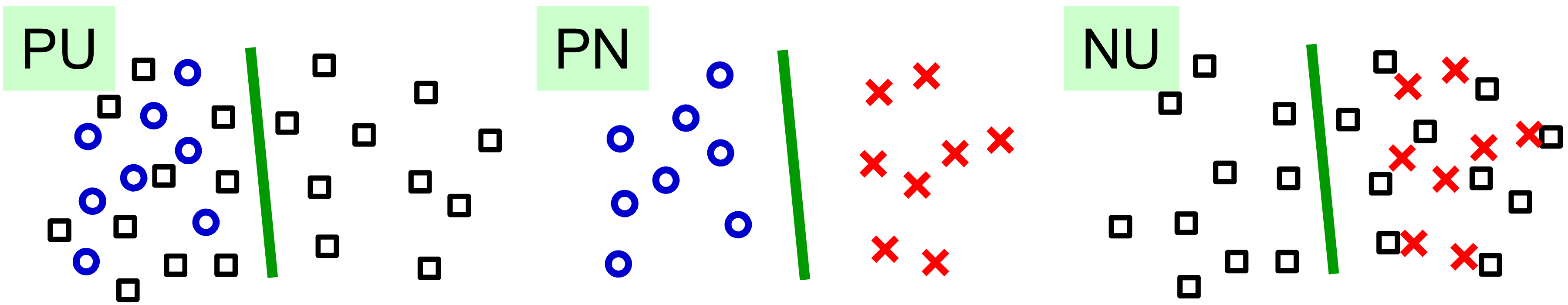
Method 2: PNU Classification (Semi-Supervised Classification)

Sakai, du Plessis, Niu & Sugiyama (ICML2017), Sakai, Niu & Sugiyama (MLJ2018)

- Let's decompose PNU into PU, PN, and NU:
  - Each is solvable.
  - Let's combine them!
- Without cluster assumptions, PN classifiers are trainable!

$1/\sqrt{n}$



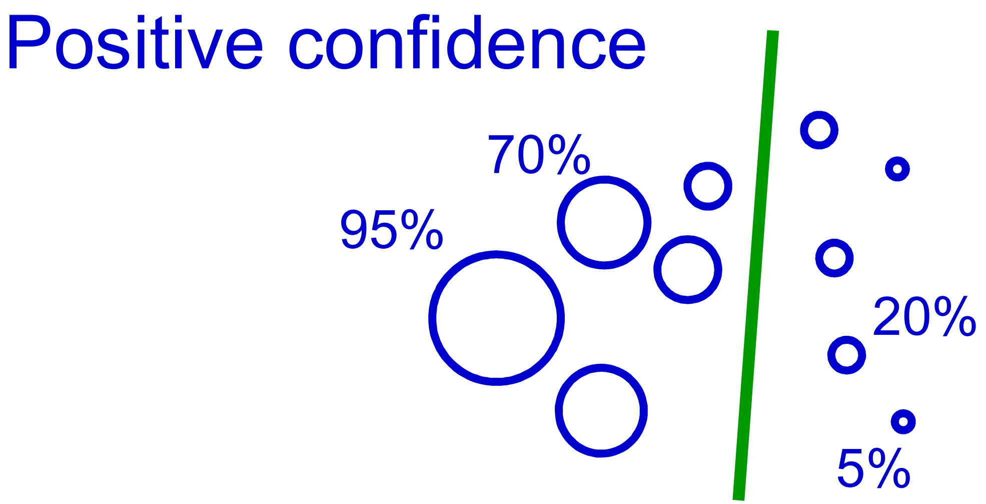


Method 3: Pconf Classification

Ishida, Niu & Sugiyama (NeurIPS2018)

- Only P data is available, not U data:
  - Data from rival companies cannot be obtained.
  - Only positive results are reported (publication bias).
- "Only-P learning" is unsupervised.
- From Pconf data, PN classifiers are trainable!

$1/\sqrt{n}$

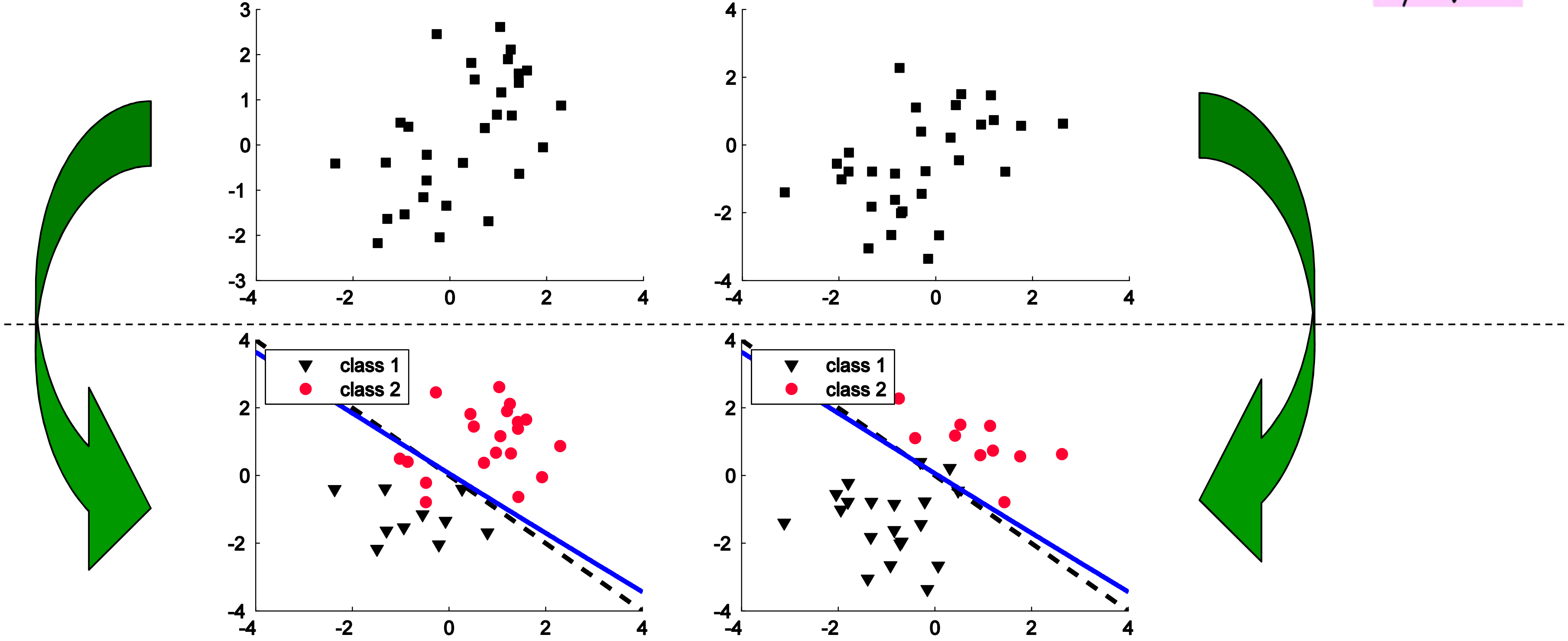


Method 4: UU Classification

du Plessis, Niu & Sugiyama (TAAI2013)
Nan, Niu, Menon & Sugiyama (ICLR2019)

- From two sets of unlabeled data with different class priors, PN classifiers are trainable!

$1/\sqrt{n}$

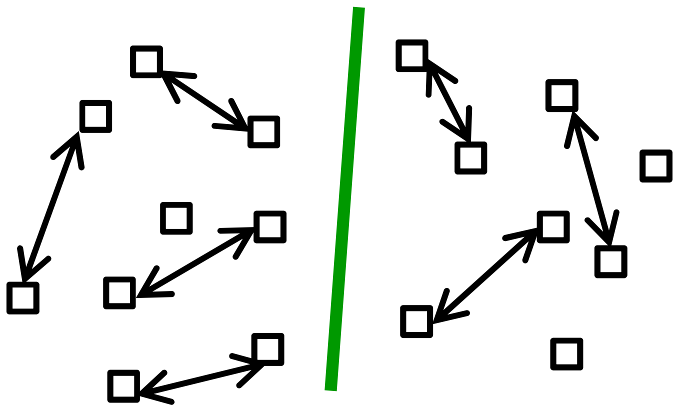


Method 5: SU Classification

Bao, Niu & Sugiyama (ICML2018)

- Delicate classification (salary, religion…):
  - Highly hesitant to directly answer questions.
  - Less reluctant to just say "same as him/her".
- From similar and unlabeled data, PN classifiers are trainable!

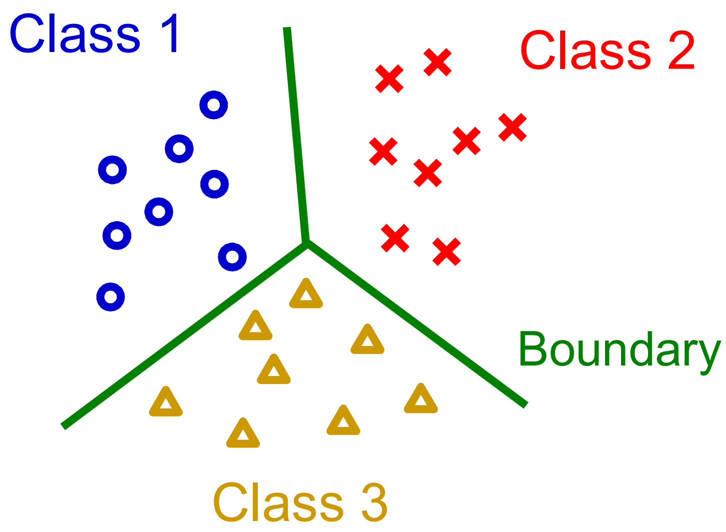
$1/\sqrt{n}$



Method 6: Comp. Classification

Ishida, Niu & Sugiyama (NIPS2017)
Ishida, Niu, Menon & Sugiyama (arXiv2018)

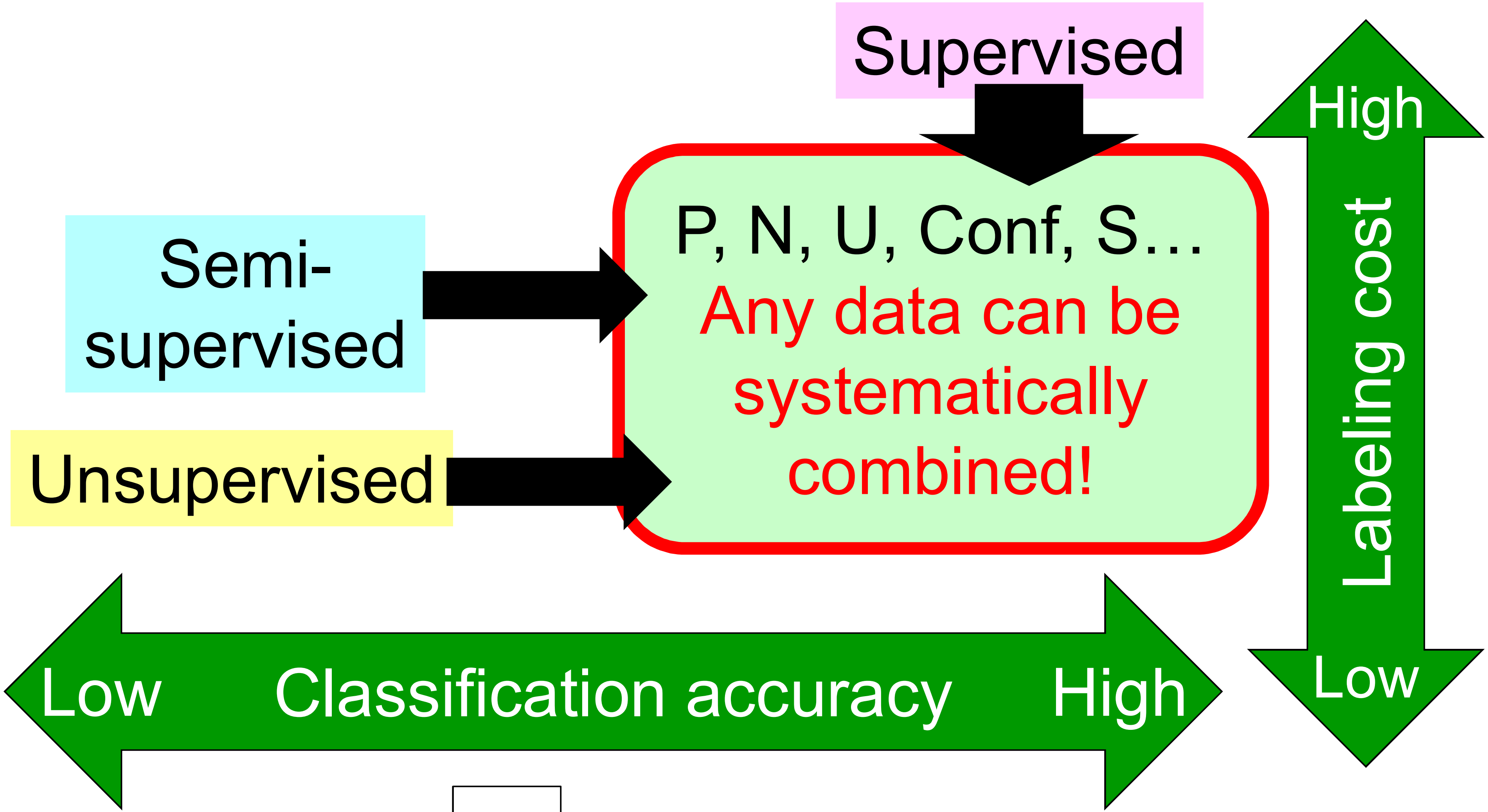
- Labeling patterns in multi-class problems:
  - Selecting a collect class from a long list of candidate classes is extremely painful.
- Complementary labels:
  - Specify a class that a pattern does not belong to.
  - This is much easier and faster to perform!



- From complementary labels, classifiers are trainable!

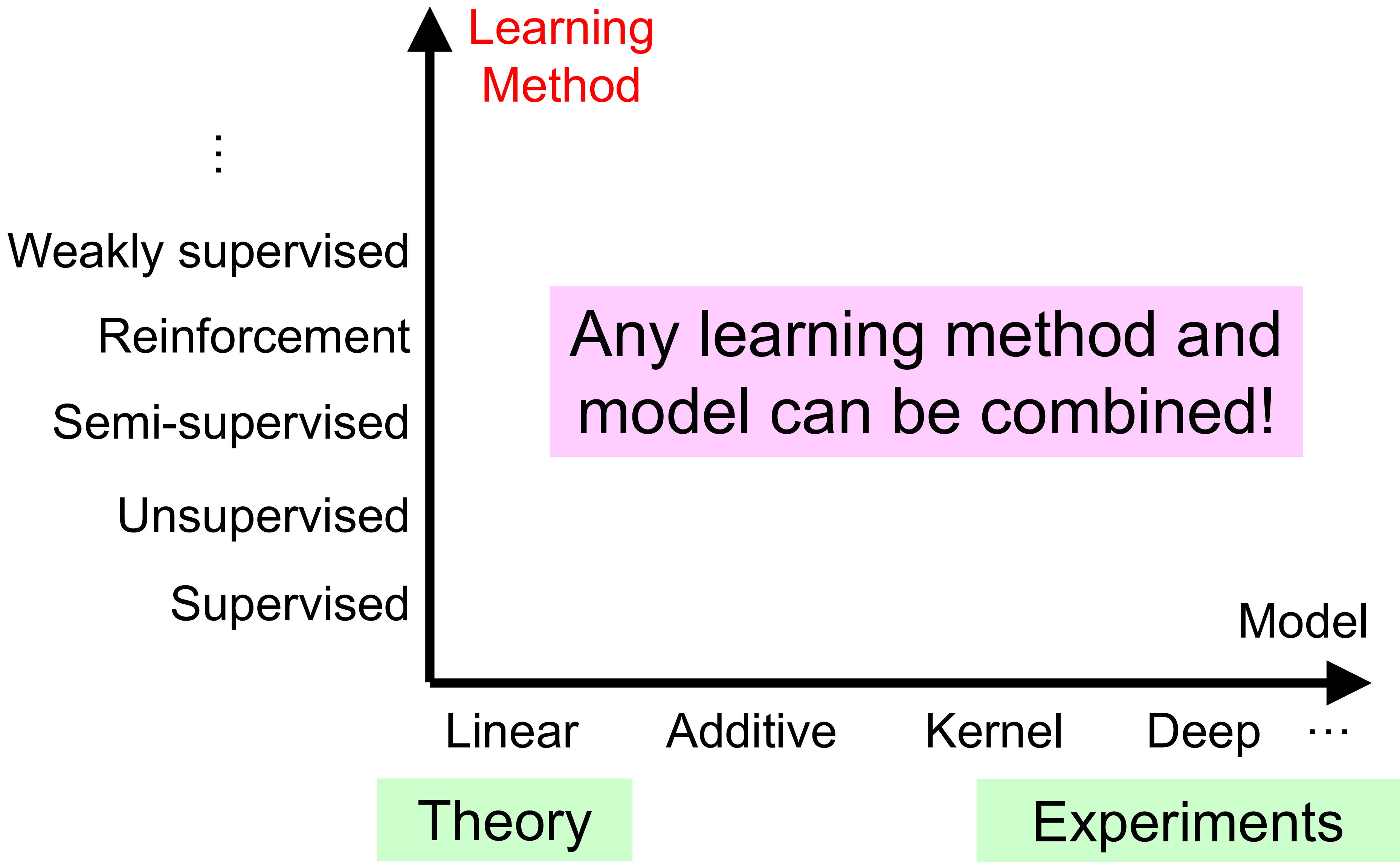
$1/\sqrt{n}$

Learning from Weak Supervision



Sugiyama, Niu, Sakai & Ishida,
Machine Learning from Weak Supervision
MIT Press, 2020 (?)

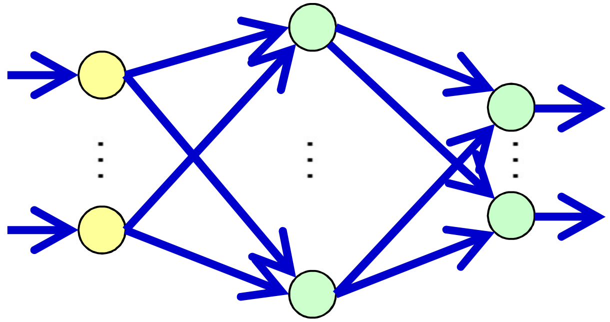
Model vs. Learning Methods



My Talk

1. Weakly supervised classification
2. Robust learning
3. More

Robustness in Deep Learning

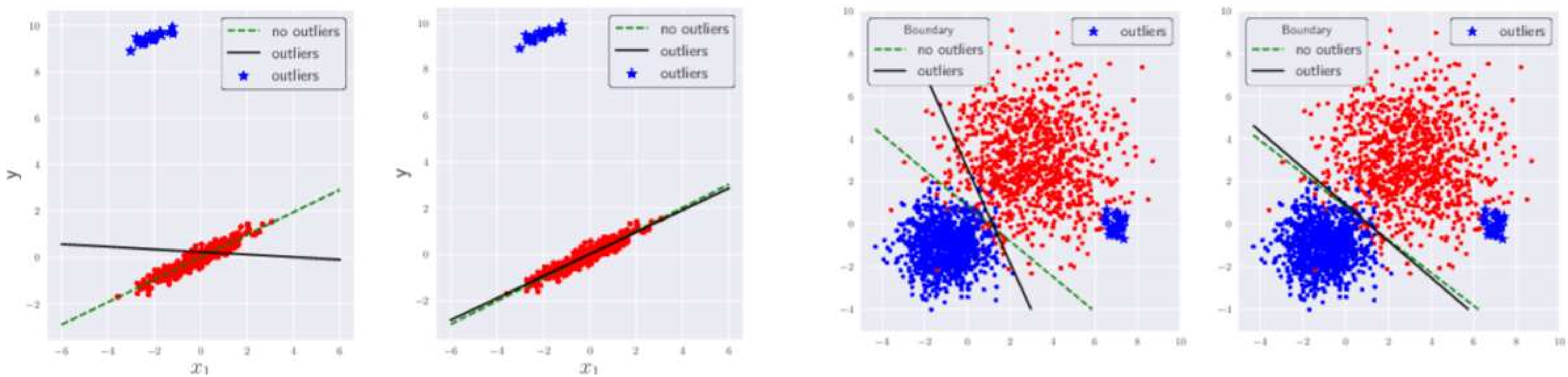


- Deep learning is successful.
- However, real-world is severe and various types of robustness is needed for reliability:
  - Robustness to noisy training data.
  - Robustness to changing environments.
  - Robustness to noisy test inputs.

Coping with Noisy Training Outputs

Futami, Sato & Sugiyama (AISTATS2018)

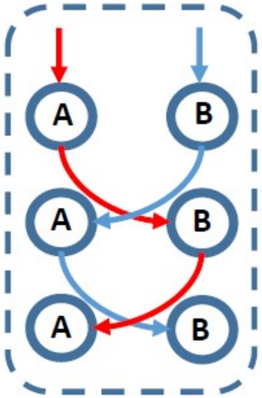
- Using a “flat” loss is suitable for robustness:
  - Ex) L^1-loss is more robust than L^2-loss.
- However, in Bayesian inference, robust loss is often computationally intractable.
- Our proposal: Not change the loss, but change the KL-div to robust-div in variational inference.



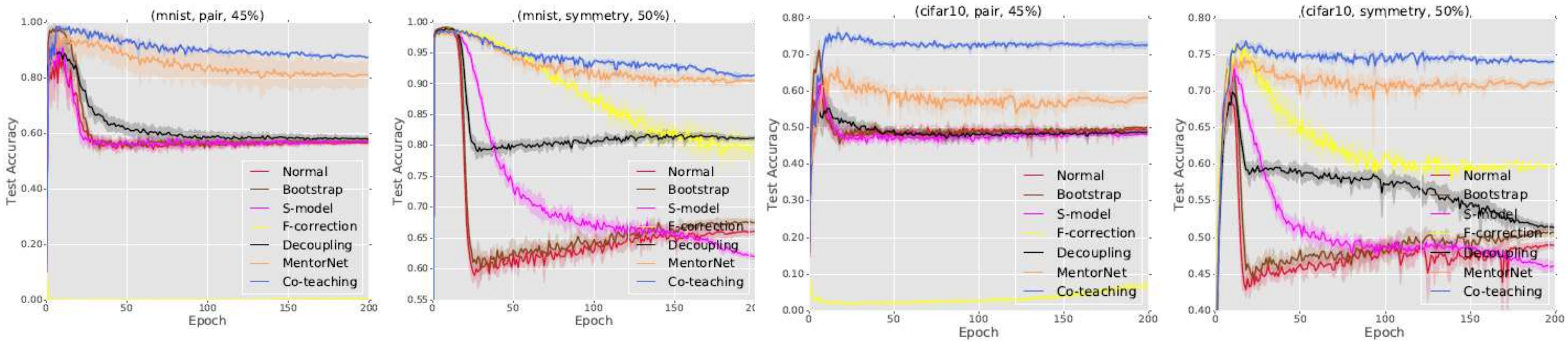
Coping with Noisy Training Outputs

Han, Yao, Yu, Niu, Xu, Hu, Tsang & Sugiyama (NeurIPS2018)

- Memorization of neural networks:
  - Empirically, clean data are fitted faster than noisy data.
- "Co-teaching" between two networks:
  - Select small-loss instances as clean data and teaches themto another network.



- Experimentally works very well!
  - But no theory.



Coping with Changing Environments

Hu, Sato & Sugiyama (ICML2018)

- **Distributionally robust supervised learning**:
  - Being robust to the worst test distribution.
  - Works well in regression.

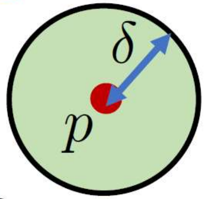
$$\min_{\theta} \sup_{\boxed{q\in\mathcal{Q}_p}} \mathbb{E}_{q(x,y)}[\ell(g_\theta(x), y)]$$

$$\mathcal{Q}_p = \{q \mid \mathrm{D}_f(q\|p) \le \delta\}$$

"f-divergence ball"

[Bagnell 2005, Ben-Tal+ 2013, Namkoong+ 2016, 2017]

E.g. KL divergence, Chi-square divergence

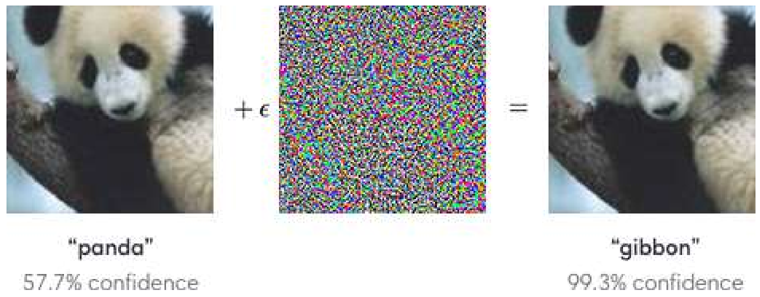


- **Our finding**: In classification, this merely results in the same non-robust classifier.
  - Since the 0-1 loss is different from a surrogate loss.
- Additional distributional assumption can help:
  - E.g., latent prior change — Storkey & Sugiyama (NIPS2007)

Coping with Noisy Test Inputs

Tsuzuku, Sato & Sugiyama (NeurIPS2018)

https://blog.openai.com/adversarial-example-research/



- **Adversarial attack** can fool a classifier.
- **Lipschitz-margin training:**

$$\forall \epsilon, \left(\|\epsilon\|_2 < c \Rightarrow t_X = \underset{i}{\operatorname{argmax}} \{ F(X+\epsilon)_i \} \right)$$

- Calculate the Lipschitz constant for each layer and derive the Lipschitz constant L_F for entire network.

$$\|F(X) - F(X+\epsilon)\|_2 \leq L_F \|\epsilon\|_2$$

- Add prediction margin to soft-labels while training.

$$M_{F,X} := F(X)_{t_X} - \max_{i \neq t_X} \{ F(X)_i \}$$

- Provable guarded area for attacks.
- Computationally efficient and empirically robust.

Coping with Noisy Test Inputs

Ni, Charoenphakdee, Honda & Sugiyama (arXiv2019)

- In severe applications, better to reject difficult test inputs and ask human to predict instead.
- Approach 1: Reject low-confidence prediction
  - Existing methods have limitation in loss functions (e.g, logistic loss), resulting in weak performance.
  - New rejection criteria for general losses with theoretical convergence guarantee.
- Approach 2: Train classifier and rejector
  - Existing methods only focuses on binary problems.
  - We show that this approach does not converge to the optimal solution in multi-class case.

My Talk

1. Weakly supervised classification
2. Robust learning
3. More

Estimation of Individual Treatment Effect

Yamane, Yger, Atif & Sugiyama (NeurIPS2018)

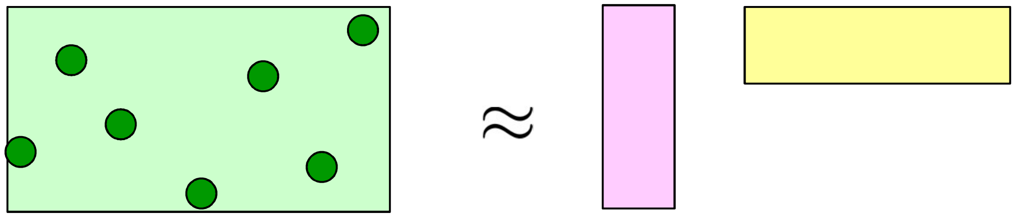
$$\mathbb{E}[y|x,t=1]-\mathbb{E}[y|x,t=-1]$$

x : subject, y: outcome, t : treatment flag

- Restriction: Due to privacy reasons, we can't have (x,y,t)-triplets, but only (x,y)- and (x,t)-pairs without correspondence in x.
- Result: Solvable if we have (x,y)- and (x,t)-pairs with two different treatment policies.
- Potential applications: Marketing/political campaign, medicine…

Sparse Matrix Completion

- Golden standard: Low-rank approximation of a matrix from its sparse observations.



- Matrix co-completion for multi-label classification with missing features and labels.

Feature | Soft labels

Xu, Niu, Han, Tsang, Zhou & Sugiyama (arXiv2018)

- Clipped matrix factorization for ceiling effect.
  - Allowing values taking beyond their upper-limits improves the recovery accuracy.

Teshima, Xu, Sato & Sugiyama (AAAI2019)

Domain Adaptation (DA)

- Unsupervised DA: source labeled and target unlabeled data
- Concern: If source- and target-data distributions are completely different, DA does not work.
  - How to measure distribution discrepancy is the key!
- Proposal: New discrepancy measures

Kuroki, Charoenphakdee, Bao, Honda, Sato & Sugiyama (AAAI2019)
Lee, Charoenphakdee, Kuroki & Sugiyama (arXiv2019)

My Talk

1. Weakly supervised classification
2. Robust learning
3. More

Summary

- Many problems are waiting to be solved!
  - Need better theory, algorithms, software, hardware, researchers, engineers, business models, ethics…
- Learning from imperfect information:
  - Weakly supervised/noisy training data
  - Reinforcement/imitation learning, bandits
- Reliable deployment of ML systems:
  - Changing environments, adversarial test inputs
  - Bayesian inference
- Versatile ML:
  - Density ratio/difference/derivative