Semi-Supervised Learning: Overview

Pilsung Kang
School of Industrial Management Engineering
Korea University

Machine Learning Categories

- Category of learning
  - ✓ According to the existence of target

Supervised Learning

A given dataset **X** & **Y**

	Var. 1	Var. 2	...	Var. d
Ins. 1	..	..	...	..
Ins. 2	..	..	...	..
...	...	...	...	...
Ins. N	..	..	..	..

$y = f(x)$

Y
..
..
...
..

Unsupervised Learning

A given dataset **X**

	Var. 1	Var. 2	...	Var. d
Ins. 1	..	..	...	..
Ins. 2	..	..	...	..
...	...	...	...	...
Ins. N	..	..	..	..

Semi-supervised Learning

A given dataset **X** & **Y**

	Var. 1	Var. 2	...	Var. d
Ins. 1	..	..	...	..
Ins. 2	..	..	...	..
...	...	...	...	...
Ins. N	..	..	..	..
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
Ins. M	..	..	..	..

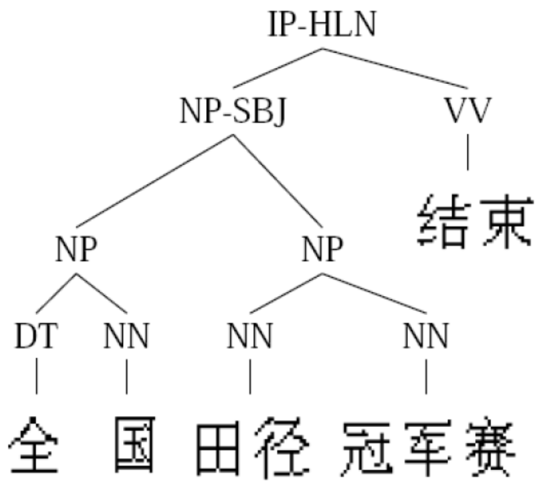
$y = f(x)$

Y
..
..
...
..

Backgrounds

Zhu (2007)

- People want better performance for free
  - ✓ Unlabeled data is cheap
  - ✓ Labeled data can be hard to get
    - human annotation is boring
    - labels may require experts
    - labels may require special devices
    - your graduate students are on vacation!!!
- Natural language parsing task (Penn Chinese Treebank)
  - ✓ 2 years for 4,000 sentences



Backgrounds

- Labeled data can be hard to get

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Backgrounds

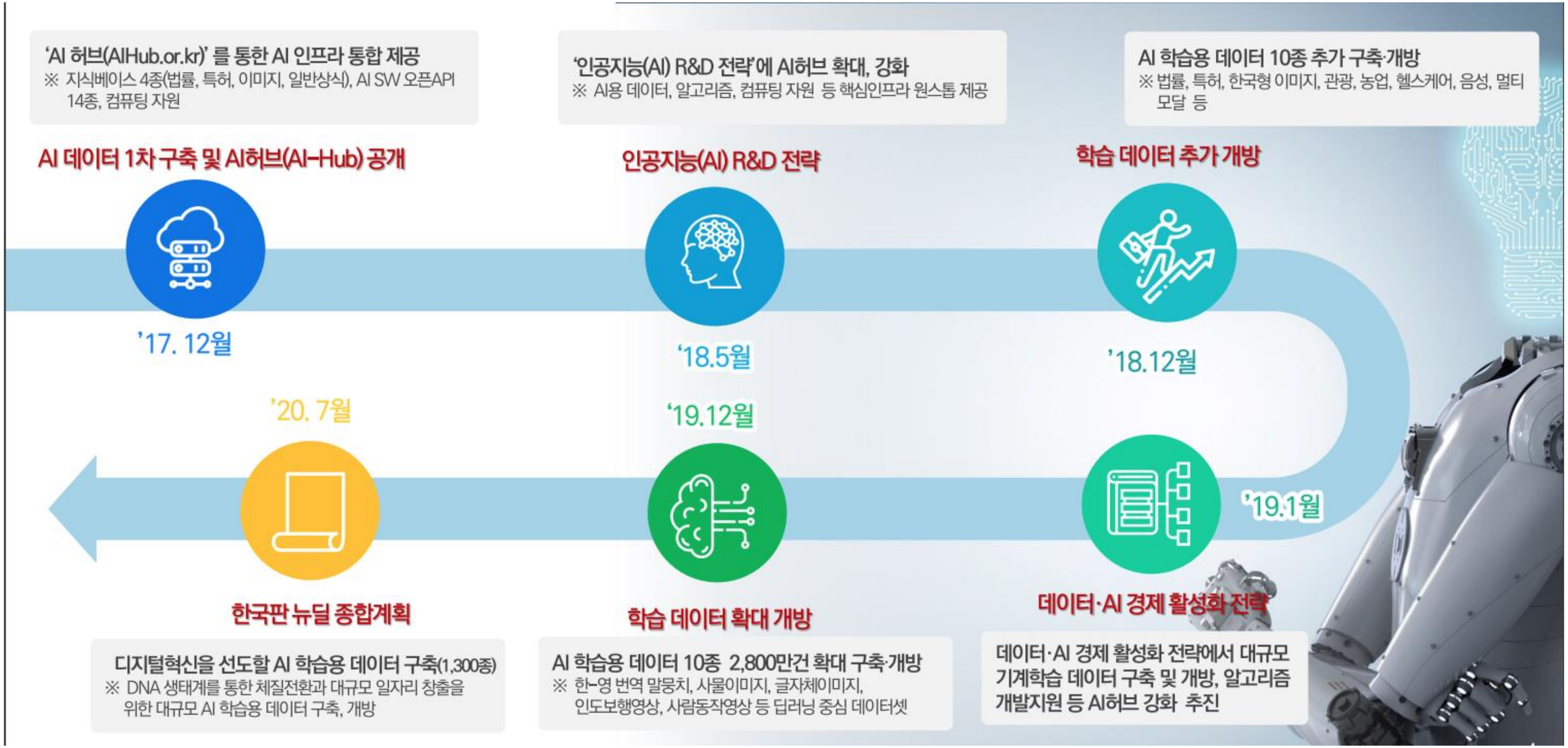
- Labeled data can be hard to get
  - 그래서 최후의 방법으로 남겨두었지만, 어쩔 수 없이 직접 labelling 하기로 하였습니다.
  - 하지만 711개 글자 class에 대하여 모두 하는 것은 비효율적일 것이라 판단하여, 글자를 모두 1개 class로 label을 할당하여 학습하기로 하였습니다.

Backgrounds

- Show me the money!

Backgrounds

- Show me the money!



Backgrounds

- Show me the money!

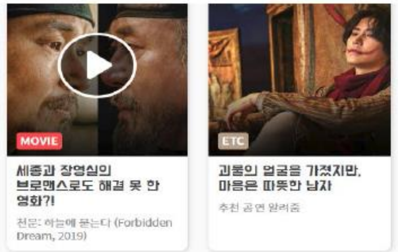
문서요약 텍스트

문서 요약 모델 및 요약 서비스 등을 위한 텍스트 AI데이터

1백단어 이상 1천 단어 이내의 길이, 1문장 이상 5문장 이내
문서 요약용 데이터 35만 개 이상

예상 서비스

문화 콘텐츠 요약 서비스,
기사 요약 서비스



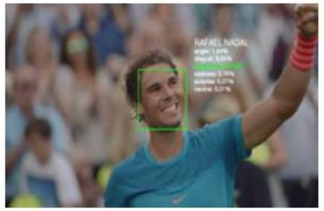
대용량 동영상 콘텐츠

자막 자동 생성을 위한 동영상 콘텐츠 AI데이터

10개 이상의 동영상 라벨링 구축, 동영상 내 1천개 이상 객체 카테고리 분류,
총 동영상 분량 500시간 이상

예상 서비스

컨텐츠 자동 검색 및 추천,
영상 분석 및 인지AI



Deep Fake 영상

가짜영상 판별을 위한 GAN기술 활용 영상합성 AI데이터

성별·나이·표정 등을 균등하게 분배한 원본 얼굴 동영상 및 변조 생성한
얼굴 동영상 20만개

예상 서비스

딥페이크 영상 탐지 및 방지
자동화

수어 영상

수어 동작(영상)과 의미(텍스트)를 결합한 영상 AI데이터

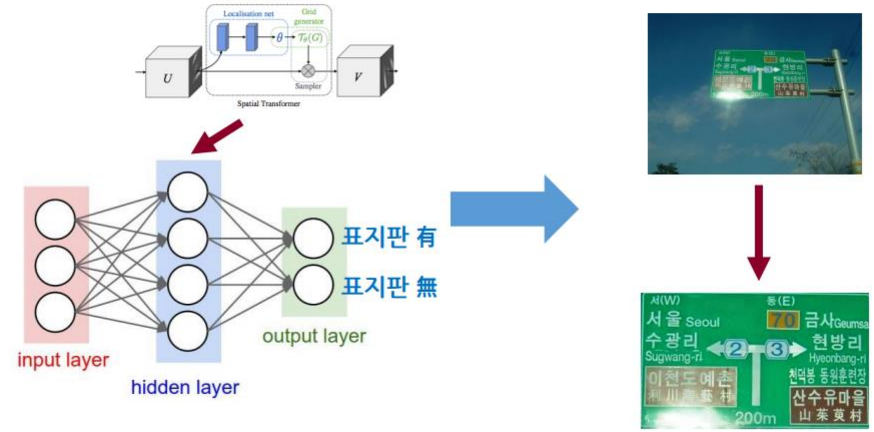
수어 4천 단어 이상, 수어 2천 문장 이상으로 구성된 3~5초 수어 동작
20만개 이상

예상 서비스

관공서 CS센터 수어 통·번역
서비스

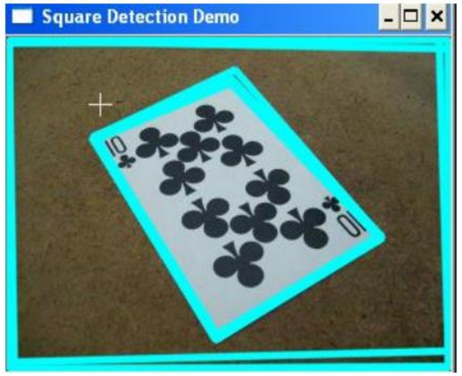
Backgrounds

- Labeled data can be hard to get
  - Spatial transformer를 이용하여 표지판 추출이 가능할까?



  - Square detection in OpenCV

OpenCV

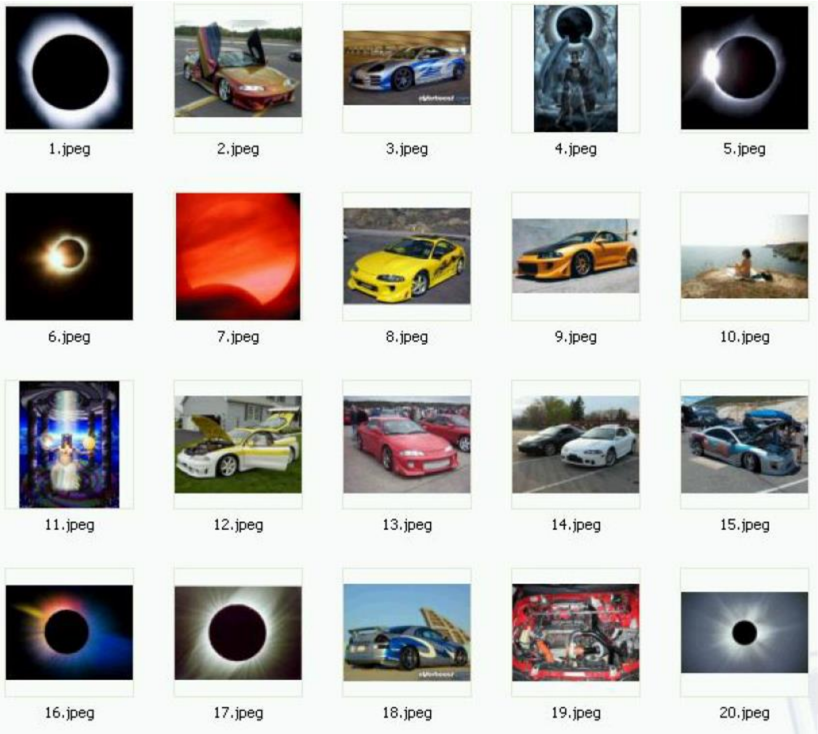


  - Square detection in OpenCV

Backgrounds

Zhu (2007)

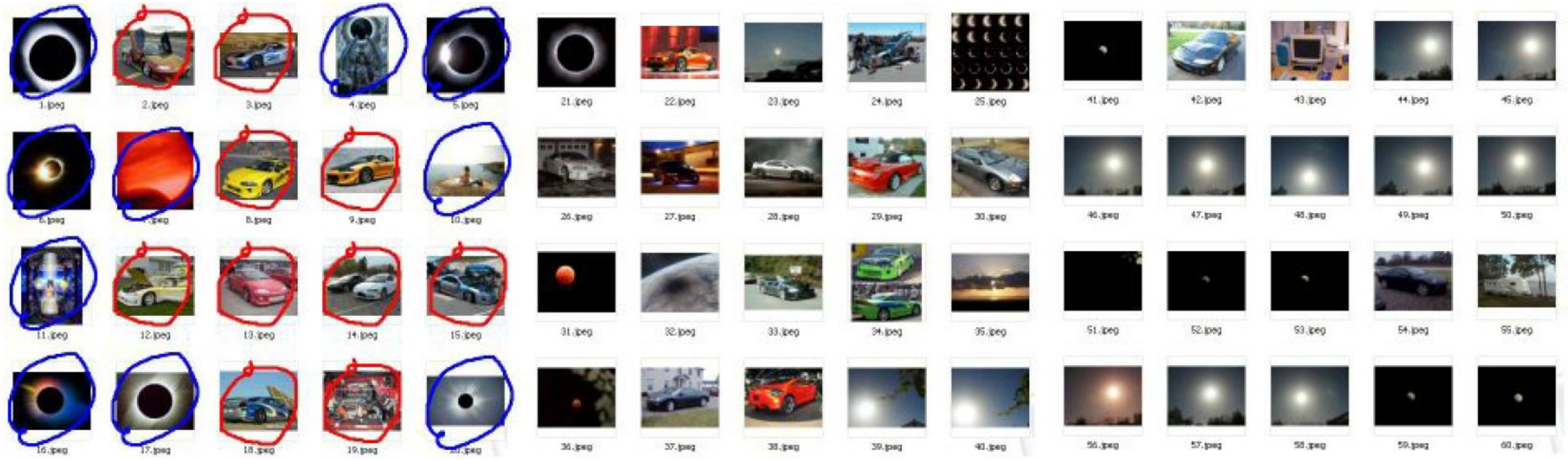
- For some tasks, it may not be too difficult to label 1000+ instances
  - ✓ Image categorization of **"eclipse"**



Backgrounds

Zhu (2007)

- Example of not-so-hard-to-get labels
  - ✓ Nonetheless…



  - ✓ We still have bunch of other images that are not labeled yet…

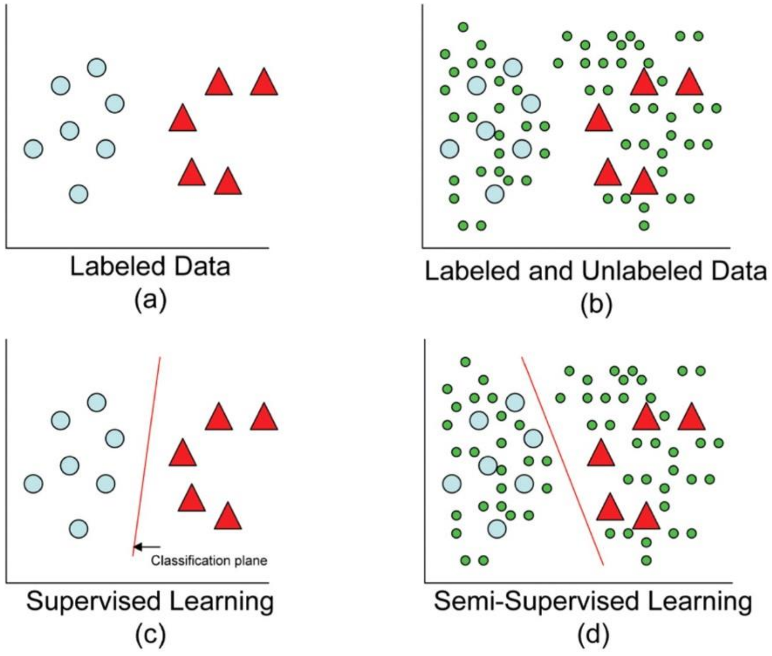
How can we utilize those unlabeled data

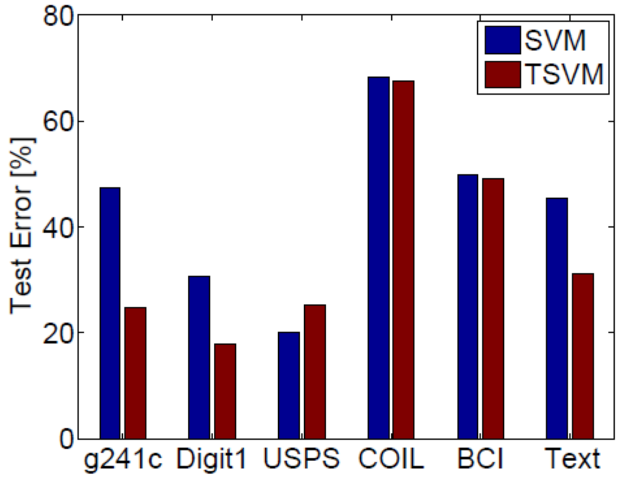
to improve the performance of supervised learning task?

Semi-Supervised Learning: Purpose

Zien (2008)

- Purpose
  - ✓ Using both labeled and unlabeled data to build better learners, than using each of alone.
- Can it be possible?





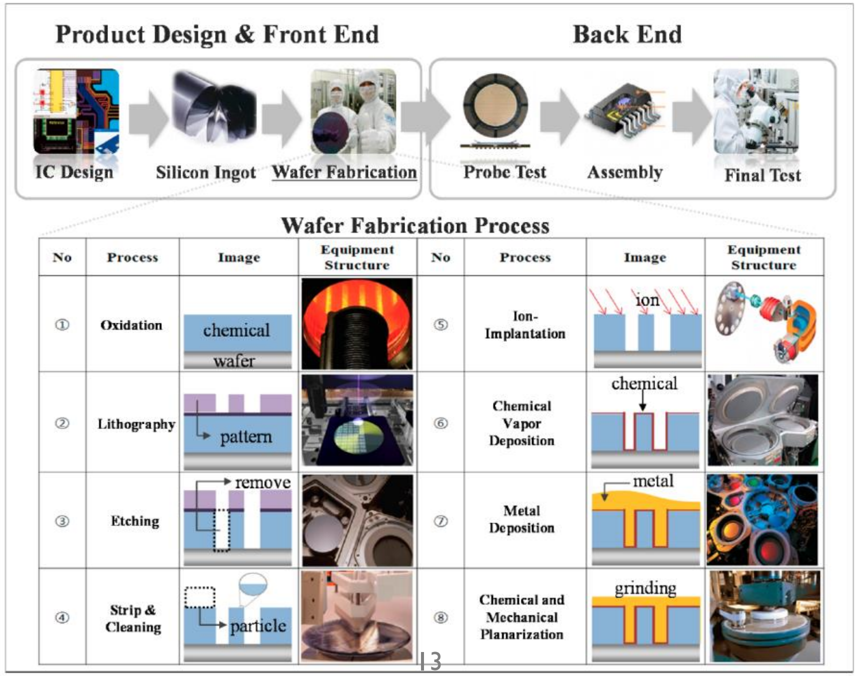
10 labeled points
~1400 unlabeled points

SVM: supervised
TSVM: semi-supervised

http://bioinformatics.oxfordjournals.org/content/24/6/783/F1.expansion.html

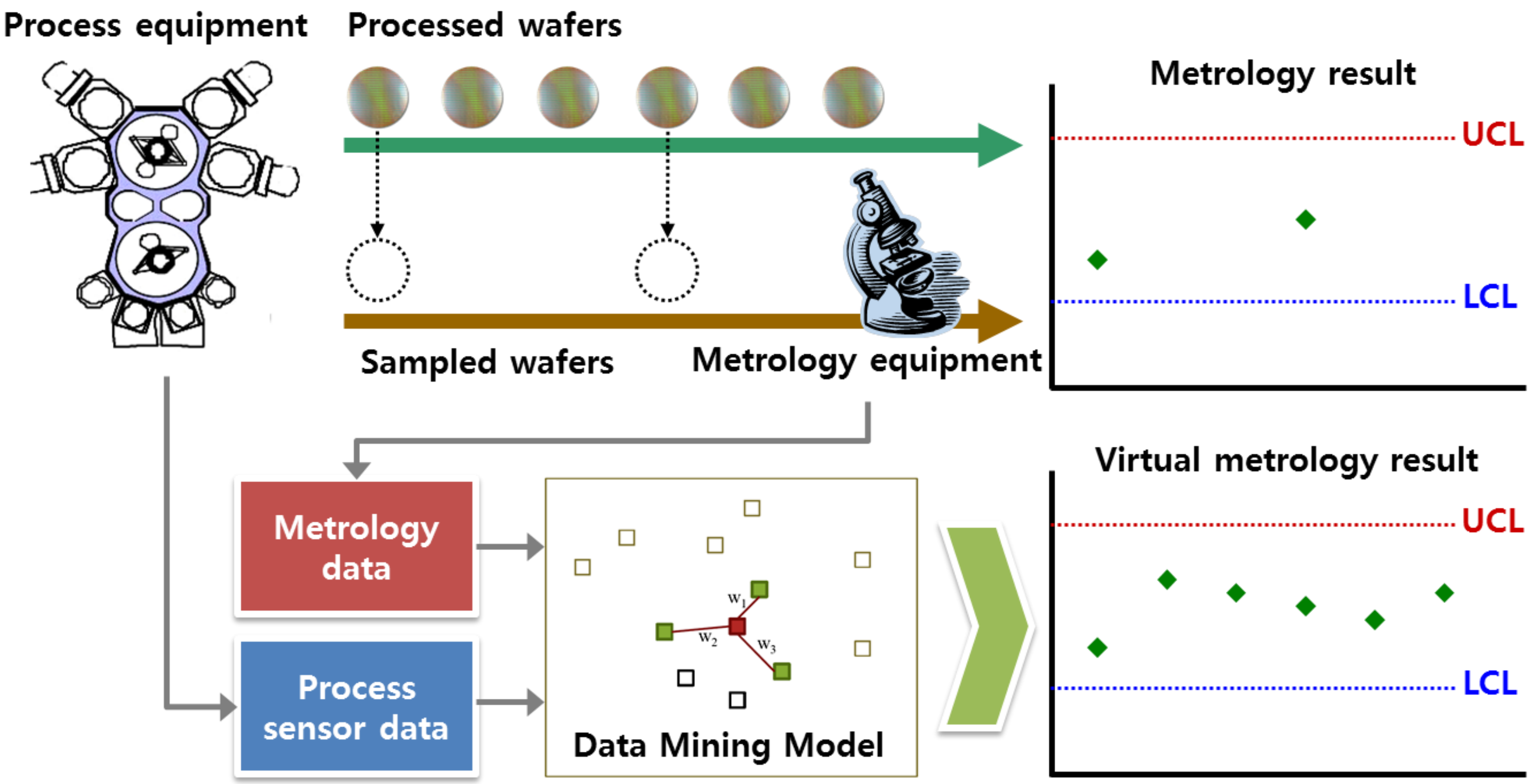
Virtual Metrology based on FDC Data

- **문제 인식**
  - ✓ 반도체 공정은 100개 이상의 세부 공정으로 이루어져 있으며, FAB-IN에서 FAB-OUT까지 평균 45일 가까운 시간이 소요 (자동차 72시간, 철강 48시간 이내)
  - ✓ 주요 공정 이후에 품질관리를 위해 계측을 수행
    - ▪ 샘플링 기반의 검사이므로 Type I/II 오류 발생
    - ▪ 계측 검사 기간에 소요되는 시간만큼의 Delay 발생



Virtual Metrology based on FDC Data

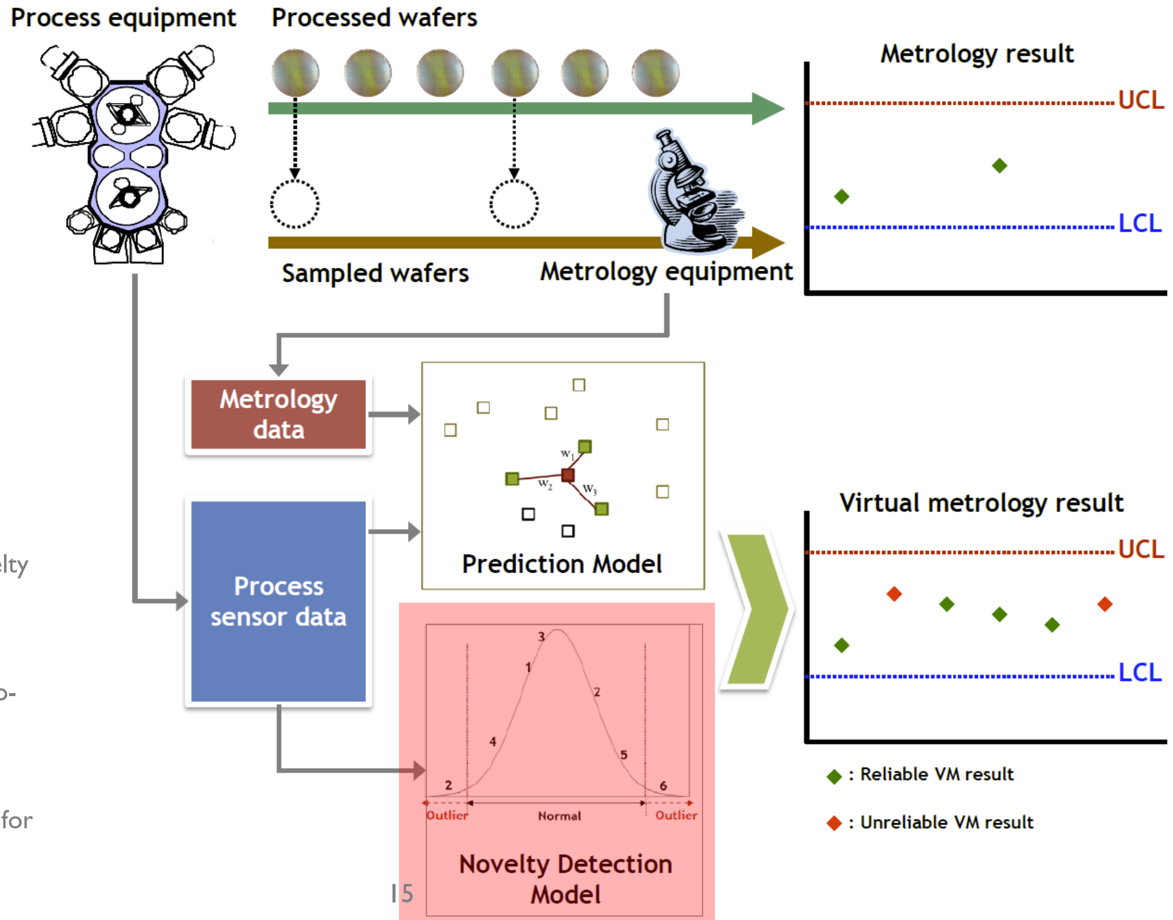
- 시도 1: FDC 데이터를 사용한 가상계측 모델 개발



Kang et. al. (2009) A virtual metrology system for semiconductor manufacturing, **Expert Systems with Applications 36**(10): 12554-12561.

Virtual Metrology with Prediction Reliability

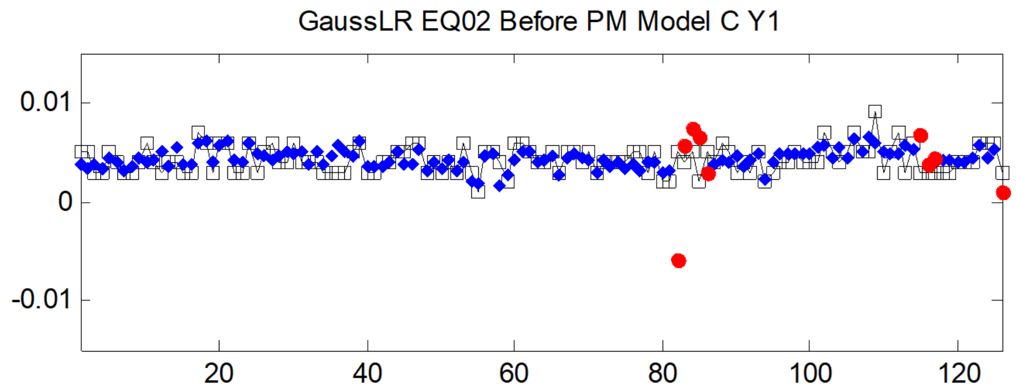
- 시도 2
  - ✓ 대안: 예측 모형의 신뢰도를 함께 제공하자
  - ✓ 높은 신뢰도를 갖는 예측 결과는 그대로 사용하고, 아닐 경우 엔지니어가 개입

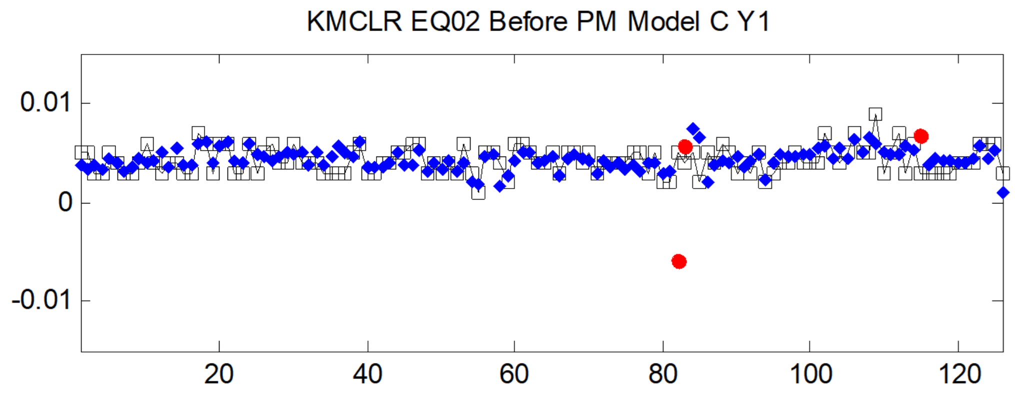


- Kim et al. (2012). Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. Expert Systems with Applications 39(4): 4075-4083.
- Kang et al. (2011). Virtual metrology for run-to-run control in semiconductor manufacturing. Expert Systems with Applications 38(3): 2508-2522.
- Kang et al. (2009). A virtual metrology system for semiconductor manufacturing. Expert Systems with Applications, 36(10): 12554-12561.

Virtual Metrology with Prediction Reliability

- 가상 계측 신뢰도 추정 예시
  - ✓ 검정색 네모: 실제 계측 값, 파란색 다이아몬드: 신뢰도가 높게 부여된 웨이퍼, 붉은색 동그라미: 신뢰도가 낮게 부여된 웨이퍼





Virtual Metrology based on FDC Data

- 데이터 관점의 이슈
  - ✓ 연구 당시(2006년~2009년)에는 Sampling 기반의 계측이 수행됨
  - ✓ 25개의 웨이퍼로 이루어진 1 Lot에서 1장의 웨이퍼에 대해서만 실계측 수행
  - ✓ 24장의 웨이퍼에 대해서는 FDC 데이터는 존재하나 계측 값은 존재하지 않음
  - ✓ Machine Learning 관점으로 보면 Input X는 존재하나 Target Y가 없는 웨이퍼들이 존재

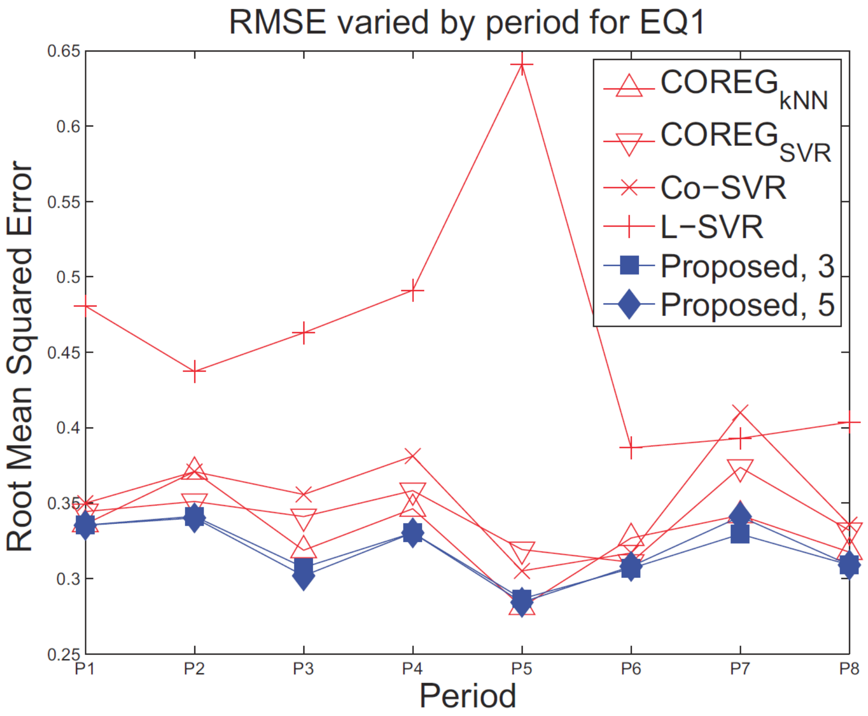
Wafer ID	X1	X2	X3	…	Xd	Y1	Y2	…	Yp

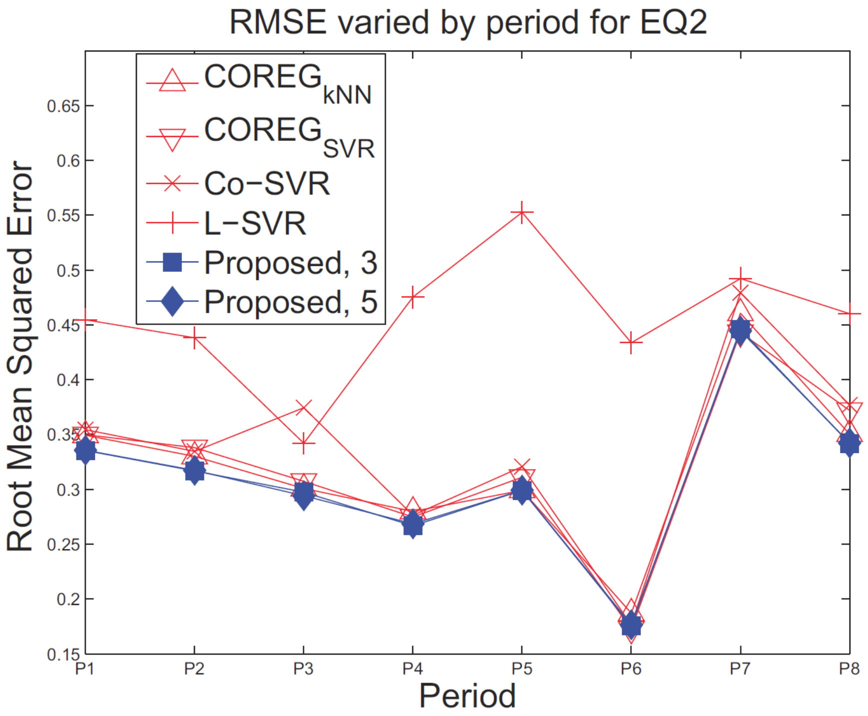
모든 웨이퍼들에 대해 FDC값 존재

일부 웨이퍼만 계측값 존재

Virtual Metrology based on FDC Data

- 방법론의 효과
  - ✓ 실계측 정보만 사용한 경우보다 대부분 예측 오차가 낮은 모형 구축 가능





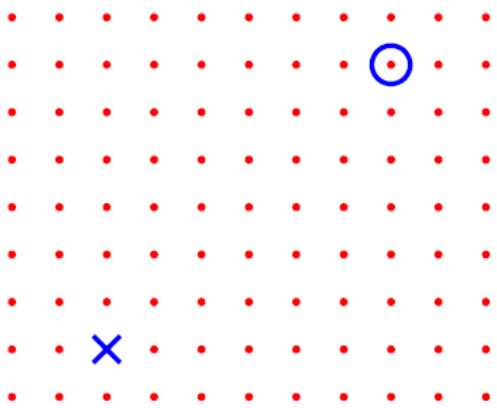
Semi-supervised Learning

Zien (2008)

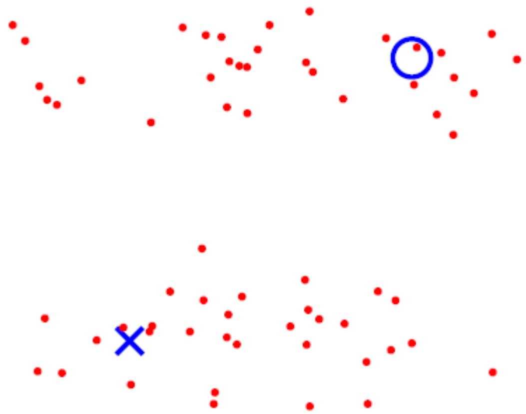
- Why would unlabeled data be useful at all?
  - ✓ Uniformly distributed data do not help
  - ✓ Must use properties of Pr(x)

Cluster Assumption

1. The data form clusters.
2. Points in the **same cluster** are likely to be of the **same class**.



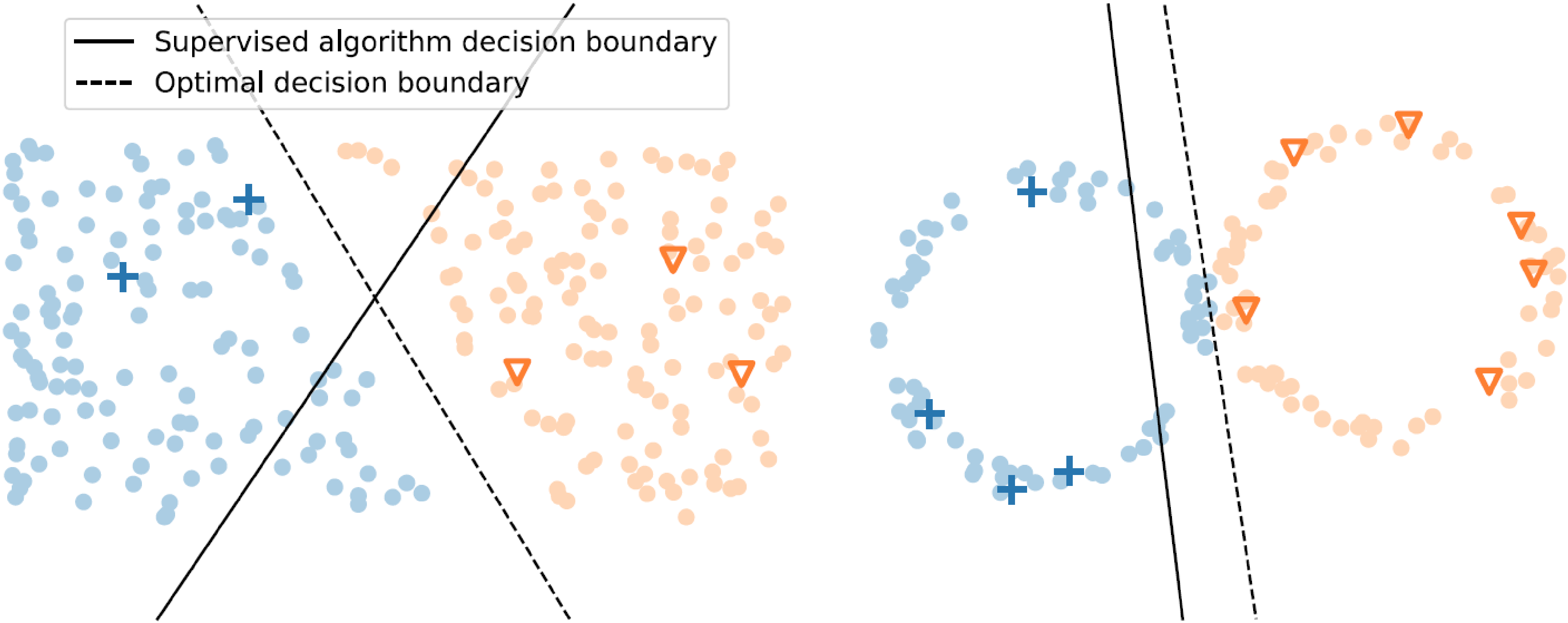
vs



Semi-supervised Learning

van Engelen and Hoos (2020)

- The effect of SSL: illustrative examples

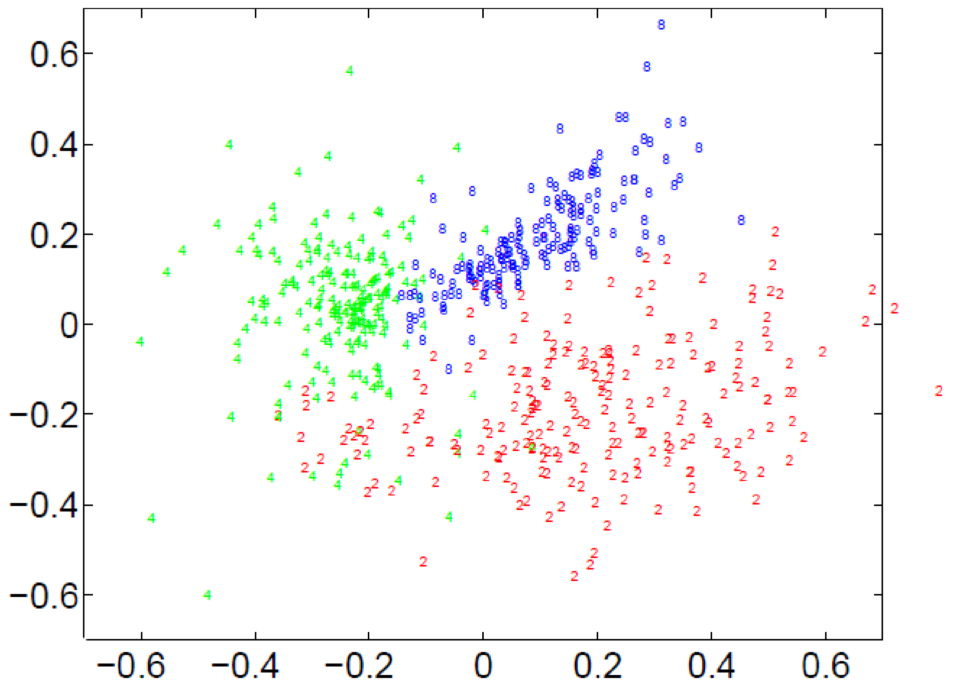


(a) Smoothness and low-density assumptions. (b) Manifold assumption.

Semi-supervised Learning

- Why would unlabeled data be useful at all?
  - ✓ The cluster assumption seems to hold for many real data sets
  - ✓ Many SSL algorithms (implicitly) make use of it

Example: 2D view on **handwritten digits** 2, 4, 8

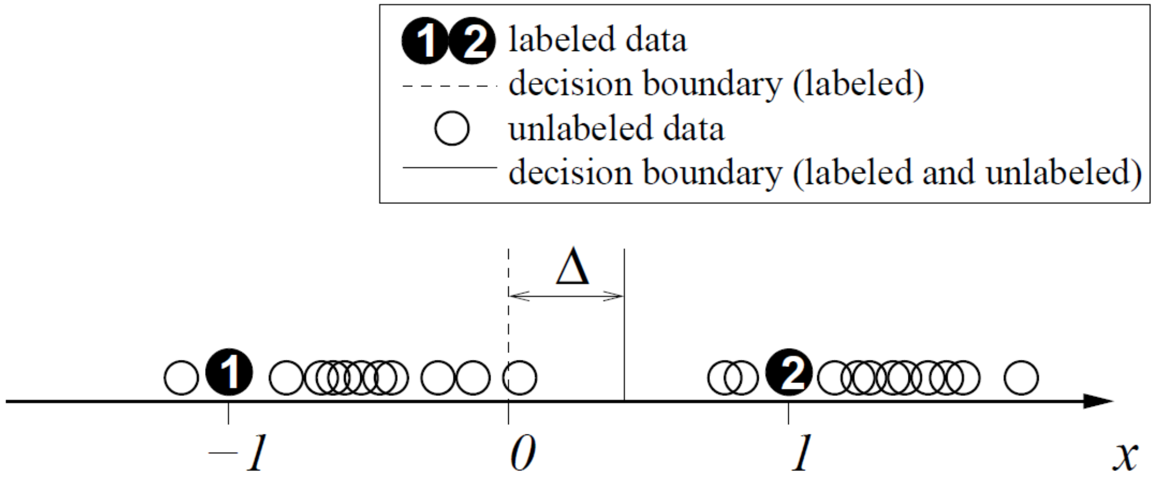


[non-linear 2D-embedding with "Stochastic Neighbor Embedding"]

Semi-supervised Learning

Zhu (2007)

- Why would unlabeled data be useful at all?
  - ✓ With and without unlabeled data: decision boundary shift



- Does unlabeled data always help?
  - ✓ Unfortunately, this is not the case, yet. (Ben-David et al. (2008) and Singl et al. (2008).)

Semi-supervised Learning

Zhu (2007)

- Notations
  - ✓ Input instance $\mathbf{x}$, label y
  - ✓ Learner $f : \mathcal{X} \mapsto \mathcal{Y}$
  - ✓ Labeled data $(\mathbf{X}_l, \mathbf{y}_l) = \{(\mathbf{x}_{1:l}, y_{1:l})\}$
  - ✓ Unlabeled data $\mathbf{X}_u = \{(\mathbf{x}_{l+1:n})\}$, available during training
  - ✓ Usually $l \ll n$
  - ✓ Test data $\mathbf{X}_{test} = \{(\mathbf{x}_{n+1:})\}$, not available during training
- SSL vs. Transductive learning

Semi-supervised learning	Transductive learning
is ultimately applied to the test data (inductive).	is only concerned with the unlabeled data.

Semi-supervised Learning

- Semi-supervised vs. Transductive

Setting	Input to the algorithm (Distribution *D*)	Algorithm output	Performance of the algorithm	Examples
Supervised Learning	labeled examples	a function that maps points to labels	expected error on an unseen point	Support Vector Machines, AdaBoost
Unsupervised Learning	unlabeled examples	a function that maps points to labels	expected error on an unseen point	Clustering
Semi-supervised Learning (SSL)	labeled & unlabeled examples	a function that maps points to labels	expected error on an unseen point	
Transductive Learning	labeled & unlabeled examples	labels of the unlabeled examples	average error on unlabeled examples	Transductive SVMs, Graph regularization

http://cs.nyu.edu/~rostami/ml/2008/rastogi-ssl.pdf

Semi-supervised Learning

Zhu (2007)

- Semi-supervised vs. Transductive

supervised learning (classification, regression) $\{(x_{1:n}, y_{1:n})\}$

$\updownarrow$

semi-supervised classification/regression $\{(x_{1:l}, y_{1:l}), x_{l+1:n}, x_{test}\}$

transductive classification/regression $\{(x_{1:l}, y_{1:l}), x_{l+1:n}\}$

$\updownarrow$

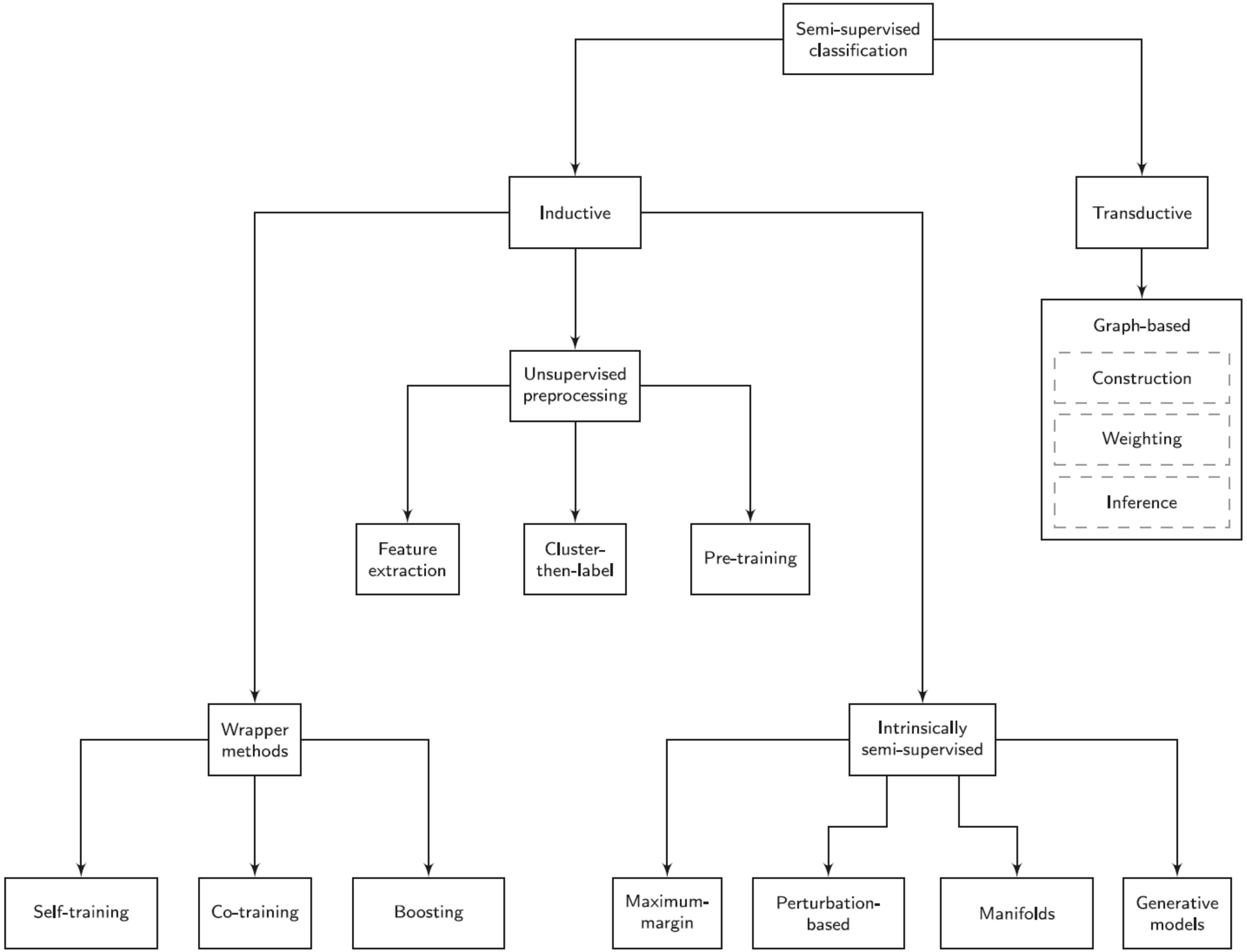
semi-supervised clustering $\{x_{1:n}, \text{must-, cannot-links}\}$

$\updownarrow$

unsupervised learning (clustering) $\{x_{1:n}\}$

Semi-supervised Learning: Taxonomy

van Engelen and Hoos (2020)




Q & A

References

Research Papers

- Ben-David, S., Lu, T., & Pál, D. (2008, July). Does Unlabeled Data Provably Help? Worst-case Analysis of the Sample Complexity of Semi-Supervised Learning. In *COLT* (pp. 33-44).
- Bennett, K., & Demiriz, A. (1999). Semi-supervised support vector machines. *Advances in Neural Information processing systems*, 368-374.
- Blum, A., & Mitchell, T. (1998, July). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory* (pp. 92-100). ACM.
- Fox-Roberts, P., & Rosten, E. (2014). Unbiased generative semi-supervised learning. *The Journal of Machine Learning Research*, *15*(1), 367-443.
- Kim, D., Seo, D., Cho, S., & Kang, P. (2017+). Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. under review.
- Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems* (pp. 3581-3589).
- Singh, A., Nowak, R., & Zhu, X. (2009). Unlabeled data: Now it helps, now it doesn't. In *Advances in neural information processing systems* (pp. 1513-1520).
- Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine Learning, 109(2), 373-440.
- Yu, S., Krishnapuram, B., Rosales, R., & Rao, R. B. (2011). Bayesian co-training. *The Journal of Machine Learning Research*, *12*, 2649-2680
- Zhou, Z. H., & Li, M. (2005, July). Semi-Supervised Regression with Co-Training. In IJCAI (Vol. 5, pp. 908-913).

References

<u>Other materials</u>

- Figures in the first page: 하상욱 단편시집 – 서울 시
- Zhu, X. (2007). Semi-Supervised Learning Tutorial. International Conference on Machine Learning (ICML 2007).
- Choi, S. (2015). Deep Learning: A Quick Overview. Deep Learning Workship. KIISE.
- Zien, A. (2008). Semi-Supervised Learning. Summer School on Neural Networks.
- Zhu, X. (2009). Tutorial on Semi-Supervised Learning. Theory and Practice of Computational Learning.