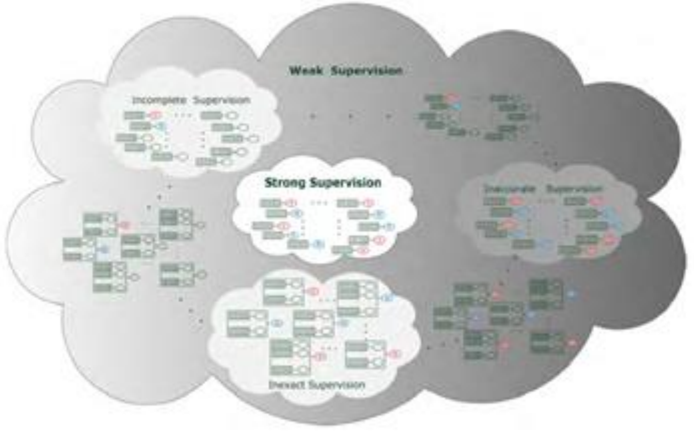
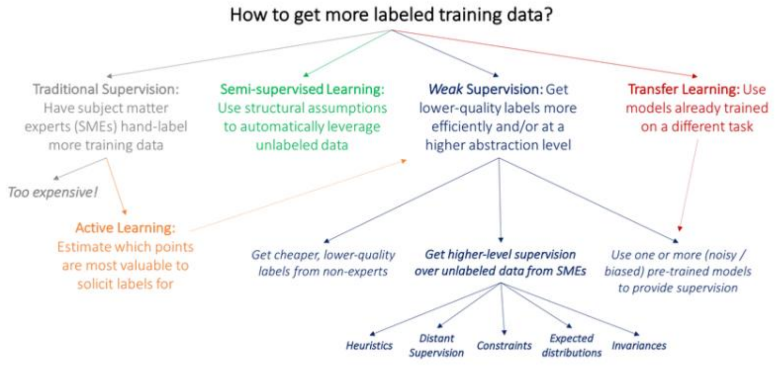
Weakly supervised learning – focus on Active Learning

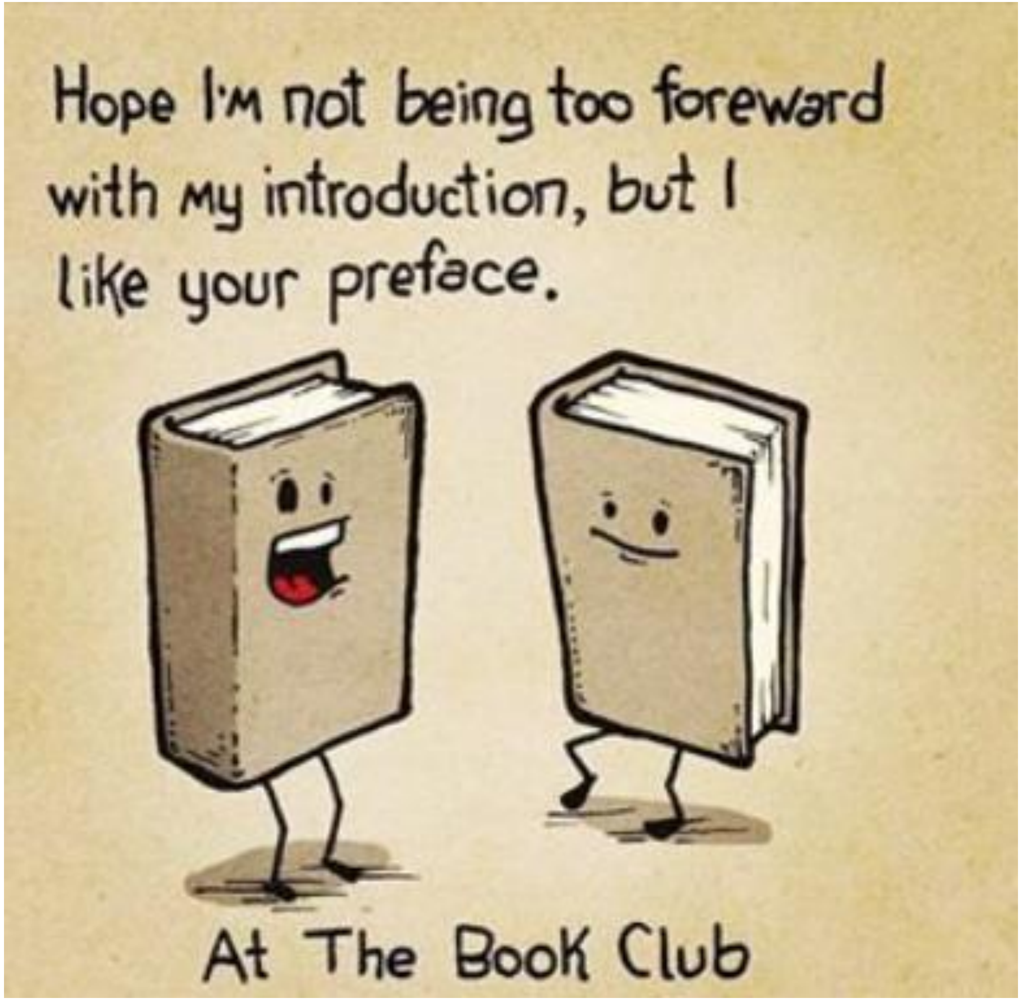




Vincent Lemaire

Figures from [...]

Preamble


Hope I'm not being too foreward with my introduction, but I like your preface.
At The Book Club

What is this talk about ?

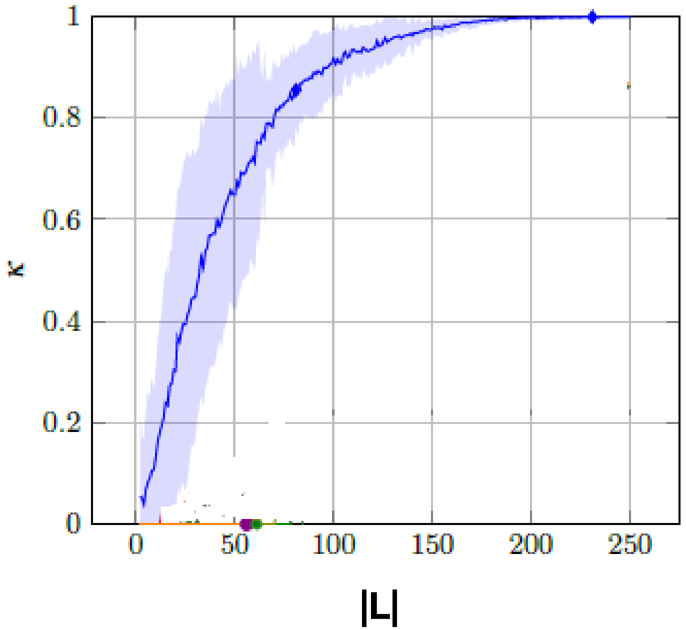
- Machine learning from big labeled data is highly successful
  - Speech recognition, image understanding, natural language translation, …
- However, there are various applications where massive labeled data is not available
  - Medicine, robots, frauds, …
- In this talk I will discuss about classification from limited information
  1. Weak data (but we assume that we have a lot of them)
  2. Small data (possible without strong (prior) domain knowledge ?)

and…

https://www.numerama.com/sciences/578203-pour-facebook-lintelligence-artificielle-sera-econome-en-donnees-ou-ne-sera-pas.html

Focus on a particular target problem : Binary Supervised Classification

Dataset	$\|\mathcal{D}\|$	$\|\mathcal{Y}\|$	#num	#nom	maj (%)	min (%)
musk	6581	2	166	1	84.55	15.45



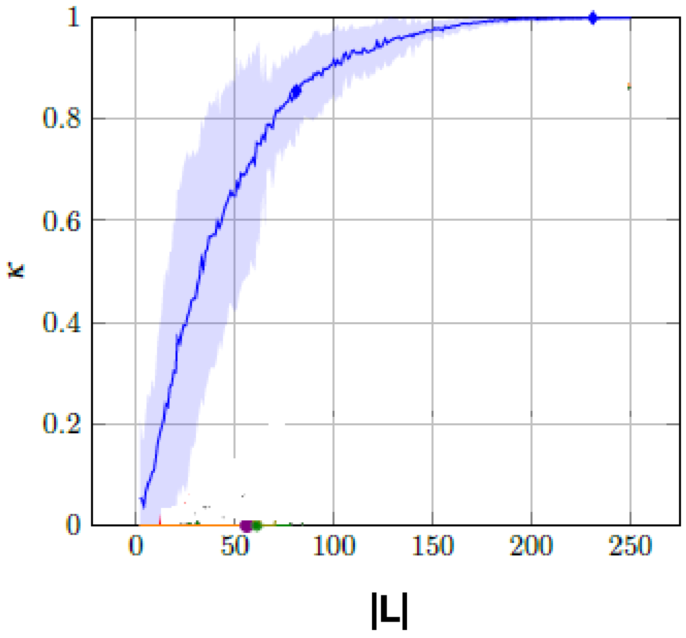
- Random Forest
- Kappa on Test

- Large amount of labeled data yields better performances
- Estimation error decreases in order 1 / √ |L|

Focus on a particular target problem : Binary Supervised Classification

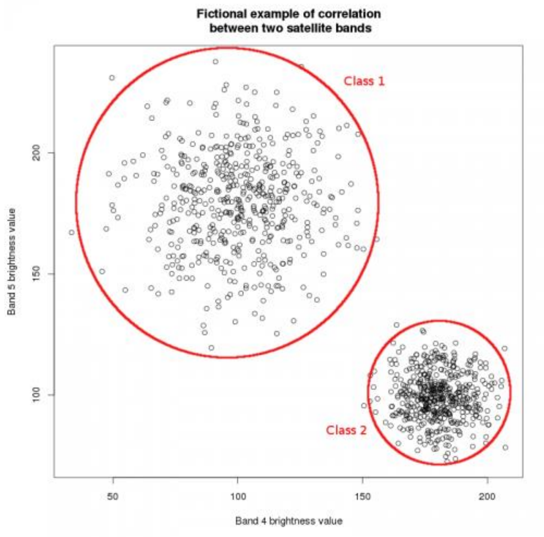
1. Unsupervised classification
2. Semi-Supervised classification
3. Supervised classification

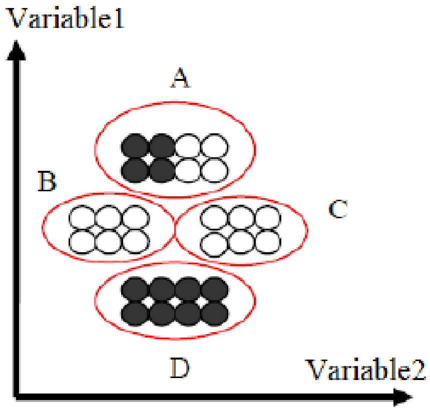
Details to come on the first two



Unsupervised Classification

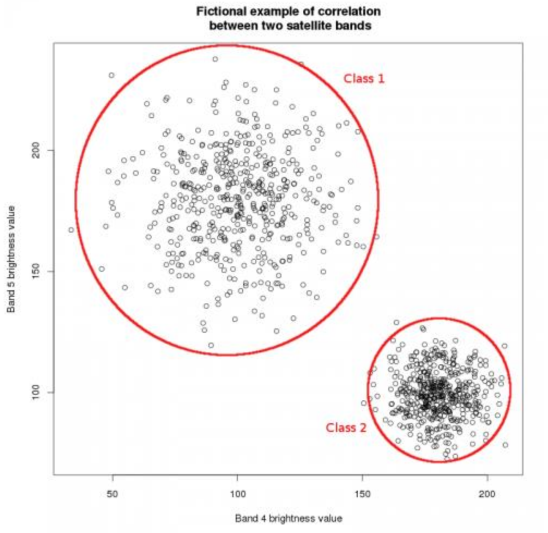
- Gathering labeled data is costly
- Try to use unlabeled data (only)
- Unsupervised Classification is typically clustering
- ‘Assumption’: each cluster corresponds to a class

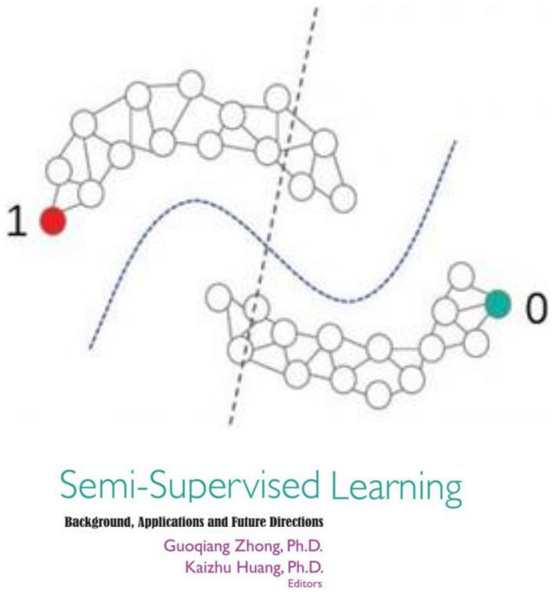




Semi-Supervised Classification

- Use:
    - a large number of unlabeled samples
    - a small number of labeled samples
- Try to find a 'boundary' (for example using labels propagation) along the cluster structure

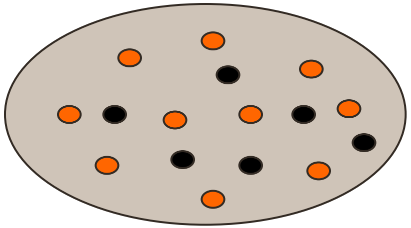




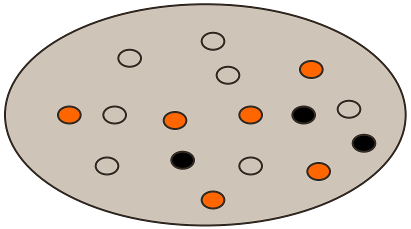
*to work well

But not only…

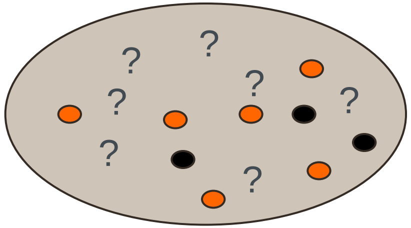
Supervised



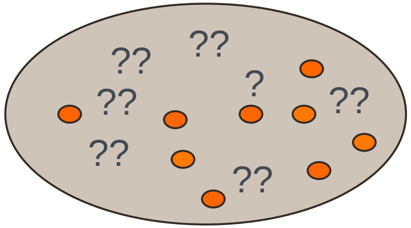
Semi Supervised



Active



Positive Unlabeled



Classification of Classification

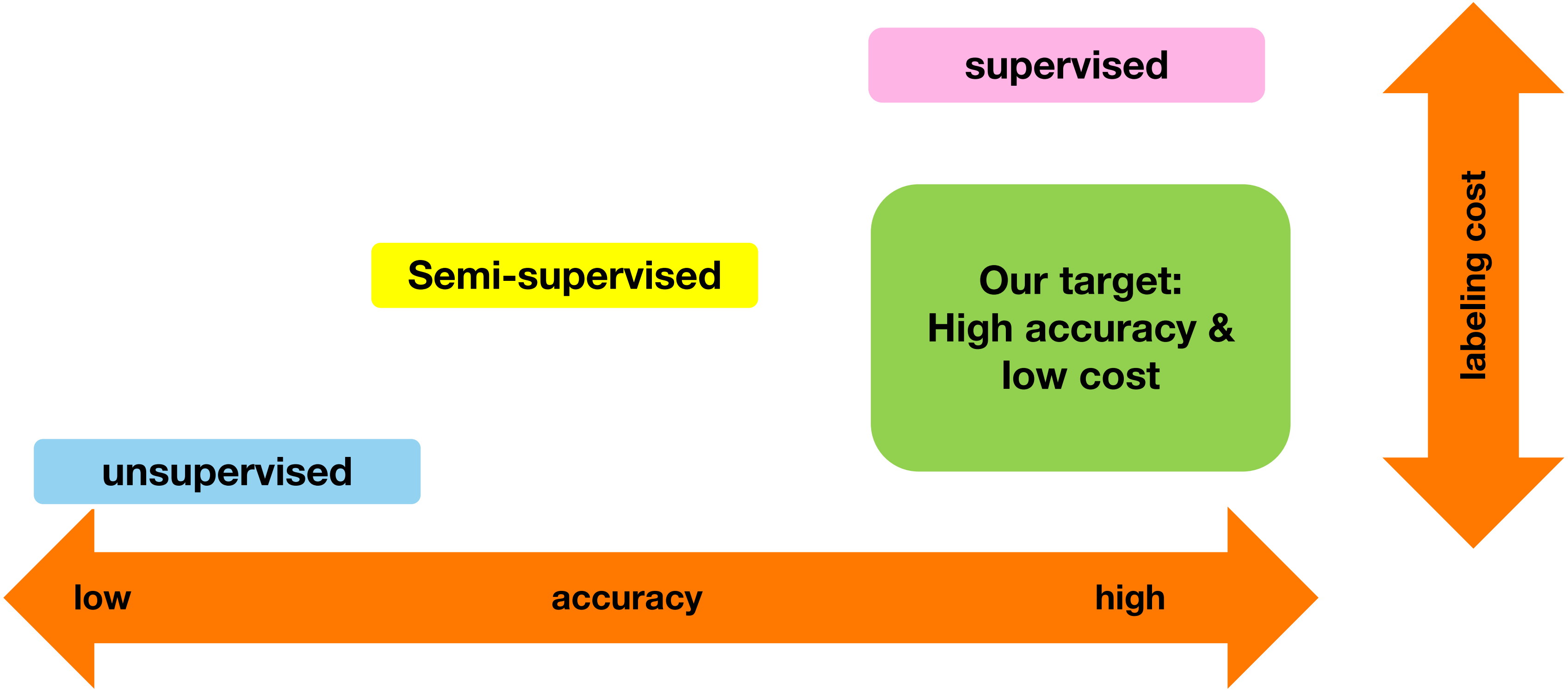
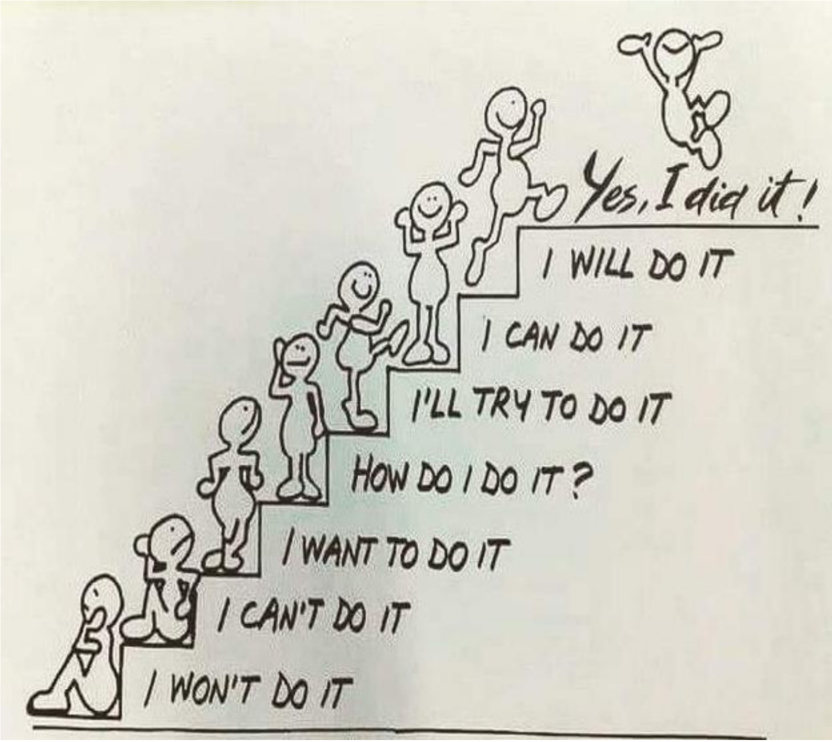
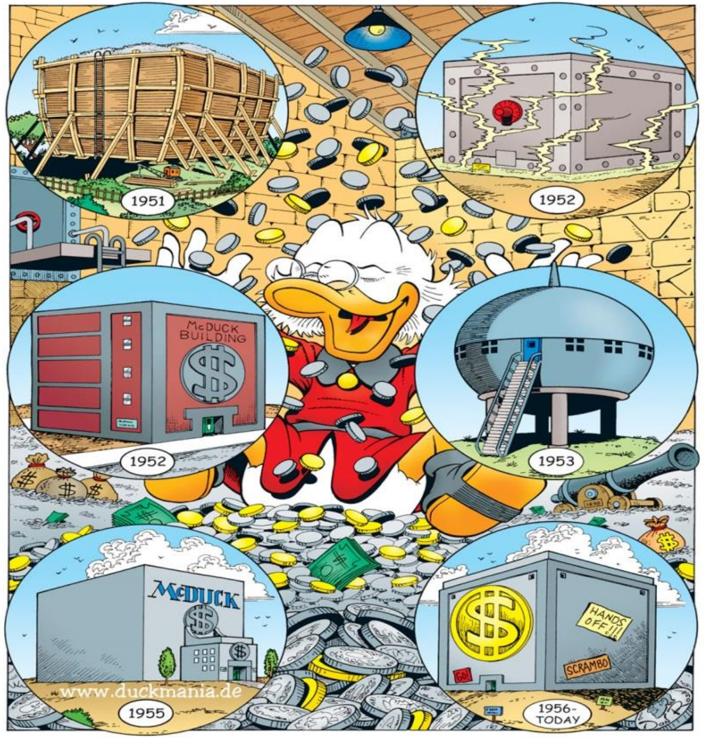


Figure from [...]

Wanting more labels or information !





But…

Insufficient quantity of labeled data

Insufficient subject-matter expertise to label data

- specific relevant expertise required
- become prohibitively expensive
  - example in medical domain

Insufficient time to label and prepare data

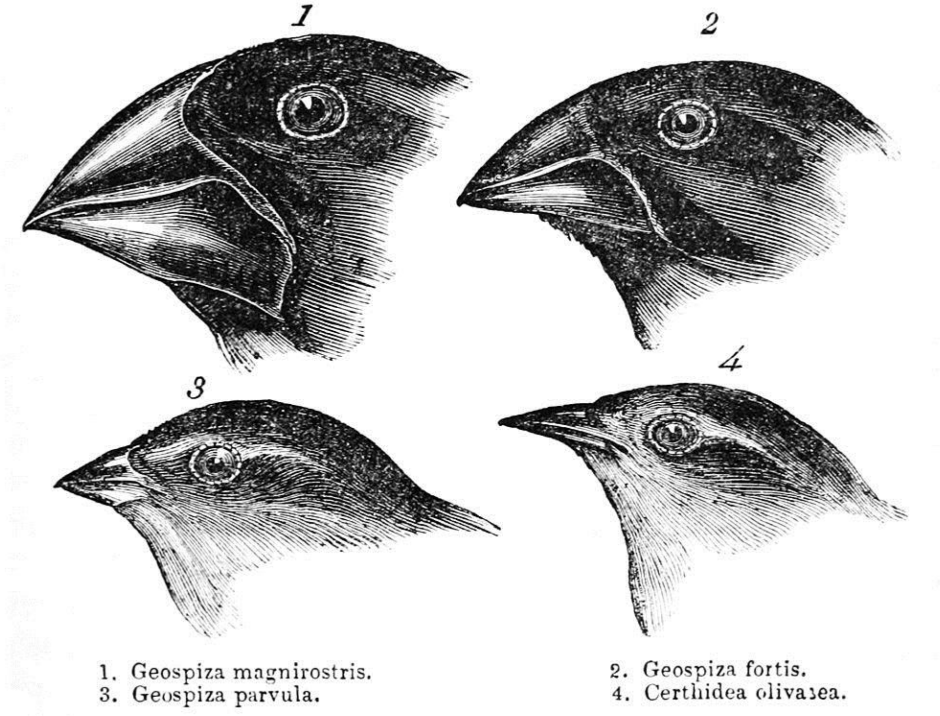
- time spent in preparing data sets
- domain by nature rapidly evolves
  - example in fraud detection or cybersecurity applications.

…

From wikipedia

Weak supervised learning

Taxonomy:
an attempt



Strong supervised learning

Weakly supervised learning

Strong versus Weak
Two aspects : Supervision, Labels

1 - Strong is strong…

many labeled examples with accurate labels

Strong versus Weak
Two aspects : Supervision, Labels

2 - Types of weak 'learning'

- Incomplete supervision:
  - a small amount of labeled data
    - but sometimes abundant unlabeled data are available
  - only labels on a 'positive class'
- Inaccurate supervision:
  - labels are not 'guaranteed' (some label information may suffer from errors)
  - labels are not 'guaranteed' (and are on 'bag of examples' (a set of keys))
- Inexact supervision:
  - labels are on 'bag of examples' (a set of keys)

Strong supervised learning

Weakly supervised learning

Strong supervised learning

Weakly supervised learning

True labels

Inaccurate labels
(label noise, …)

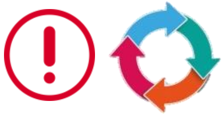
Weakly supervised learning
Inaccurate Labels versus True labels

Zhi-Hua Zhou, 2017
“A Brief Introduction to Weakly Supervised Learning”

B. Frénay and M. Verleysen. “Classification in the presence of label noise: A survey”. IEEE Trans. Neural Networks and Learning Systems, 25(5):845–869, 2014.

- Inaccurate or imprecise labels
  - labels are on ‘bag of examples’
  - labels are not ‘guaranteed’ , noisy labels:
    - learning with label noise
      - use an algorithm robust to the label noise (if noise marginal)
      - try to model the labels and the noise (with assumption on the noise)
      - filter the noisy training set to have a clean one

- True labels but incomplete supervision (incomplete information)
  - Few labels are available
  - Only true labels on one class
  - Labels at or not at the right ‘proxy’



sometimes weak on both sides

Strong supervised learning

3 criterion on labels:
- **Quantity ?**
- **Quality ?**
- **Adapted ?**

Weakly supervised learning

true labels

Innacurate label
(label noise, …)

Strong supervised learning

3 criterion on labels:

- **Quantity ?**
- **Quality ?**
- **Adapted ?**

Weakly supervised learning

true labels

Innacurate label
(label noise, …)

labels at the right
'proxy'

labels not at the right
'proxy'

Proxy ?

Inexact supervision

- concerns about the situation where some supervision information is given, but not
  - as exact as desired or at the right proxy or labels* are on subsets of the data
  - example 1: is there an protest ?
    - detect people, how many people, distance between people, …

(a) (b)

Image from “A Method for Counting People in Crowded Scenes” – AVSS 2010

*but that could be noisy and may conflict

*general: multiple noisy labeling functions can conflict and have dependencies

Proxy ?

Inexact supervision

- concerns about the situation where some supervision information is given, but not
    - as exact as desired or at the right proxy or labels* are on subsets of the data
        - transfer learning
        - multi-instance learning
        - build ‘labels” (Snuba, Snorkel, …)

*but that could be noisy and may conflict
*general: multiple noisy labeling functions can conflict and have dependencies

Strong supervised learning

3 criterion on labels:

- **Quantity ?**
- **Quality ?**
- **Adapted ?**

Weakly supervised learning

true labels

Innacurate label
(label noise, …)

labels at the right
‘proxy’

labels not at the right
‘proxy’

Strong supervised learning

3 criterion on labels:

- **Quantity ?**
- **Quality ?**
- **Adapted ?**

Weakly supervised learning

true labels

Innacurate label
(label noise, …)

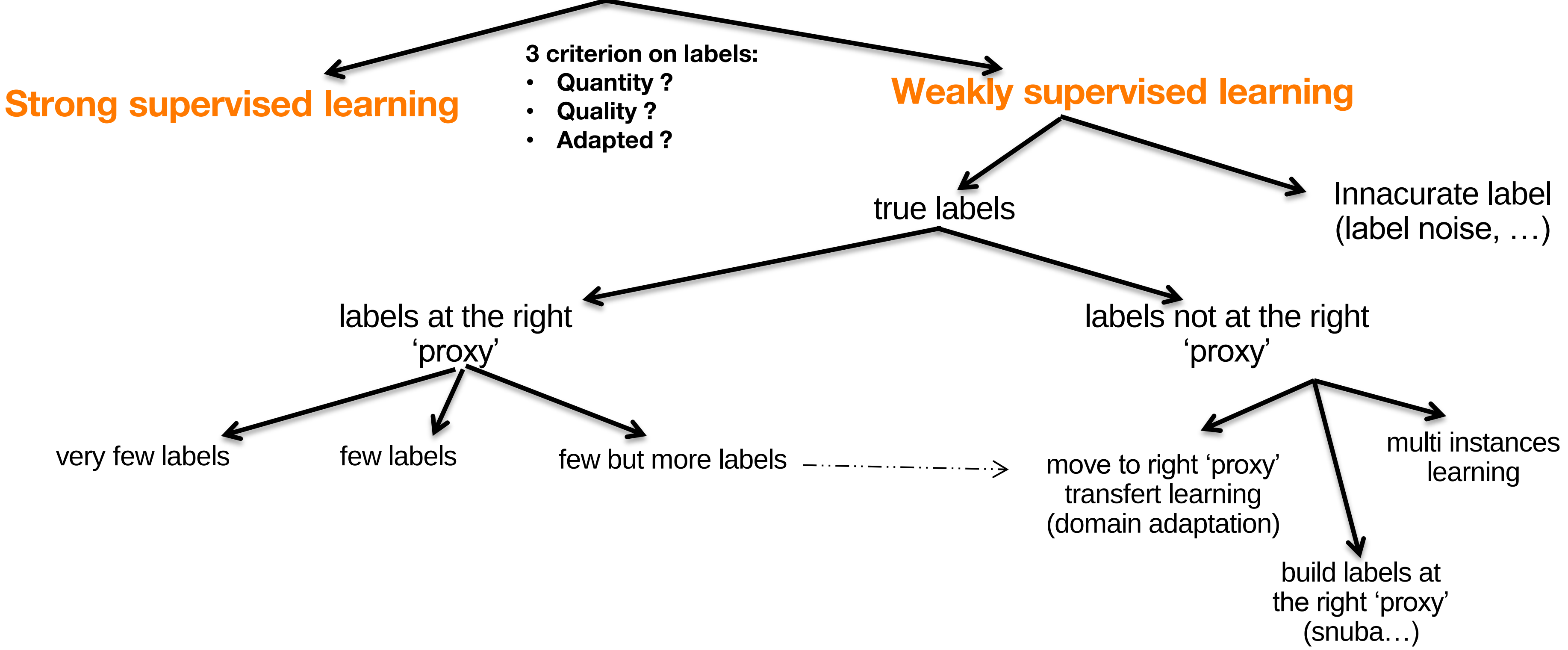
labels at the right
'proxy'

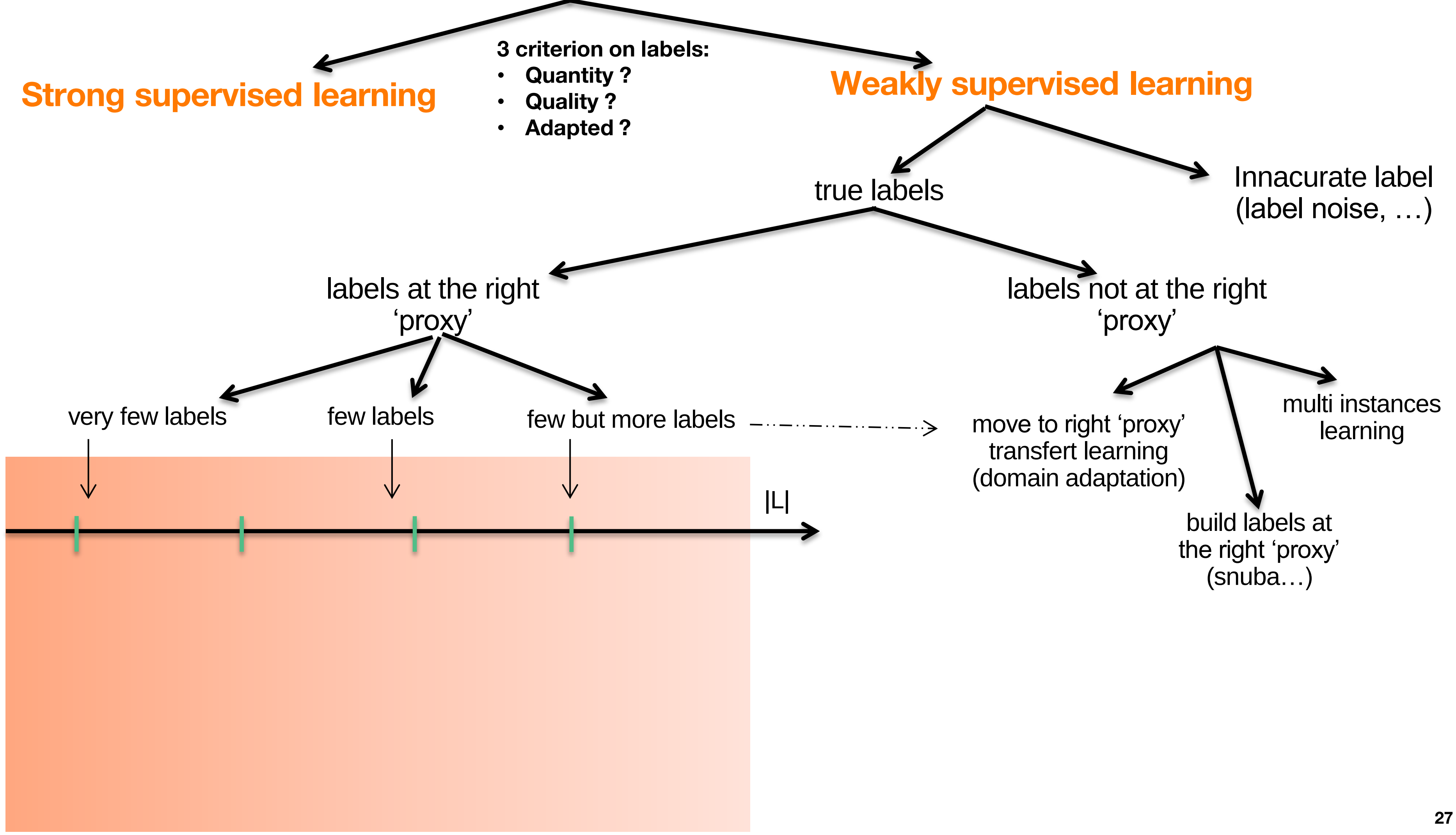
labels not at the right
'proxy'

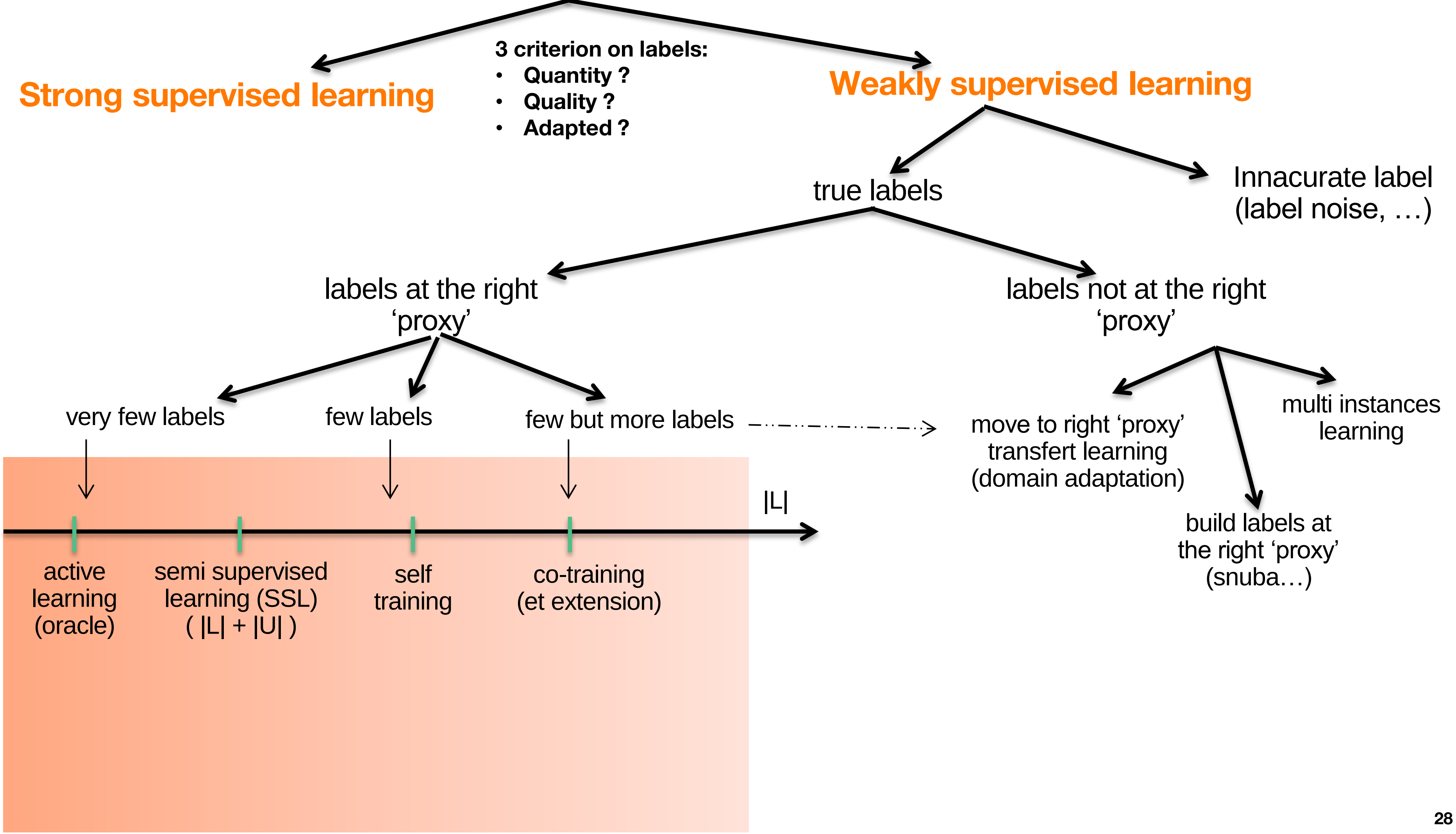
move to right 'proxy'
transfert learning
(domain adaptation)

multi instances
learning

build labels at
the right 'proxy'
(snuba…)


Strong supervised learning
3 criterion on labels:
• Quantity ?
• Quality ?
• Adapted ?
Weakly supervised learning
true labels
Innacurate label
(label noise, …)
labels at the right
'proxy'
labels not at the right
'proxy'
very few labels
few labels
few but more labels
move to right 'proxy'
transfert learning
(domain adaptation)
multi instances
learning
build labels at
the right 'proxy'
(snuba…)


Strong supervised learning
3 criterion on labels:
• Quantity ?
• Quality ?
• Adapted ?
Weakly supervised learning
true labels
Innacurate label
(label noise, …)
labels at the right
'proxy'
labels not at the right
'proxy'
very few labels
few labels
few but more labels
move to right 'proxy'
transfert learning
(domain adaptation)
multi instances
learning
build labels at
the right 'proxy'
(snuba…)
|L|


3 criterion on labels:
• Quantity ?
• Quality ?
• Adapted ?
Strong supervised learning
Weakly supervised learning
true labels
Innacurate label (label noise, …)
labels at the right 'proxy'
labels not at the right 'proxy'
very few labels
few labels
few but more labels
move to right 'proxy' transfert learning (domain adaptation)
multi instances learning
build labels at the right 'proxy' (snuba…)
|L|
active learning (oracle)
semi supervised learning (SSL) (|L| + |U|)
self training
co-training (et extension)

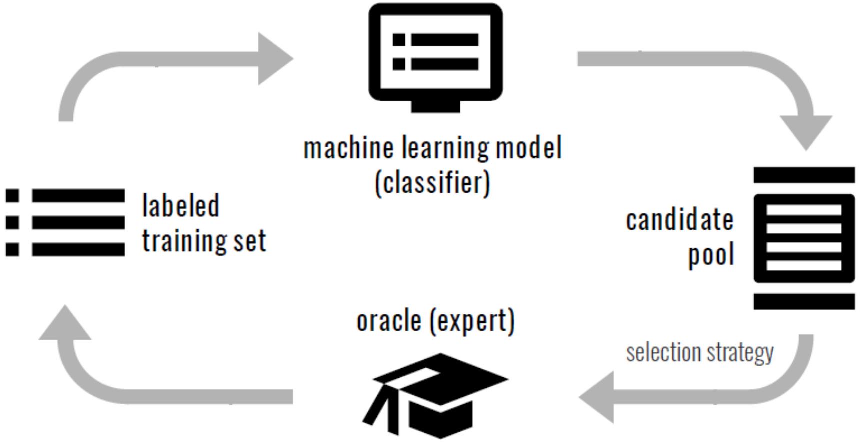
Active Learning (principle)

Definition:

- Active Component: ask queries to an oracle
- Improve the performance of a classier
- Minimizing the cost of obtaining labeled data

Conclusion:

- Active Learning optimizes a performance which is induced by a classifier through selecting the most beneficial unlabeled instances to be labeled by an oracle to build the training basis.



Semi-supervised learning

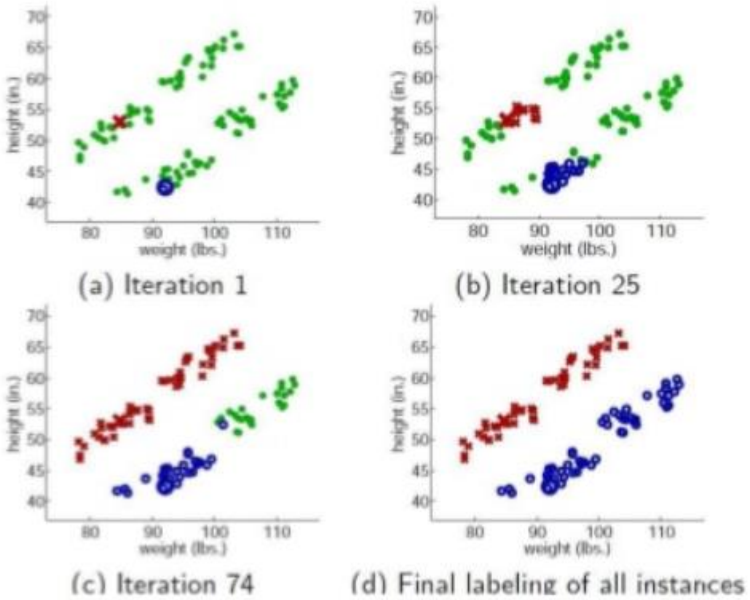
Semi supervised learning attempts to automatically exploit unlabeled data in addition to labeled data to improve learning performance, where no “human” intervention is assumed

- generative models
- low-density separation
- graph-based methods
- heuristic approaches
    - self training
    - co-training
    - …

« Semi-Supervised Learning », Chapelle et al. The MIT Press 2010

Self training

Idea : Train, predict, re-train using classifier's best predictions, repeat

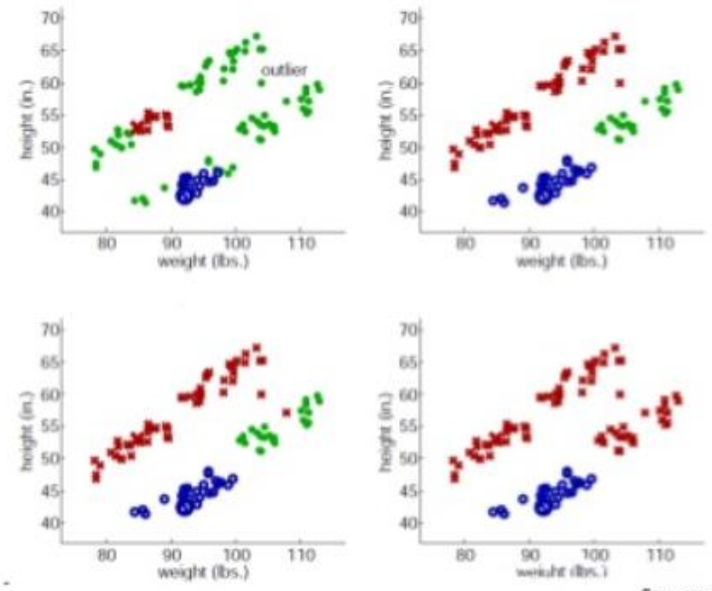


Source: https://tinyurl.com/y98dzxb

1-NN good case

Self training

Idea : Train, predict, re-train using classifier's best predictions, repeat

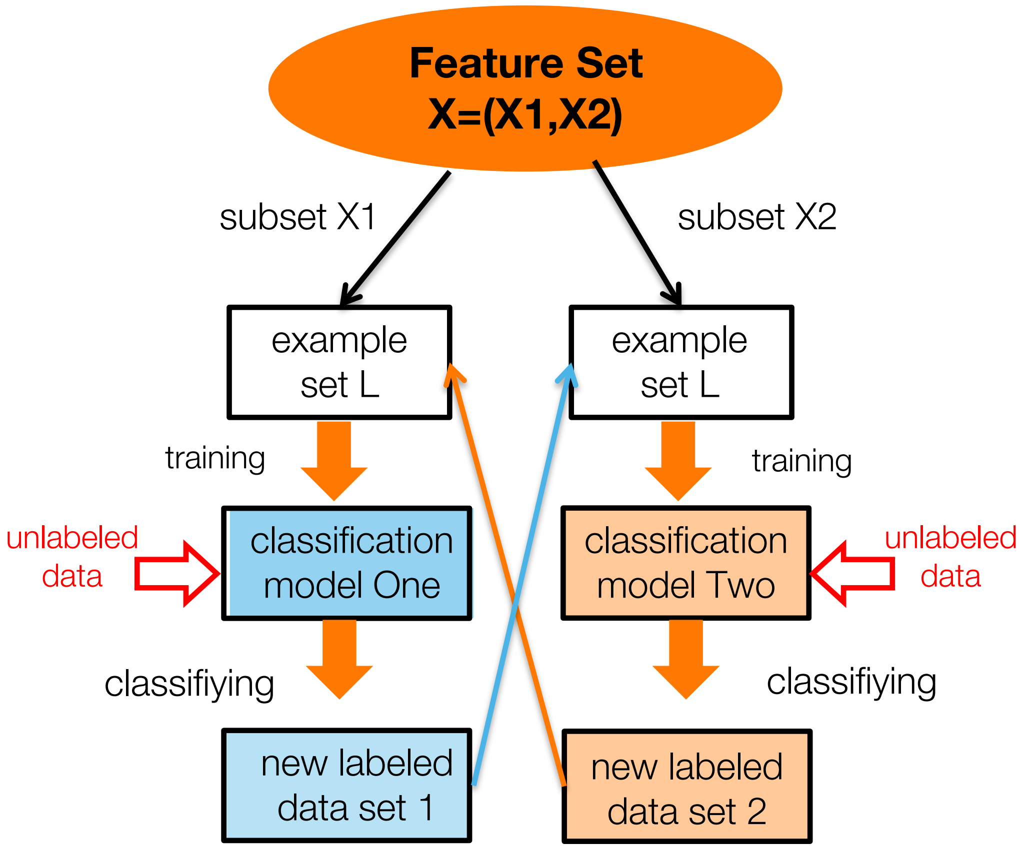


Source: https://tinyurl.com/y98dzxb

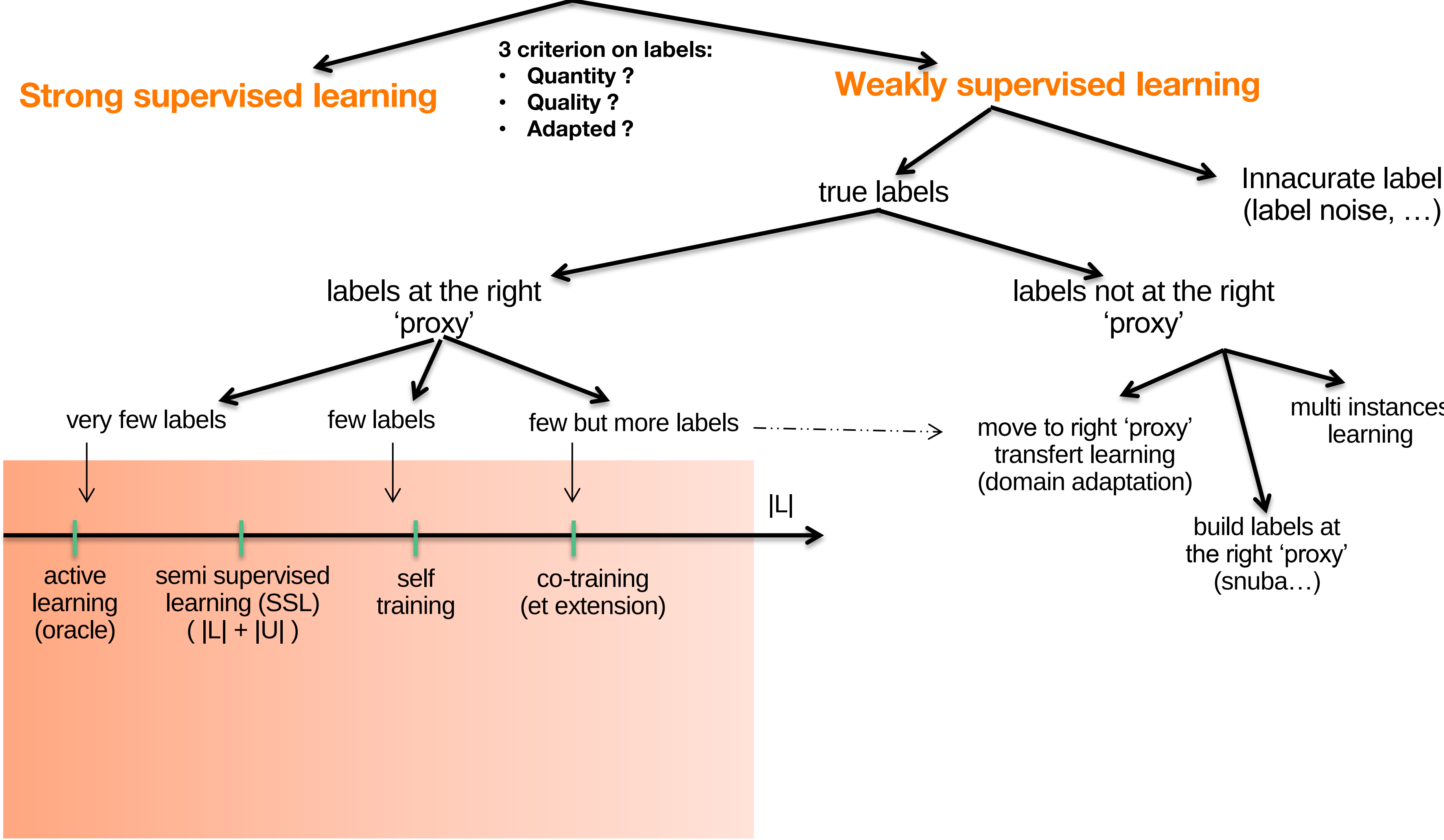
1-NN bad case

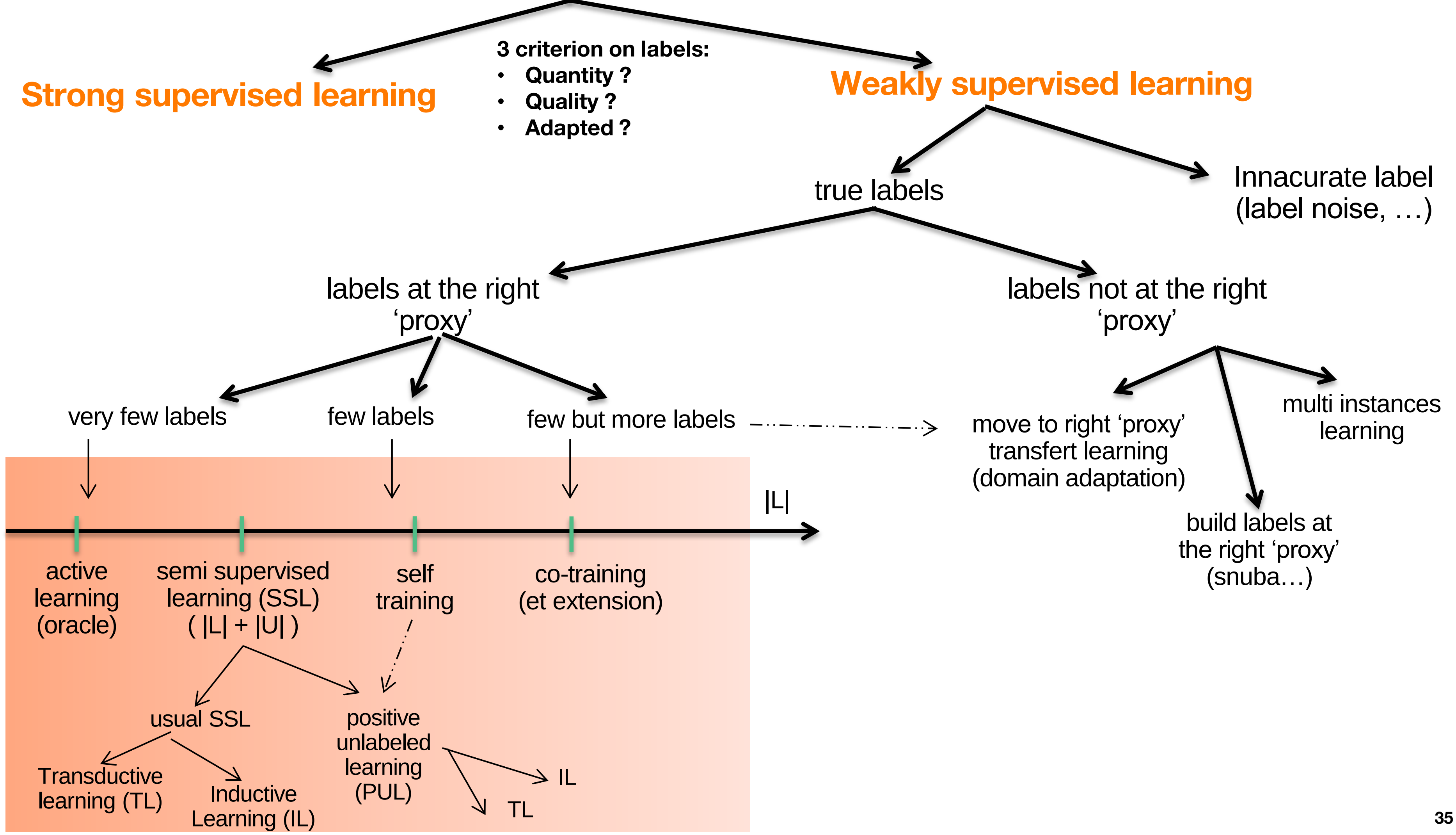
Co-Training

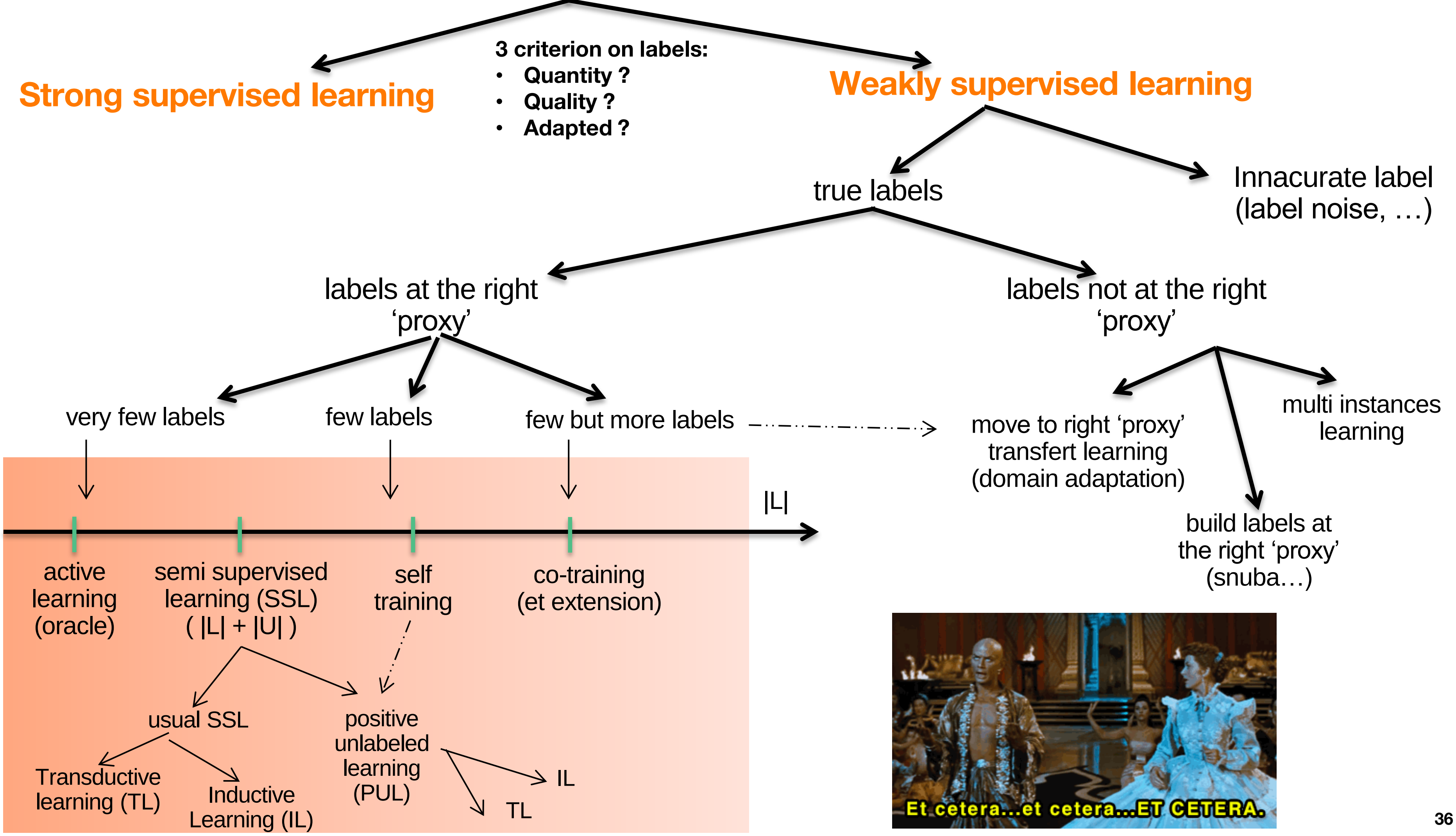
- Each instance has “two- (independent)-views”
- Each view should provide a “good classifier”
- Each view teach the other view (by providing labeled instances)



Blum, A., Mitchell, T. Combining labeled and unlabeled data with co-training. *COLT: Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann, 1998, p. 92-100.


3 criterion on labels:
• Quantity ?
• Quality ?
• Adapted ?
Strong supervised learning
Weakly supervised learning
true labels
Innacurate label (label noise, …)
labels at the right 'proxy'
labels not at the right 'proxy'
very few labels
few labels
few but more labels
move to right 'proxy' transfert learning (domain adaptation)
multi instances learning
build labels at the right 'proxy' (snuba…)
|L|
active learning (oracle)
semi supervised learning (SSL) (|L| + |U|)
self training
co-training (et extension)


3 criterion on labels:
• Quantity ?
• Quality ?
• Adapted ?
Strong supervised learning
Weakly supervised learning
true labels
Innacurate label (label noise, …)
labels at the right 'proxy'
labels not at the right 'proxy'
very few labels
few labels
few but more labels
move to right 'proxy' transfert learning (domain adaptation)
multi instances learning
build labels at the right 'proxy' (snuba…)
|L|
active learning (oracle)
semi supervised learning (SSL) (|L| + |U|)
self training
co-training (et extension)
usual SSL
positive unlabeled learning (PUL)
Transductive learning (TL)
Inductive Learning (IL)
IL
TL


3 criterion on labels:
• Quantity ?
• Quality ?
• Adapted ?
Strong supervised learning
Weakly supervised learning
true labels
Innacurate label (label noise, …)
labels at the right 'proxy'
labels not at the right 'proxy'
very few labels
few labels
few but more labels
move to right 'proxy' transfert learning (domain adaptation)
multi instances learning
build labels at the right 'proxy' (snuba…)
|L|
active learning (oracle)
semi supervised learning (SSL) (|L| + |U|)
self training
co-training (et extension)
usual SSL
positive unlabeled learning (PUL)
Transductive learning (TL)
Inductive Learning (IL)
IL
TL
Et cetera...et cetera...ET CETERA.

Strong supervised learning

Weakly supervised learning

3 criterion on labels:

- **Quantity ?**
- **Quality ?**
- **Adapted ?**

Inaccurate label
(label noise, …)

True labels

labels at the right 'proxy'

labels not at the right 'proxy'

very few labels

few labels

few but more labels

move to right 'proxy'
transfer learning
(domain adaptation)

multi instances learning

build labels at the right 'proxy'
(snuba…)

|L|

active learning (oracle)

semi supervised learning (SSL)
(|L| + |U|)

self training

co-training
(et extension)

usual SSL

positive unlabeled learning (PUL)

Transductive learning (TL)

Inductive Learning (IL)

IL

TL

Ok but more formalism required…

Example

Model Label Noise

Tutorial ACML 2019 : Ivor W Tsang, Bo Han "Towards Noisy Supervision: Problems, Theories, and Algorithms"

(1) Random Classification Noise (RCN):

$$\rho_Y(X) = P(\tilde{Y}|Y,X) = P(\tilde{Y}|Y);\ \rho_{+1}(X) = \rho_{-1}(X) = \rho.$$

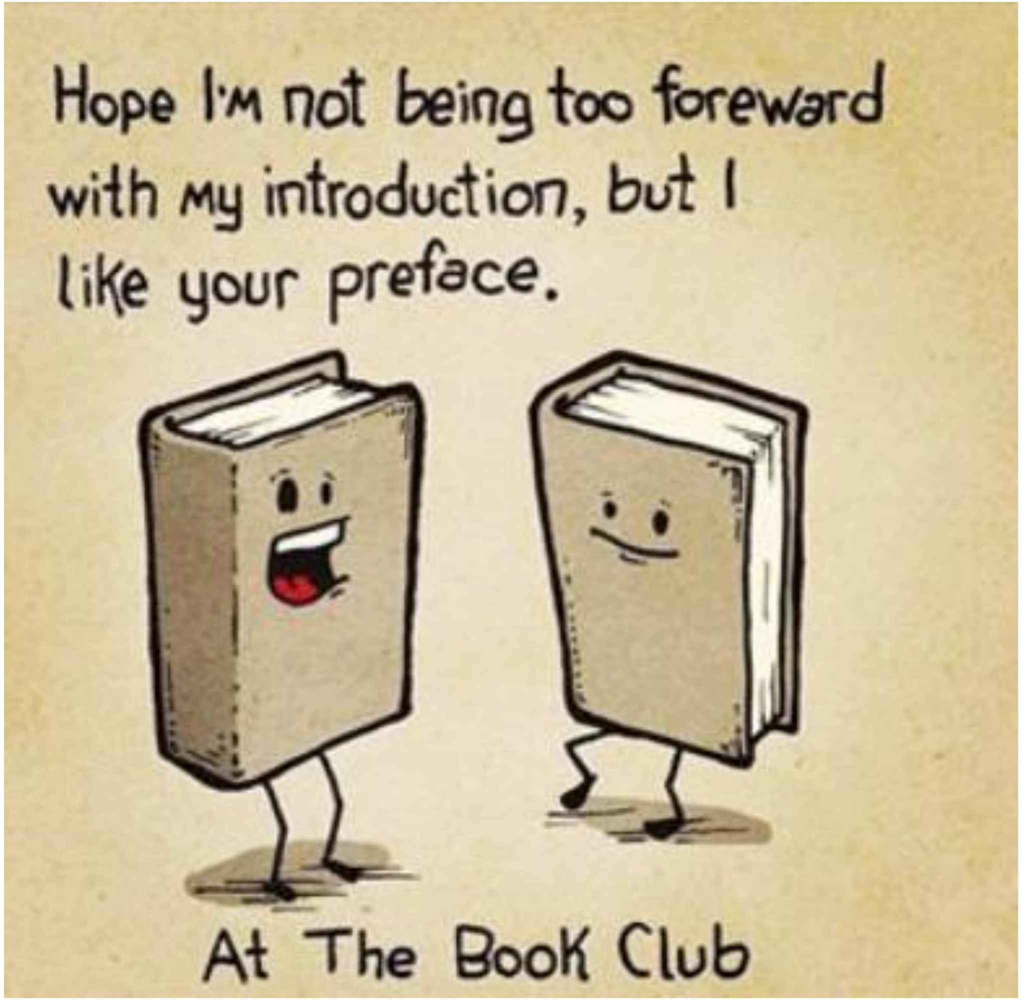
(2) Class-Dependent Noise (CCN):

$$\rho_Y(X) = P(\tilde{Y}|Y,X) = P(\tilde{Y}|Y);\ \rho_{+1}(X) = \rho_{+1}, \rho_{-1}(X) = \rho_{-1}.$$

(3) Instance- and Label-Dependent Noise (ILN):

$$\rho_Y(X) = P(\tilde{Y}|Y,X).$$

Preamble - End



Active learning – outline

- Topic 1: Selection Strategies (or not)
- Topic 2: Evaluation of Pool-based Active Learning
- Topic 3: Software Framework
- Application: Sorting Robot

Topic 1: Selection Strategies (or not)

Active Learning

[1] Charles C. Bonwell and James A. Eison. Active learning: Creating excitement in the classroom. ASHE-ERIC Higher Education Report, 1, 1991.

From Education . . .

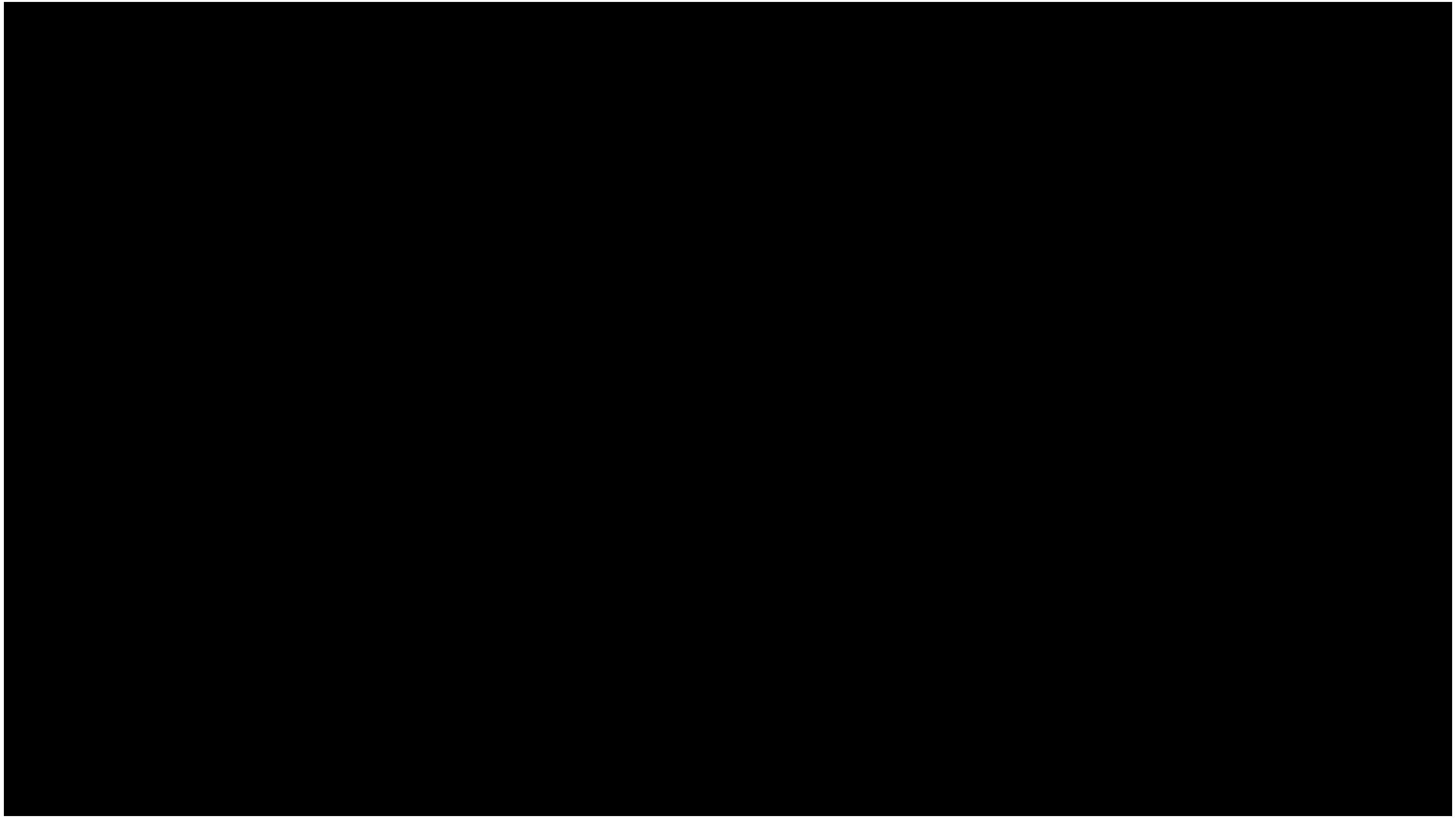
C. Bonwell and J. Eison [1]: In active learning, students participate in the process and students participate when they are doing something besides passively listening. It is a model of instruction or an education action that gives the responsibility of learning to learners themselves.

. . . to Machine Learning:

Settles [2, p.5]: Active learning systems attempt to overcome the labeling bottleneck by asking queries in the form of unlabeled instances to be labeled by an oracle. In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data.

[2] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, Madison, Wisconsin, USA, 2009.

Active Learning From Education . . .. . . to Machine Learning:



Active Learning

Setting

- Some information is costly (some not)
- Active learner controls selection process

Objective

- Select the most valuable information
- Baseline: Random selection

Historical Remarks

- Optimal experimental design
    - Valerii V. Fedorov. “Theory of Optimal Experiments Design”, Academic Press, 1972.
- Learning with queries/query synthesis
    - Dana Angluin. “Queries and concept learning”, Machine Learning, 2:319{342, 1988.
- Selective sampling
    - David Cohn, L. Atlas, R. Ladner, M. El-Sharkawi, R. II Marks, M. Aggoune, and D. Park. “Training connectionist networks with queries and selective sampling”, In Advances in Neural Information Processing Systems (NIPS). Morgan Kaufmann, 1990.

Selective Data Acquisition Tasks

Active Learning Scenarios

- Query synthesis: example generated upon query
- Pool U of unlabeled data: static, repeated access
- Stream: sequential arrival, no repeated access

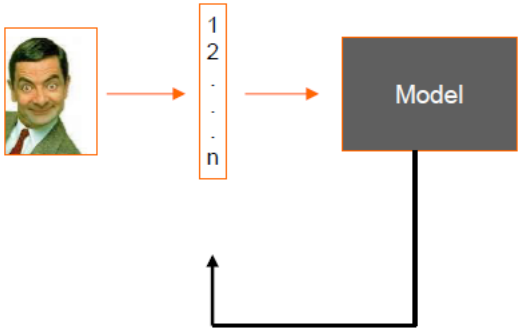
Type of Selected Information

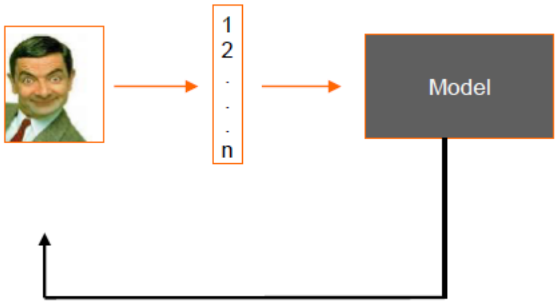
- Active label acquisition
- Active feature (value) acquisition
- Active class selection, also denoted
  Active class-conditional example acquisition
- . . .

Selective Data Acquisition Tasks

A short diverticula

from pool of unlabelled data





Active Sampling
(inductive learning)

main assumption : obtaining an
unlabeled instance is not free

Selective Sampling
(transductif learning)

this talk

main assumption : obtaining an
unlabeled instance is free

Combine?
Link with “Active class selection”?

Definition of Active Learning

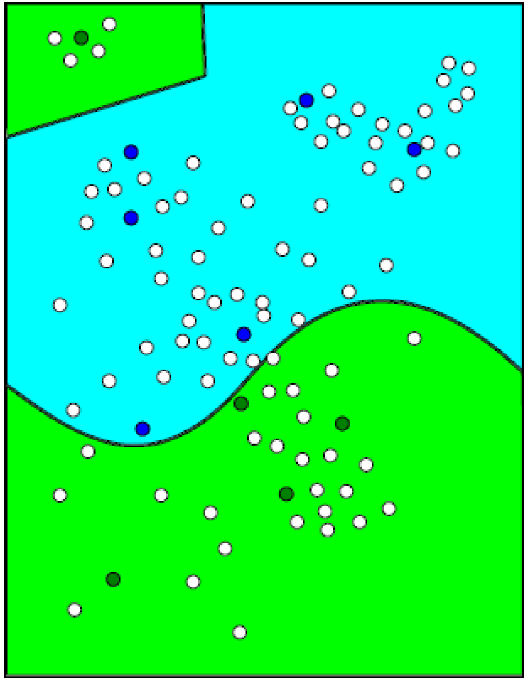
Definition:

- Active Component: ask queries to an oracle
- Improve the performance of a classifier
- Minimizing the cost of obtaining labeled data

Conclusion:

Active Learning optimizes a **performance** which is induced by a **classifier** through selecting the most beneficial **unlabeled instances** to be labeled by an **oracle** to build the **training basis**.

Visualization

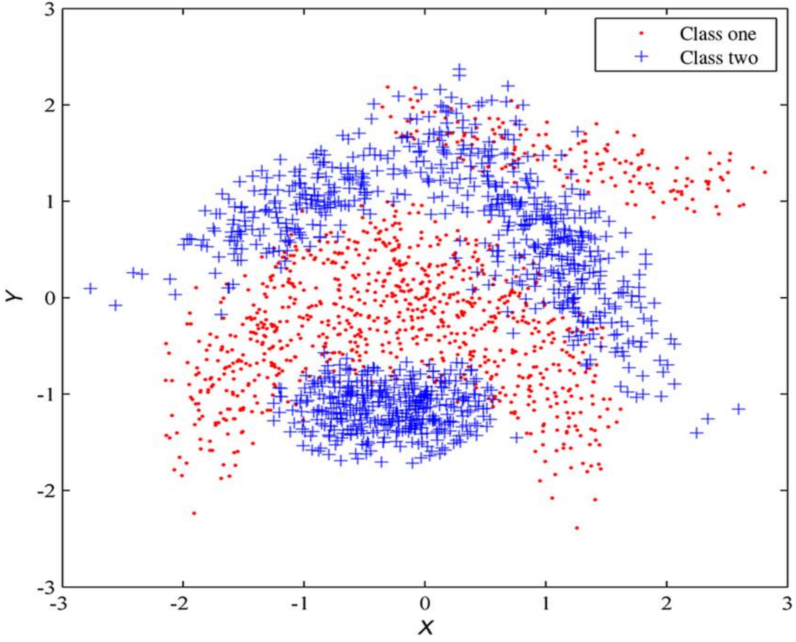


What factors influence the decision?

- Density (improve the classifier, where decisions are important)
- Decision boundary (be specific, where change is expected)
- Label density (explore unexplored regions)

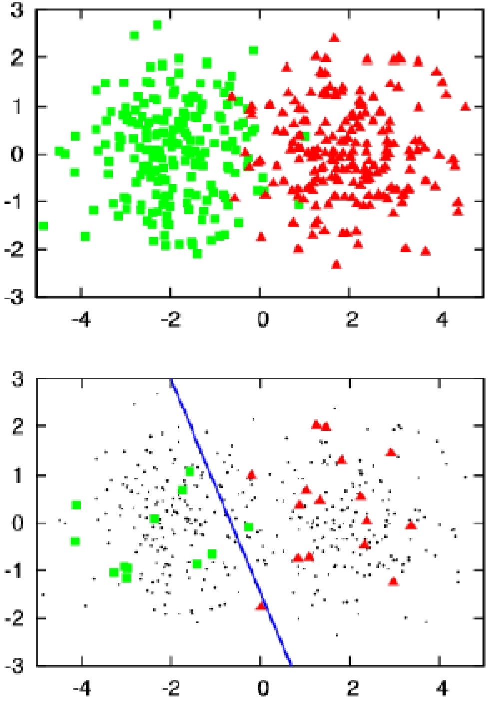
Random sampling

- Also called passive sampling
- Selects instances randomly for labeling
- Competitive approach
- Standard baseline
- Free of heuristics
- Performs very well on the 'banana dataset'



Uncertainty sampling

“Training connectionist networks with queries and selective sampling”.
David Cohn, L. Atlas, R. Ladner, M. El-Sharkawi, R. Il Marks, M. Aggoune, and D. Park.
In Advances in Neural Information Processing Systems (NIPS). Morgan Kaufmann, 1990.



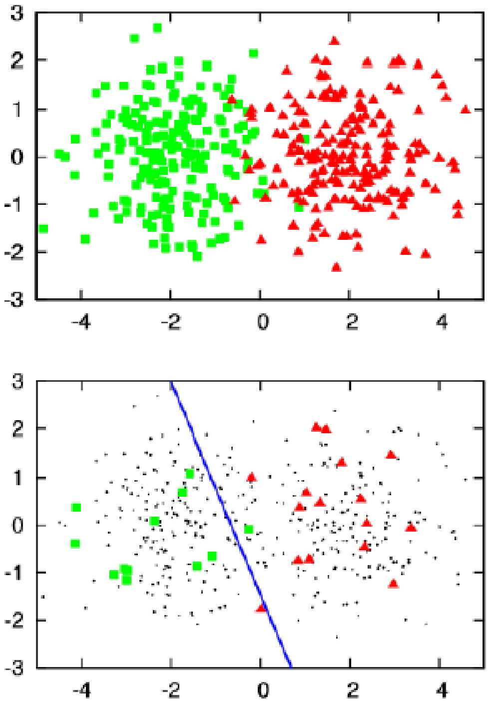
Idea

- Select those instances where we are least certain about the label

Approach

- 3 labels preselected
- Linear classifier
- Use distance to the decision boundary as uncertainty measure

Uncertainty sampling



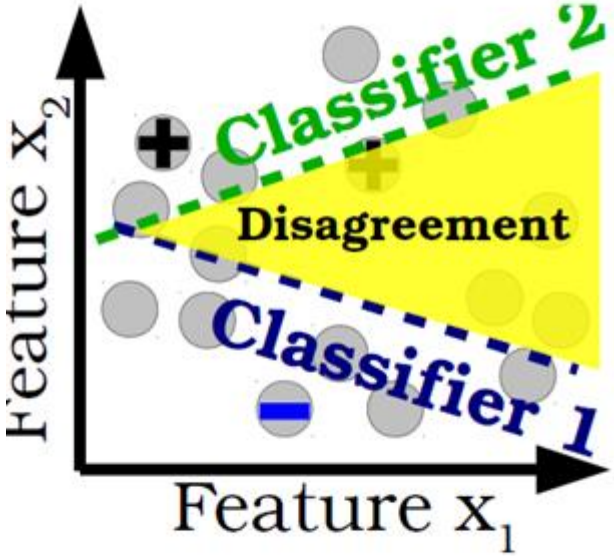
✚ easy to implement
✚ fast

-- no exploration (often combined with random sampling)
-- impact not considered (density weighted extensions exist)
-- problem with complex structures (performance can be even worse than random)

Pure exploitation, does not explore
Can get stuck in regions with high Bayesian error

Ensemble-Based Strategy

"Query by committee", H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Fifth workshop on computational learning theory. Morgan Kaufmann, 1992.



Idea

Use disagreement between base classifiers

Approach

1. Get an initial set of labels
2. Split that set into (overlapping) subsets
3. On each subset, train a different base-classifier
4. Repeat until stop
5. On each unlabeled instance do
6. Apply all base-classifiers
7. Request label, if base-classifiers disagree
8. Update all base-classifiers
9. Go to step 4

Expected Error Reduction

- Simulates the acquisition of each label candidate and each possible outcome (class)

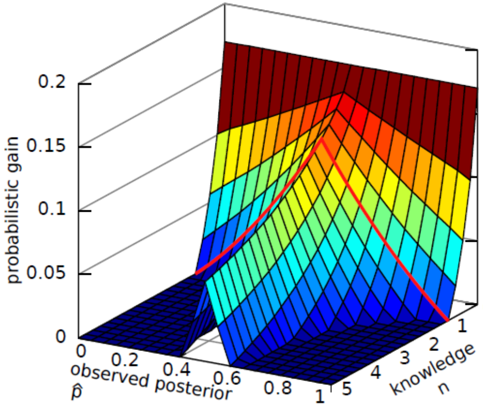
- Calculates the generalization error of the simulated new model
- Chooses the label with lowest generalization error

$$x^* = \operatorname{argmin}_x \sum_{i \in \{1,\ldots,C\}} P_\theta(y_i \mid x) \left(\sum_{x' \in \mathcal{U}} 1 - P_{\theta+(x,y_i)}(\hat{y} \mid x') \right)$$

+ decision theoretic model
- long execution time (closed form solutions for specific classifiers, approximations for speed up)

Probabilistic Active Learning

"Optimized probabilistic active learning (OPAL) for fast, non-myopic, cost-sensitive active classification", Georg Krempl, Daniel Kottke, and Vincent Lemaire. In Machine Learning, 100(2), 2015.



- Models the true posterior as being Beta-distributed
  - variance of posterior is correlated with the number of local observations
  - thereby omit the complex simulation of expected error reduction
- Calculates the performance improvement of the model

\+ decision theoretic model
\+ fast w.r.t. expected error reduction

\- local number of labels required

DUAL

“Dual strategy active learning.”, Pinar Donmez, JaimeG. Carbonell, and Paul N. Bennett, In Machine Learning: ECML 2007

- combination of density weighted uncertainty sampling and standard (uniform) uncertainty sampling
- adaptive weights

4DS

- Uses four different scores for a classifier based on Gaussian mixtures (CMM):
  - distance, density, diversity, distribution
  - automatically weighted

“Let us know your decision: Pool-based active training of a generative classifier with the selection strategy 4DS”, Tobias Reitmaier and Bernhard Sick, in Information Sciences, 2013

One-by-one vs. Batch Acquisition

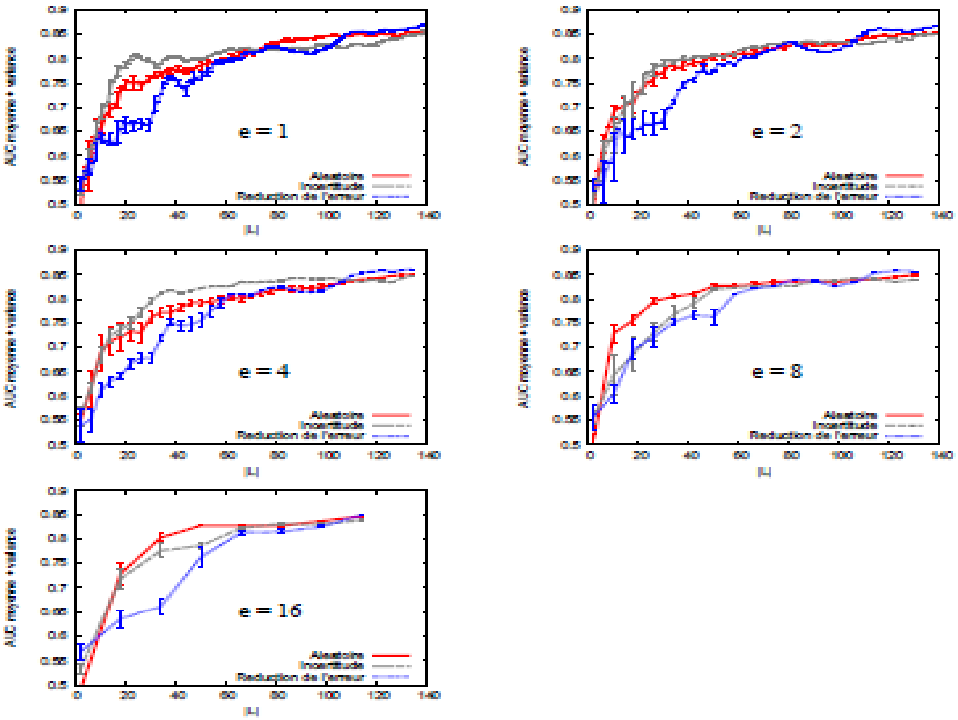


FIG. D.g – Résultats complets pour le jeu de données "Glass"

- Definition:
  - One-by-one: subsequently selecting instances
  - Batch: selects a specific number of labeling candidates for labeling at one time
- Batch-Acquisition:
  - Problem: most approaches would select very similar instances
  - Approach: diversity score

Strategy vs. Classifiers ?

Stratégie	RF	SVM	5NN	GNB	C4.5	RL	VFDT
Margin	[A]						
Entropy						[B]	
SGmulti				[C]			[C]
ATU			[A, C]		[C]		
OER		[A]		[A]	[A]		

[A] D. Pereira-Santos et al., «Empirical investigation of active learning strategies», Neurocomputing, 2019
[B] Y. Yang et M. Loog, «A benchmark and comparison of active learning for logistic regression», Pattern Recognition, 2018
[C] D. Pereira-Santos et al., «Comparison of active learning strategies and proposal of a multiclass hypothesis space search», in Proceedings of HAIS2014, Springer, 2014

A new age ?

Combining all of these (heuristics strategies) ?

“Active Learning by Learning”, Hsu et al. in AAAI 2015.

Active Learning By Learning (ALBL) algorithm is a meta active learn algorithm designed to solve this problem. ALBL considers multiple existing active learning algorithms and adaptively learns a querying strategy based on the performance of these algorithms.

by DFID - UK Department for International Development; licensed under CC BY-SA 2.0 via Wikimedia Commons

Strategy 1: ask most confused question

Strategy 2: ask most frequent question

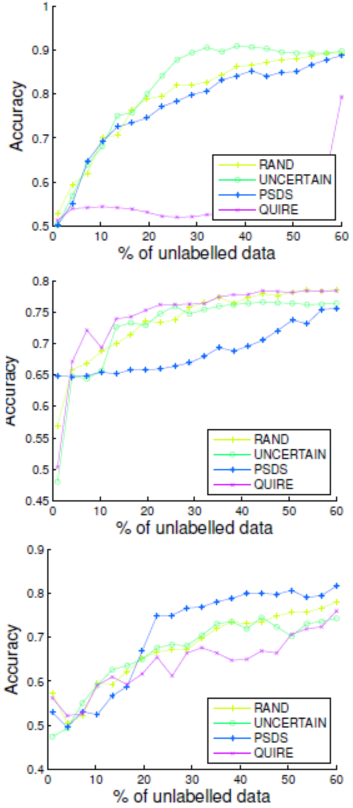
Strategy 3: ask most helpful question

Do you use a fixed strategy in practice?

Rather fixing a strategy learning a strategy ?

Combining all of these (heuristics strategies) ?

“Active Learning by Learning”, Hsu et al. in AAAI 2015.



Strategy 1: ask most confused question
→ uncertainty

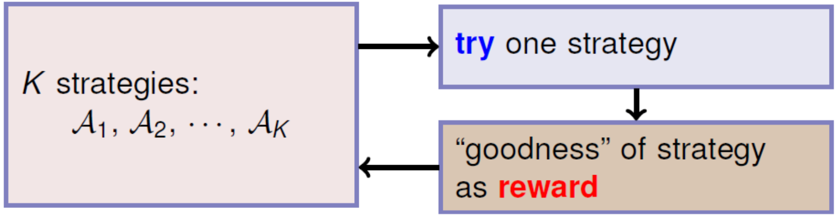
Strategy 2: ask most frequent question
→ representative

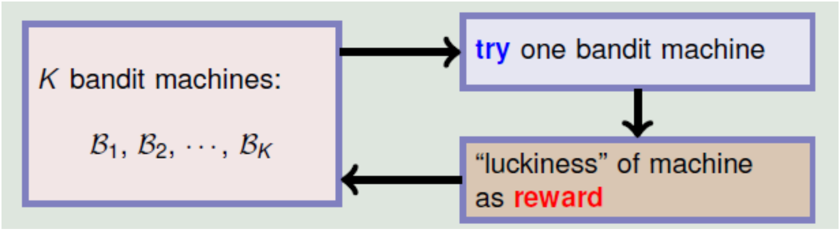
Strategy 3: ask most helpful question
→ exp.-err. Reduction

Choosing one single strategy is non-trivial

Combining all of these (heuristics strategies) ?

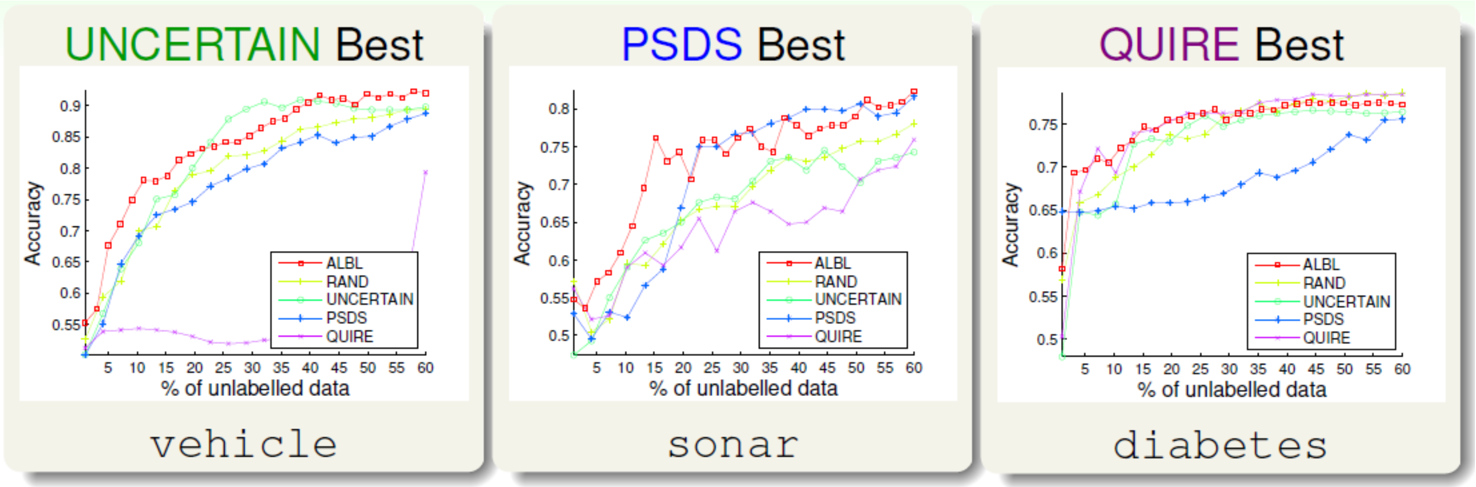
"Active Learning by Learning", Hsu et al. in AAAI 2015.





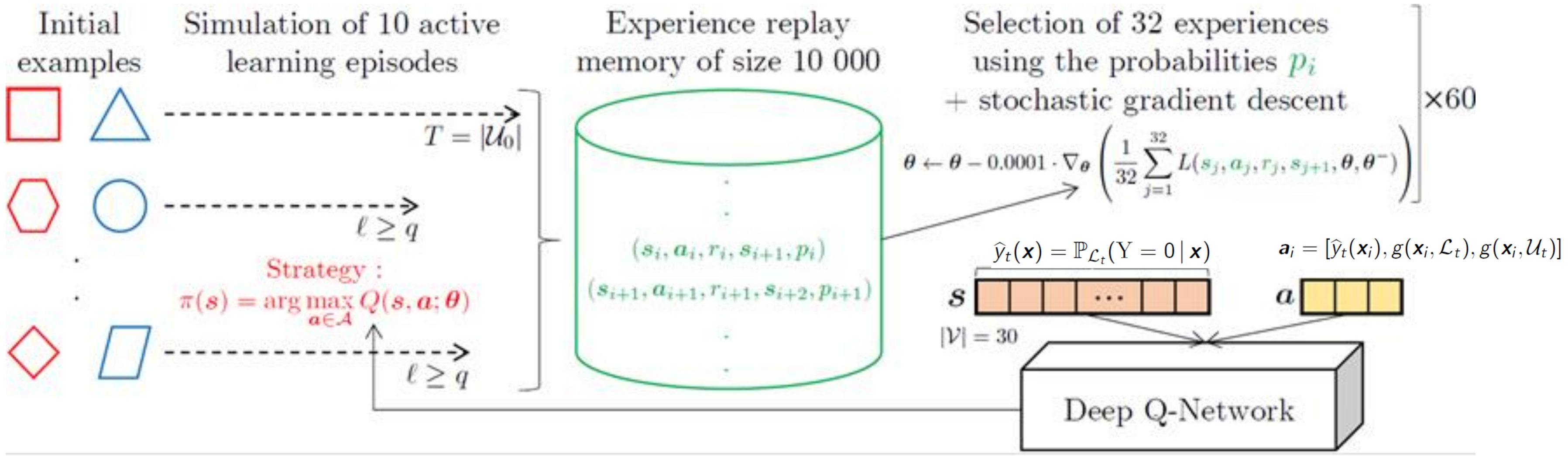
one well-known probabilistic bandit learner (EXP4.P)

two issues: **try** and **reward**



Rather fixing a strategy learning a strategy ?

Discovering General-Purpose Active, Learning Strategies Konyushkova, K., Sznitman, R., Fua, P, in arXiv:1810.04114 (2019)



“Learning active learning: an evaluation”, L. Desreumaux, V. Lemaire submited to Intelligent Data Analysis (IDA) 2020

Where are we ?

Which method used (or recommend) in an industrial context ?

orange™

Active Learning

A: Strategies

So many (heuristics) strategies suggested in the literature:

- random
- uncertainty
- error reduction
- density based
- …

B: Learn how to combine strategies

"Active Learning by Learning", Hsu et al. in AAAI 2015.

Active Learning By Learning (ALBL) algorithm is a meta active learn algorithm designed to solve this problem. ALBL considers multiple existing active learning algorithms and adaptively learns a querying strategy based on the performance of these algorithms.

Claim : A ≤ B

C: Learn (and transfer) a strategy

Discovering General-Purpose Active, Learning Strategies Konyushkova, K., Sznitman, R., Fua, P, in arXiv:1810.04114 (2019)

Claim : A ≤ B ≤ C

D. Pereira-Santos et al., «Empirical investigation of active learning strategies», Neurocomputing, 2019

Best(A): RF+Margin

What did we compare ?

- **RF + Random**

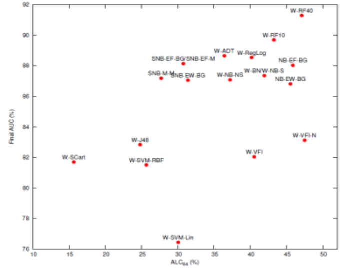
  Salperwyck, C. et V. Lemaire (2011). Learning with few examples : An empirical study on leading classifiers. In International Joint Conference on Neural Networks, pp. 1010–1019.

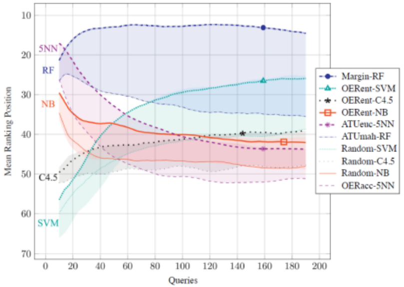
- **RF + Margin**

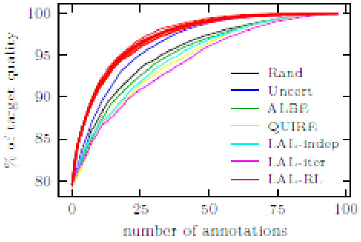
  Pereira-Santos, D., R. B. C. Prudêncio, et A. C. de Carvalho (2019). Empirical investigation of active learning strategies. Neurocomputing 326–327, 15–27.

- **Deep QL**

  Konyushkova, K., R. Sznitman, et P. Fua (2019). Discovering General-Purpose Active Learning Strategies. arXiv:1810.04114







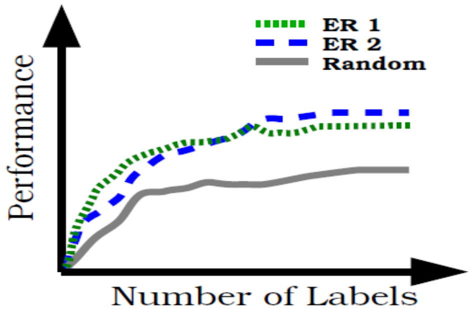
Results

Jeu de données	Rnd/LR	Margin/LR	LAL/LR	Rnd/RF	Margin/RF	LAL/RF	maj (%)
adult	77.93	78.91	78.97	80.17	**81.27**	**81.21**	76.06
banana	53.03	**57.39**	53.12	**80.24**	73.81	73.58	55.16
bank-marketing-full	86.85	**87.62**	**87.72**	88.19	88.34	88.49	88.30
climate-simulation-craches	87.22	**89.13**	**88.62**	91.15	91.14	91.13	91.48
eeg-eye-state	56.08	55.32	56.11	65.53	**67.58**	64.42	55.12
hiva	64.43	**70.84**	**71.80**	96.32	**96.47**	**96.44**	96.50
ibn-sina	84.77	**88.58**	**88.90**	90.53	**93.41**	**92.75**	76.18
magic	76.49	**77.93**	**77.64**	78.05	**80.79**	**79.68**	65.23
musk	83.73	82.34	81.95	89.55	**96.18**	95.35	84.55
nomao	89.45	**91.43**	**91.37**	89.41	**92.32**	**92.07**	69.40
orange-fraud	76.70	**81.74**	74.26	89.15	**90.66**	**90.48**	63.75
ozone-onehr	92.90	94.26	**95.06**	96.61	96.83	96.89	97.11
qsar-biodegradation	80.98	**82.62**	**83.53**	80.34	**82.76**	**82.40**	66.35
seismic-bumps	90.87	**92.59**	**92.14**	92.48	**92.92**	**93.02**	93.41
skin-segmentation	77.05	**82.69**	**83.21**	91.51	**95.70**	**95.77**	71.51
statlog-german-credit	70.76	**72.12**	**72.34**	72.25	72.93	72.78	70.00
thoracic-surgery	83.76	83.93	82.72	83.51	**84.41**	**84.18**	85.11
thyroid-hypothyroid	97.21	**97.99**	**97.97**	97.75	**98.77**	**98.71**	95.43
wilt	93.53	**95.18**	92.87	94.86	**97.23**	**97.02**	94.67
zebra	86.40	90.31	**91.36**	94.71	**95.54**	95.25	95.42
Moyenne	80.51	**82.65**	82.08	87.12	**88.45**	88.08	79.53
win/tie/loss	0/5/15	**3/15/2**	2/15/3	1/4/15	**3/16/1**	0/16/4	

- the choice of model is decisive
- using margin sampling with this model allows a significant performance improvement.
- LAL: a good active learning strategy has been learned
- but the learned strategy is no
- better than margin sampling
- and not always better than random
- hard to beat the majority vote in case of very imbalanced problems

Evaluation

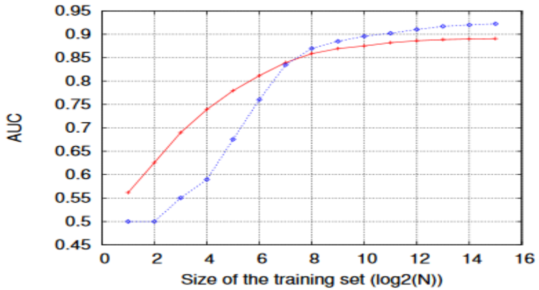
Learning Curves



- Plot performance against the number of requested labels
- Expected behavior:
  - Performance increases with number of labels
  - Convergence: after ∞ label requests,
    all strategies should have the same performance

Evaluation

Degrees of Freedom



Performance measure (Application dependent):

- Accuracy
- Misclassification loss
- F1-score
- Area under the ROC curve

Evaluation points (Application dependent):

- Final performance
- Mean performance (area under the learning curve)
- Each step
- Learning stages
- Pair-wise comparisons (win percentage)

How to interpret the results of a learning curve?

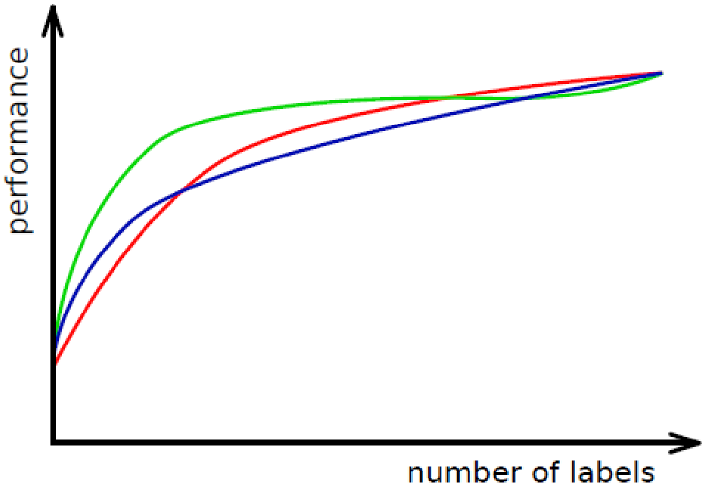
- converging as fast as possible
- converging to the highest overall value

How to summarize results from a learning curve?

- Table at specific time points (early, mid, late)
- Area under the learning curve, mean (depends on stopping point)
- deficiency
- data utilization rate
- comparison of score differences

Area Under the Learning Curve (AULC)

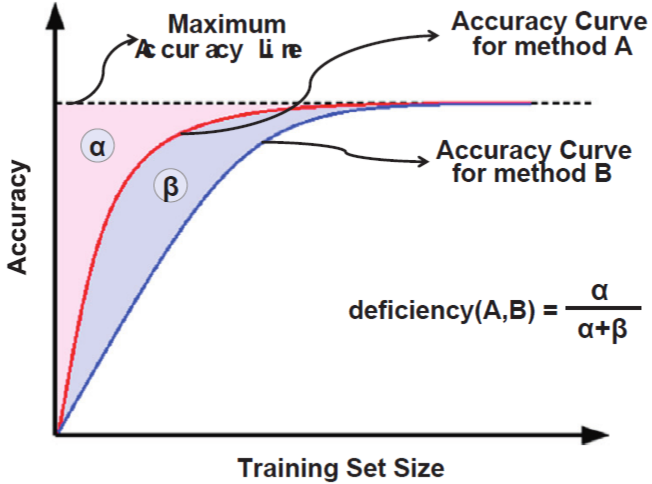
“Active learning to maximize area under the roc curve”, Matt Culver, Deng Kun, and Stephen Scott, in Sixth International Conference on Data Mining (ICDM'06)



- AULC above that of a random-sampling learner
- Calculated for maximum budget, thus sensitive to budget
- Negative value indicates worse-than-random performance
- Note: all strategies should pass through the same |L| values

Deficiency

“Active learning for sketch recognition”, Erelcan Yanik and Tevk Metin Sezgin, in Computers and Graphics, 2015.

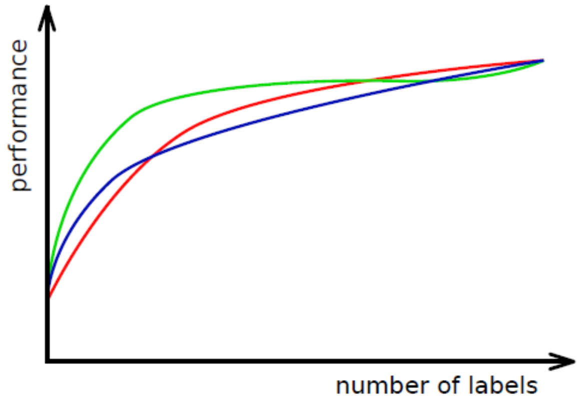


Deficiency as ratio of area between accuracy of a method and maximum accuracy line.

Illustration from “Online choice of active learning algorithms”, Yoram Baram, Ran El Yaniv, and Kobi Luz, in Journal of Machine Learning Research, 2004.

Data Utilization Rate (DUR)

“Active learning to maximize area under the roc curve”, Matt Culver, Deng Kun, and Stephen Scott, In Sixth International Conference on Data Mining (ICDM'06),



- The minimum number of samples needed to reach a target accuracy, divided by the number of samples needed by a random sampling learner
- Indication of efficiency for selecting of data
- Sensitive to choice of target accuracy,
  ignores performance changes at other points

Evaluation

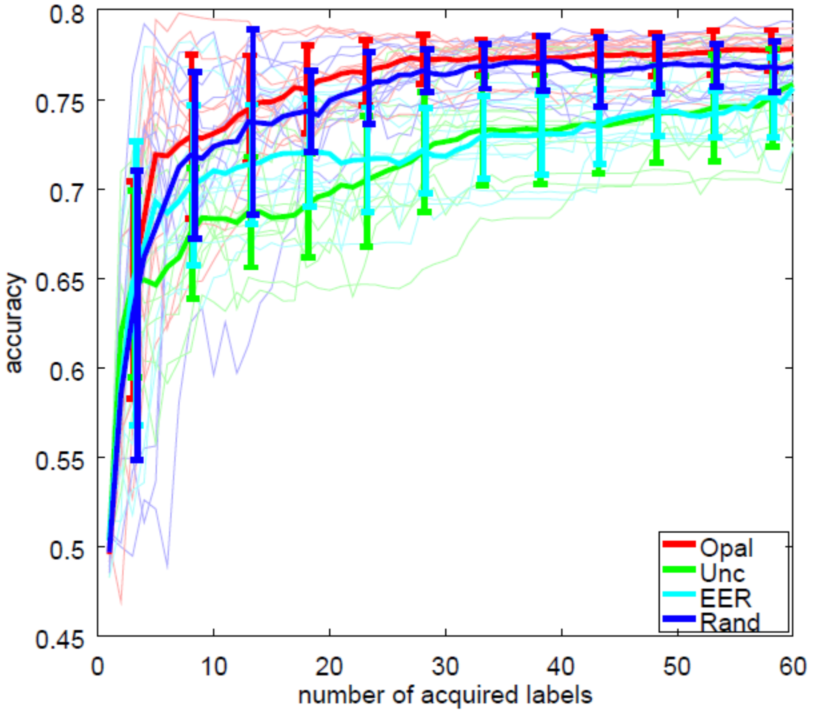
The evaluation methodology should be

1. reliable
   - robust to varying seeds or shuffling data
   - reproducible (well-described, availability of data)
2. realistic
   - valid assumptions for real applications
3. comparable
   - development of a standardized active learning evaluation gold standard to compare algorithms without reimplementing

“Challenges of reliable, realistic and comparable active learning evaluation”,
Daniel Kottke, Denis Huseljic, Adrian Calma, Georg Krempl, and Bernhard Sick,
in Proc. of the Workshop and Tutorial on Interactive Adaptive Learning, 2017.

How many repetitions are required?

Comparison of algorithms using 5-fold cross validation



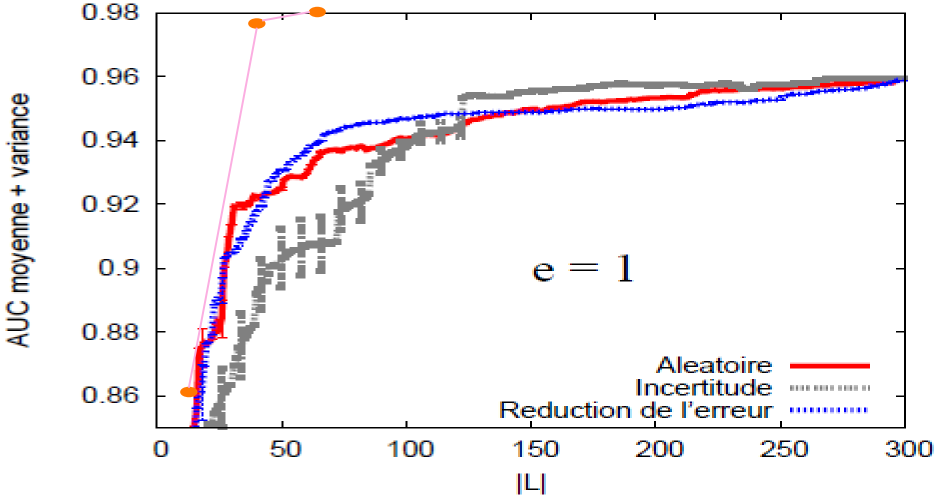
- Which values to compare?
  - not across label acquisitions (highly correlated) but across multiple repetitions
  - at which point in time?
- Statistical tests
  - t-Test cmp. mean (assumes that mean is normal distributed)
  - Wilcoxon Signed Rank Test cmp. tendency (parameter-free test)
- always present results with statistical significance and effect size

Parameters

- tuning instances should be considered in the number of acquisitions
- how many instances should be used for tuning? (many classifiers are sensitive to the number of instances)
- normally, no instances for supervised parameter tuning available
- tuning parallel to sampling may be complicated
- no test set !

Learning ‘speed’

"Learning with few examples: an empirical study on leading classifiers“, Christophe Salperwyck and Vincent Lemaire, in International Joint Conference on Neural Networks (IJCNN), 2011)



Real applications often are more challenging

- Often highly specialized (hard to transfer approaches to related domains)
- Imperfect labelers (experts might be wrong)
- In real-world only one shot (mean results are not representative)
- Labels are not always available (in time and space)
- Performance guarantees (cmp. random sampling)
- Assess online performance of an actively trained classifier
- Different costs for different annotations or classes
- Impossible to tune the ‘user’ parameters of the classifier
- Ground truth might not be available (no test set…)

Laurent Candillier and Vincent Lemaire. Design and Analysis of the Nomao Challenge - Active Learning in the Real-World. In: Proceedings of the ALRA : Active Learning in Real-world Applications, Workshop ECML-PKDD 2012, Friday, September 28, 2012, Bristol, UK.

Thanks

- Daniel Kottke (University of Kassel, Germany) – slides AL
- Georg Krempl (University of Utrecht, Netherlands) –slides AL
- Alexis Bondu (Orange Labs, France) – beta test of this talk
- Antoine Cornuéjols (AgroParisTech, France) – beta test of this talk and discussion to create the “weakly supervised taxonomy”
- Pierre Nodet (Orange Labs, France) – beta test of this talk and discussion to create the “weakly supervised taxonomy”
- Bruno Kauffman (Orange Labs, France) – beta test of this talk