Promises and Pitfalls of Threshold-based Auto-labeling

Harit Vishwakarma

hvishwakarma@cs.wisc.edu

Ph.D. Student

Dept. of Computer Sciences
University of Wisconsin-Madison

Appearing in NEURAL INFORMATION PROCESSING SYSTEMS 2023

Spotlight

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Heguang Lin
hglin@seas.upenn.edu

Frederic Sala
fredsala@cs.wisc.edu

Ramya Korlakai Vinayak
ramya@ece.wisc.edu

Roadmap

What & Why auto-labeling?

Data labeling problem

Wide adoption of auto-labeling

How does it work?

Workflow of TBAL

Finding the auto-labeling region

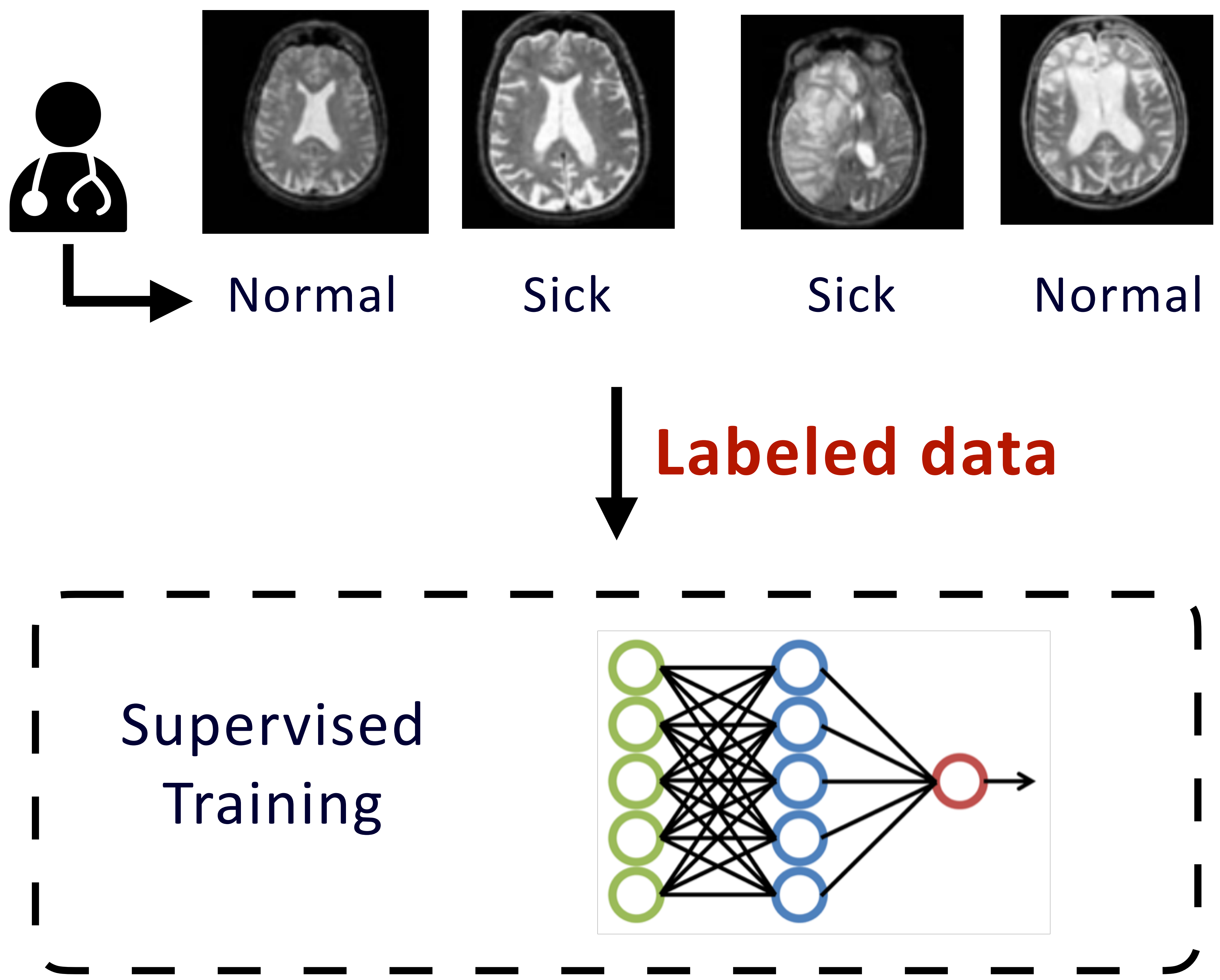
Analysis & Results

Conditions when TBAL works.

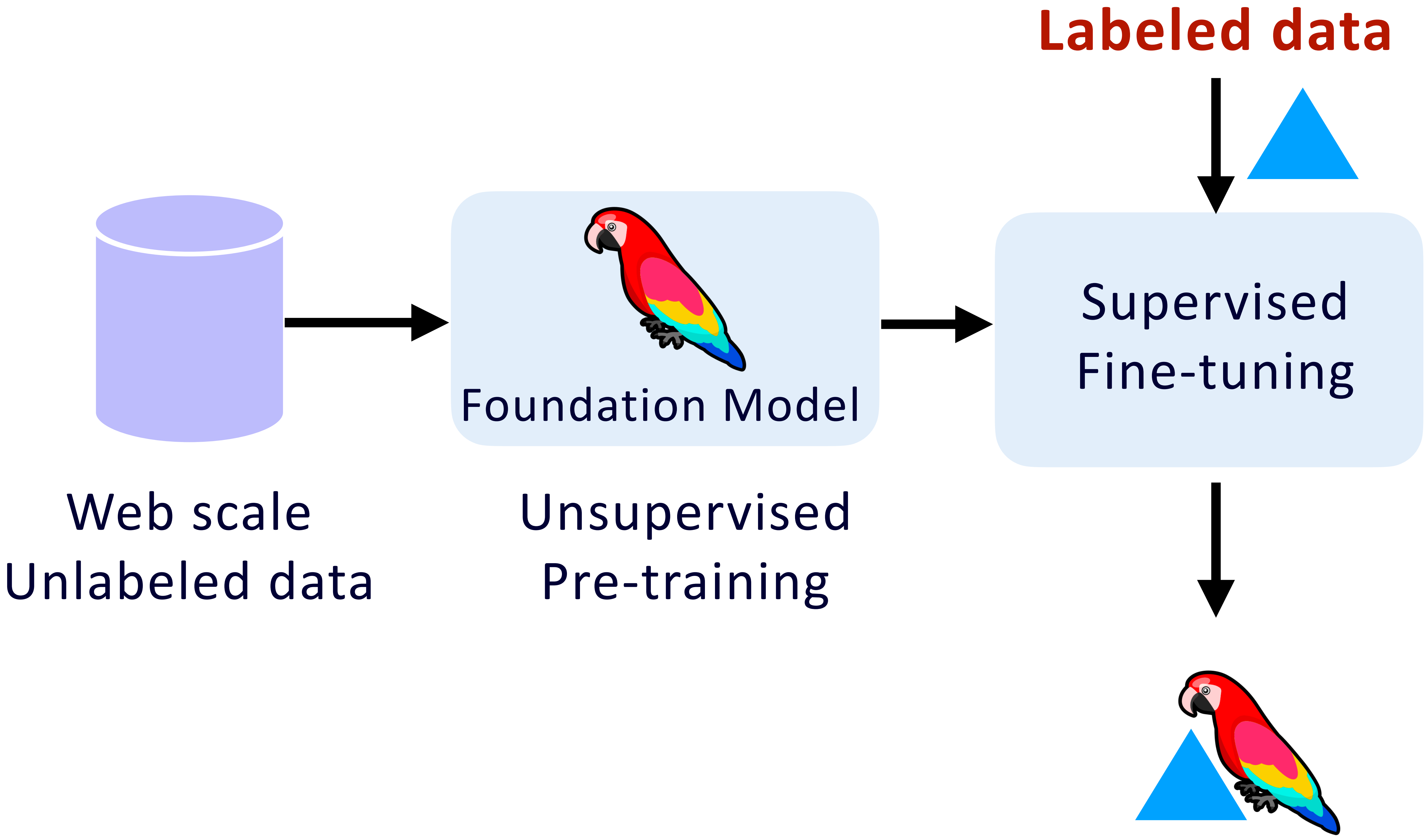
Comparison with Active Learning, Selective Classification

We need labeled data and often a lot of it!

Diagnosing a novel disease using brain scans



Fine-tuning Foundation models or Aligning LLMs



Data Labeling costs a lot of time and money

IMAGENET Deng et. Al. 2009

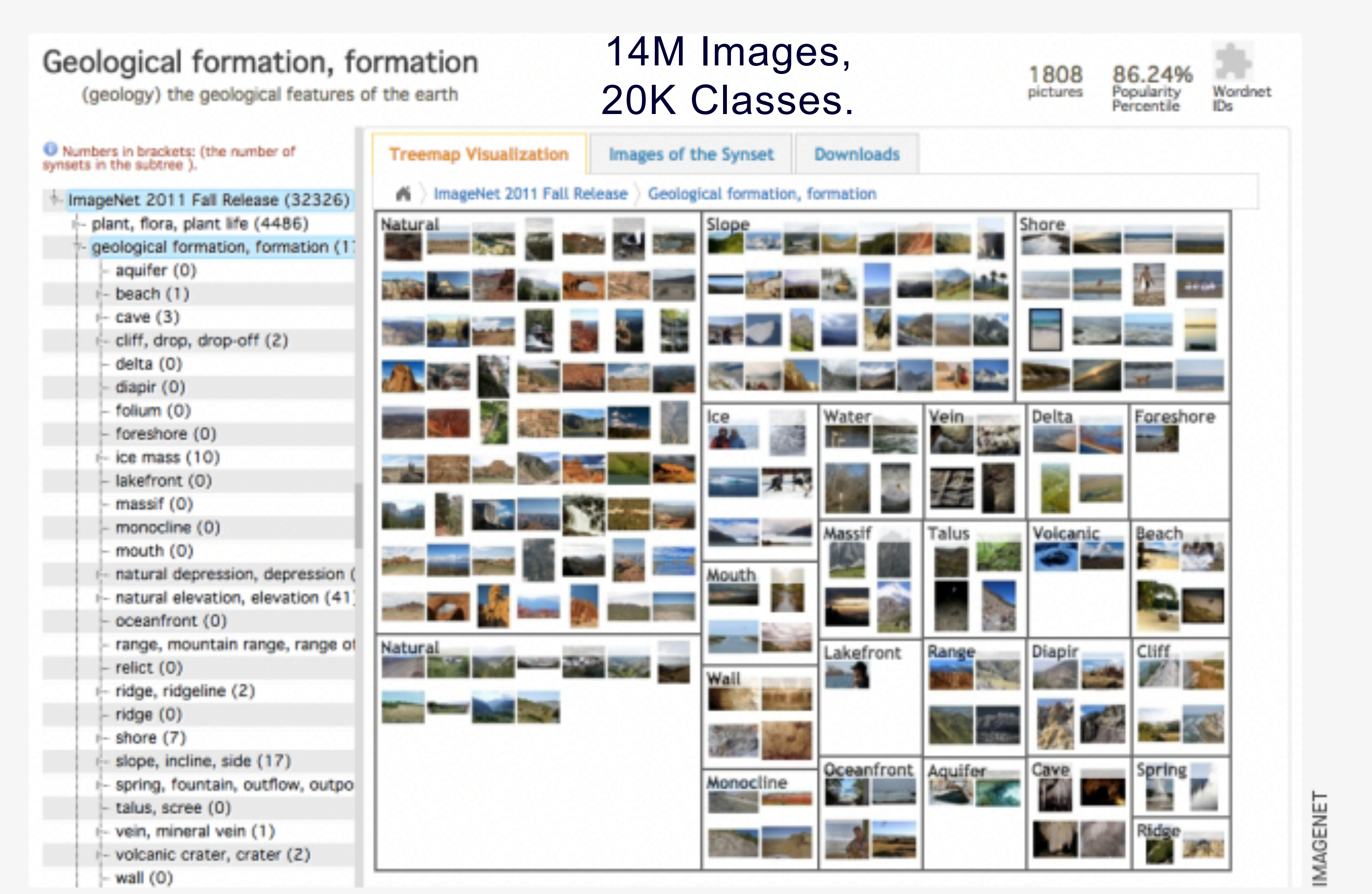
Crowdsourcing is widely used
to get labels

Wisdom of Crowd

amazon mechanical turk

and many others…

Took multiple years and a lot of human effort

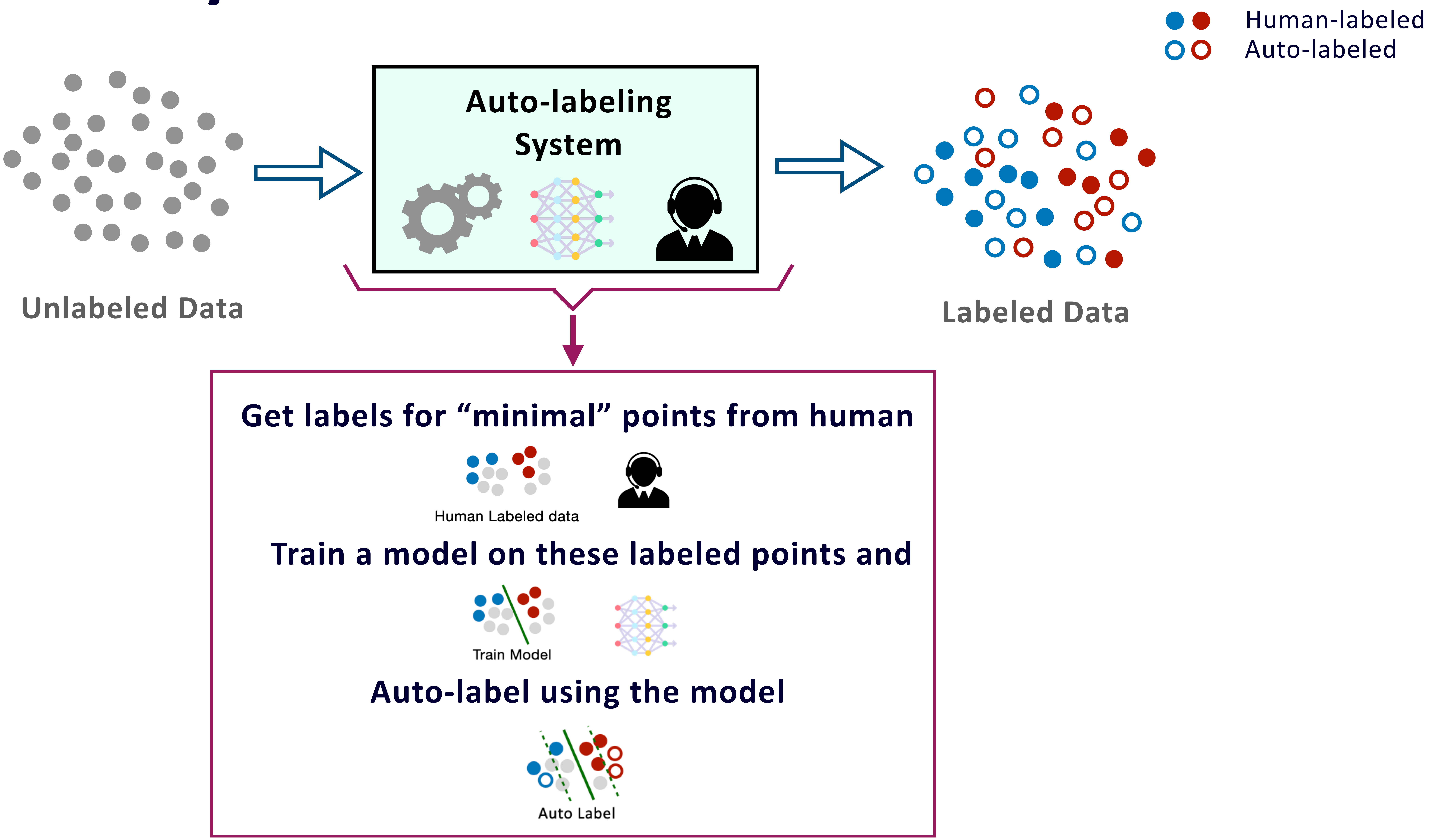


A screenshot of the ImageNet database online

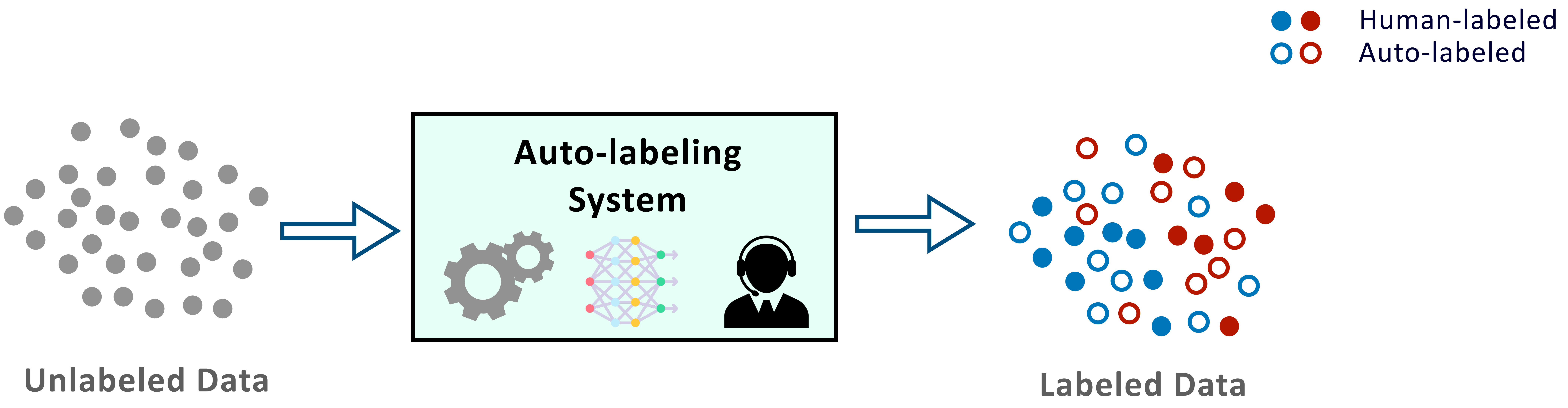
Takes a lot of time and money
to get labels.

How do we get **accurately labeled** data, while spending **less time and money**?

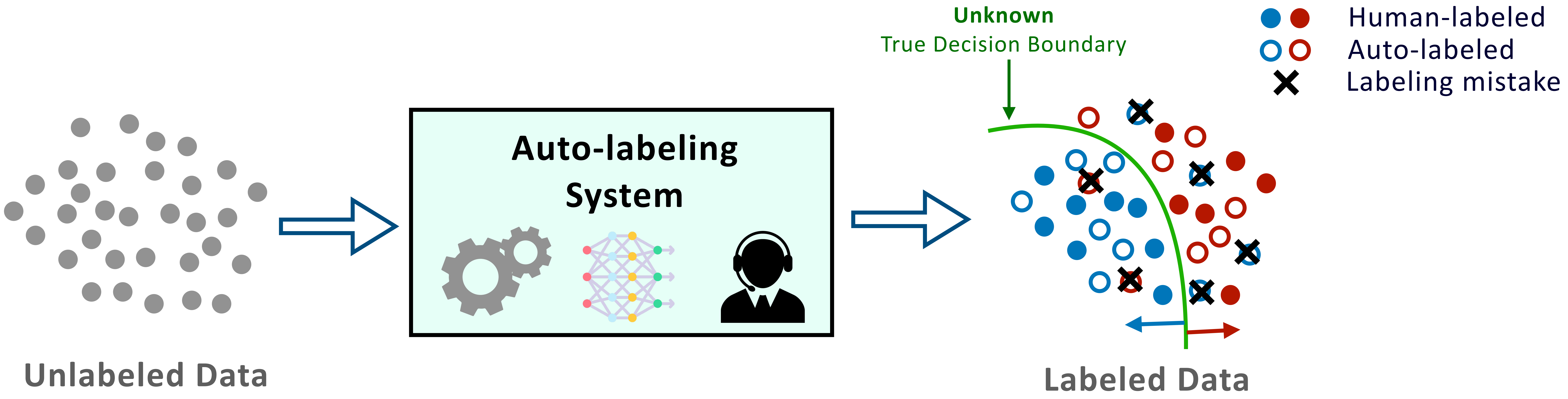
Automatically label datasets with minimal human feedback



Auto-Labeling Errors and Their Impact



Auto-Labeling Errors and Their Impact

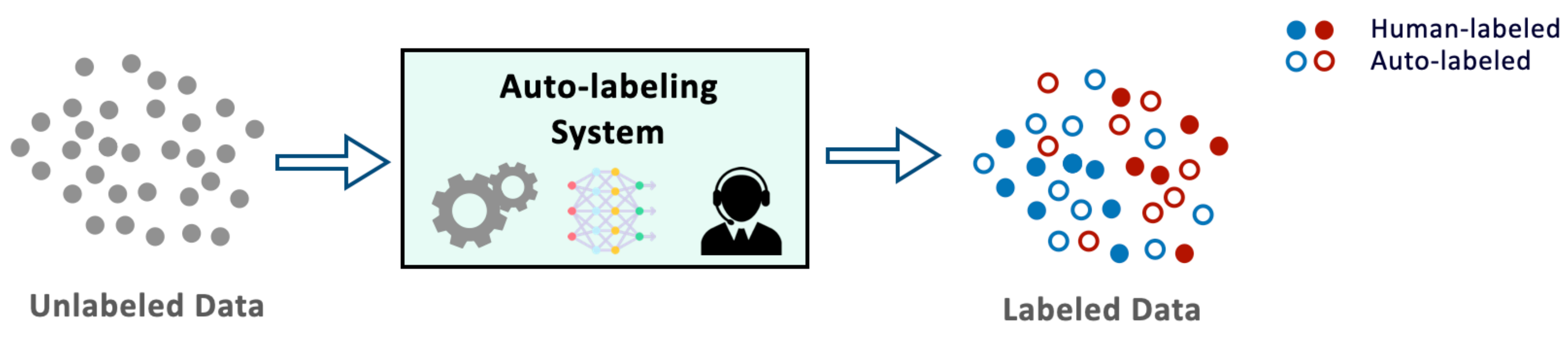


1. The output dataset may have labeling errors

2. The impact of errors in datasets is more severe

a) Multiple downstream applications
b) Longer shelf-life than models.

Auto-labeling systems are widely used



Auto-labeling Platforms

aws Amazon SageMaker Ground Truth
Snorkel
Labelbox
ENCORD
Superb AI
Toloka
SAMSUNG
SAMSUNG SDS
cloudfactory
V7
Open for Innovation KNIME

Auto-labeling is <u>heavily used commercially.</u>

Even in high risk applications

health care, telecom, recruiting…

So we need to understand them.

Despite wide adoption, our **understanding of auto-labeling systems is limited!**

Despite wide adoption, our **understanding of auto-labeling systems is limited!**

To address this gap we **develop a theoretical understanding** of auto-labeling systems.

Auto-labeling systems are widely used

Auto-labeling Platforms

Amazon SageMaker Ground Truth

Snorkel

Labelbox

ENCORD

Superb AI

Toloka

SAMSUNG

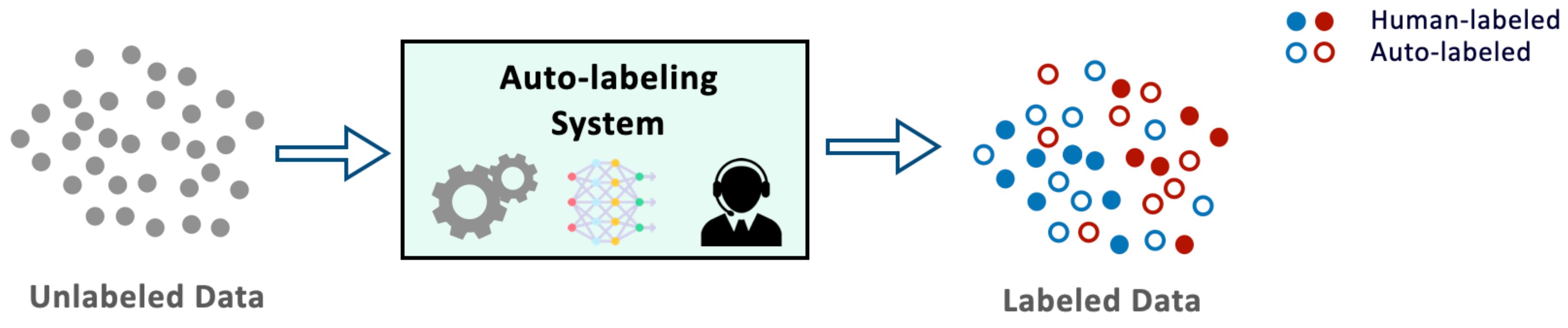
SAMSUNG SDS

cloudfactory

V7

Open for Innovation KNIME

Study a w/f inspired from it.



Auto-labeling is <u>heavily used commercially.</u>

Even in high risk applications

health care, telecom, recruiting…

So we need to understand them.

Roadmap

What & Why auto-labeling?

Data labeling problem

Adoption of auto-labeling

How does it work?

Workflow of TBAL

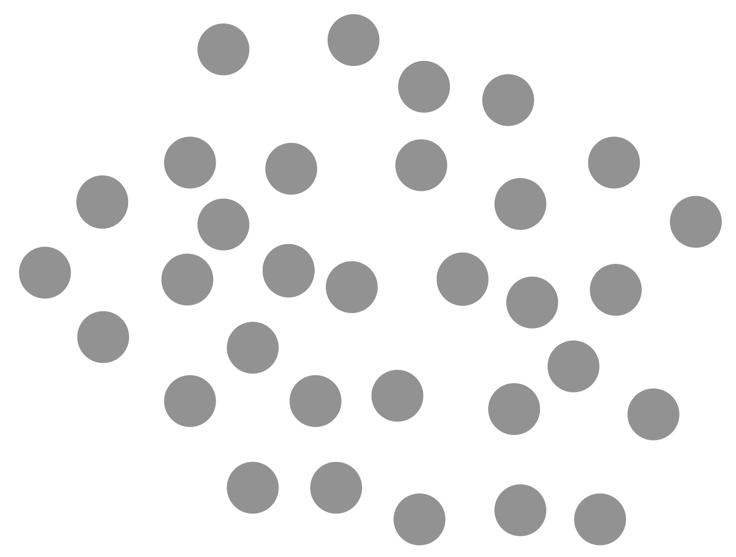
Finding the auto-labeling region

Analysis & Results

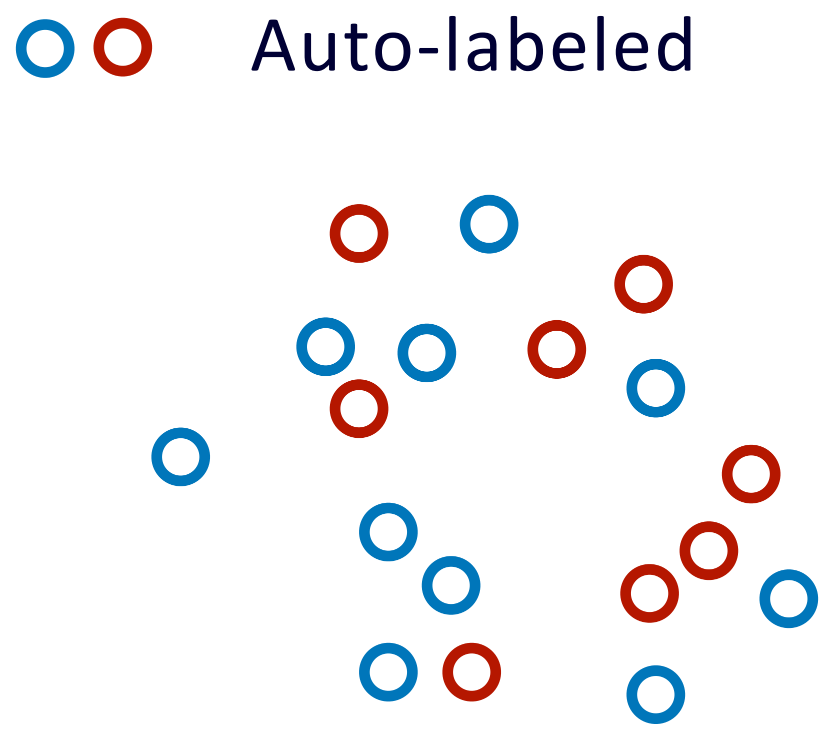
Conditions when TBAL works.

Comparison with Active Learning, Selective Classification

Quality and Quantity of Auto-labeled Data

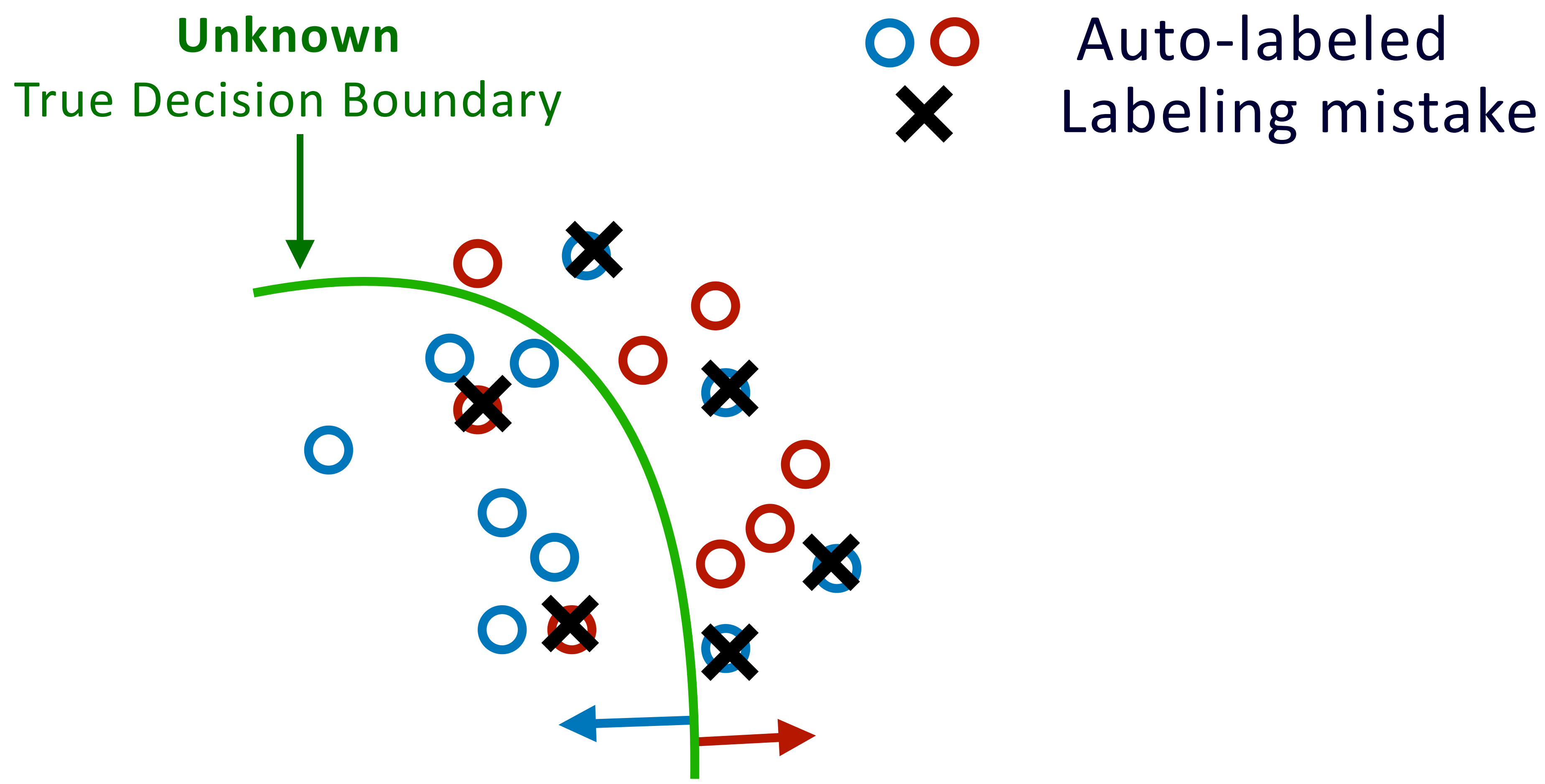


N Number of **unlabeled** points



A Set of **auto-labeled** points

N_a Number of auto-labeled points



M_a Number of labeling mistakes

Quantity

Auto-labeling Coverage

$$\hat{\mathcal{P}} = \frac{N_a}{N}$$

Good Stuff
maximize this

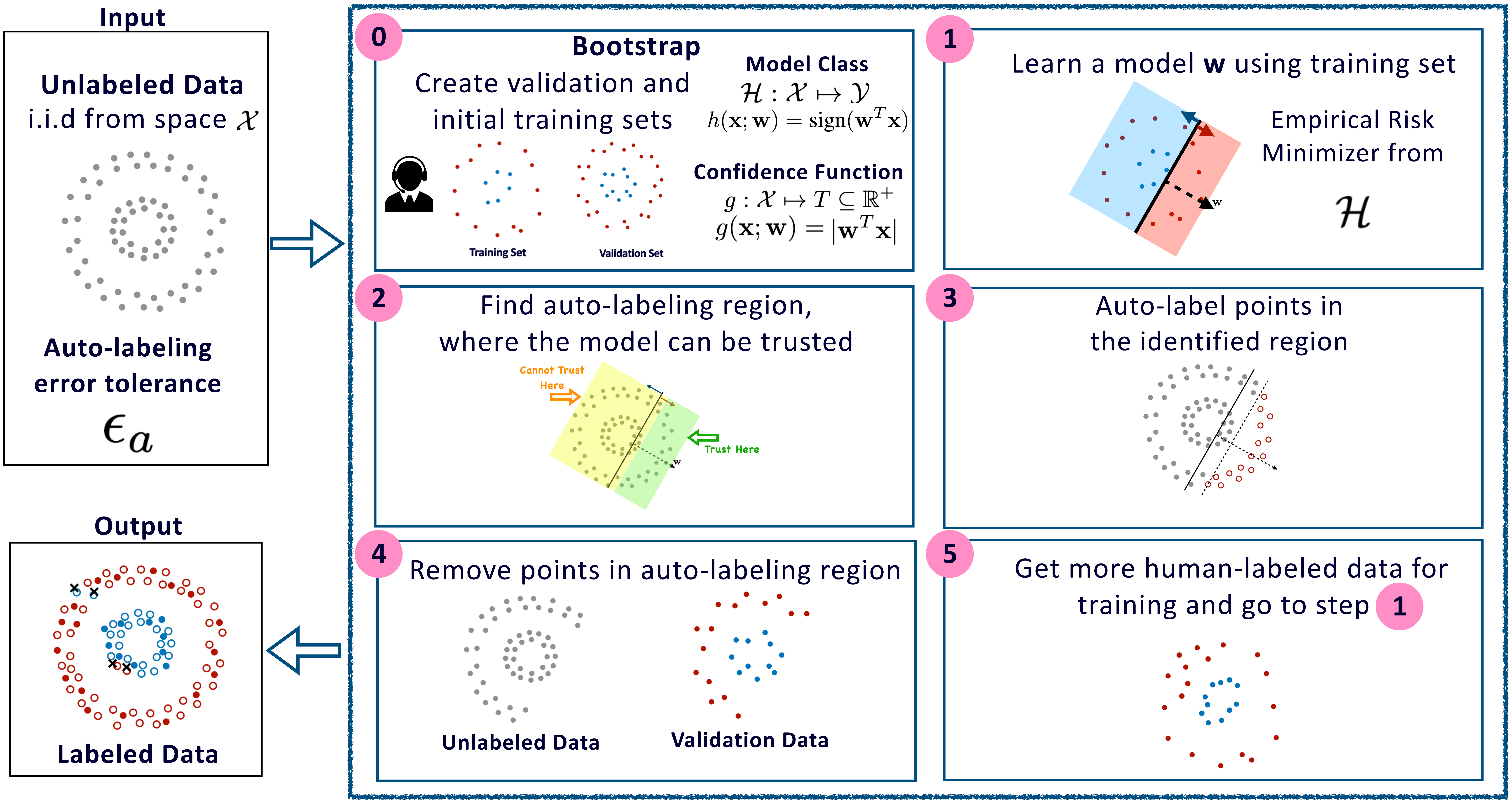
Quality

Auto-labeling Error

$$\widehat{\mathcal{E}} = \frac{M_a}{N_a}$$

Bad Stuff
minimize this

Threshold-based Auto-labeling Workflow(TBAL)



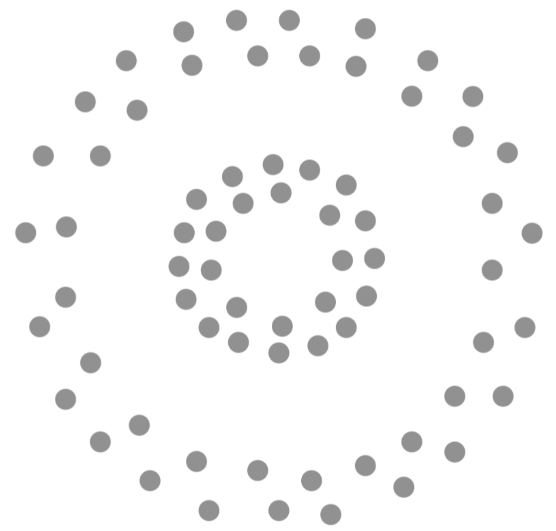
~~Pretend~~ we are LLMs and

Let’s think step by step with an example

Threshold-based Auto-labeling Workflow(TBAL)

Input

Unlabeled Data
i.i.d from space $\mathcal{X}$



Auto-labeling error tolerance

ϵ_a

0 **Bootstrap**

Create validation and initial training sets

Model Class

$\mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$

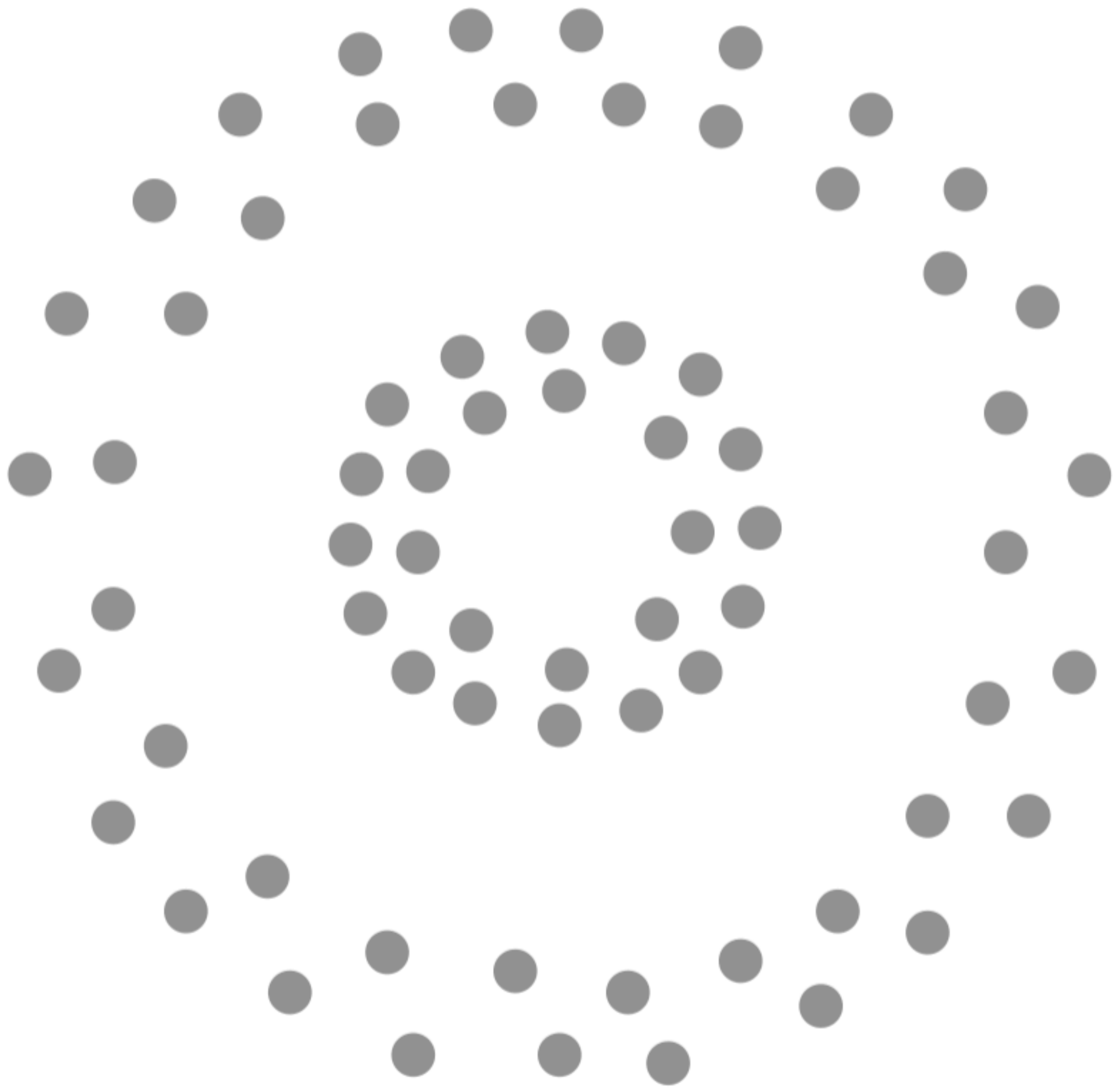
$h(\mathbf{x}; \mathbf{w}) = \text{sign}(\mathbf{w}^T \mathbf{x})$

Confidence Function

$g : \mathcal{X} \mapsto T \subseteq \mathbb{R}^+$

$g(\mathbf{x}; \mathbf{w}) = |\mathbf{w}^T \mathbf{x}|$

Input

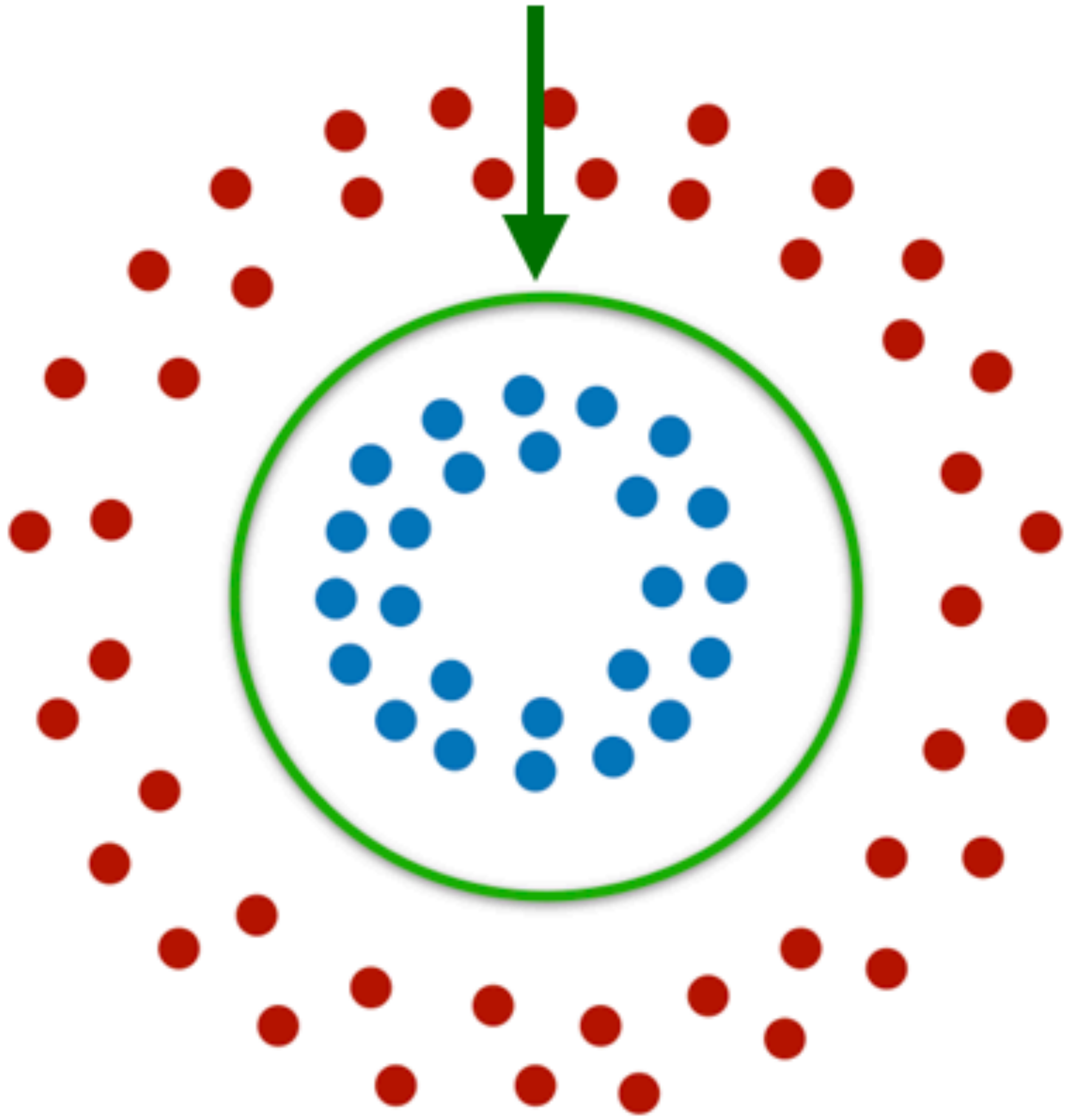


Unlabeled Data

i.i.d from space $\mathcal{X}$

Unknown

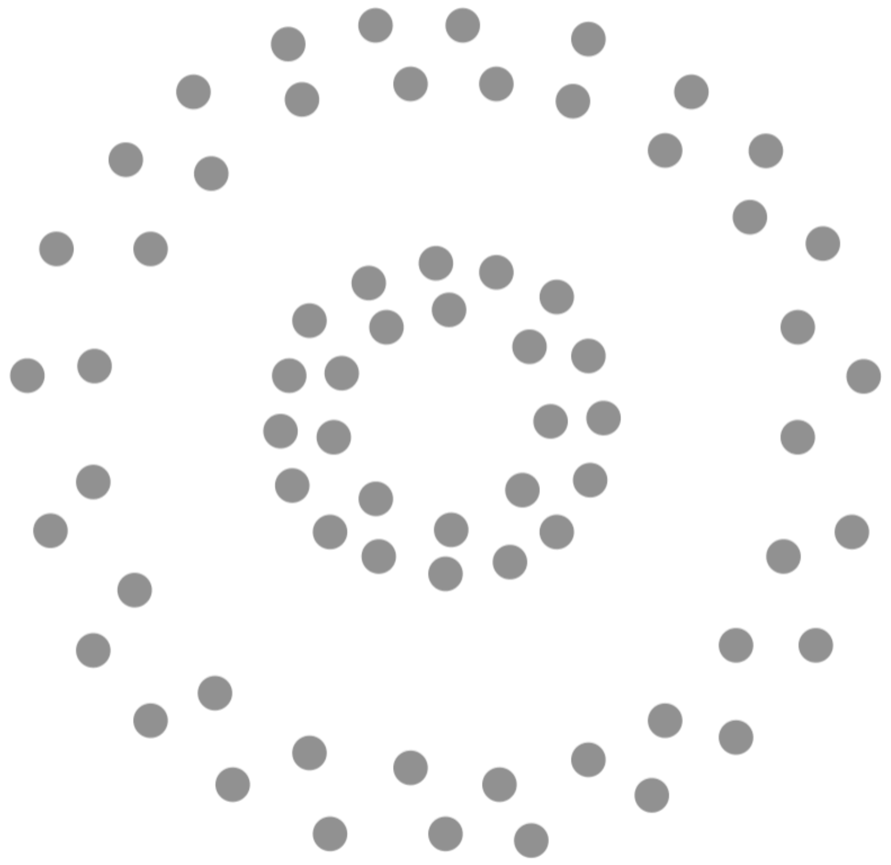
True Decision Boundary $f^\star$



Known to the oracle

Learning $f^\star$ is NOT the goal.

Input

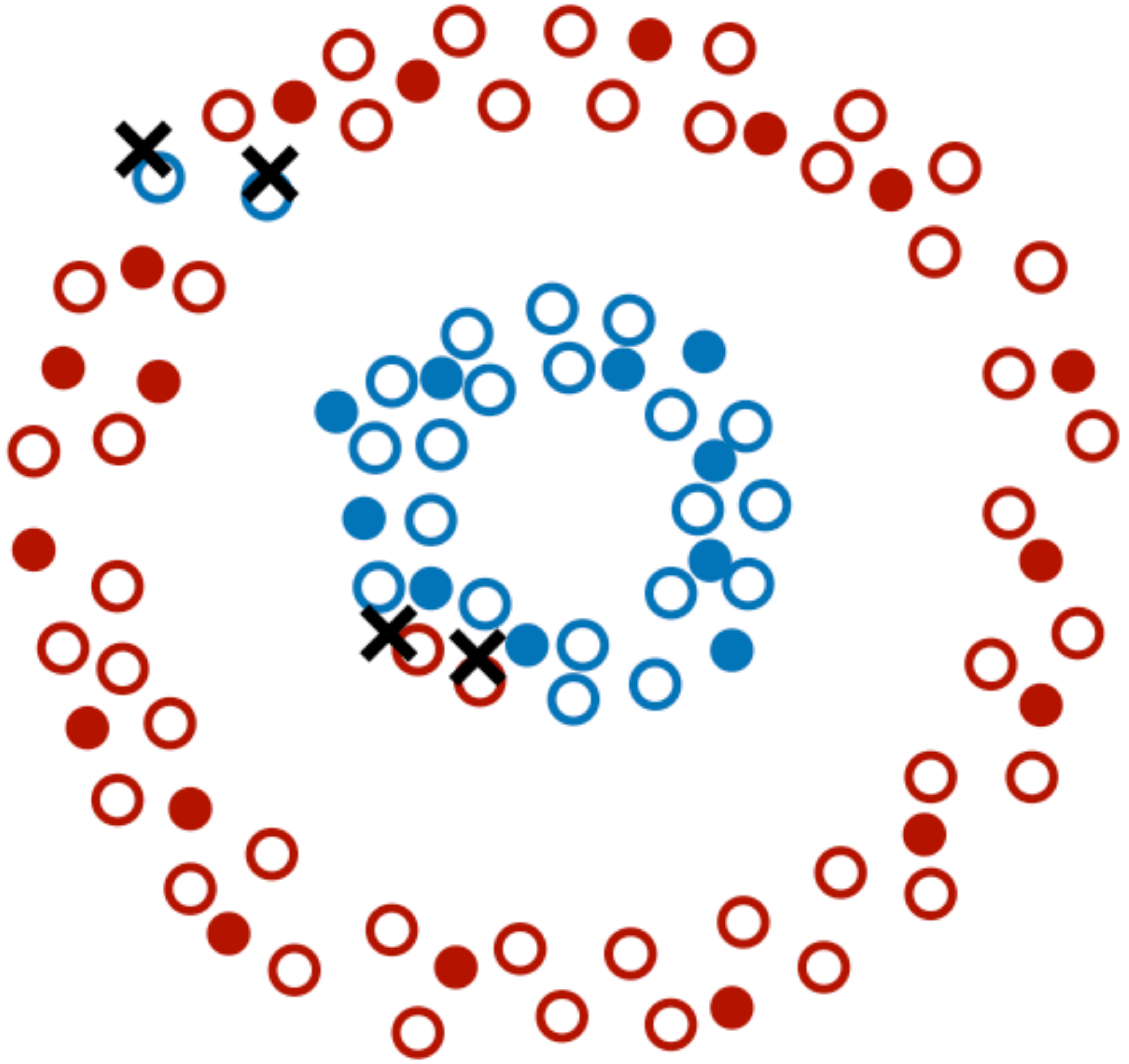


Unlabeled Data

Auto-labeling error tolerance

ϵ_a

Expected Output



Labeled Data

● ● Human-labeled
○ ○ Auto-labeled
✕ Labeling mistake

Auto-labeling Error

$$\widehat{\mathcal{E}} = \frac{M_a}{N_a} = \frac{\#\times}{\#\bigcirc + \#\bigcirc} \leq \epsilon_a$$

Coverage

$$\hat{\mathcal{P}} = \frac{N_a}{N} = \frac{\#\bigcirc + \#\bigcirc}{\#\bullet}$$

TBAL Workflow : Bootstrap (Step 0)

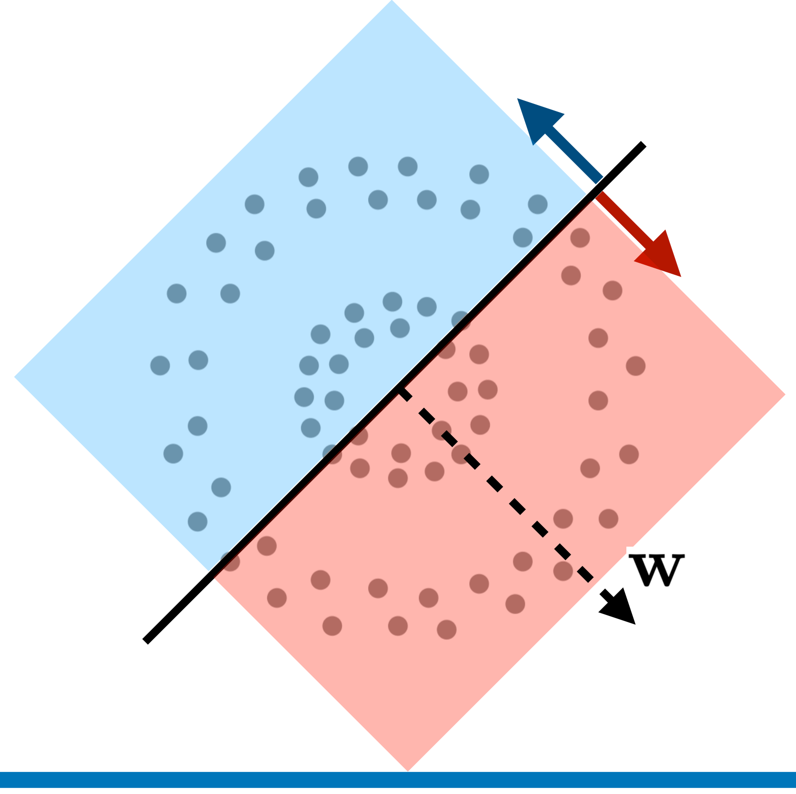
Pick a Model class and Confidence function

Model/Hypothesis Class

$$\mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$$

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^2 : ||\mathbf{x}||_2 \leq 1\}$$

$$\mathcal{Y} = \{-1, +1\}$$



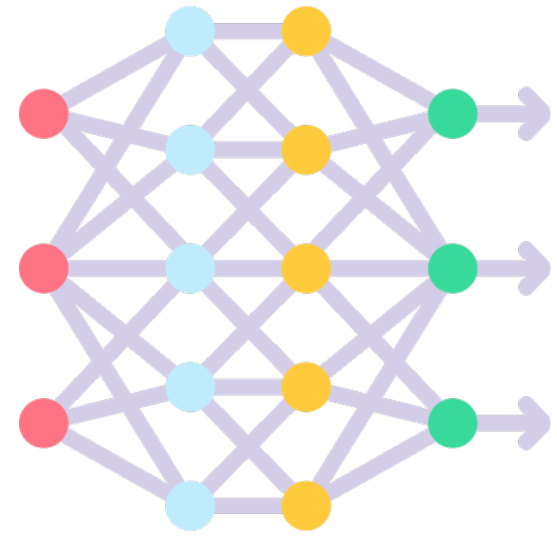
Linear Classifiers

$$\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^2 : ||\mathbf{w}||_2 \leq 1\}$$

$$h(\mathbf{x}; \mathbf{w}) = \text{sign}(\mathbf{w}^T \mathbf{x})$$

$$\boxed{f^\star \notin \mathcal{H}}$$

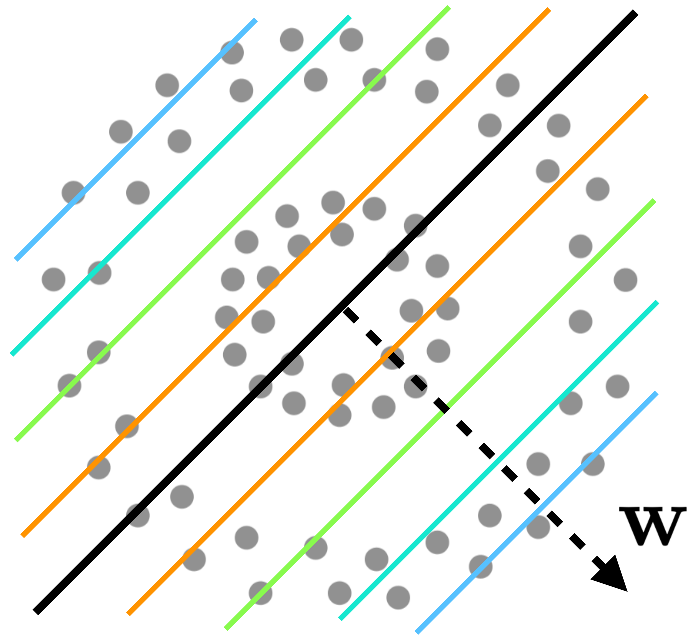
Neural Nets



Confidence/Scoring Function

$$g : \mathcal{X} \mapsto T \subseteq \mathbb{R}^+$$

$$T = [0, 1]$$



Linear Confidence Function

$$g(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + e^{-|\mathbf{w}^T \mathbf{x}|}}$$

$$\equiv |\mathbf{w}^T \mathbf{x}|$$

Softmax Score

0.2	0.1	**0.7**

TBAL Workflow : Bootstrap (Step 0)

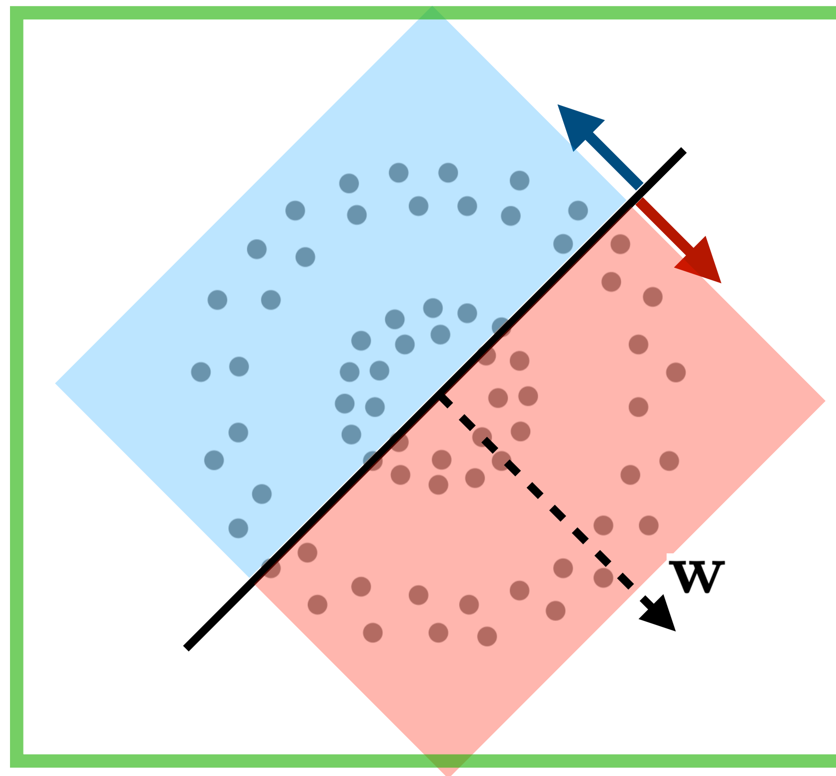
Pick a Model class and Confidence function

Model/Hypothesis Class

$$\mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$$

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^2 : ||\mathbf{x}||_2 \leq 1\}$$

$$\mathcal{Y} = \{-1, +1\}$$



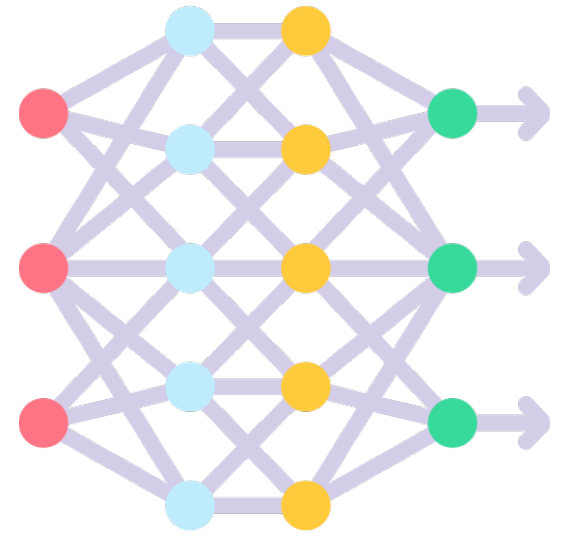
Linear Classifiers

$$\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^2 : ||\mathbf{w}||_2 \leq 1\}$$

$$h(\mathbf{x}; \mathbf{w}) = \text{sign}(\mathbf{w}^T\mathbf{x})$$

$$\boxed{f^\star \notin \mathcal{H}}$$

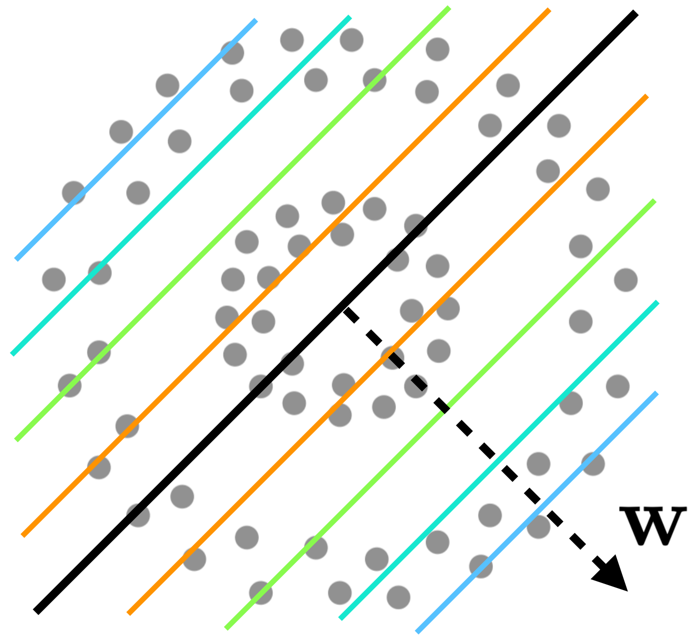
Neural Nets



Confidence/Scoring Function

$$g : \mathcal{X} \mapsto T \subseteq \mathbb{R}^+$$

$$T = [0, 1]$$



Linear Confidence Function

$$g(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + e^{-|\mathbf{w}^T\mathbf{x}|}}$$

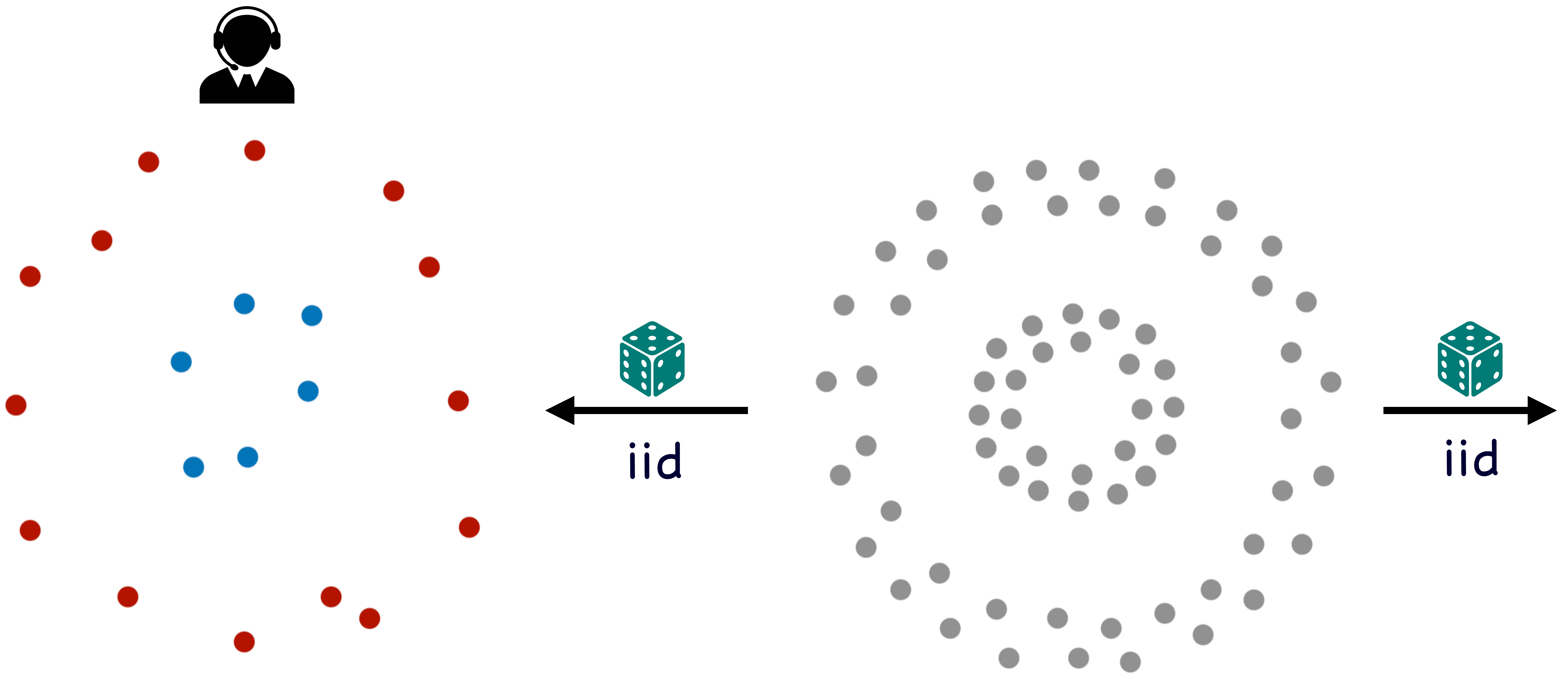
$$\equiv |\mathbf{w}^T\mathbf{x}|$$

Softmax Score

0.2	0.1	**0.7**

TBAL Workflow : Bootstrap (Step 0)

Get some labeled data for training and validation



Training Set **Unlabeled Set** **Validation Set**

$$D_{train} = \{(\mathbf{x}_i, y_i) : i \in I_{train}\}$$

Start small and gradually add more

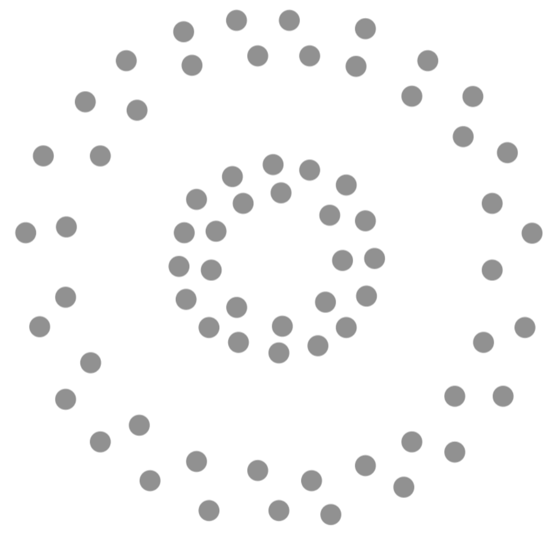
$$D_{val} = \{(\mathbf{x}_i, y_i) : i \in I_{val}\}$$

Get “sufficiently” large amount of it.

Threshold-based Auto-labeling Workflow(TBAL)

Input

Unlabeled Data
i.i.d from space $\mathcal{X}$



Auto-labeling error tolerance
ϵ_a

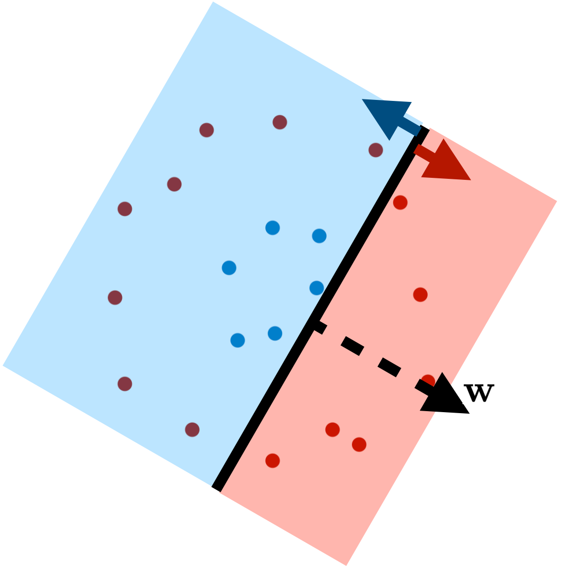
0 **Bootstrap**

Create validation and initial training sets

Model Class
$\mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$
$h(\mathbf{x}; \mathbf{w}) = \mathrm{sign}(\mathbf{w}^T \mathbf{x})$

Confidence Function
$g : \mathcal{X} \mapsto T \subseteq \mathbb{R}^+$
$g(\mathbf{x}; \mathbf{w}) = |\mathbf{w}^T \mathbf{x}|$

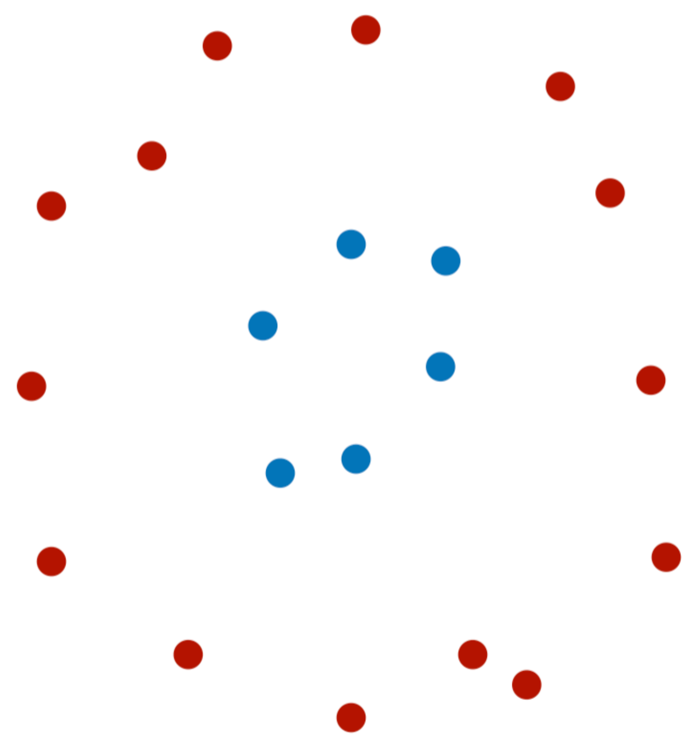
1 Learn a model **w** using training set



Empirical Risk Minimizer from

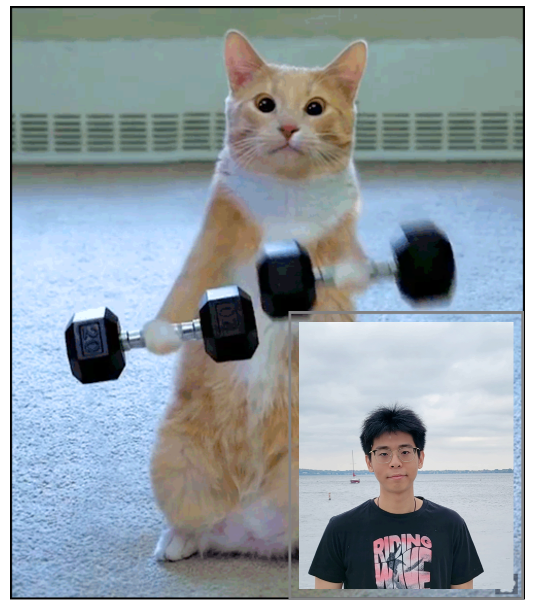
$\mathcal{H}$

TBAL Workflow : Step 1 Model training

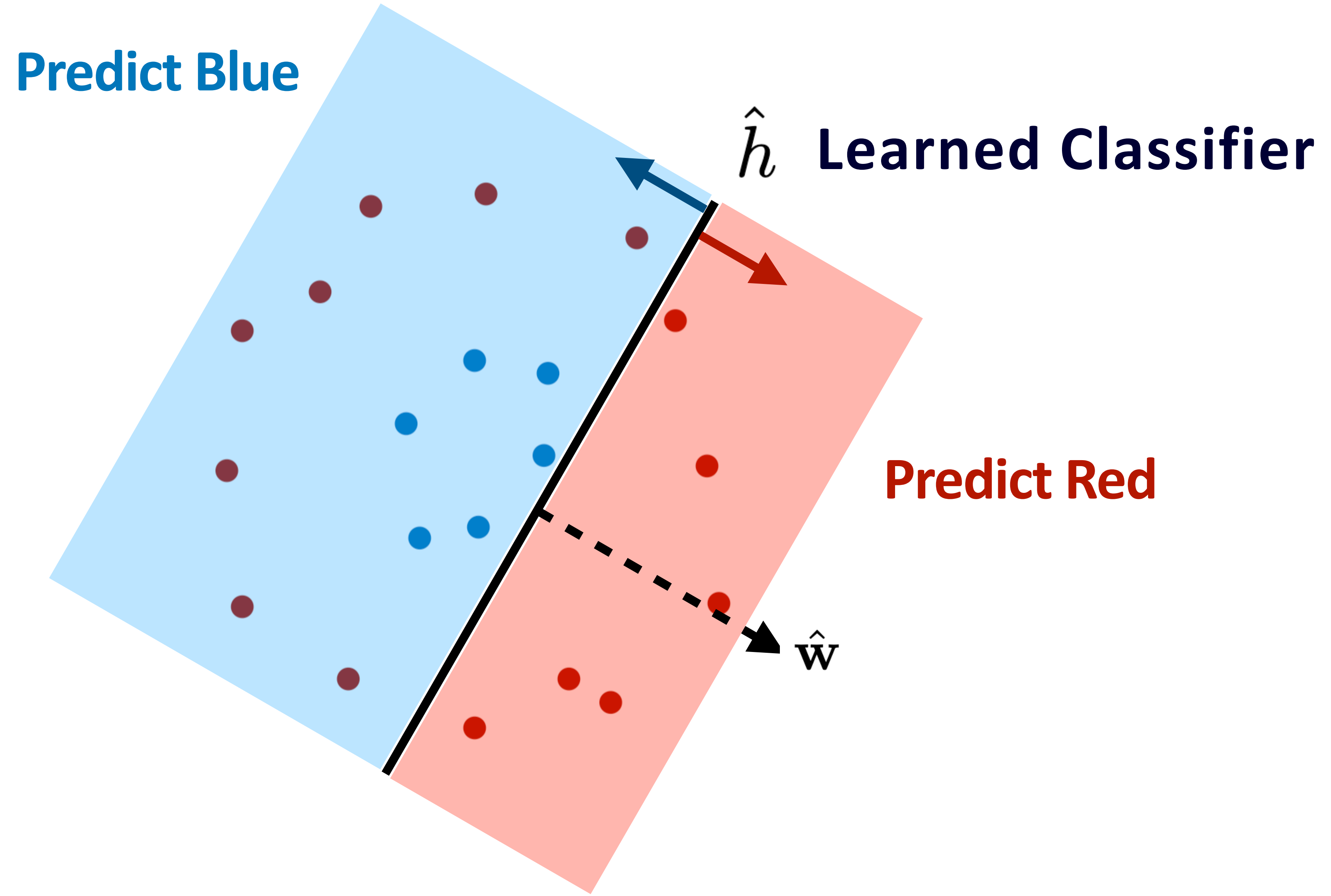


Training Set

Train a model



Lin



$$\hat{h} = \texttt{EmpiricalRiskMinimizer}(\mathcal{H}, D_{train})$$

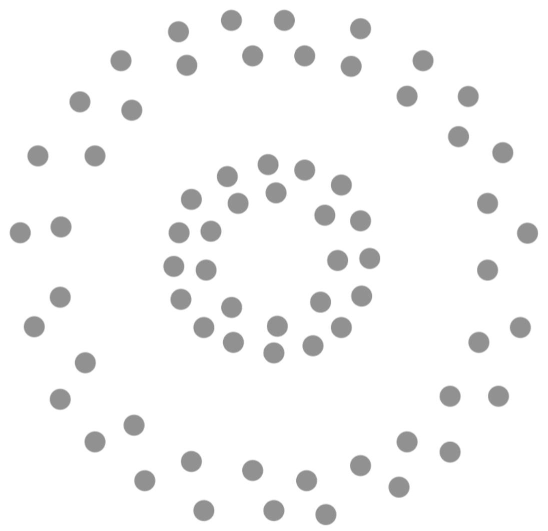
$$\hat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{|D_{train}|} \sum_{(\mathbf{x}_i, y_i) \in D_{train}} \mathbb{1}\{h(\mathbf{x}_i) \neq y_i\}$$

In practice, usually some surrogate loss is minimized

Threshold-based Auto-labeling Workflow(TBAL)

Input

Unlabeled Data
i.i.d from space $\mathcal{X}$



Auto-labeling error tolerance

ϵ_a

0 **Bootstrap**

Create validation and initial training sets

Model Class

$\mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$

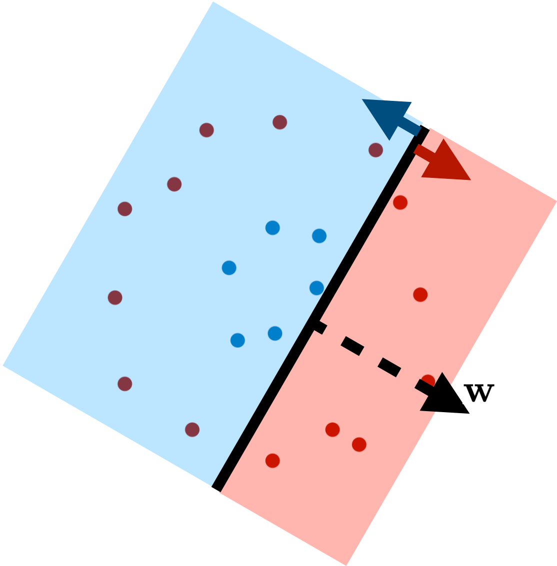
$h(\mathbf{x}; \mathbf{w}) = \mathrm{sign}(\mathbf{w}^T \mathbf{x})$

Confidence Function

$g : \mathcal{X} \mapsto T \subseteq \mathbb{R}^+$

$g(\mathbf{x}; \mathbf{w}) = |\mathbf{w}^T \mathbf{x}|$

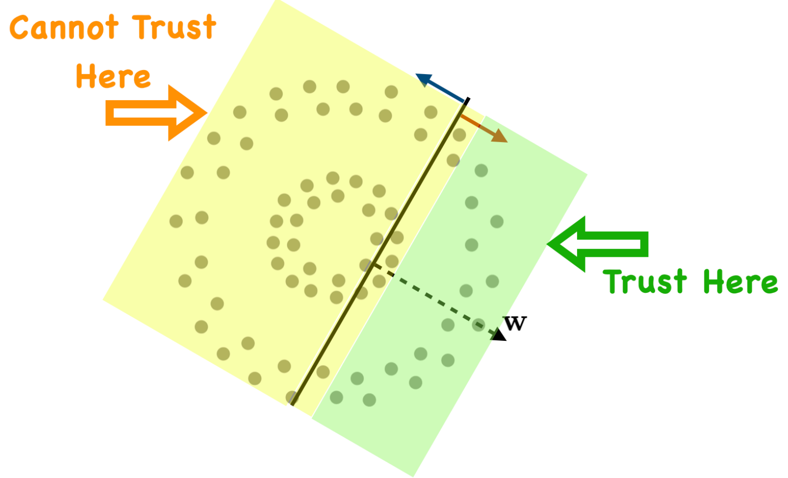
1 Learn a model **w** using training set



Empirical Risk Minimizer from

$\mathcal{H}$

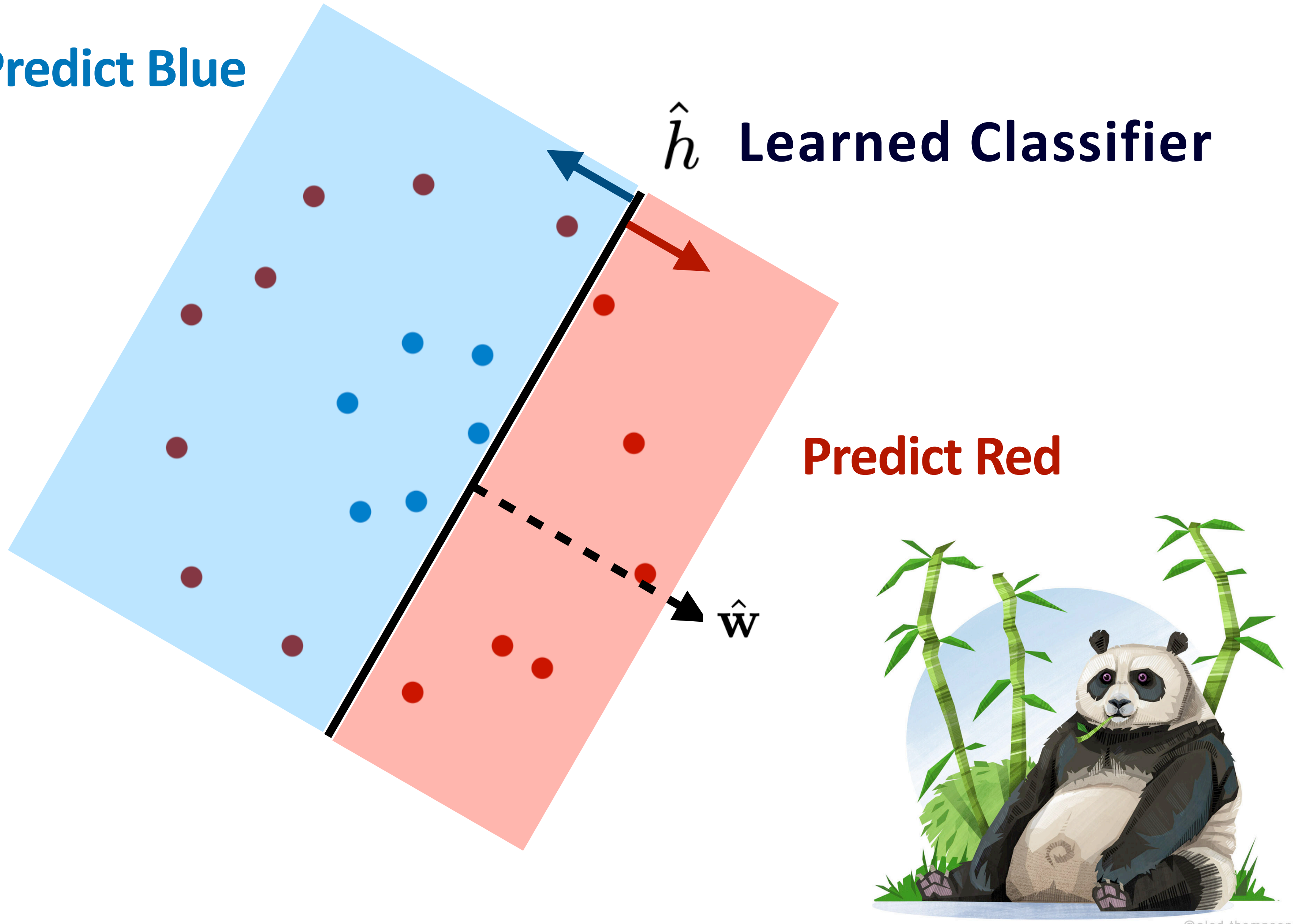
2 Find auto-labeling region, where the model can be trusted



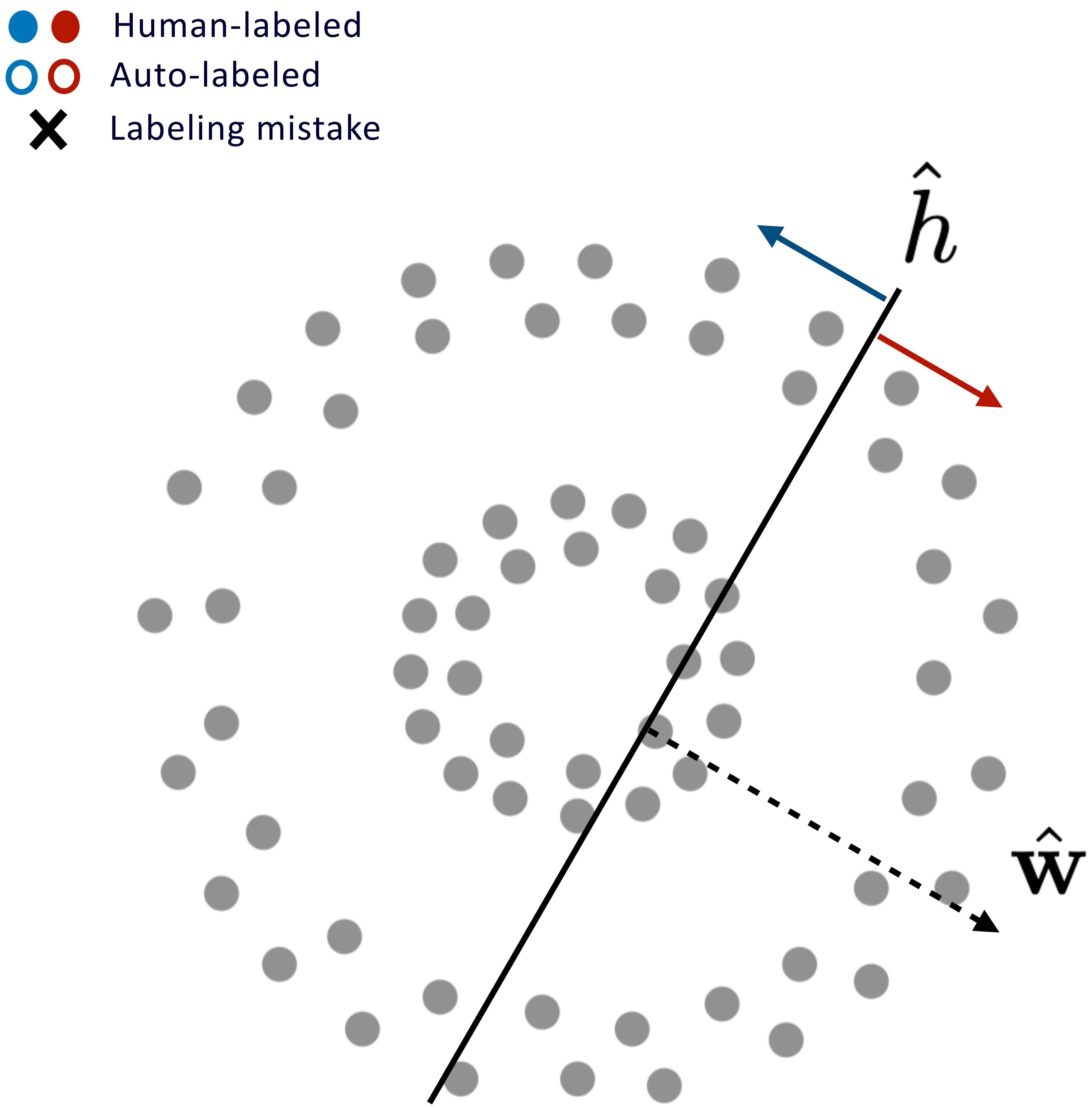
TBAL Workflow: Step 2

Find the Auto-labeling region

Idea 1: Auto-label everywhere.



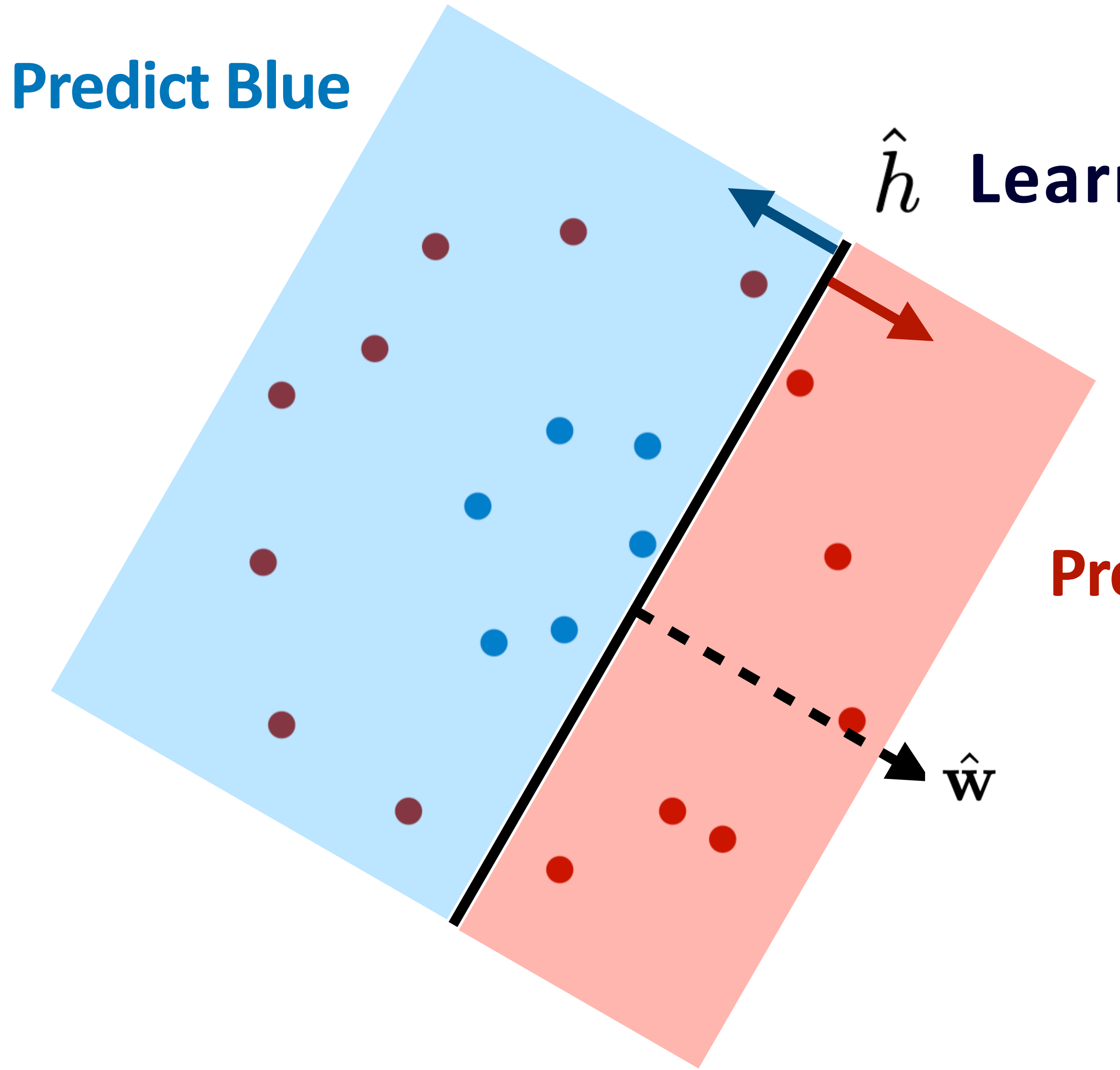
I am lazy and I trust Lin (the model)!
I will auto-label everywhere



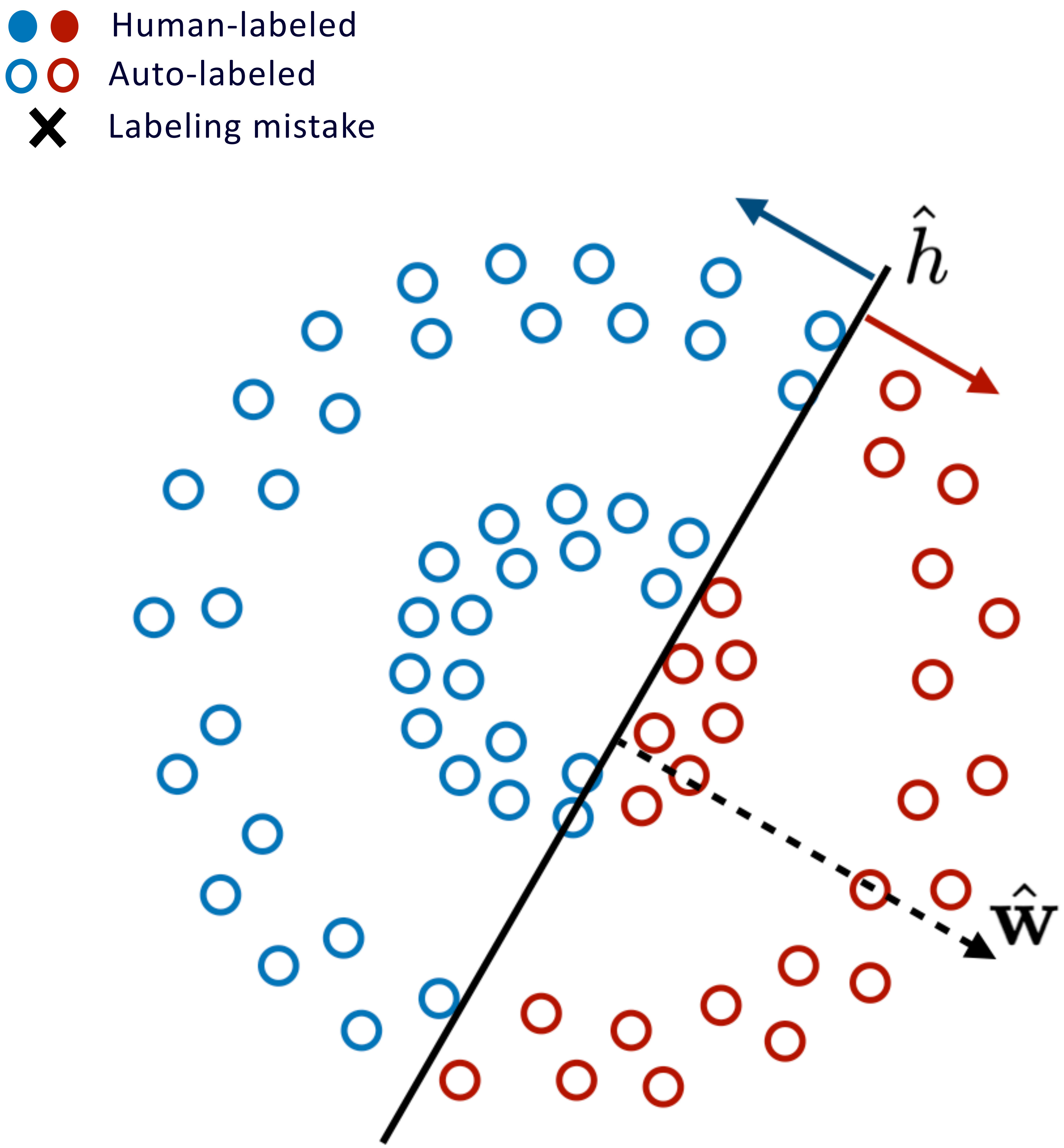
TBAL Workflow: Step 2

Find the Auto-labeling region

Idea 1: Auto-label everywhere.



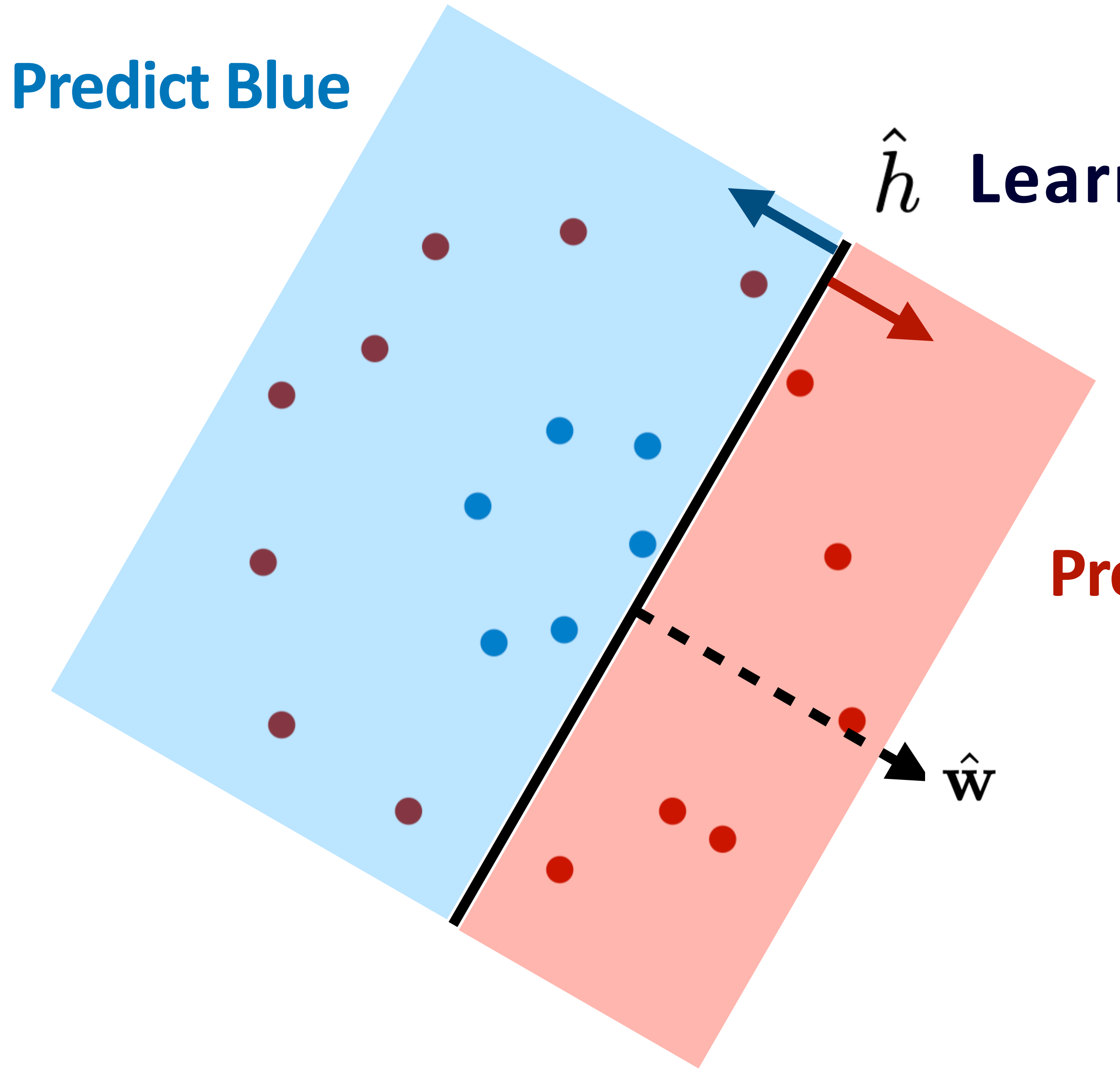
I am lazy and I trust Lin (the model)!
I will auto-label everywhere



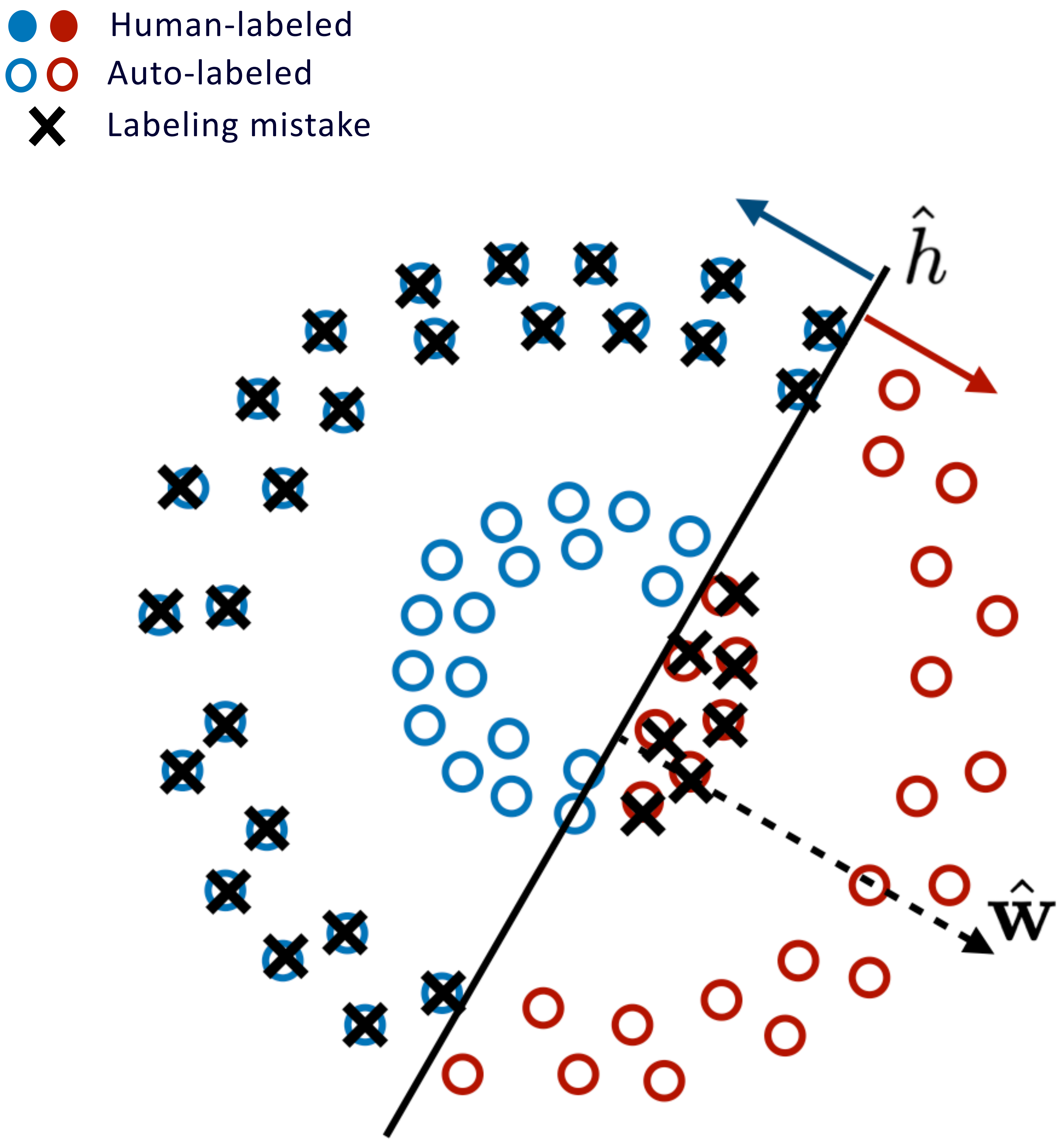
TBAL Workflow: Step 2

Find the Auto-labeling region

Idea 1: Auto-label everywhere.



I am lazy and I trust Lin (the model)!
I will auto-label everywhere



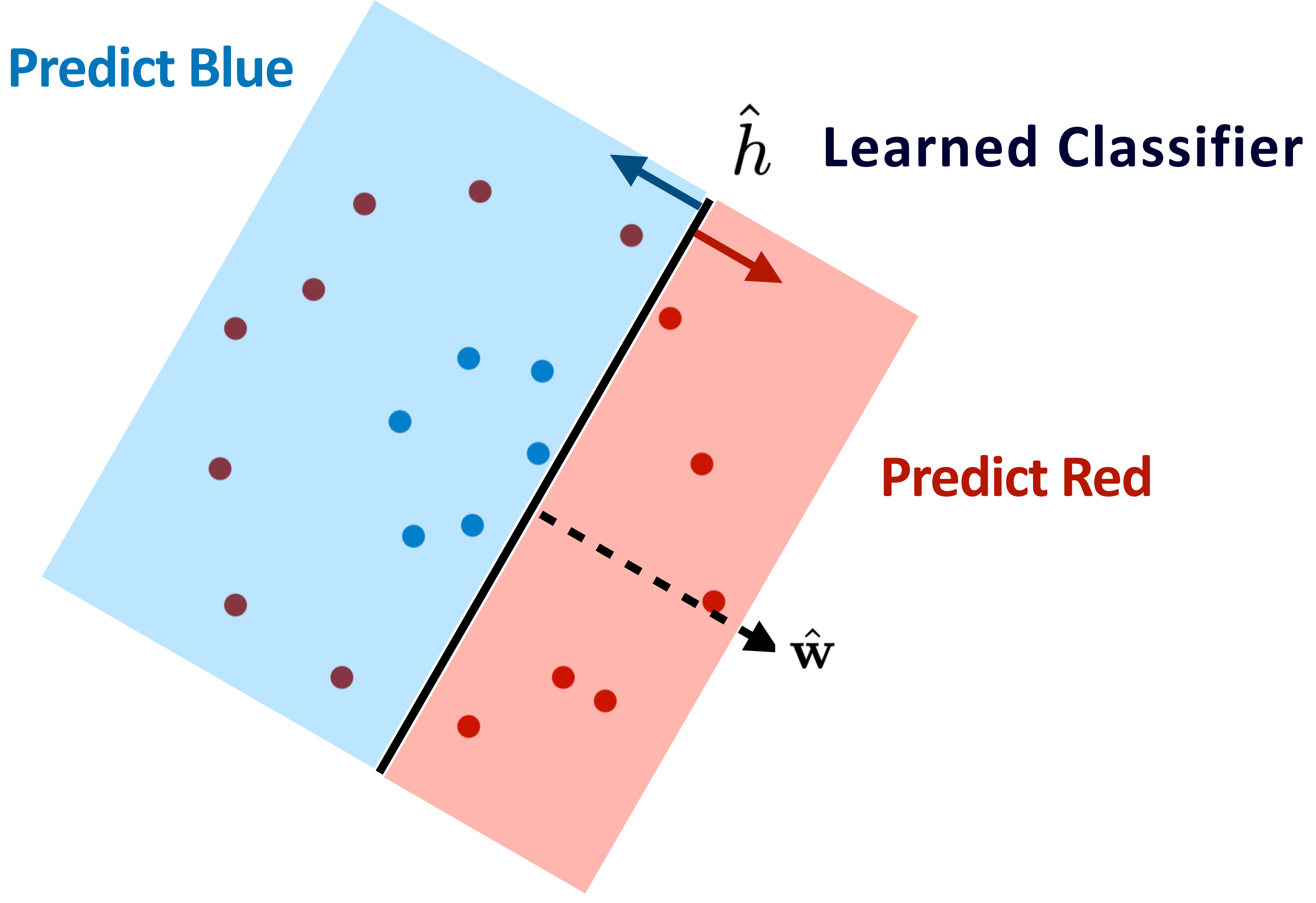
Could lead to high auto-labeling errors!

Panda's strategy does not work,
he goes to Master Shifu for advice.

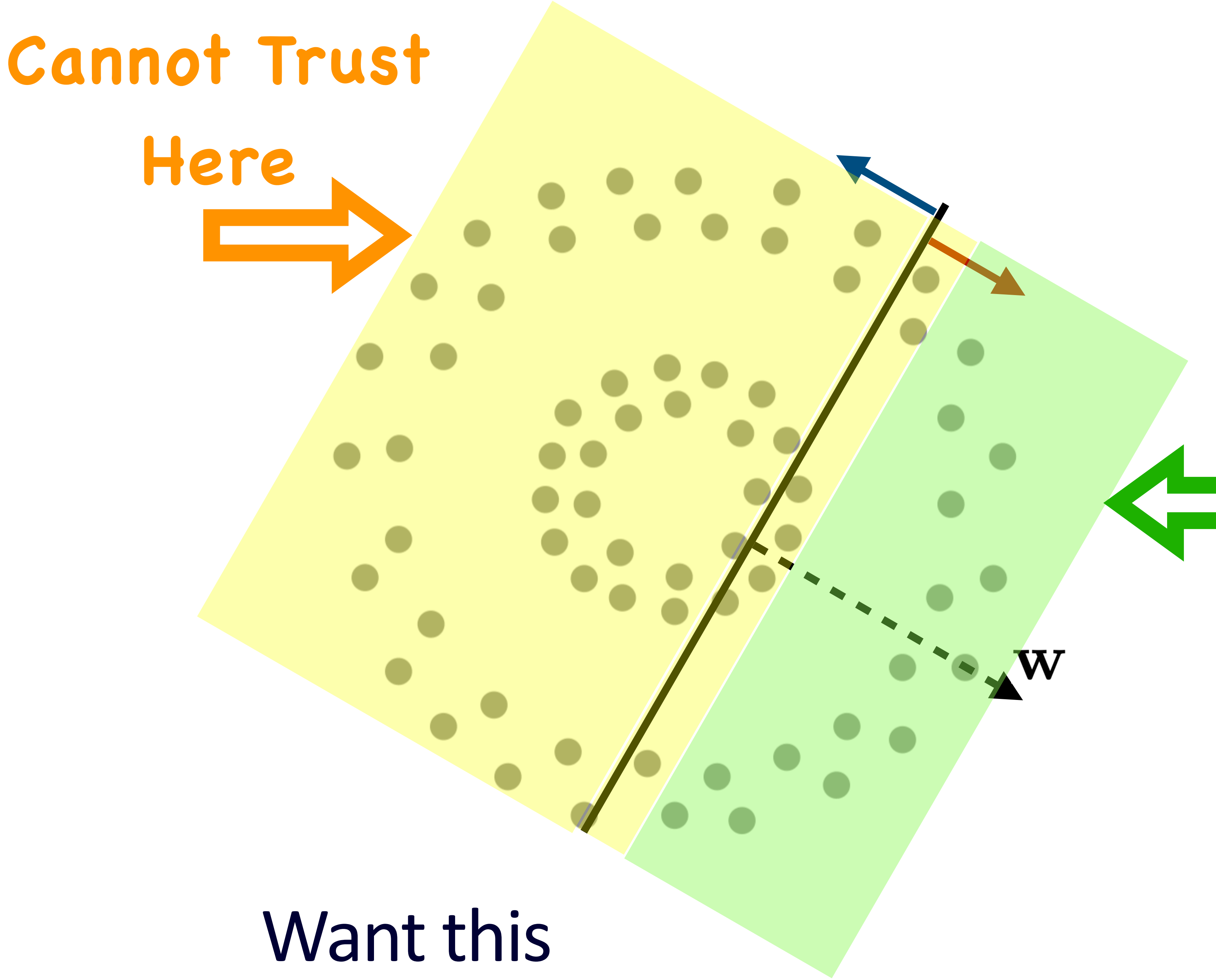
TBAL Workflow: Step 2

Find the Auto-labeling region

Idea 2: Auto-label where the model is accurate (or trustworthy?)



Learned Model



Want this

Master Shifu
The advisors

How to find the yellow and green regions?

TBAL Workflow: Step 2

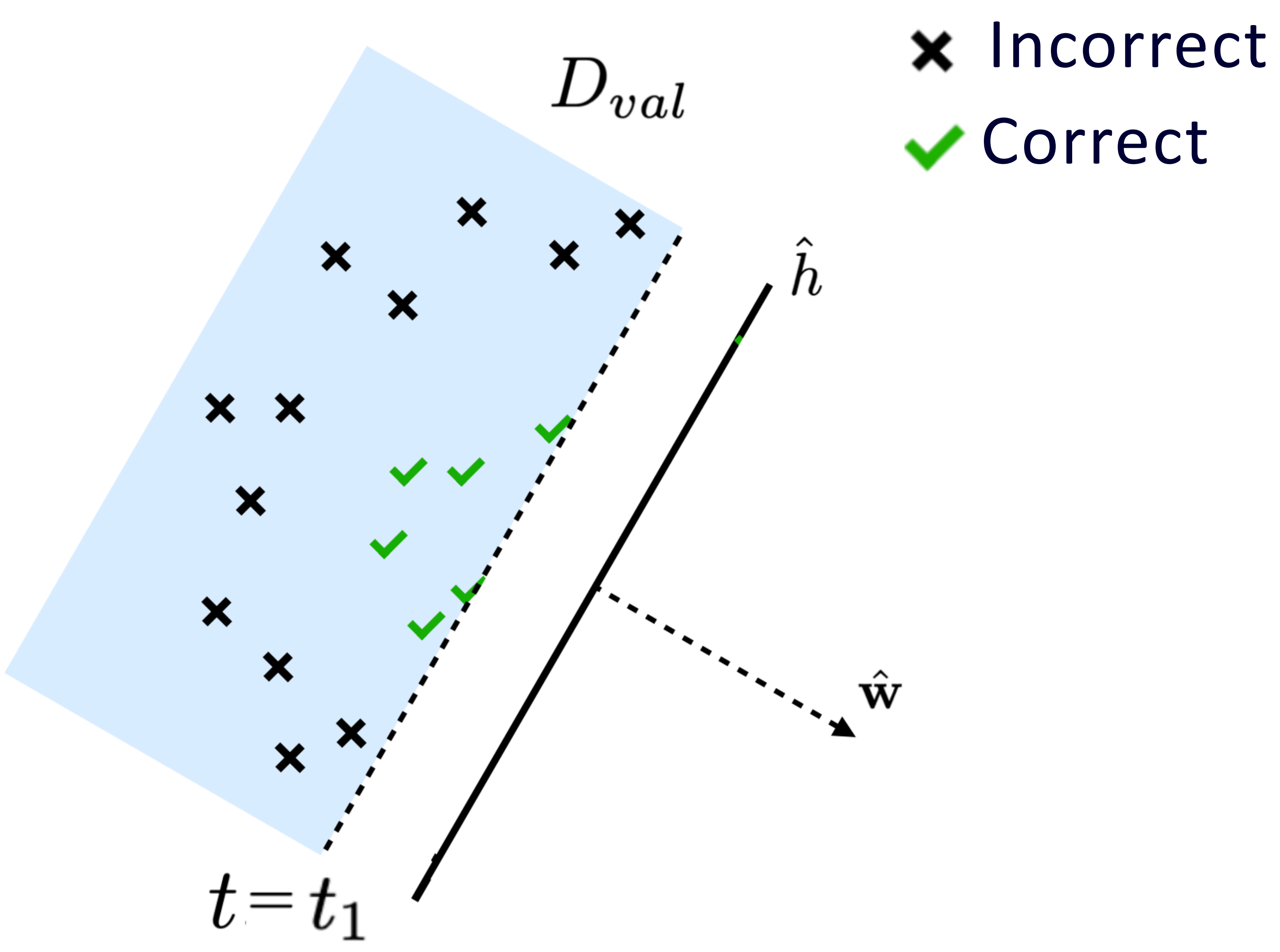
Find the Auto-labeling region

Use the **validation data** to find the region where the classifier can be trusted

TBAL Workflow: Step 2

Find the Auto-labeling region

$$g(\mathbf{x}; \hat{\mathbf{w}}) = |\hat{\mathbf{w}}^T \mathbf{x}|$$



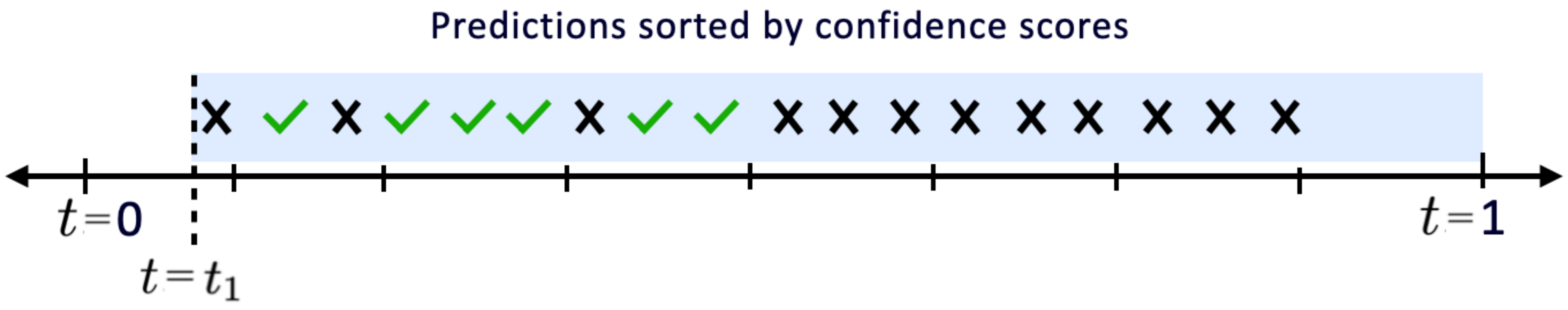
Regions defined by the confidence function

$$A_v(\hat{\mathbf{w}}, t, y) = \{\mathbf{x} \in X_v : \boxed{g(\mathbf{x}; \hat{\mathbf{w}}) \geq t}, \hat{h}(\mathbf{x}, \hat{\mathbf{w}}) = y\}$$

Auto-labeling Error estimation in these regions

$$\widehat{\mathcal{E}}_v(\hat{\mathbf{w}} | t, y) = \frac{1}{|A_v(\hat{\mathbf{w}}, t, y)|} \sum_{\mathbf{x} \in A_v(\hat{\mathbf{w}}, t, y)} \mathbb{1}\{\hat{h}(\mathbf{x}; \hat{\mathbf{w}}) \neq f^\star(\mathbf{x})\}$$

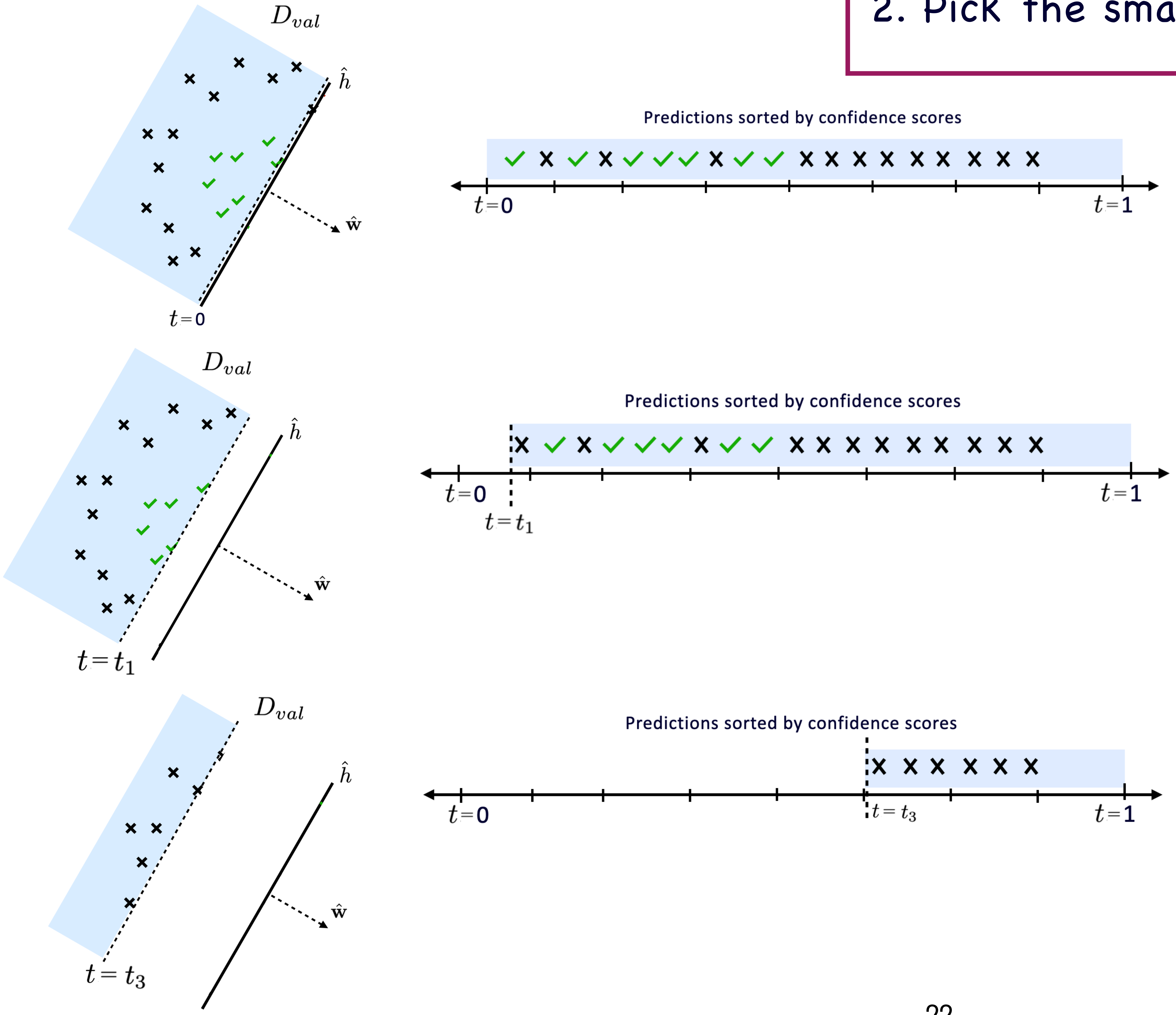
$$\blacktriangle = \frac{\# \times}{\# \checkmark + \# \times}$$

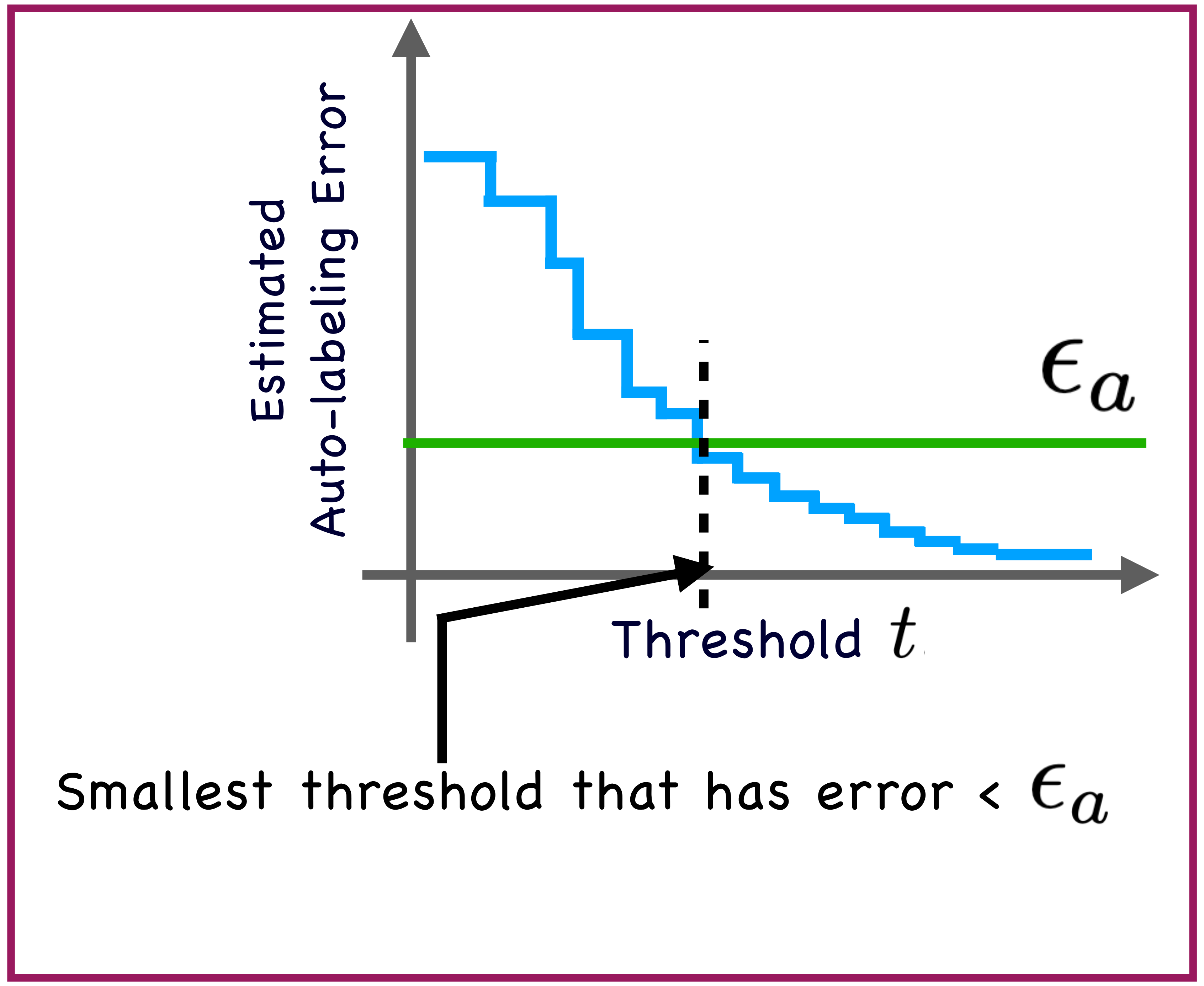


TBAL Workflow: Step 2

Find the Auto-labeling region

1. Estimate the auto-labeling error at several thresholds
2. Pick the smallest threshold having error at most ϵ_a





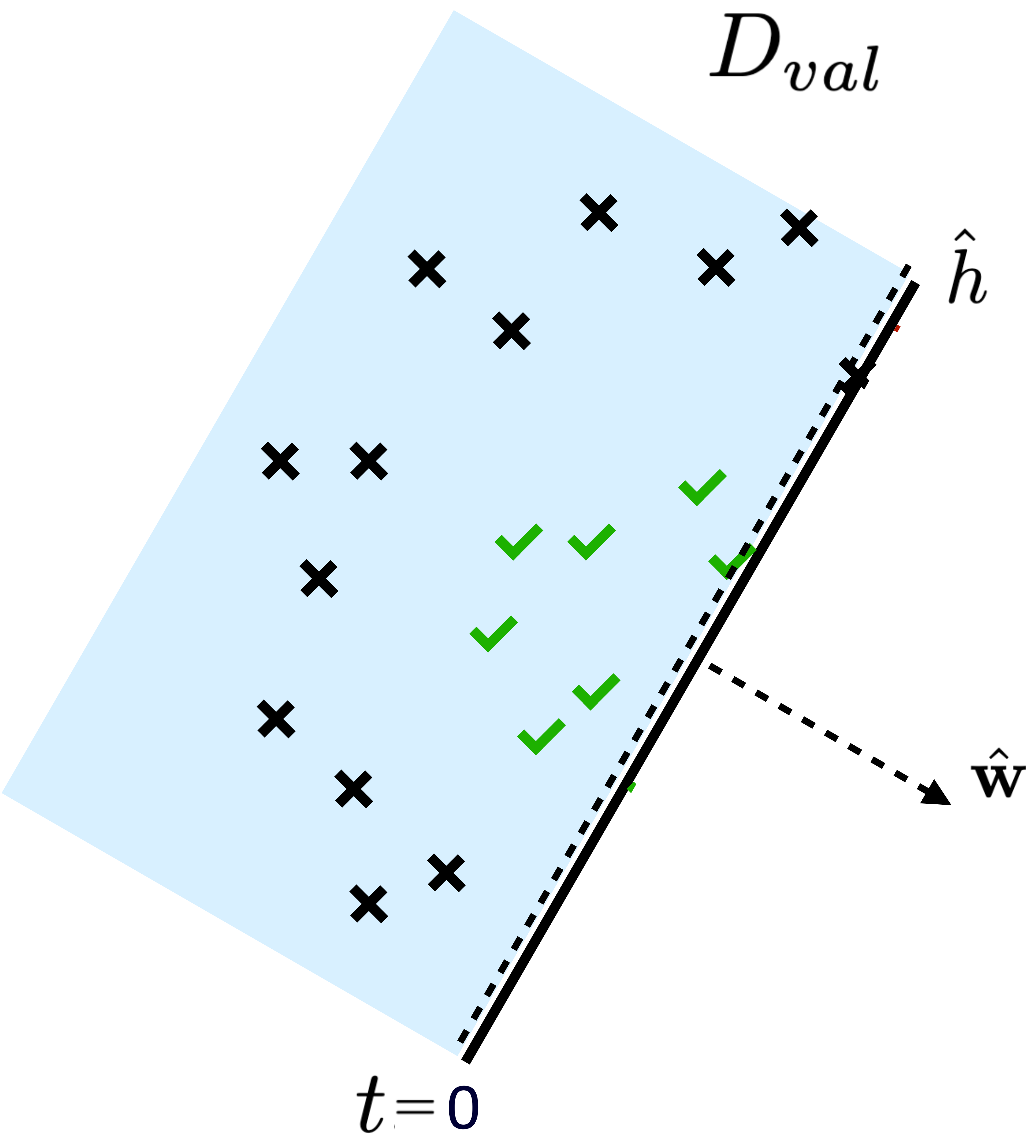
The hope

TBAL Workflow: Step 2

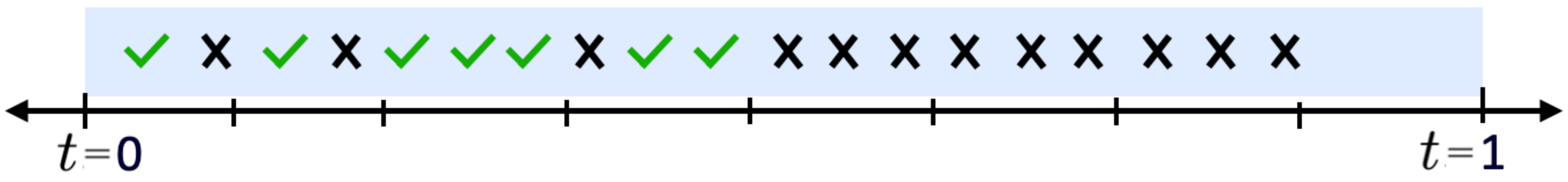
Find the Auto-labeling region

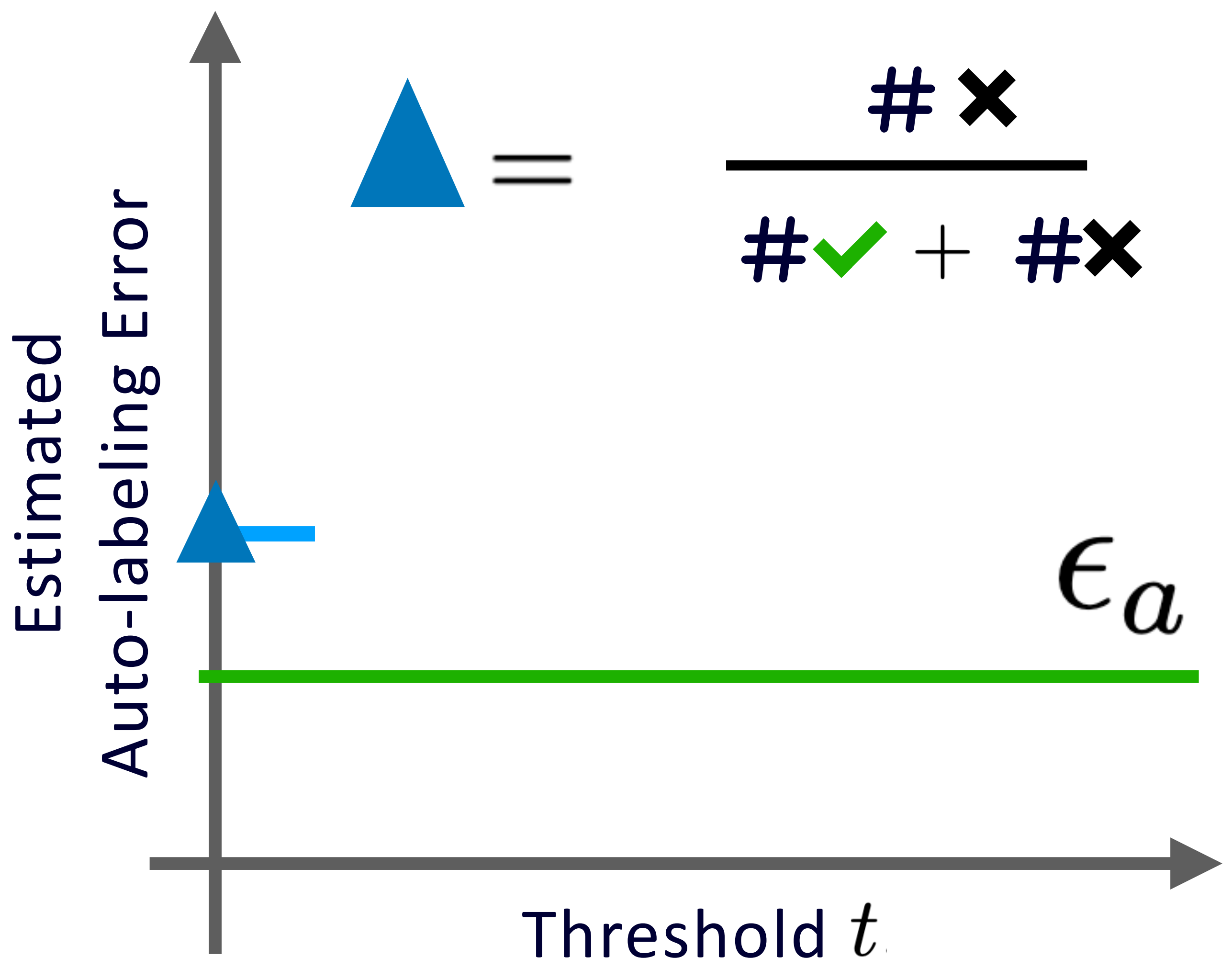
$$g(\mathbf{x}; \hat{\mathbf{w}}) = |\hat{\mathbf{w}}^T \mathbf{x}|$$

$$A_v(\hat{\mathbf{w}}, t, y) = \{\mathbf{x} \in X_v : g(\mathbf{x}; \hat{\mathbf{w}}) \geq t, \hat{h}(\mathbf{x}, \hat{\mathbf{w}}) = y\}$$



Predictions sorted by confidence scores



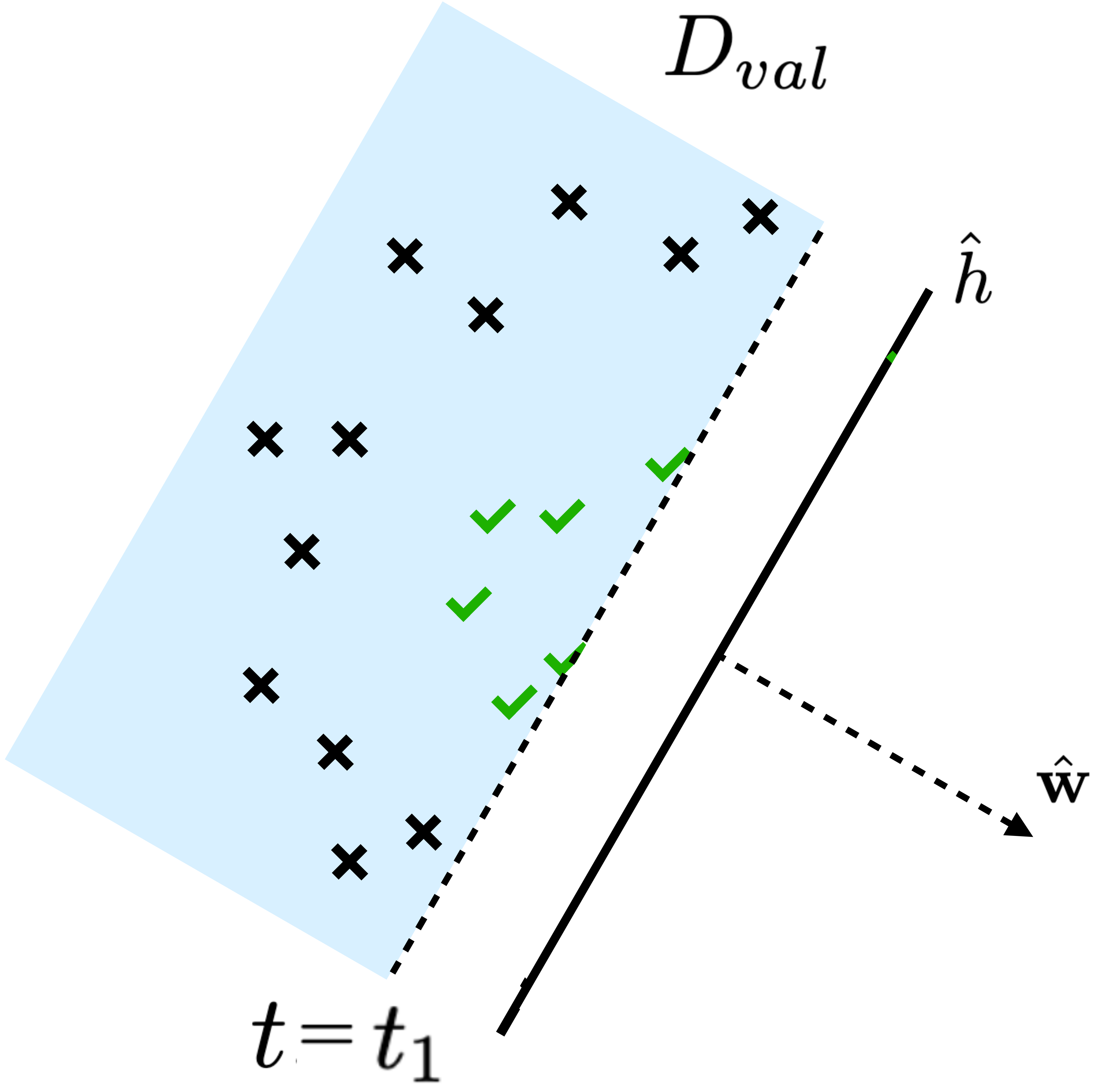


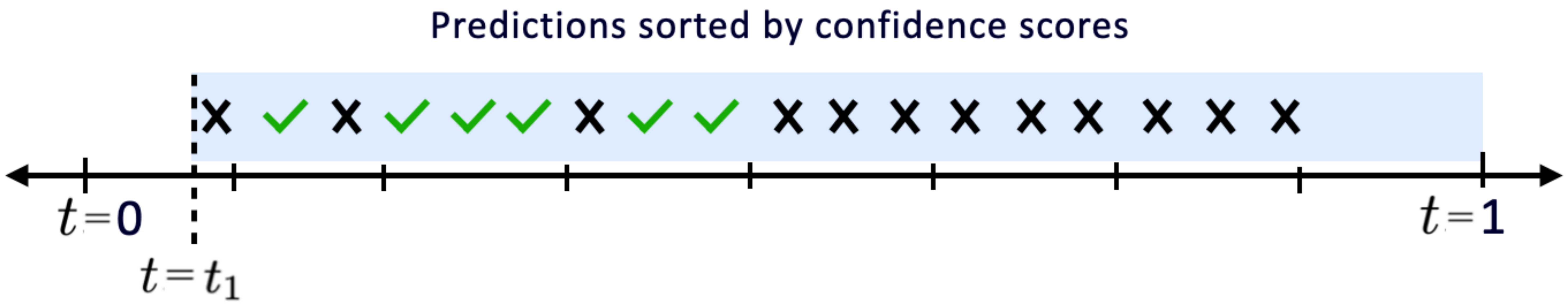
TBAL Workflow: Step 2

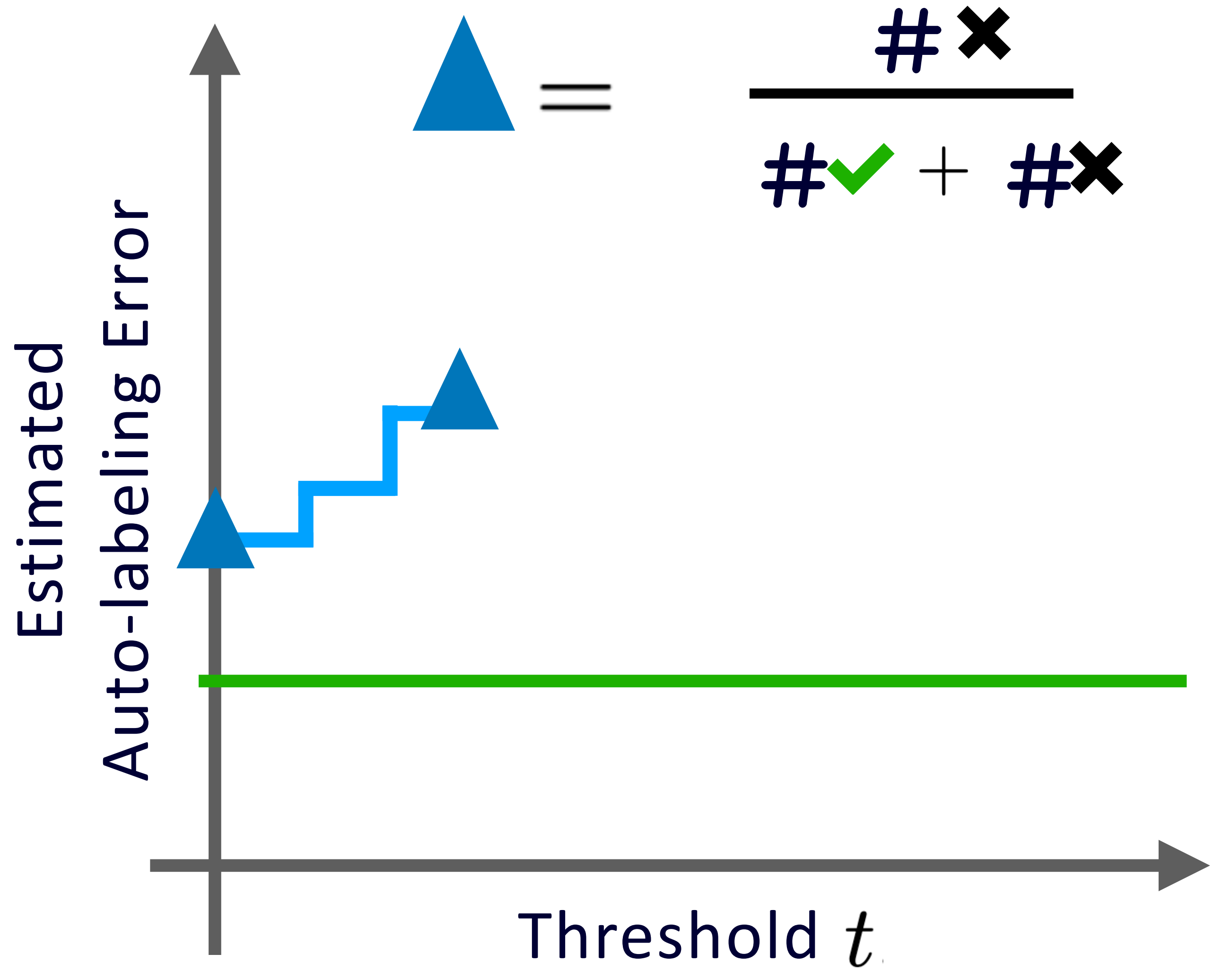
Find the Auto-labeling region

$$g(\mathbf{x}; \hat{\mathbf{w}}) = |\hat{\mathbf{w}}^T \mathbf{x}|$$

$$A_v(\hat{\mathbf{w}}, t, y) = \{\mathbf{x} \in X_v : g(\mathbf{x}; \hat{\mathbf{w}}) \geq t, \hat{h}(\mathbf{x}, \hat{\mathbf{w}}) = y\}$$





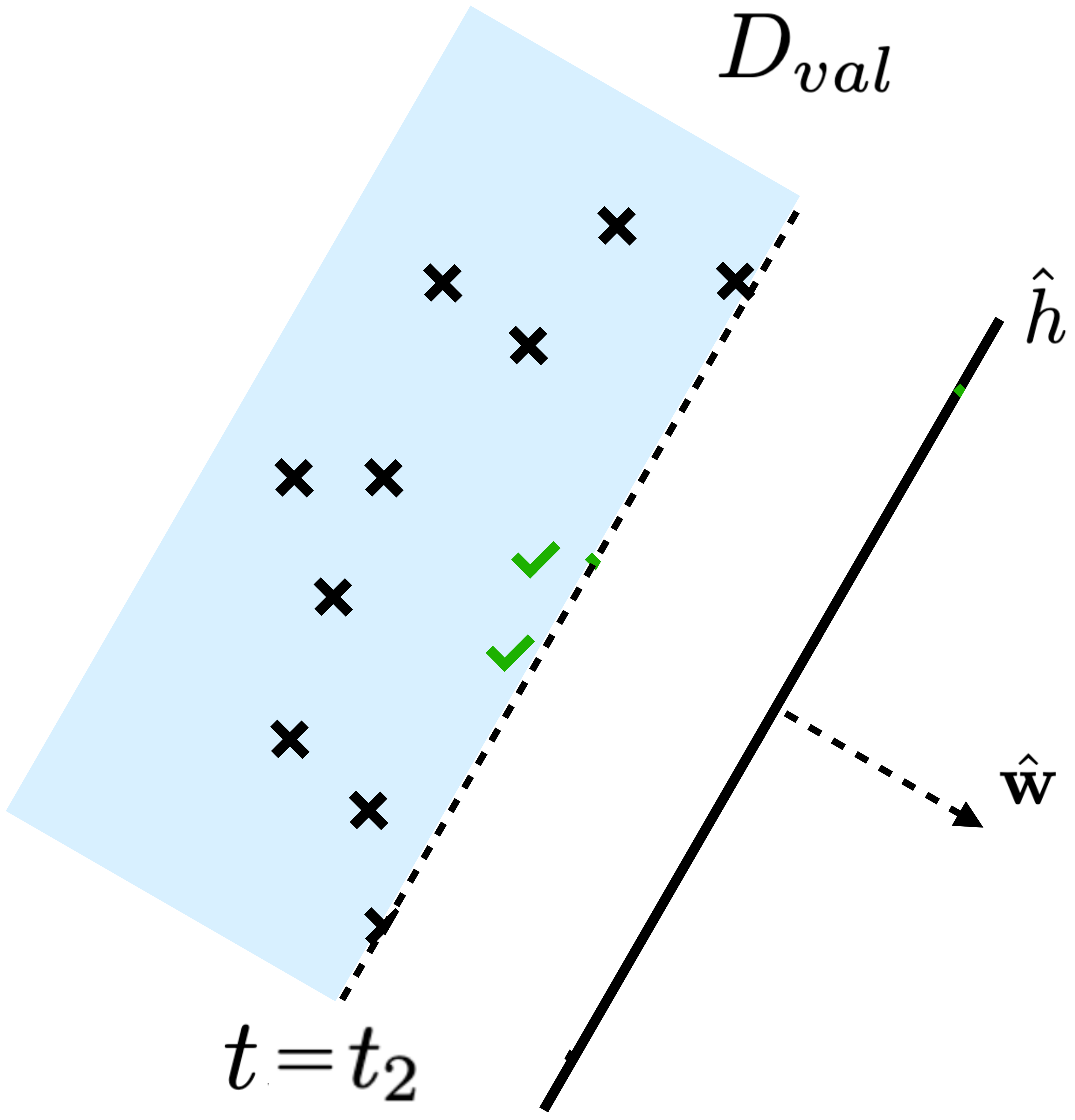


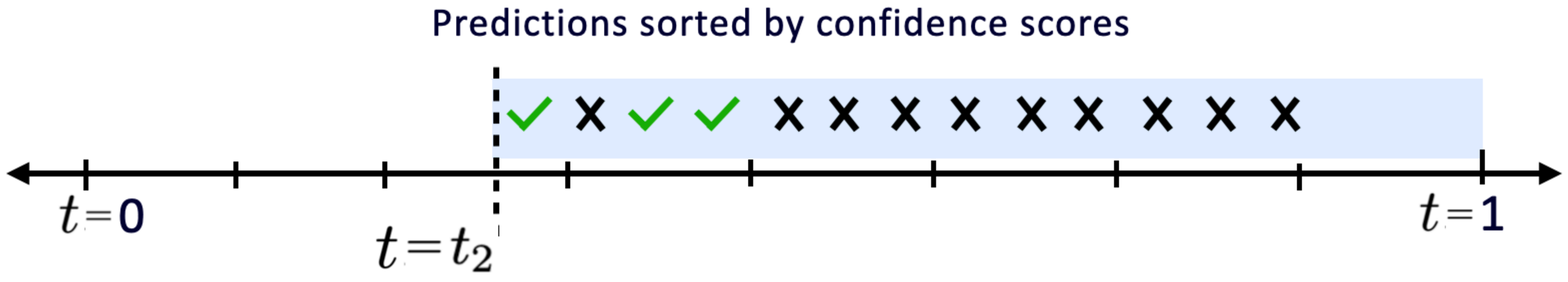
TBAL Workflow: Step 2

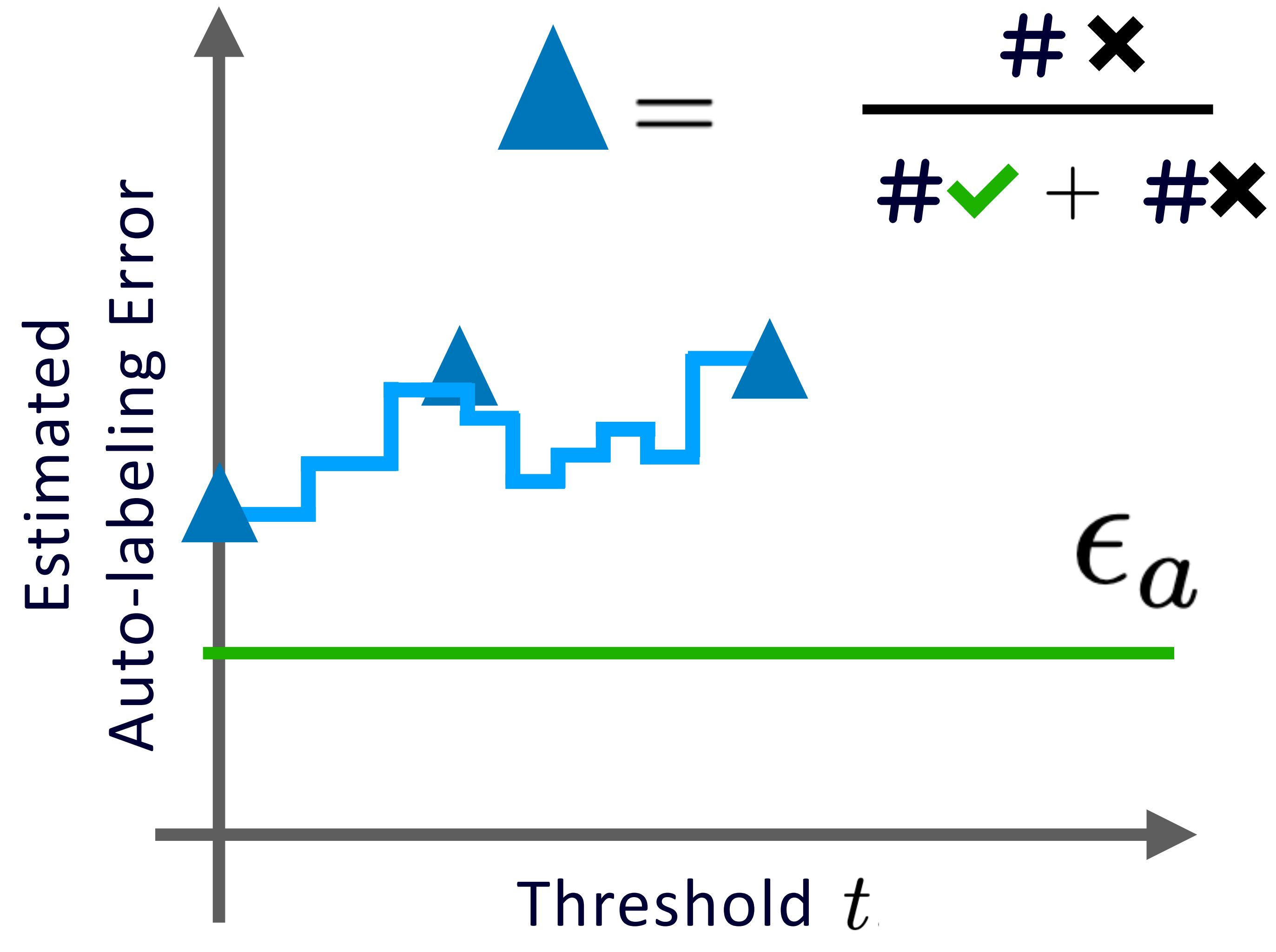
Find the Auto-labeling region

$$g(\mathbf{x}; \hat{\mathbf{w}}) = |\hat{\mathbf{w}}^T \mathbf{x}|$$

$$A_v(\hat{\mathbf{w}}, t, y) = \{\mathbf{x} \in X_v : g(\mathbf{x}; \hat{\mathbf{w}}) \geq t, \hat{h}(\mathbf{x}, \hat{\mathbf{w}}) = y\}$$





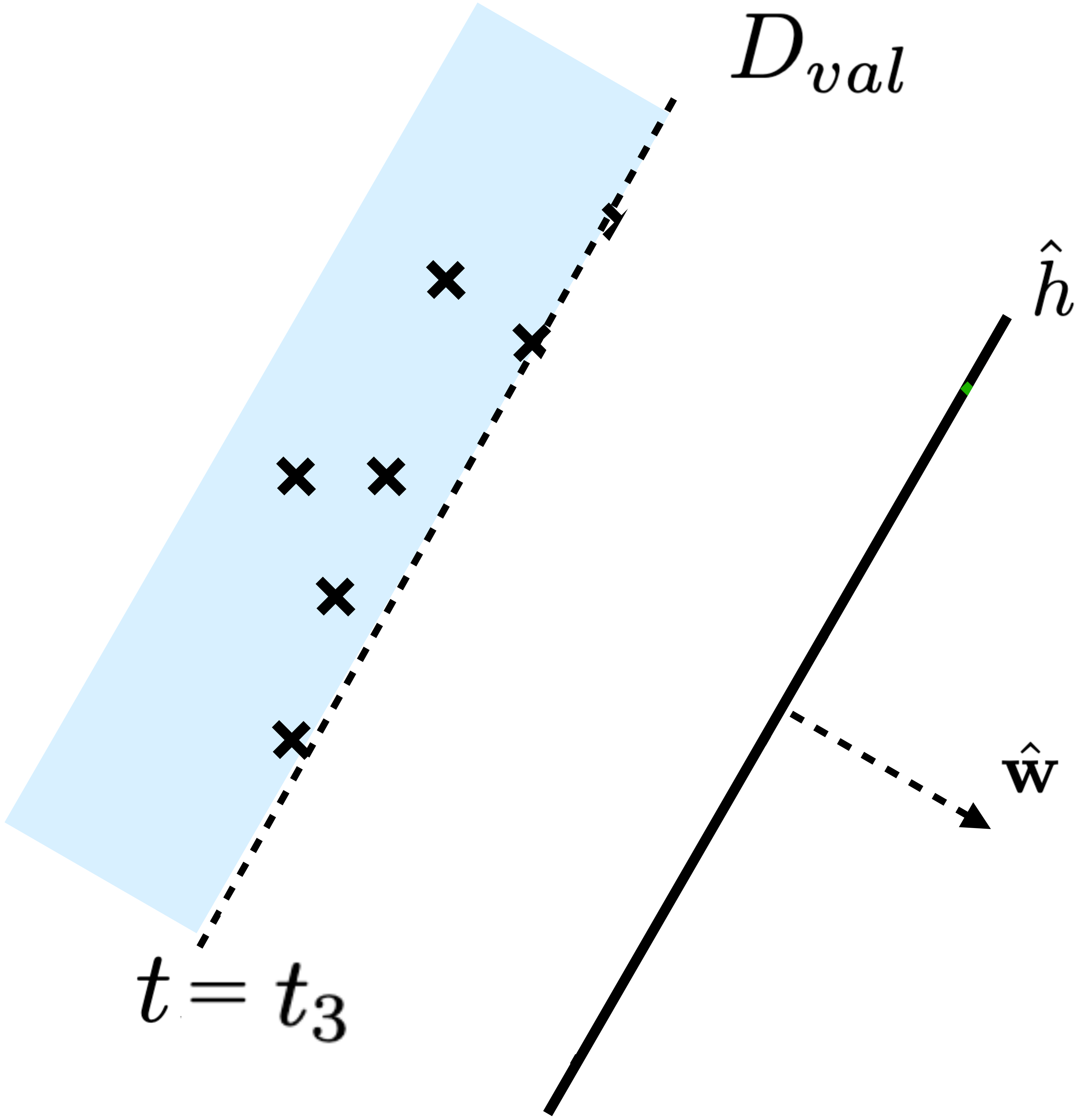


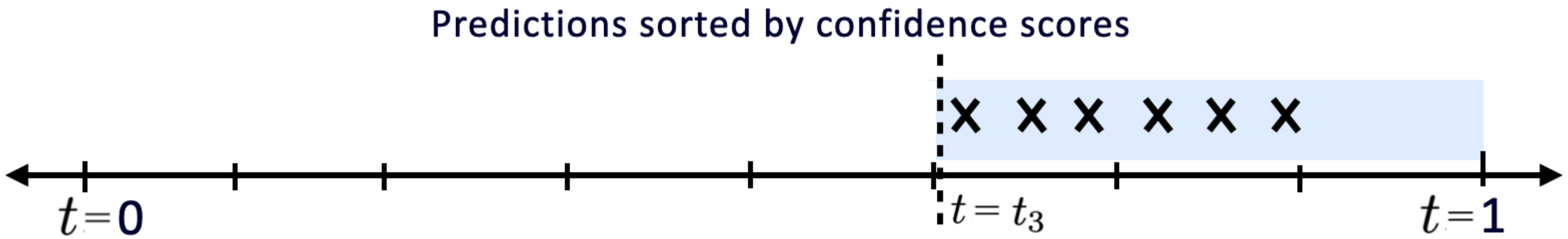
TBAL Workflow: Step 2

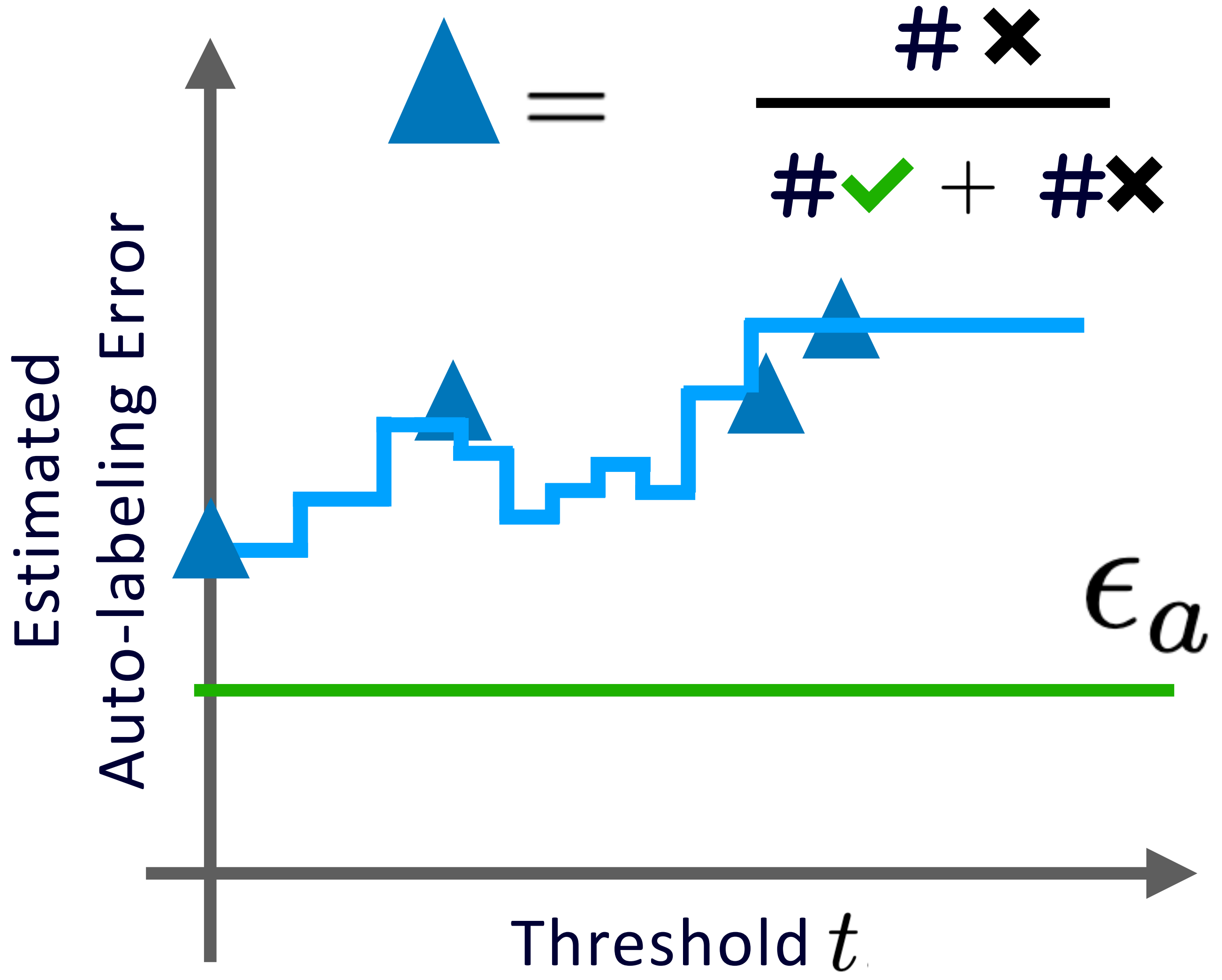
Find the Auto-labeling region

$$g(\mathbf{x}; \hat{\mathbf{w}}) = |\hat{\mathbf{w}}^T \mathbf{x}|$$

$$A_v(\hat{\mathbf{w}}, t, y) = \{\mathbf{x} \in X_v : g(\mathbf{x}; \hat{\mathbf{w}}) \geq t, \hat{h}(\mathbf{x}, \hat{\mathbf{w}}) = y\}$$





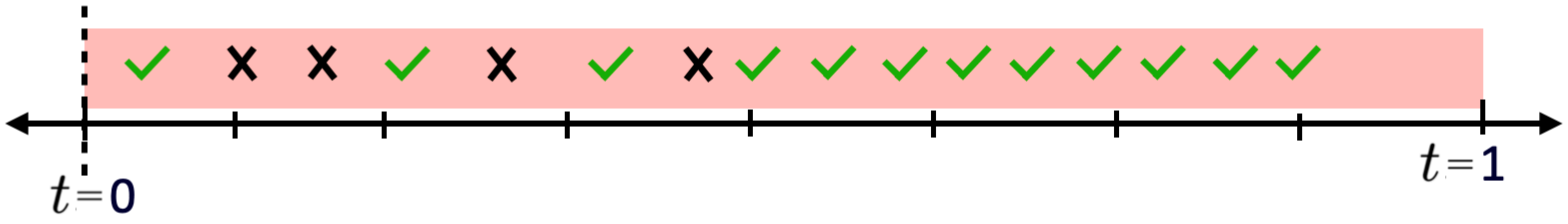


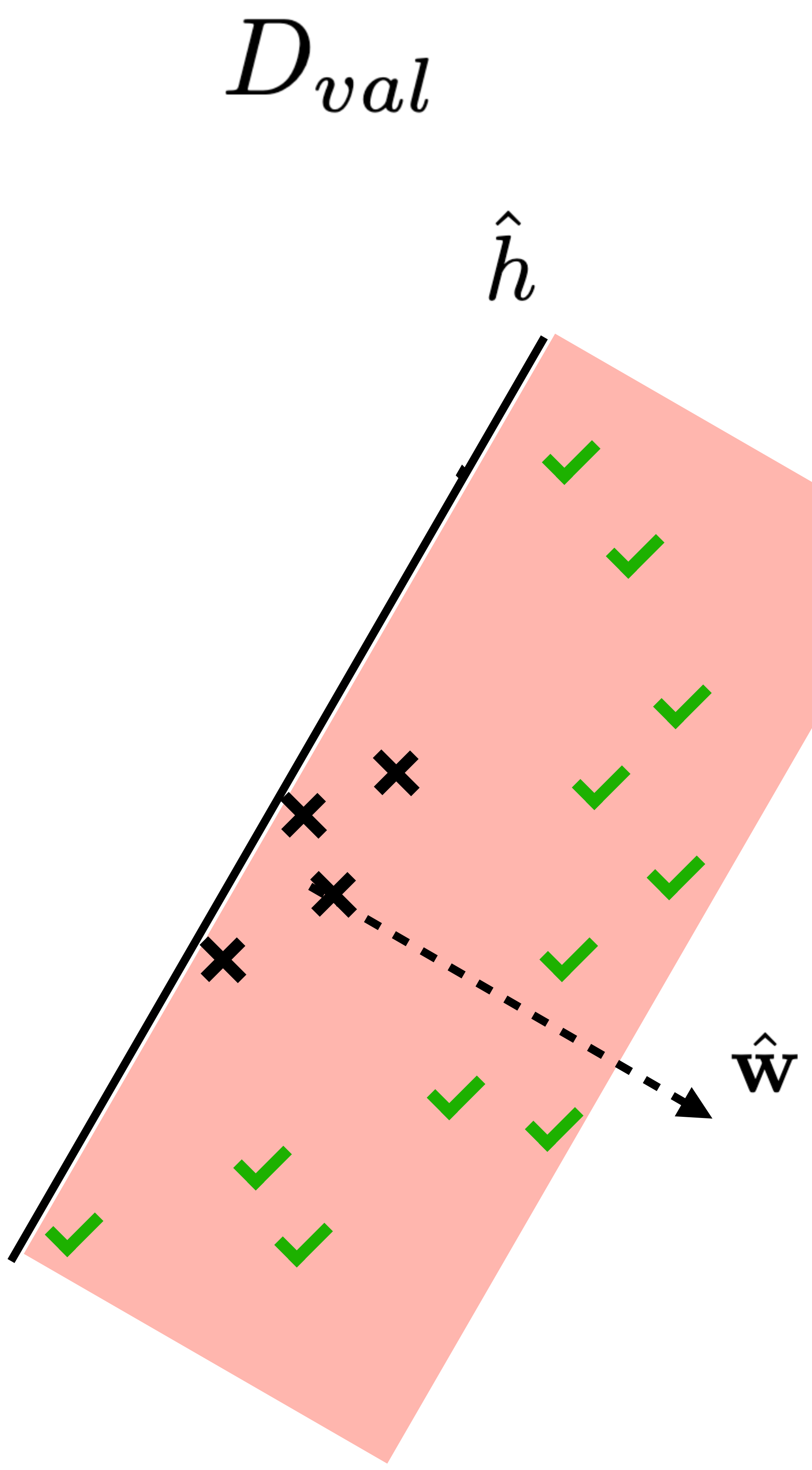
Cannot find a threshold on this side.

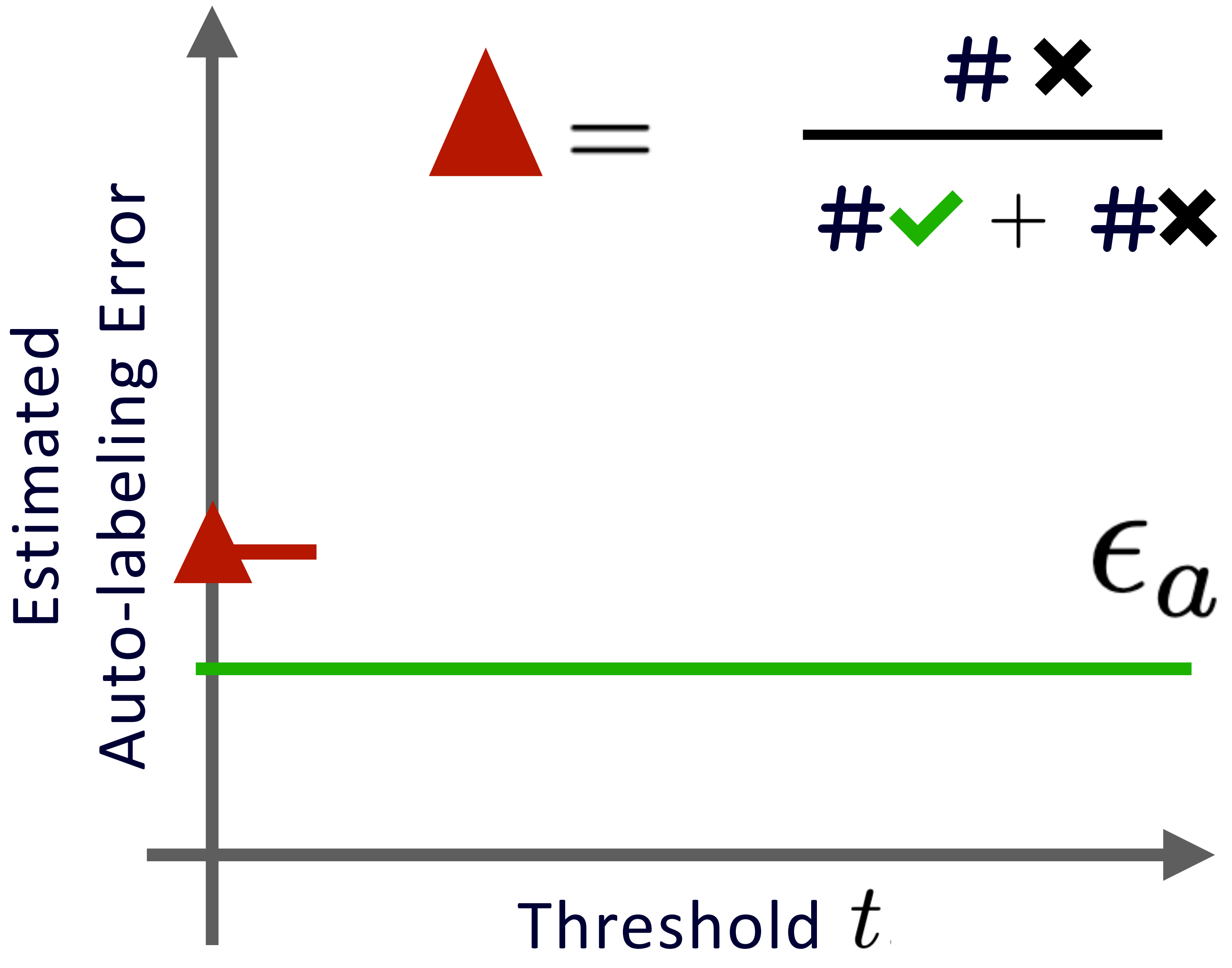
TBAL Workflow: Step 2
Find the Auto-labeling region

$$g(\mathbf{x}; \hat{\mathbf{w}}) = |\hat{\mathbf{w}}^T \mathbf{x}|$$

$$A_v(\hat{\mathbf{w}}, t, y) = \{\mathbf{x} \in X_v : g(\mathbf{x}; \hat{\mathbf{w}}) \geq t, \hat{h}(\mathbf{x}, \hat{\mathbf{w}}) = y\}$$





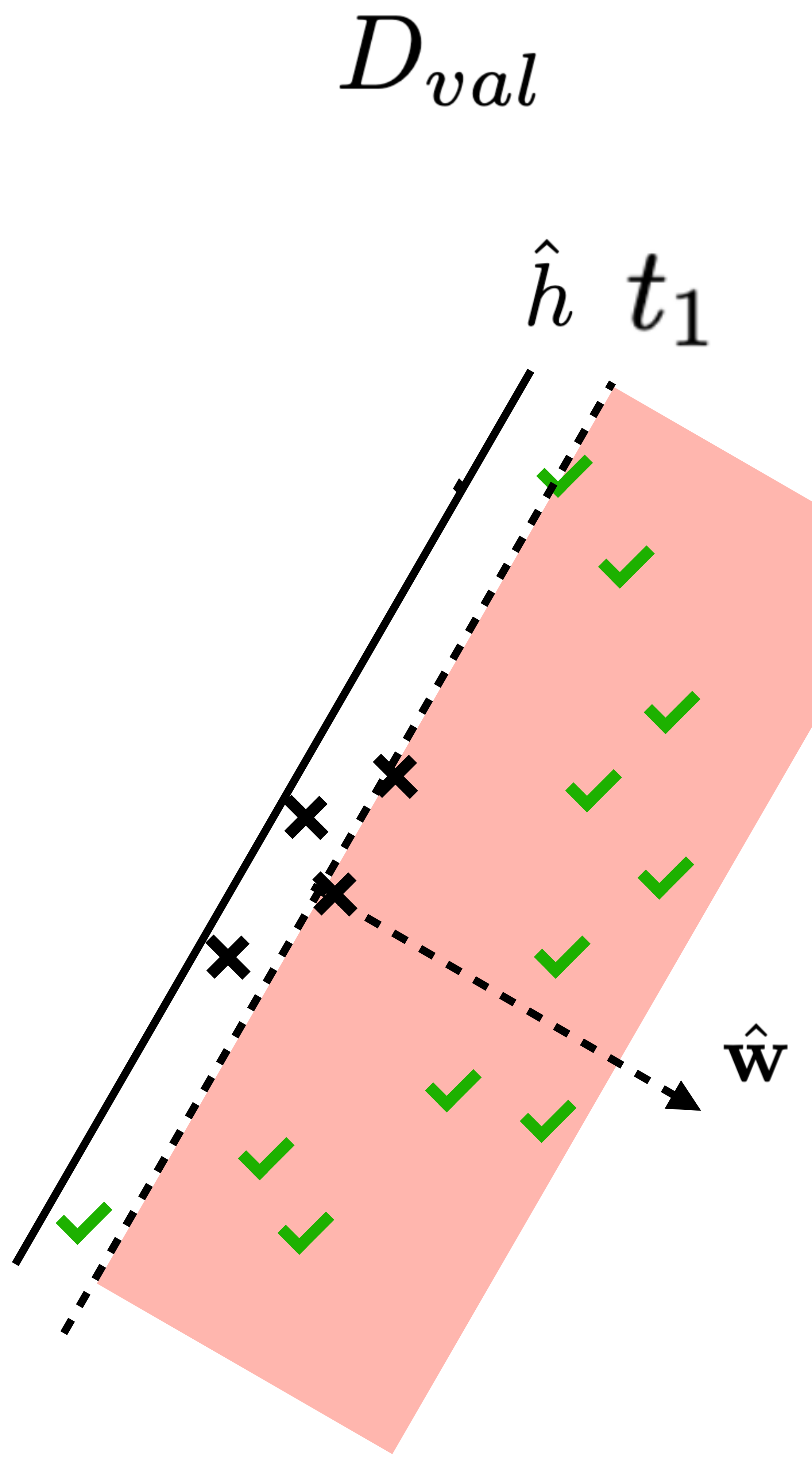


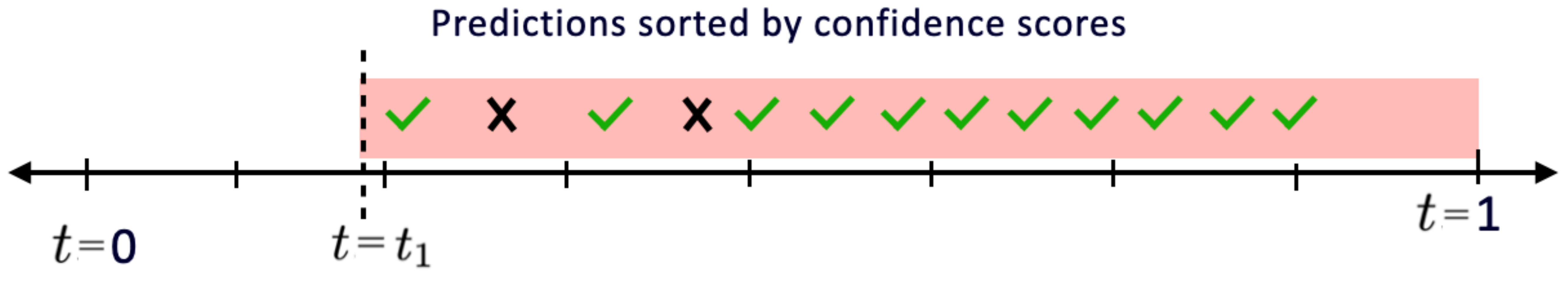
TBAL Workflow: Step 2

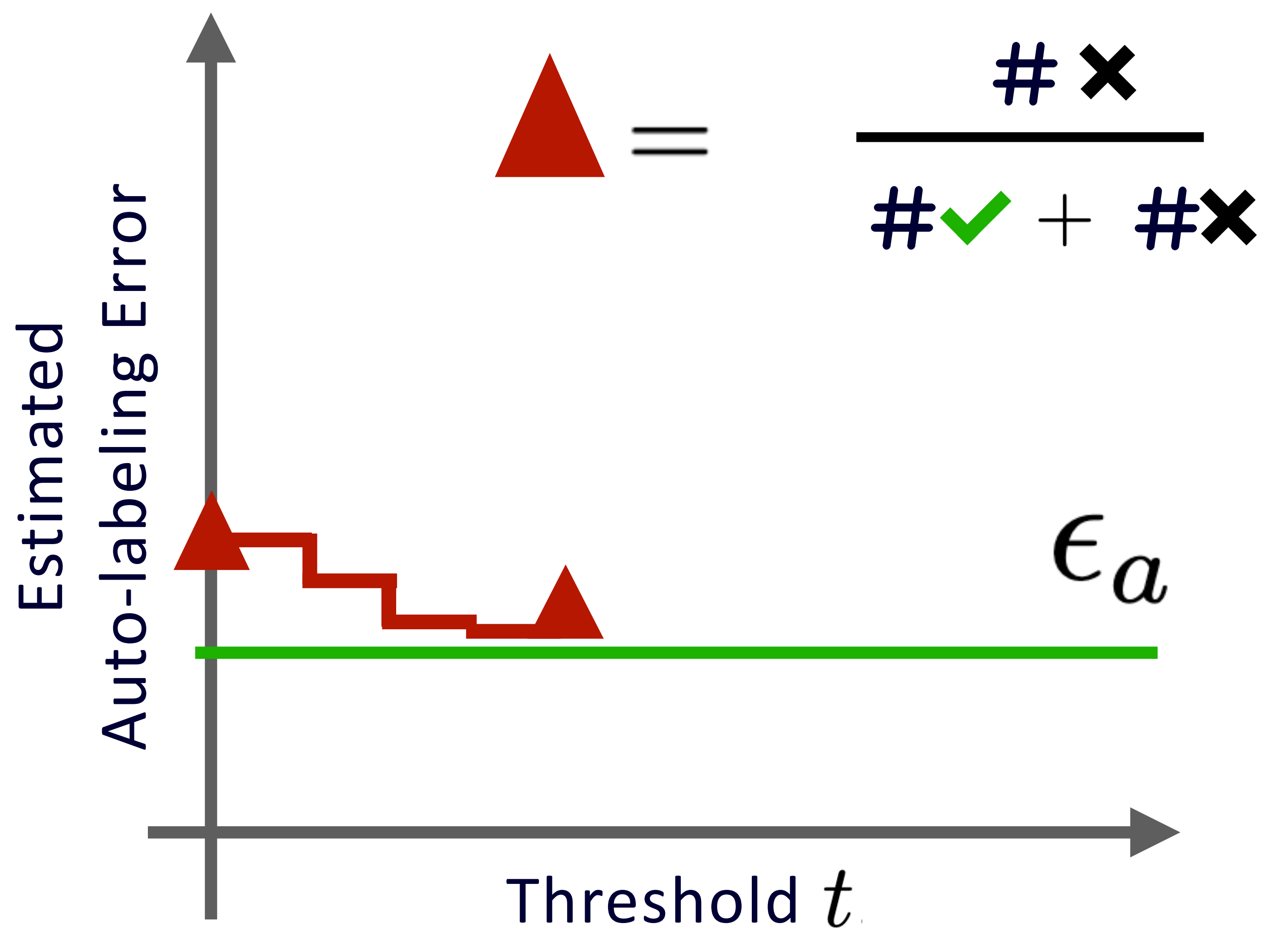
Find the Auto-labeling region

$$g(\mathbf{x}; \hat{\mathbf{w}}) = |\hat{\mathbf{w}}^T \mathbf{x}|$$

$$A_v(\hat{\mathbf{w}}, t, y) = \{\mathbf{x} \in X_v : g(\mathbf{x}; \hat{\mathbf{w}}) \geq t, \hat{h}(\mathbf{x}, \hat{\mathbf{w}}) = y\}$$







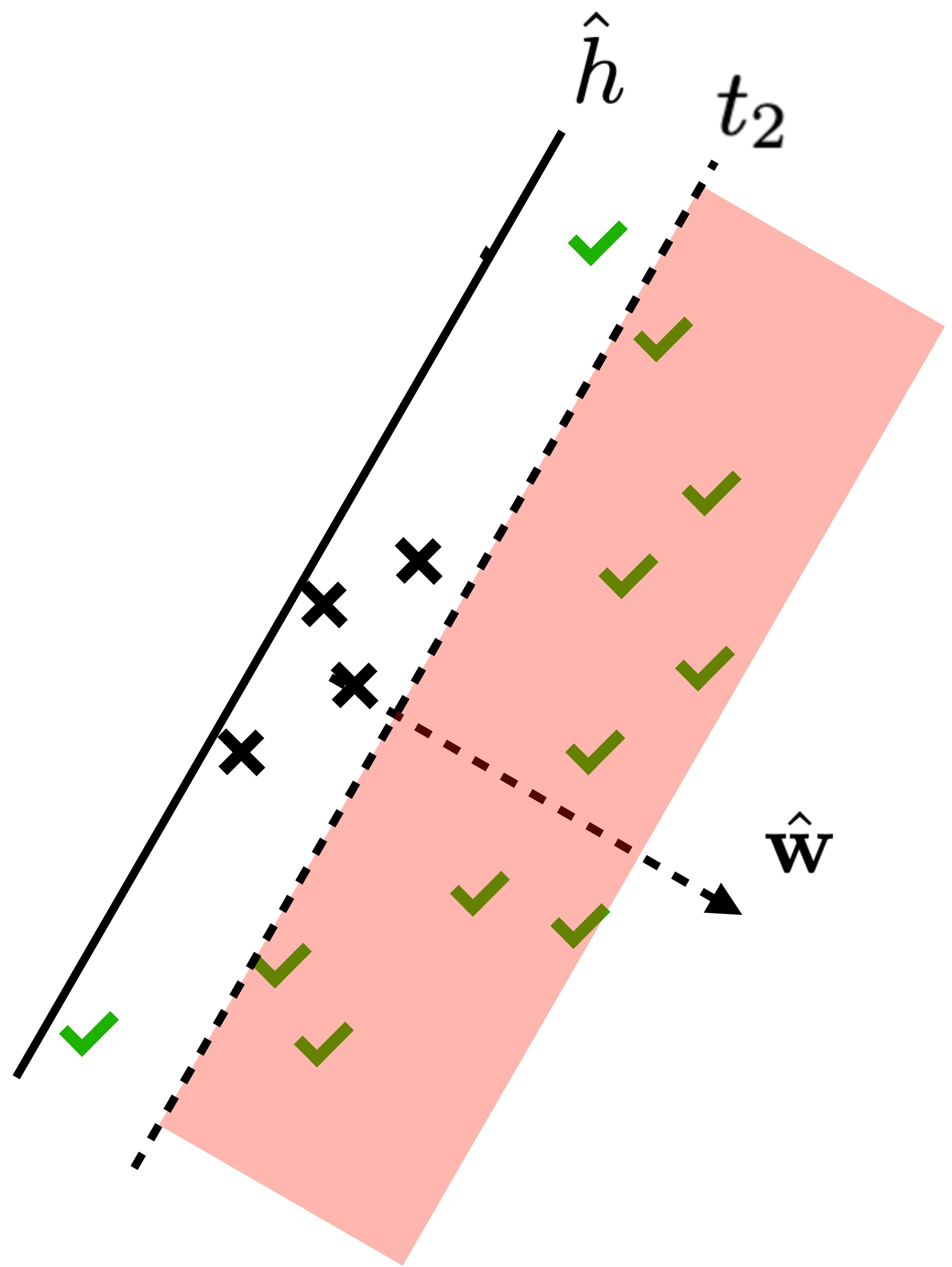
TBAL Workflow: Step 2

Find the Auto-labeling region

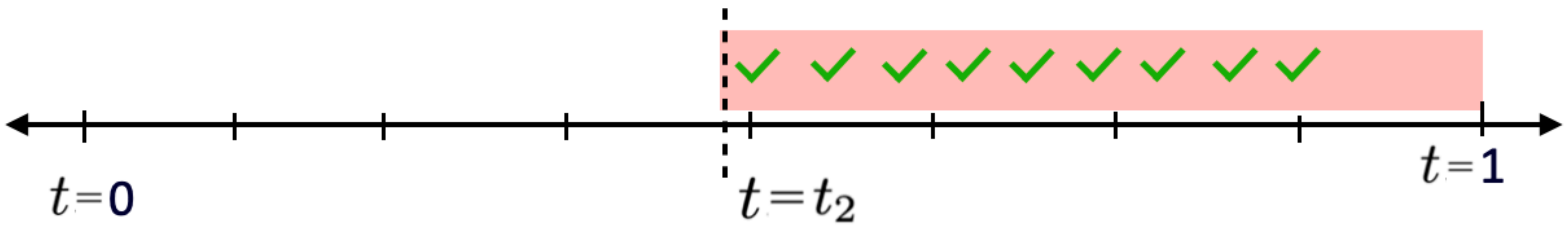
$$g(\mathbf{x}; \hat{\mathbf{w}}) = |\hat{\mathbf{w}}^T \mathbf{x}|$$

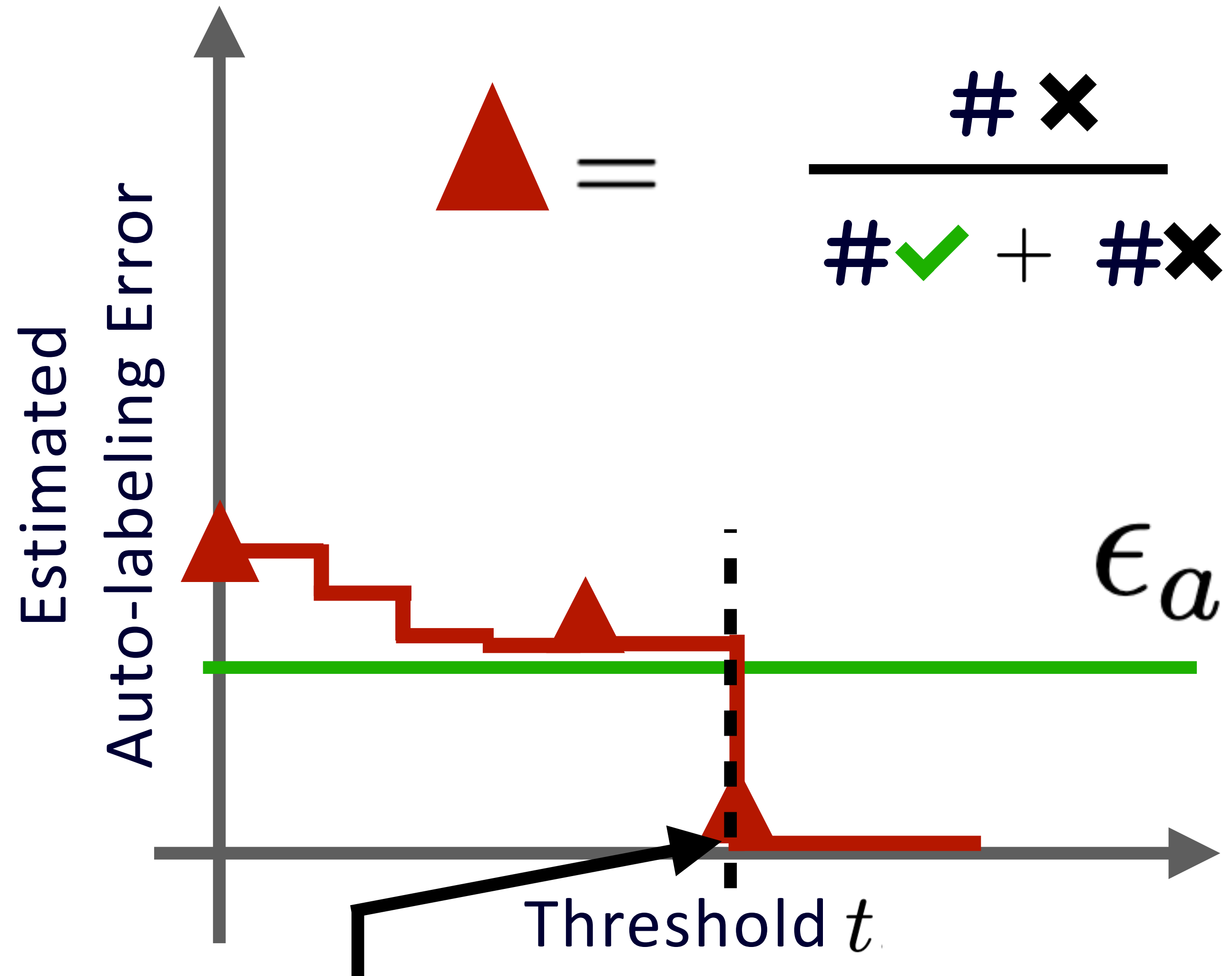
$$A_v(\hat{\mathbf{w}}, t, y) = \{\mathbf{x} \in X_v : g(\mathbf{x}; \hat{\mathbf{w}}) \geq t, \hat{h}(\mathbf{x}, \hat{\mathbf{w}}) = y\}$$

D_{val}



Predictions sorted by confidence scores



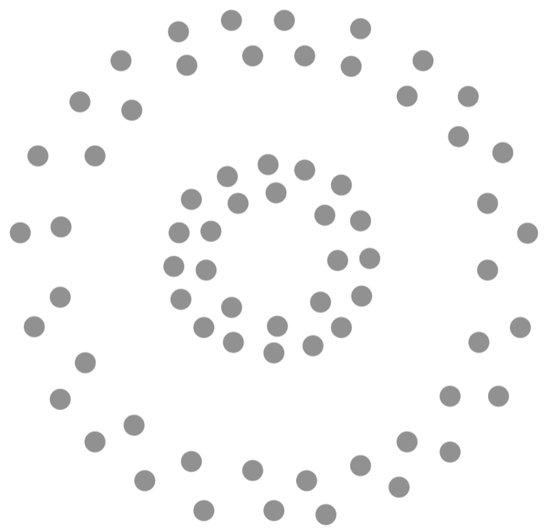


We found a threshold that has error < ϵ_a

Threshold-based Auto-labeling Workflow(TBAL)

Input

Unlabeled Data
i.i.d from space $\mathcal{X}$



Auto-labeling error tolerance
ϵ_a

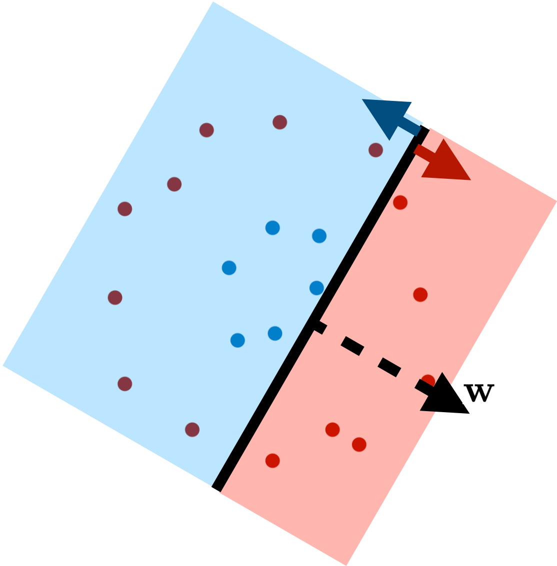
0 **Bootstrap**

Create validation and initial training sets

Model Class
$\mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$
$h(\mathbf{x}; \mathbf{w}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x})$

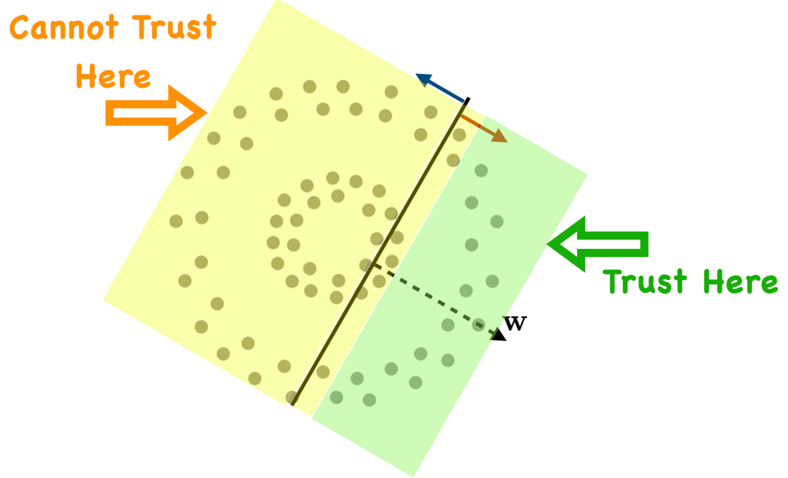
Confidence Function
$g : \mathcal{X} \mapsto T \subseteq \mathbb{R}^+$
$g(\mathbf{x}; \mathbf{w}) = |\mathbf{w}^T\mathbf{x}|$

1 Learn a model **w** using training set

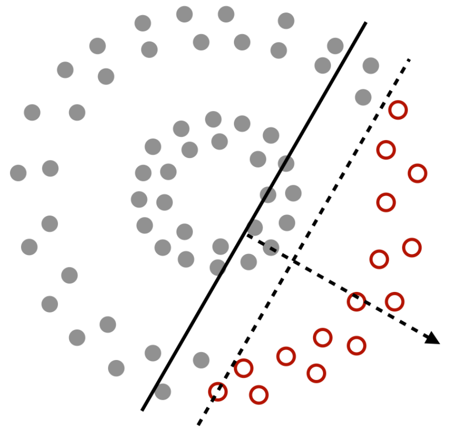


Empirical Risk Minimizer from
$\mathcal{H}$

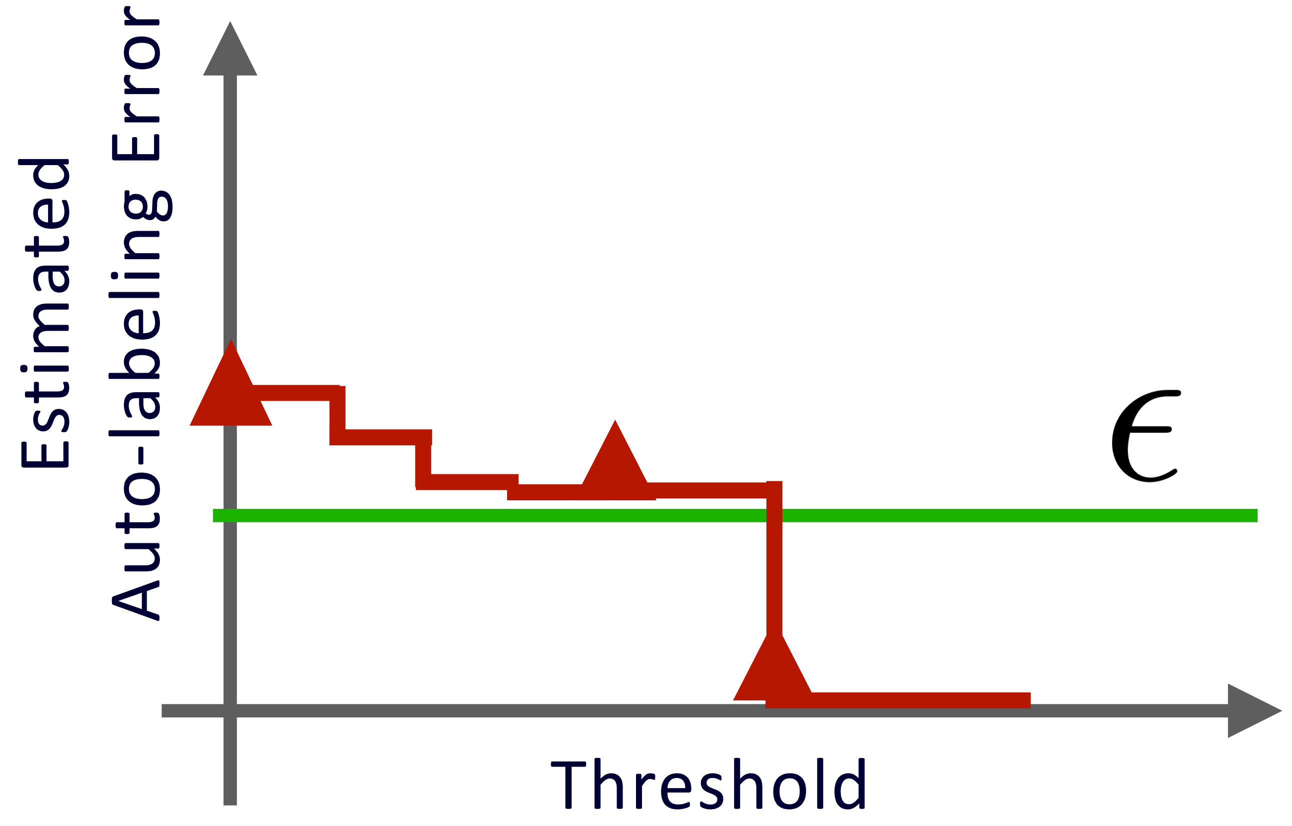
2 Find auto-labeling region, where the model can be trusted



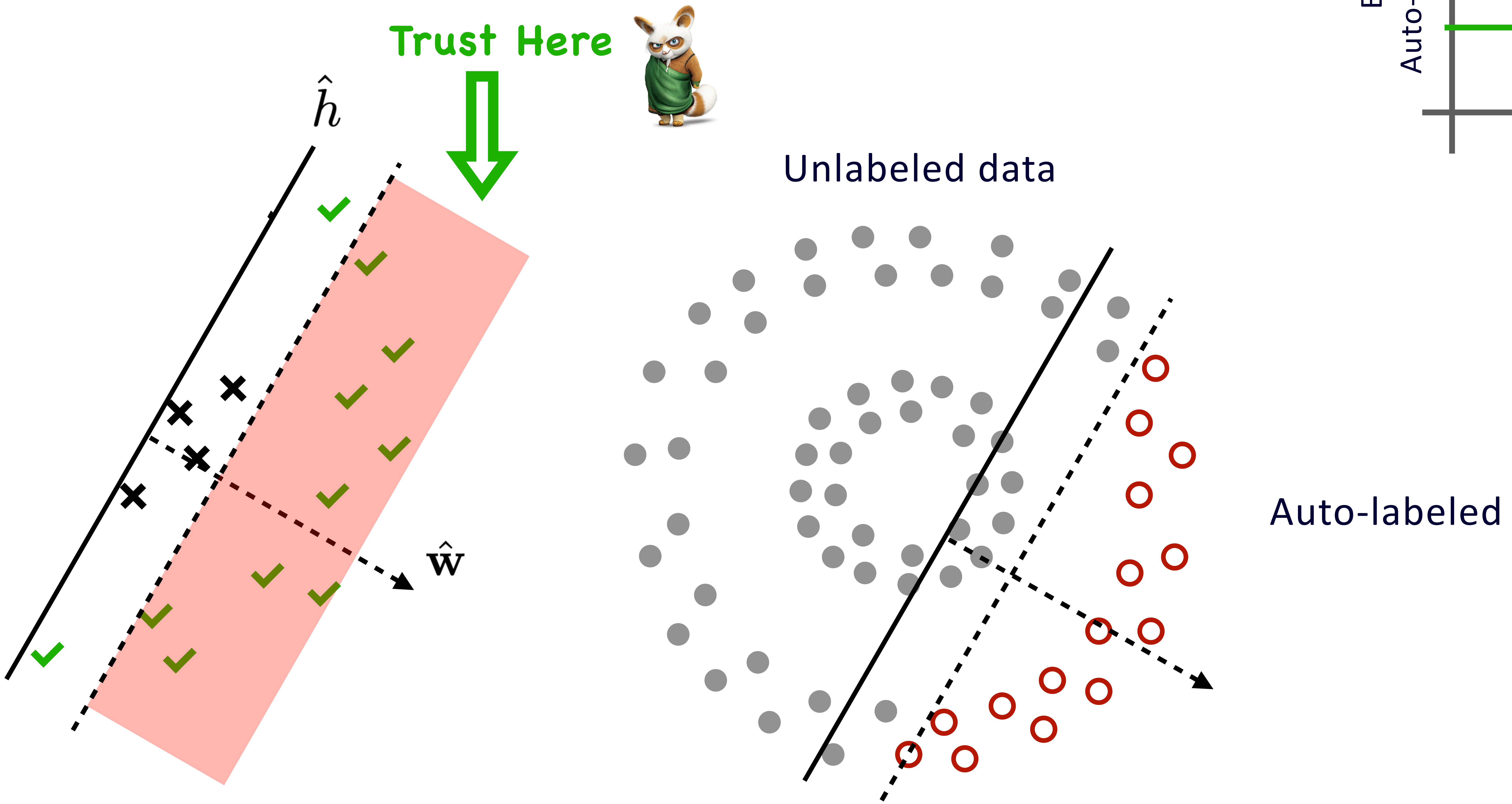
3 Auto-label points in the identified region



TBAL Workflow: Step 3 Auto-label points in the identified region



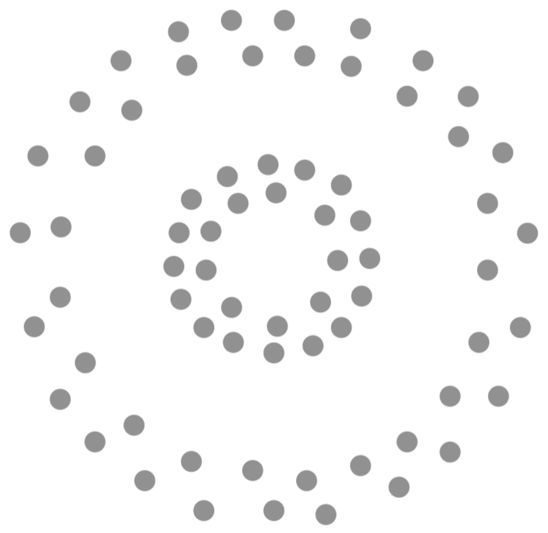
We found a threshold that has error < ϵ_a



Threshold-based Auto-labeling Workflow(TBAL)

Input

Unlabeled Data
i.i.d from space $\mathcal{X}$



Auto-labeling error tolerance

ϵ_a

0 **Bootstrap**

Create validation and initial training sets

Training Set **Validation Set**

Model Class

$\mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$

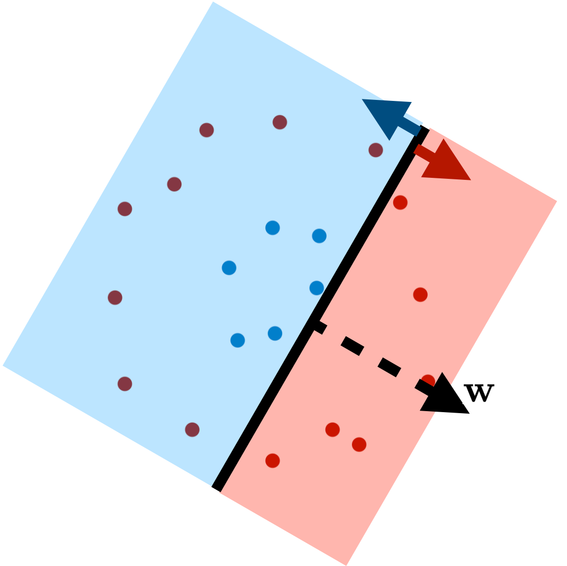
$h(\mathbf{x}; \mathbf{w}) = \mathrm{sign}(\mathbf{w}^T \mathbf{x})$

Confidence Function

$g : \mathcal{X} \mapsto T \subseteq \mathbb{R}^+$

$g(\mathbf{x}; \mathbf{w}) = |\mathbf{w}^T \mathbf{x}|$

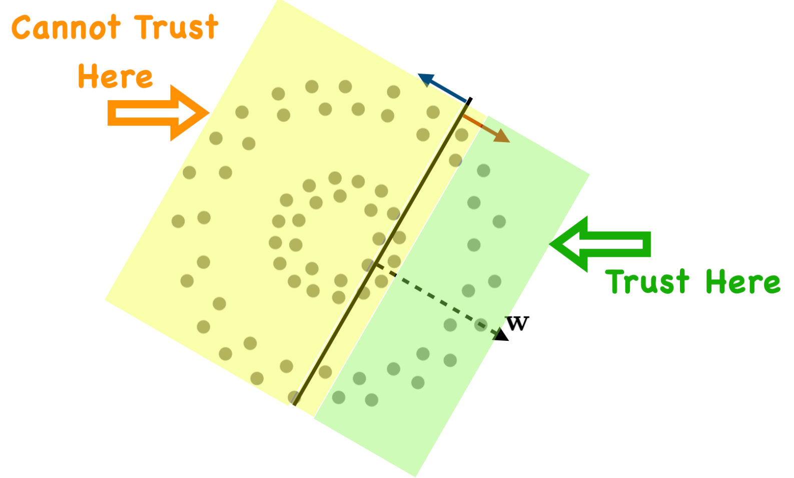
1 Learn a model **w** using training set



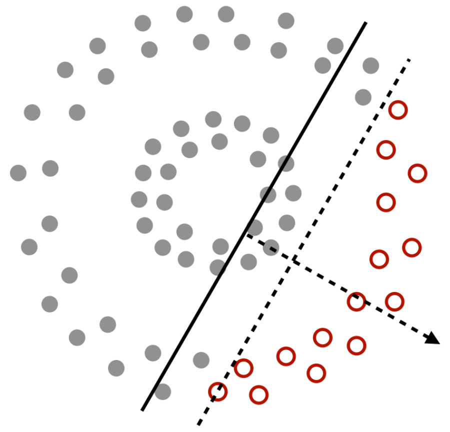
Empirical Risk Minimizer from

$\mathcal{H}$

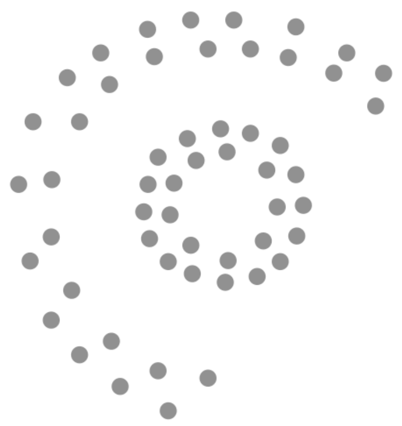
2 Find auto-labeling region, where the model can be trusted



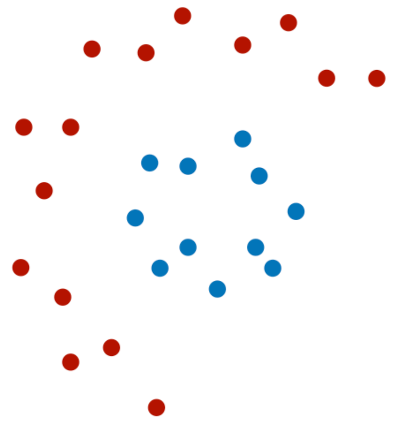
3 Auto-label points in the identified region



4 Remove points in auto-labeling region



Unlabeled Data

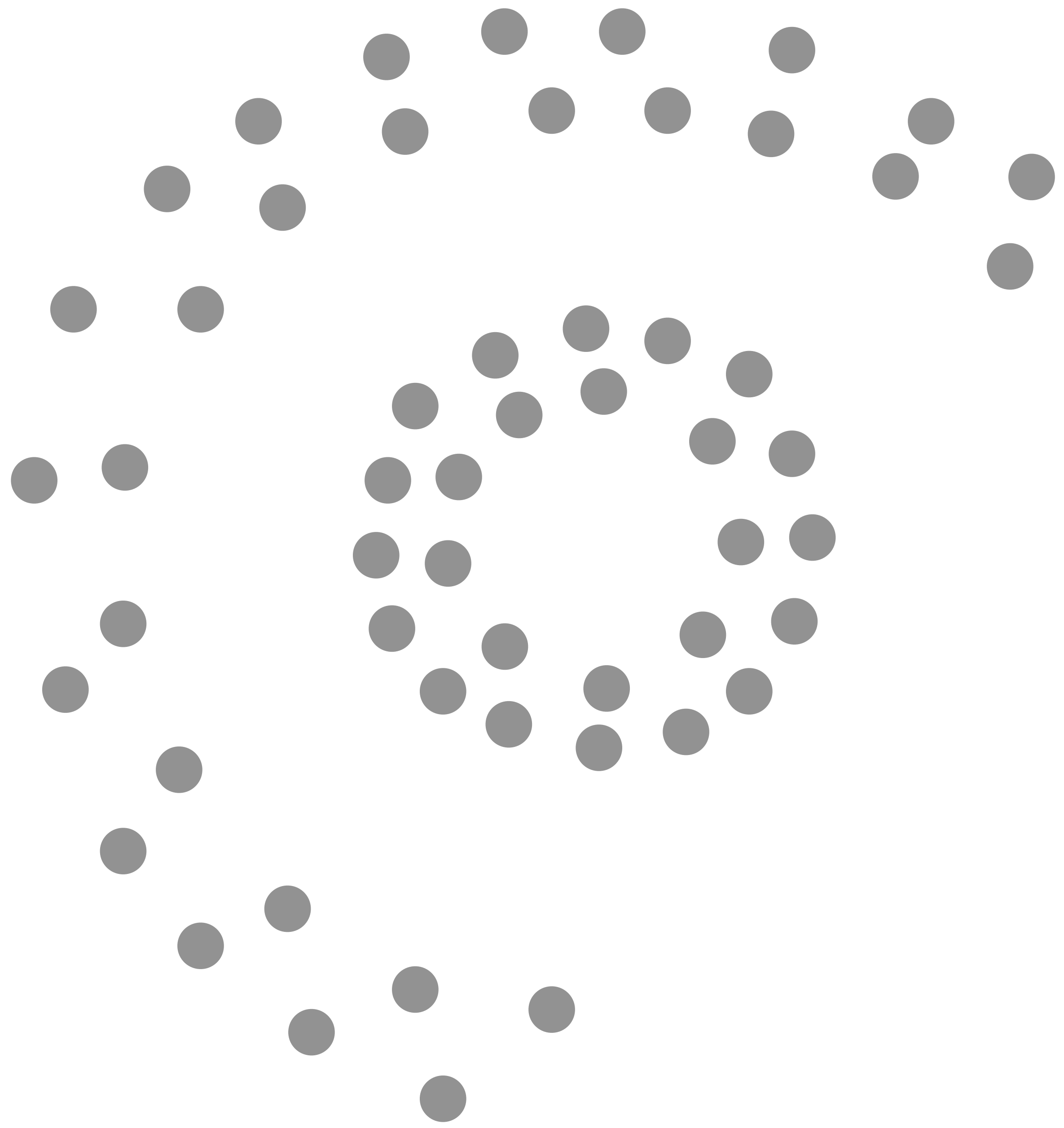


Validation Data

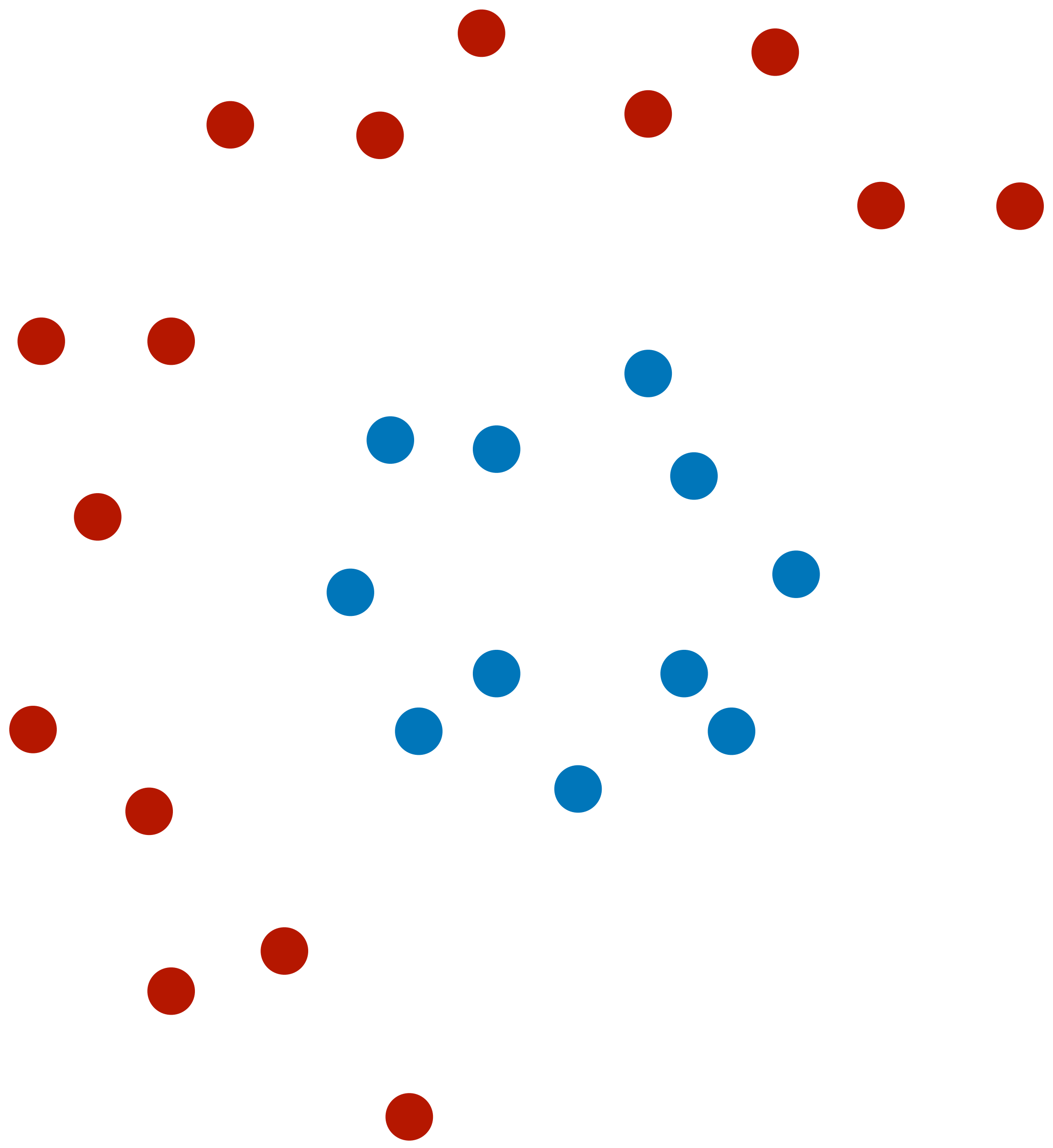
TBAL Workflow: Step 4 Prepare for the next round

Remove auto-labeled points
from the pool.

Remove points from the validation set
Falling in the auto-labeling region.



Remaining unlabeled data

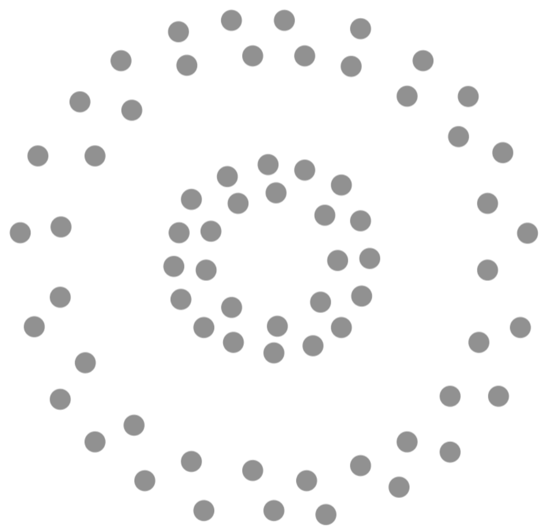


Remaining validation data

Threshold-based Auto-labeling Workflow(TBAL)

Input

Unlabeled Data
i.i.d from space $\mathcal{X}$



Auto-labeling error tolerance
ϵ_a

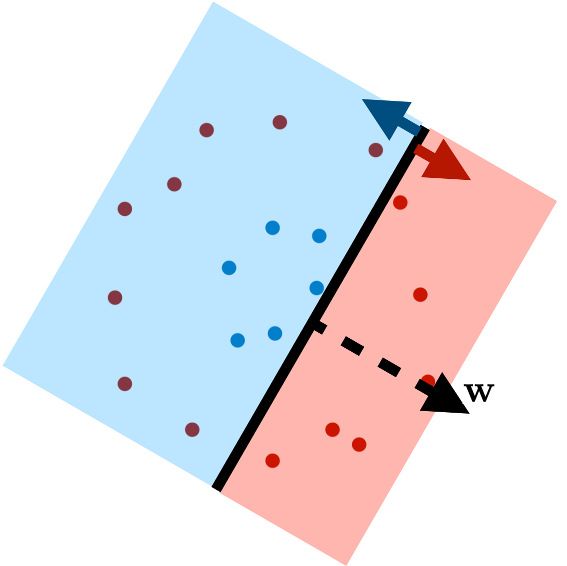
0 **Bootstrap**

Create validation and initial training sets

Model Class
$\mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$
$h(\mathbf{x}; \mathbf{w}) = \text{sign}(\mathbf{w}^T \mathbf{x})$

Confidence Function
$g : \mathcal{X} \mapsto T \subseteq \mathbb{R}^+$
$g(\mathbf{x}; \mathbf{w}) = |\mathbf{w}^T \mathbf{x}|$

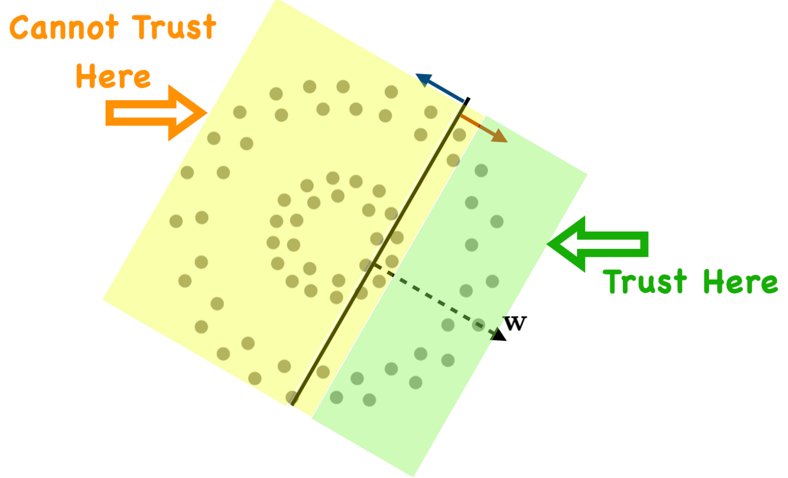
1 Learn a model **w** using training set



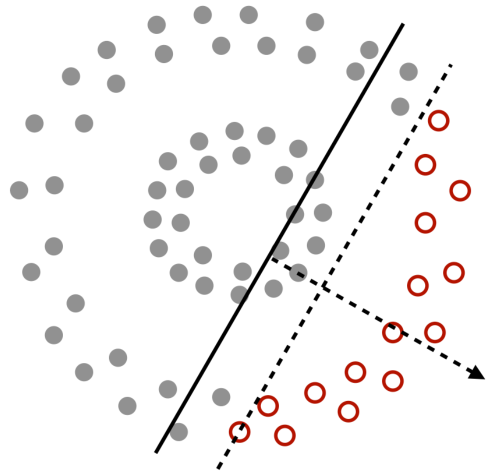
Empirical Risk Minimizer from

$\mathcal{H}$

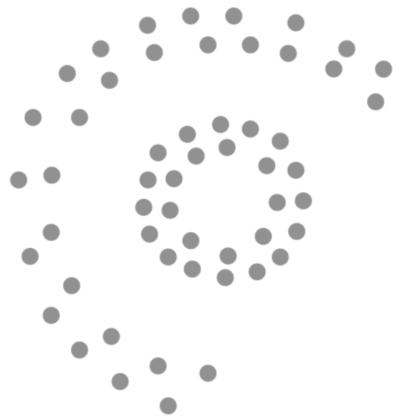
2 Find auto-labeling region, where the model can be trusted



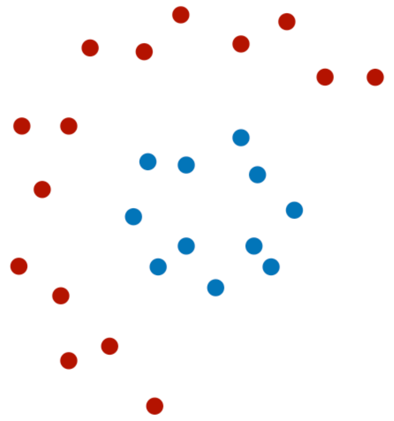
3 Auto-label points in the identified region



4 Remove points in auto-labeling region

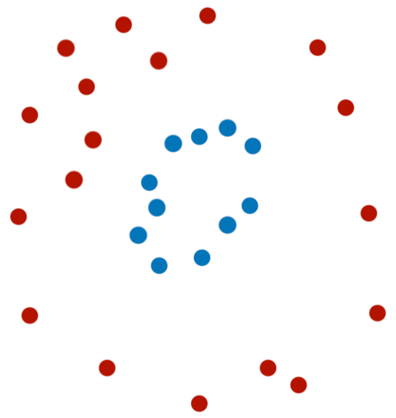


Unlabeled Data



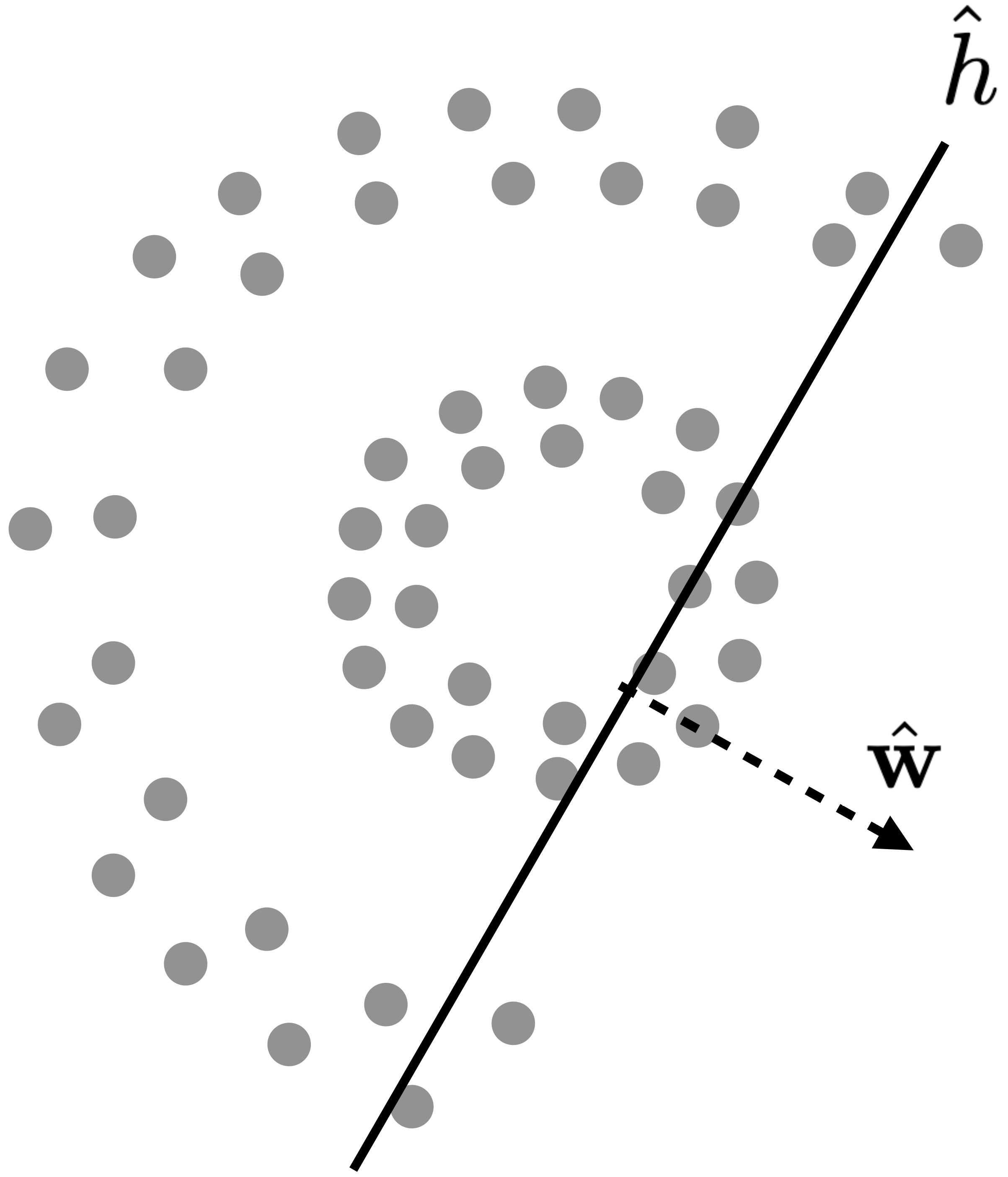
Validation Data

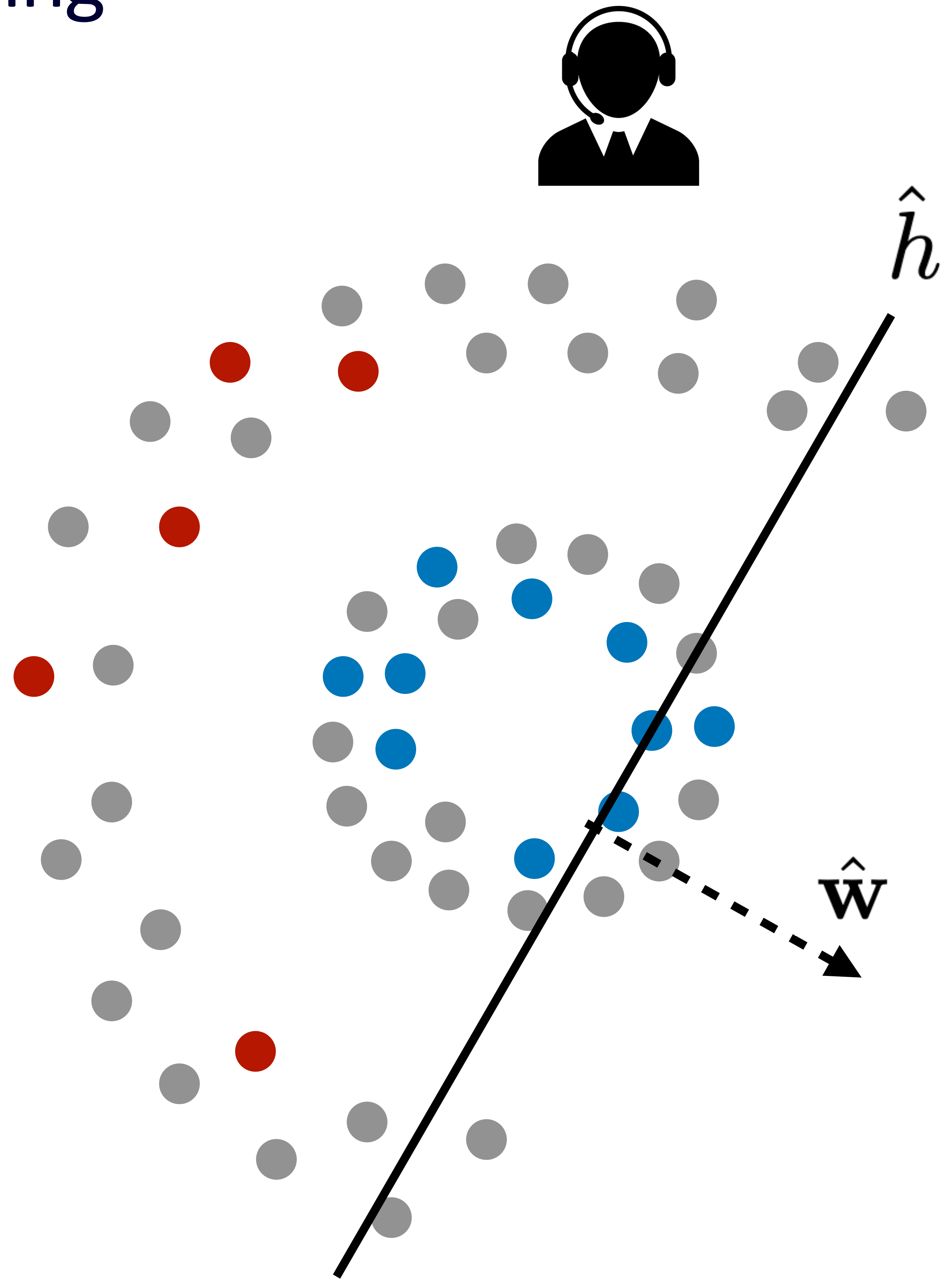
5 Get more human-labeled data for training and go to step **1**



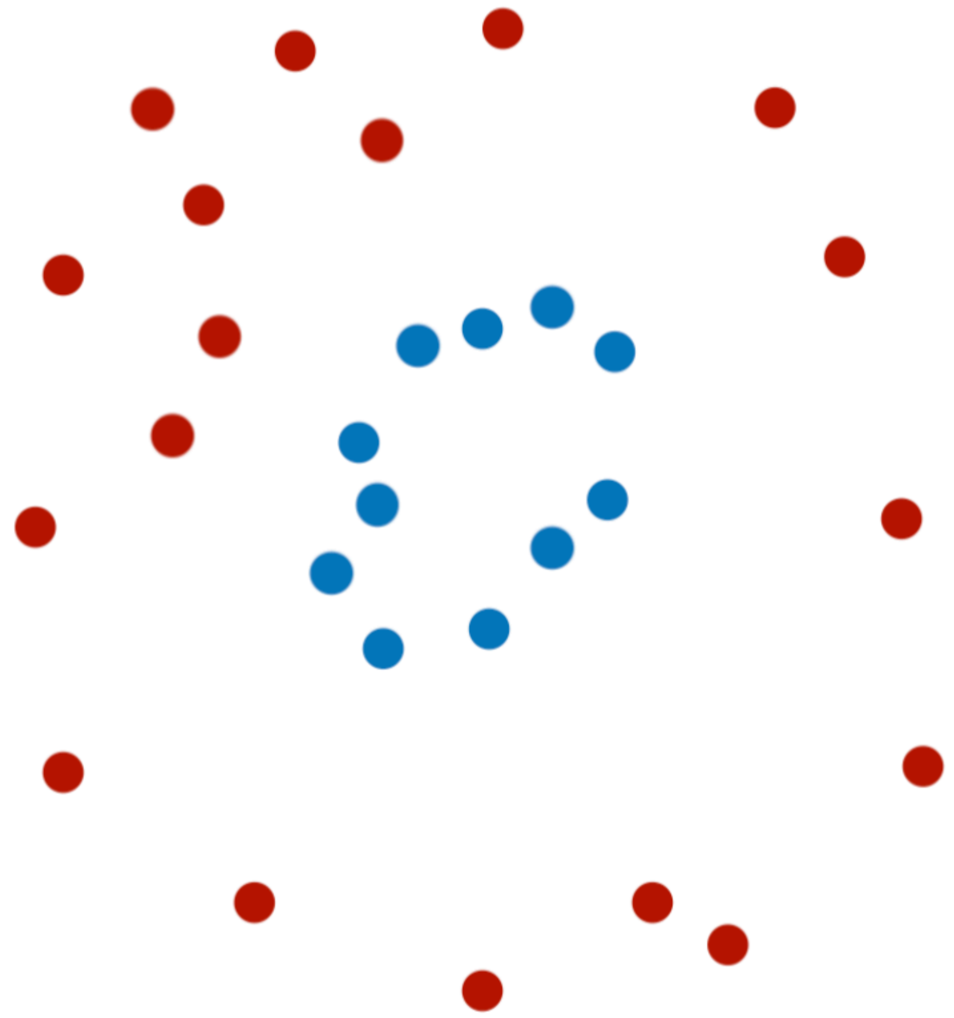
Step 5: Query next batch of human-labeled data for training

Use some active querying strategy
example: uncertainty sampling





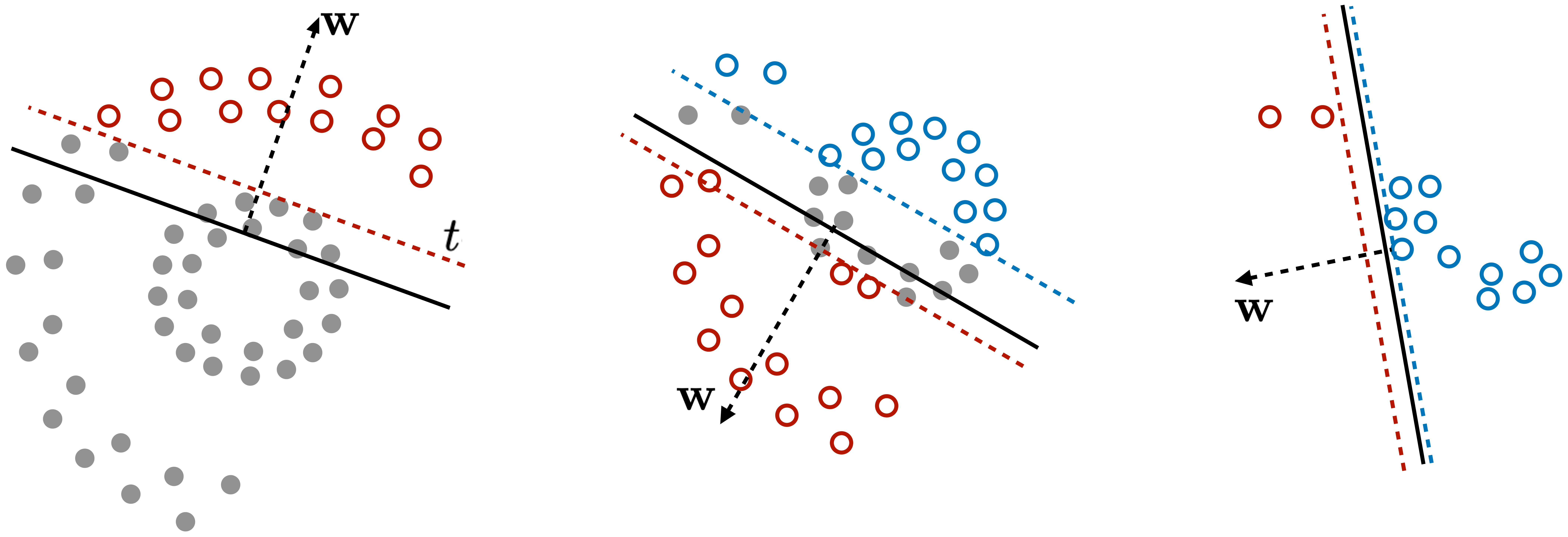
Next round's training data



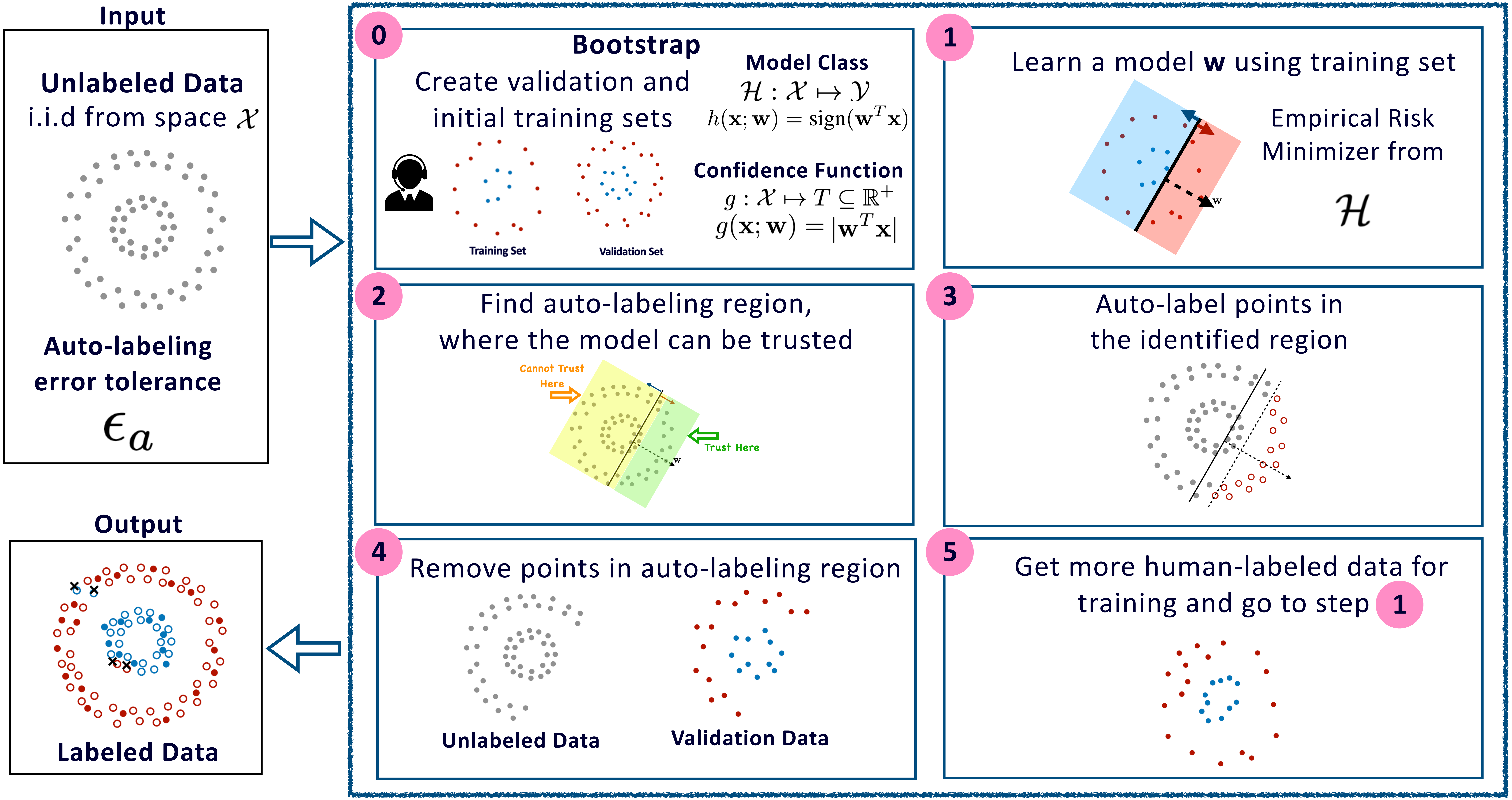
If there is unlabeled data left

Go to Step 1

Intermediate Rounds Output



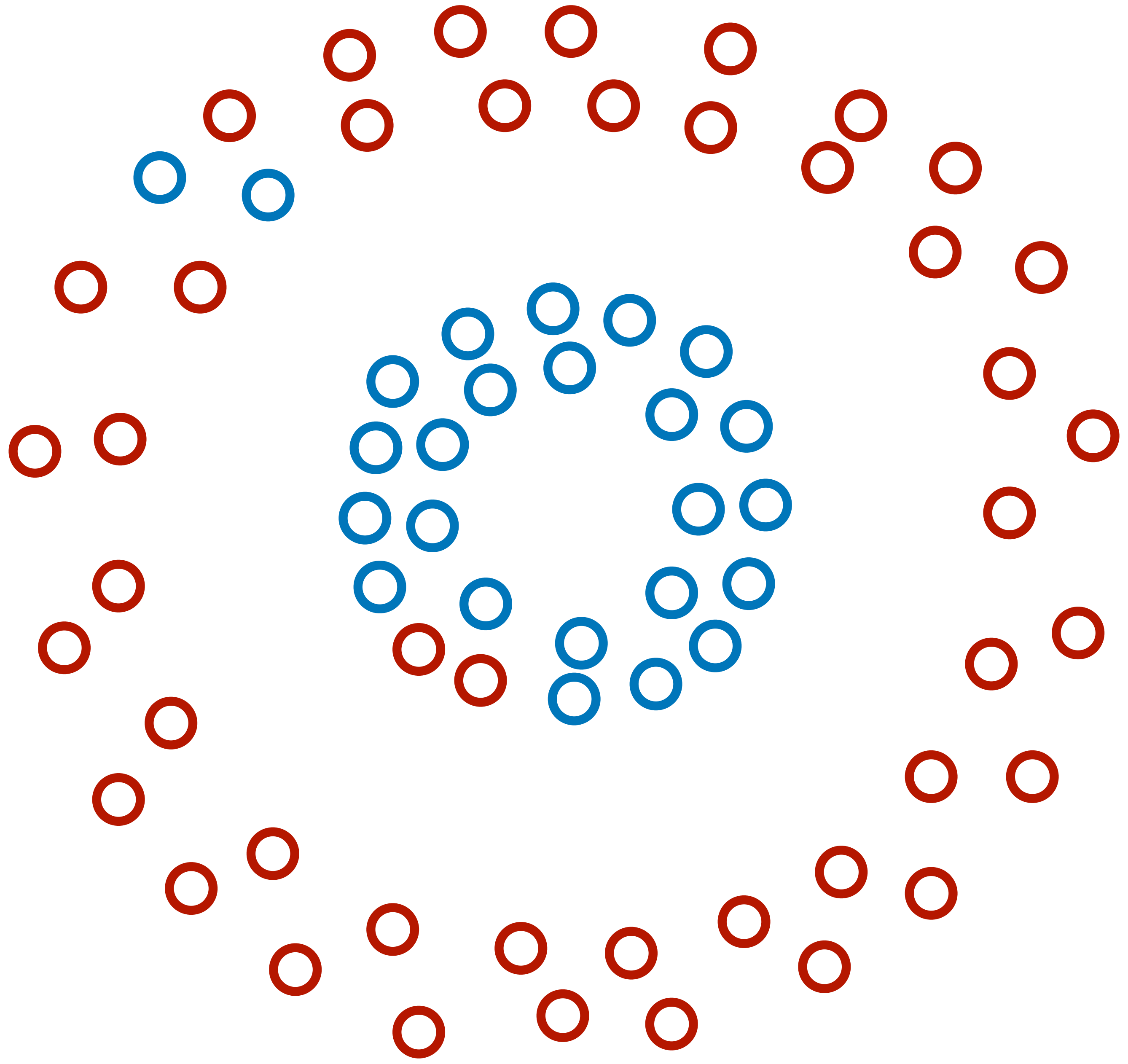
Threshold-based Auto-labeling Workflow(TBAL)



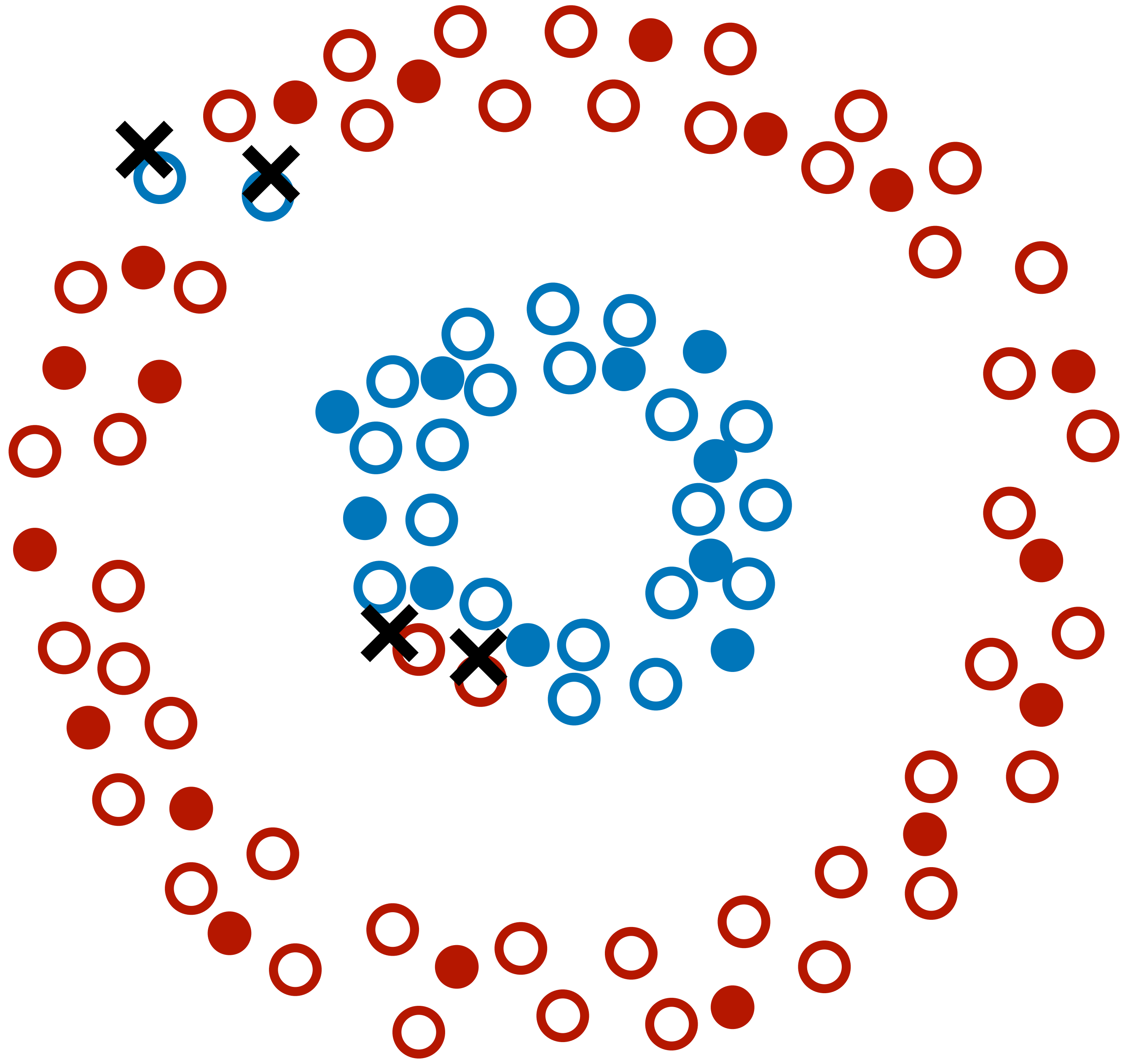
Final Output

Human-labeled
Auto-labeled
Labeling mistake

Auto-labeled data in the end



Output Labeled Dataset



Error and Coverage

Auto-labeling Error < 1%

Coverage > 95%

Roadmap

What & Why auto-labeling?

Data labeling problem

Adoption of auto-labeling

How does it work?

Workflow of TBAL

Finding the
auto-labeling region

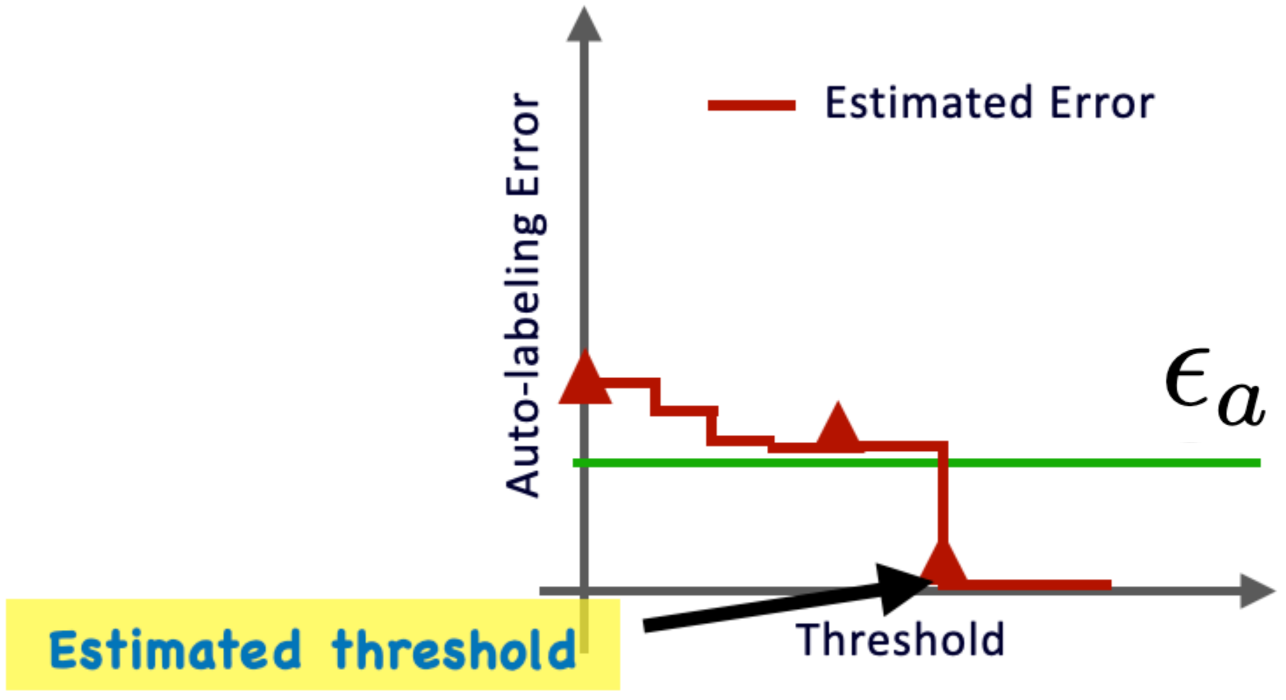
Analysis & Results

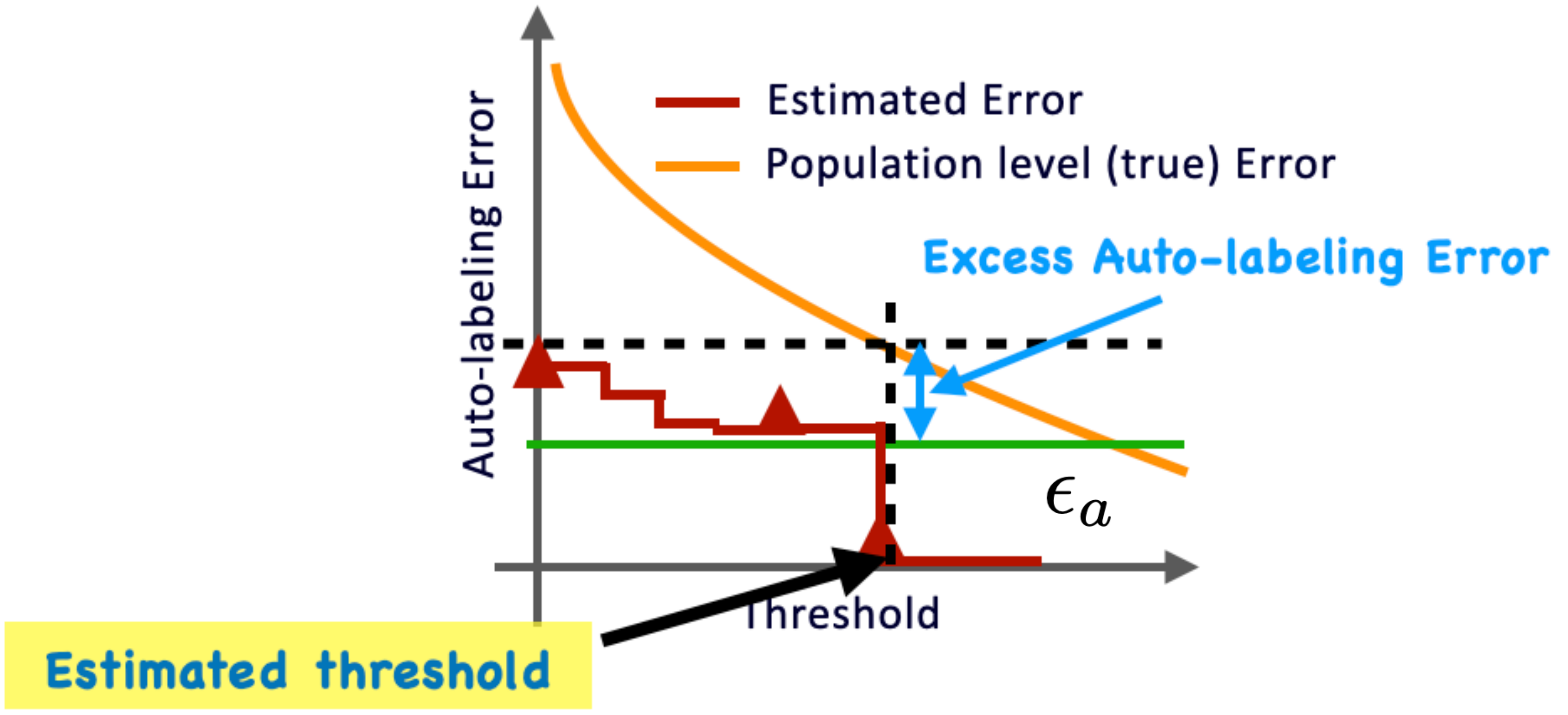
Conditions when TBAL works.

Comparison with
Active Learning, Selective
Classification

Theoretical Results

Conditions on the validation data for accurate auto-labeling





In the general setup: No assumptions on data distribution and function classes

Upper bound on excess auto-labeling error

$$\mathcal{O}\left(\frac{1}{\sqrt{N_v}} + \mathfrak{R}_{N_v}(\mathcal{H}^{T,g})\right)$$

N_v

Validation points

$$\mathcal{H}^{T,g} := \mathcal{H} \times T \quad (h,t) \in \mathcal{H}^{T,g}$$

$$(h,t)(\mathbf{x}) := \begin{cases} h(\mathbf{x}) & \text{if } g(h,\mathbf{x}) \geq t \\ \text{abstain} & \text{o.w.} \end{cases}$$

Lower bound on number of validation samples to ensure auto-labeling error is below ϵ_a

$$\Omega\left(\frac{1}{\epsilon_a^2}\right)$$

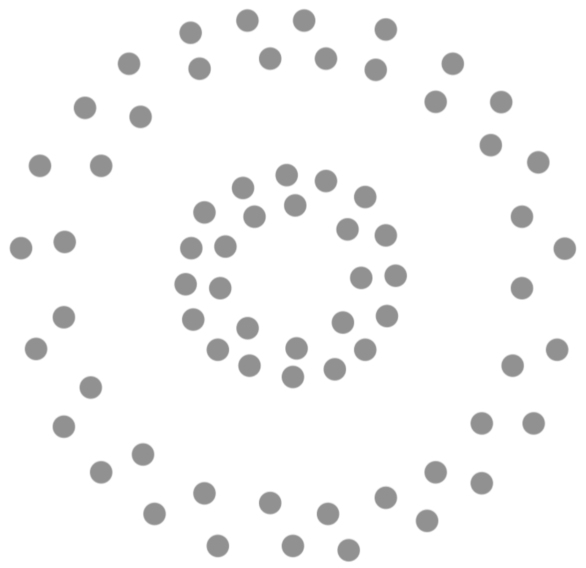
Instantiate the upper bound for uniform distribution on unit-ball in $\mathbb{R}^d$ with homogeneous linear separators

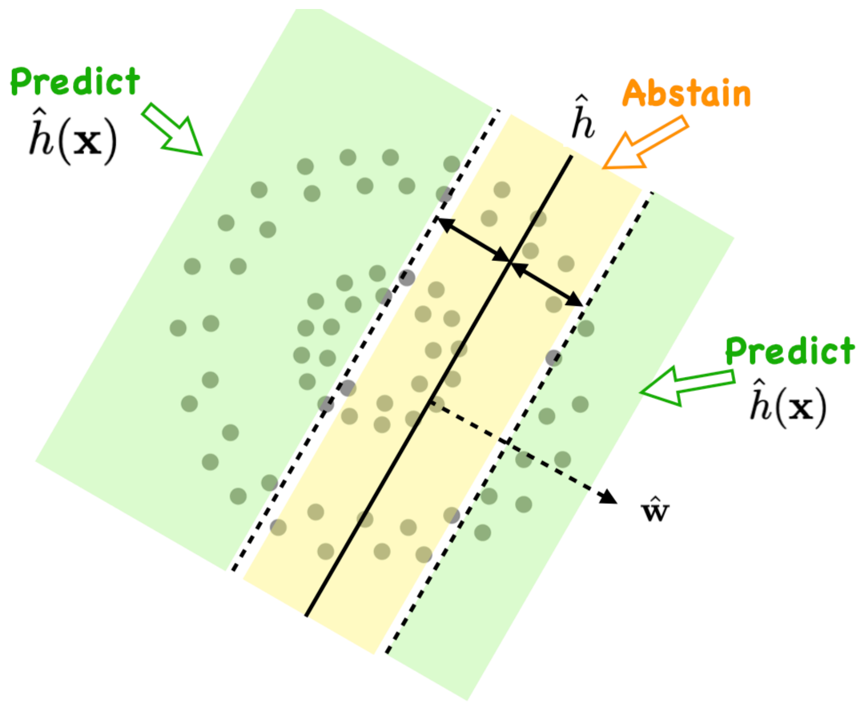
Proof Sketch

With Finite Samples

$$A_v(h,t) = \{\mathbf{x} \in X_v : g(\mathbf{x}; h) \geq t\}$$

$$\widehat{\mathcal{E}}_v(h|t) = \frac{1}{|A_v(h,t)|} \sum_{\mathbf{x} \in A_v(h,t)} \mathbb{1}\{h(\mathbf{x}) \neq f^\star(\mathbf{x})\}$$

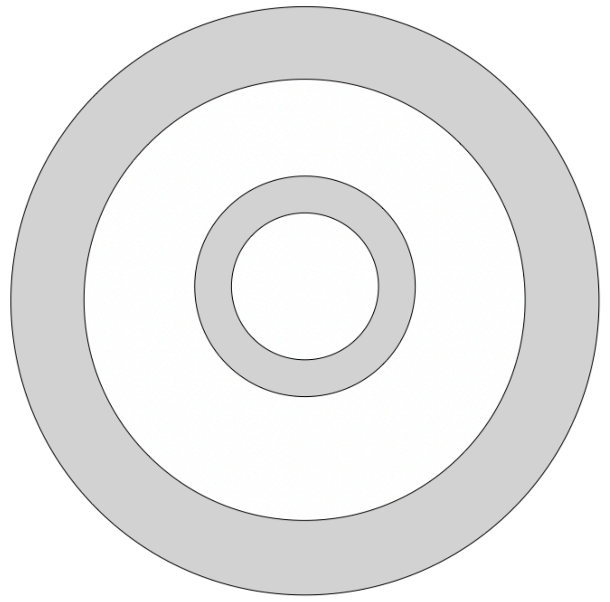


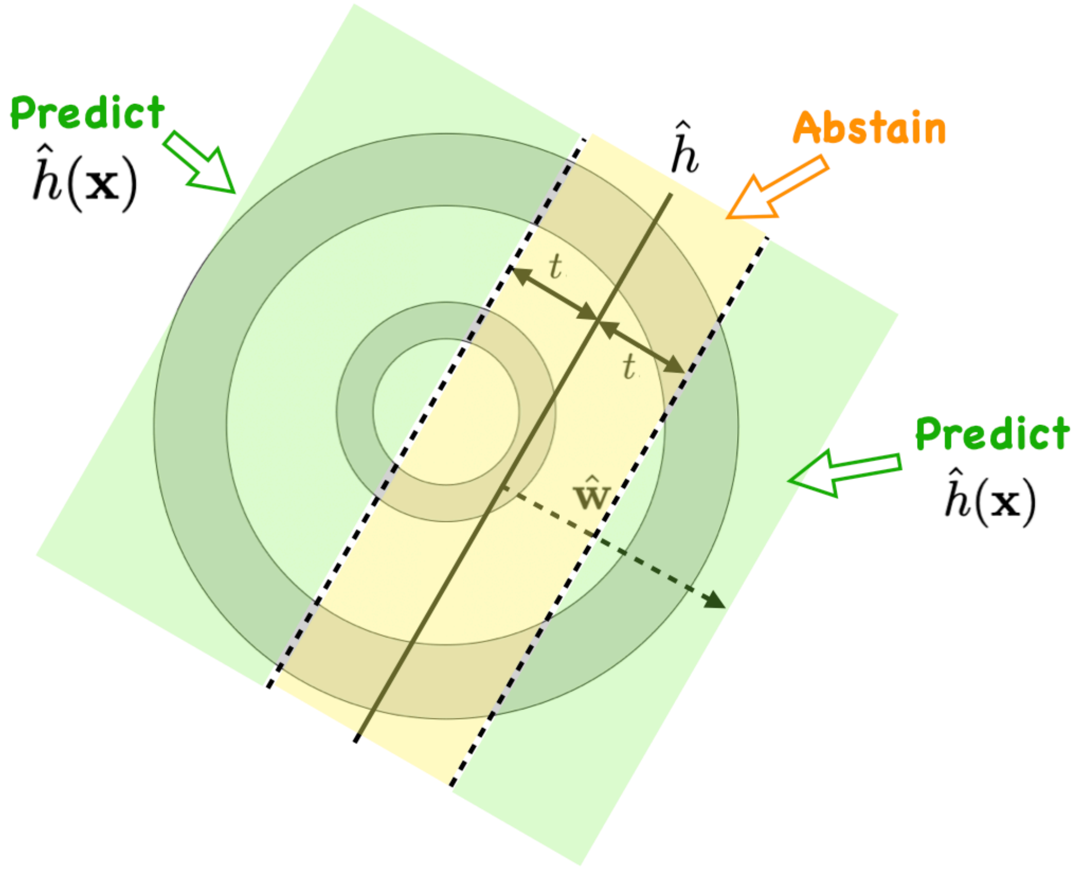


Population Level

$$\mathcal{A}(h,t) = \{\mathbf{x} \in \mathcal{X} : g(\mathbf{x}; h) \geq t\}$$

$$\mathcal{E}(h|t) = \mathbb{E}_{x|\mathcal{A}(h,t)}[\mathbb{1}\{h(\mathbf{x}) \neq f^\star(\mathbf{x})\}]$$





Want this

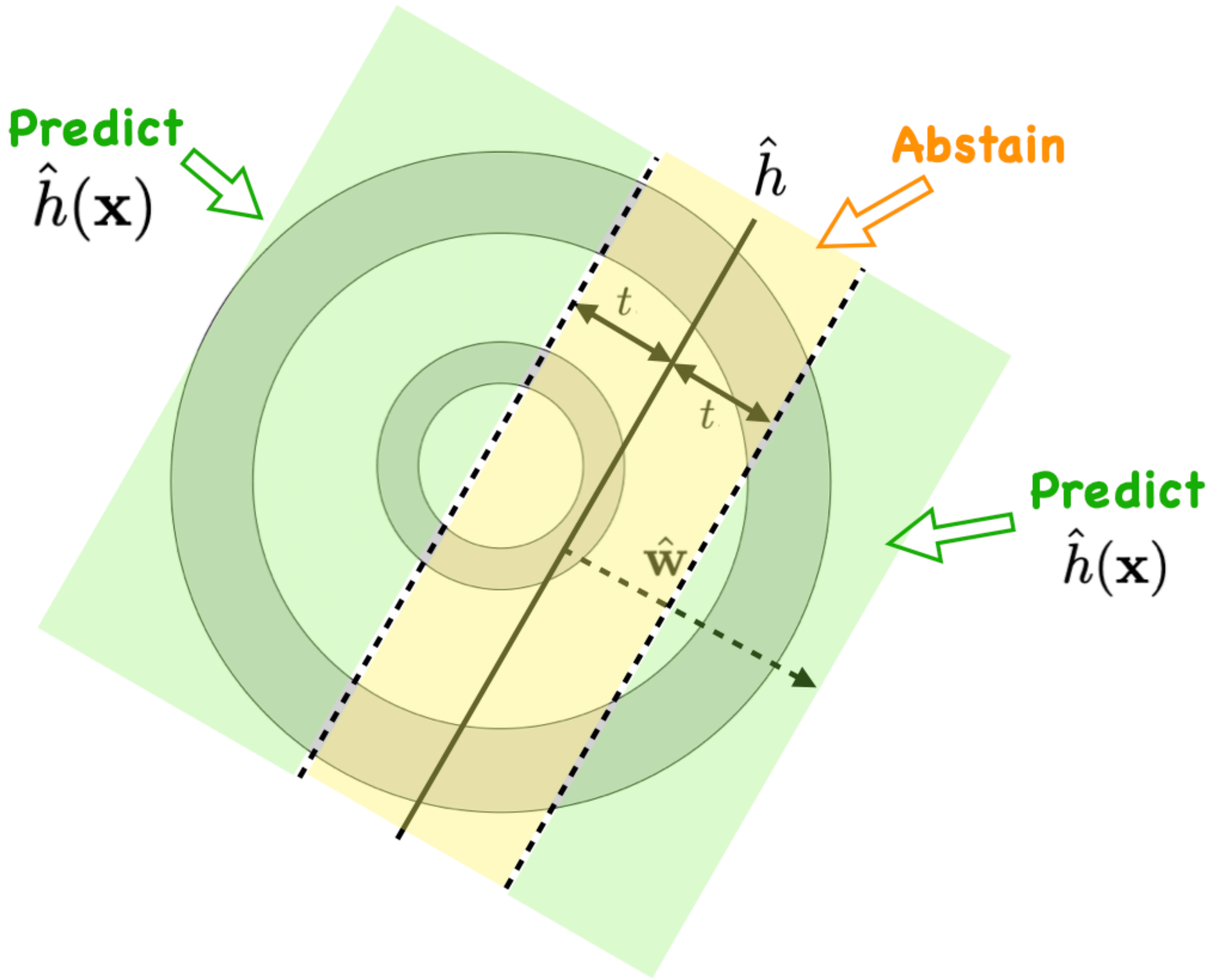
$$\text{w.p.} \quad 1 - \delta$$

$$\mathcal{E}(h|t) \leq \widehat{\mathcal{E}}_v(h|t) + \psi(N_v, \delta, \mathcal{H}, g, T) \qquad \forall h \in \mathcal{H}, \forall t \in T$$

Proof Sketch

$$\mathcal{H}^{T,g} := \mathcal{H} \times T \quad (h,t) \in \mathcal{H}^{T,g}$$

$$(h,t)(\mathbf{x}) := \begin{cases} h(\mathbf{x}) & \text{if } g(h,\mathbf{x}) \geq t \\ \text{abstain} & \text{o.w.} \end{cases}$$



$$\mathcal{E}(h,t) = \mathbb{E}_{\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f^\star(\mathbf{x})\} \wedge \mathbb{1}\{g(\mathbf{x}) \geq t\}]$$

$$\mathbb{P}(h,t) = \mathbb{E}_{\mathbf{x}}[\mathbb{1}\{g(\mathbf{x}) \geq t\}] \qquad \mathcal{E}(h|t) = \frac{\mathcal{E}(h,t)}{\mathbb{P}(h,t)}$$

$$\widehat{\mathcal{E}}_v(h,t) = \frac{1}{N_v} \sum_{\mathbf{x}_i \in X_v} \mathbb{1}\{h(\mathbf{x}_i) \neq f^\star(\mathbf{x}_i)\} \wedge \mathbb{1}\{g(\mathbf{x}_i) \geq t\}$$

$$\hat{P}_v(h,t) = \frac{1}{N_v} \sum_{\mathbf{x}_i \in X_v} \mathbb{1}\{g(\mathbf{x}_i) \geq t\} \qquad \widehat{\mathcal{E}}_v(h|t) = \frac{\widehat{\mathcal{E}}_v(h,t)}{\hat{P}_v(h,t)}$$

Uniform convergence results

$$|\mathcal{E}(h,t) - \widehat{\mathcal{E}}_v(h,t)| \leq \mathcal{O}\left(\mathfrak{R}_{N_v}(\mathcal{H}^{T,g}) + \sqrt{\frac{1}{N_v}\log(\frac{1}{\delta})}\right)$$

$$|\mathbb{P}(h,t) - \hat{\mathbb{P}}_v(h,t)| \leq \mathcal{O}\left(\mathfrak{R}_{N_v}(\mathcal{H}^{T,g}) + \sqrt{\frac{1}{N_v}\log(\frac{1}{\delta})}\right)$$

$$\Downarrow$$

w.p. $1-\delta$

$$\mathcal{E}(h|t) \leq \widehat{\mathcal{E}}_v(h|t) + \psi(N_v, \delta, \mathcal{H}, g, T) \quad \forall h \in \mathcal{H}, \forall t \in T$$

Experiments

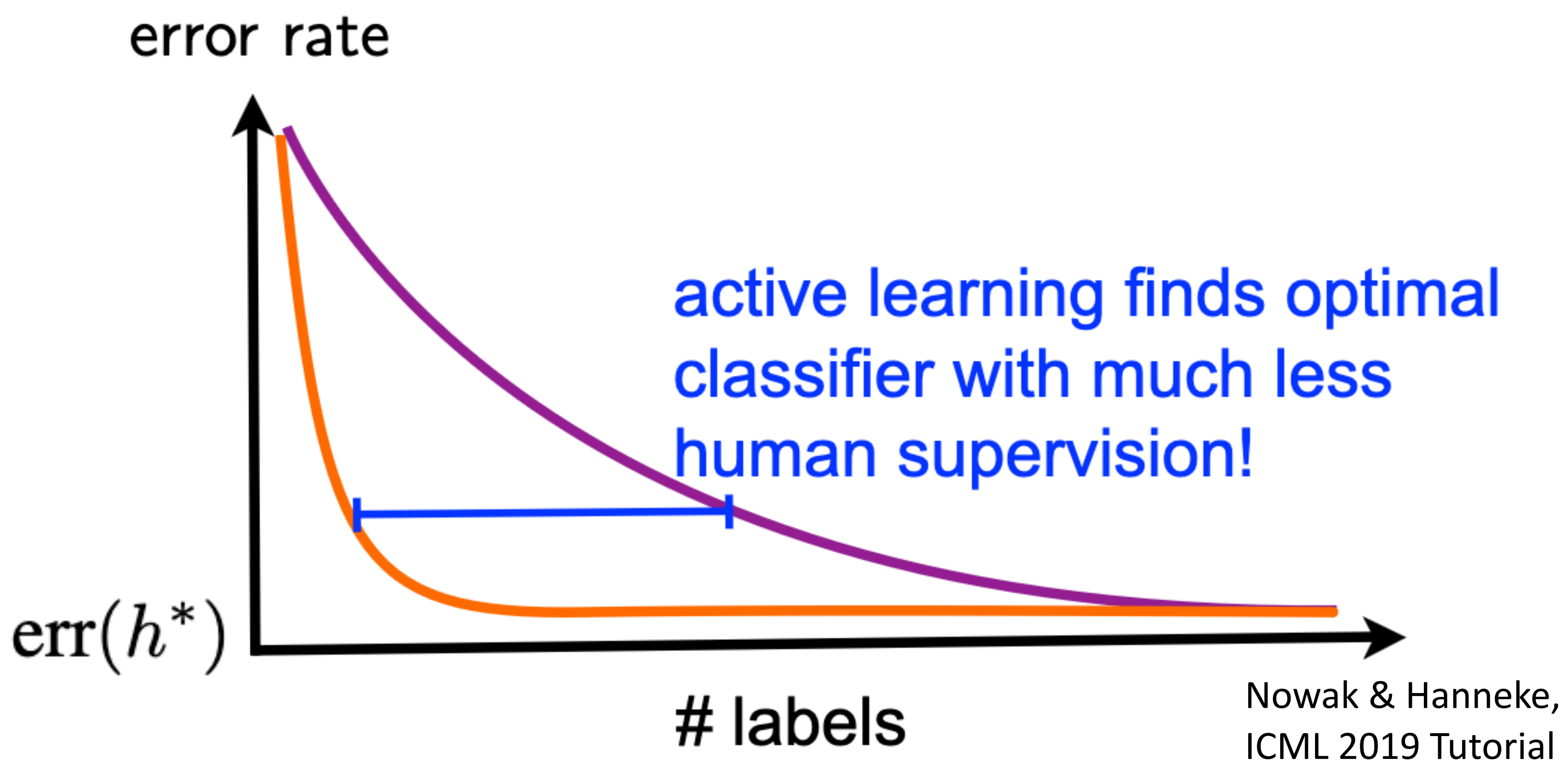
Active Learning and Selective Classification

Active Learning (AL)

$$\mathrm{err}(h) = \mathbb{E}_{\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq y\}]$$

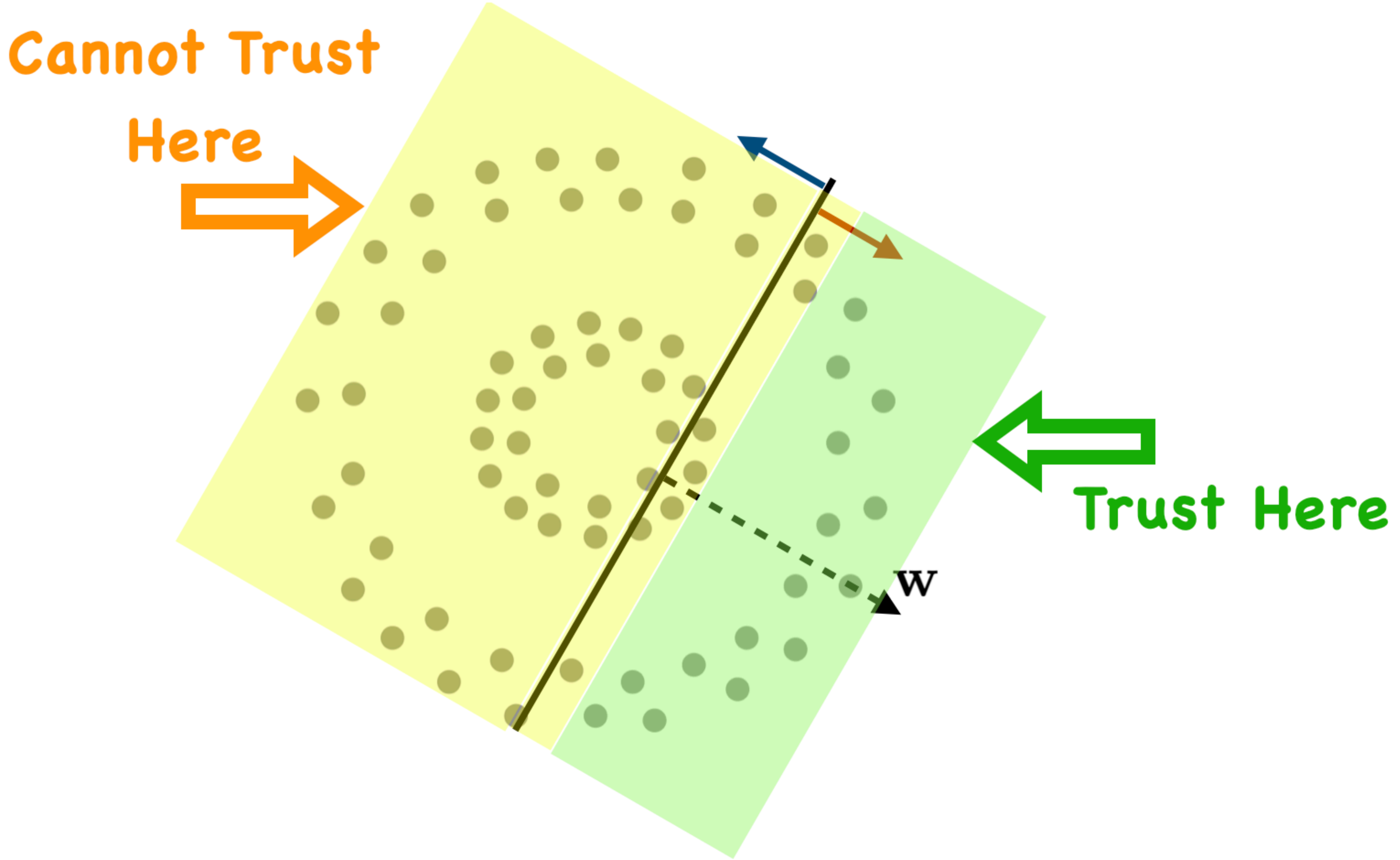
$$h^* \in \arg\min_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq y\}]$$

$$\mathrm{err}(\hat{h}) - \mathrm{err}(h^*) \to 0$$



Cohn et al. 1994;
Balcan, Dasgupta, Nowak, Zhu, Hanneke, Jamieson, Chaudhury…. (Over the last 3 decades)

Selective Classification (SC)

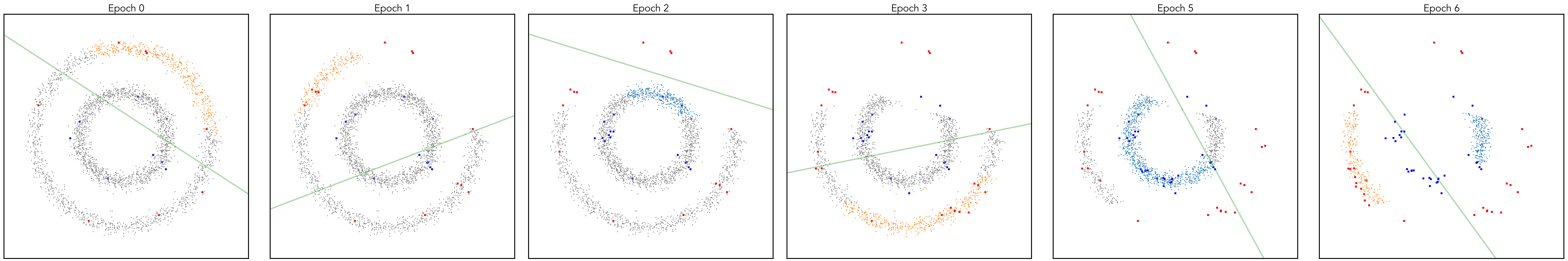


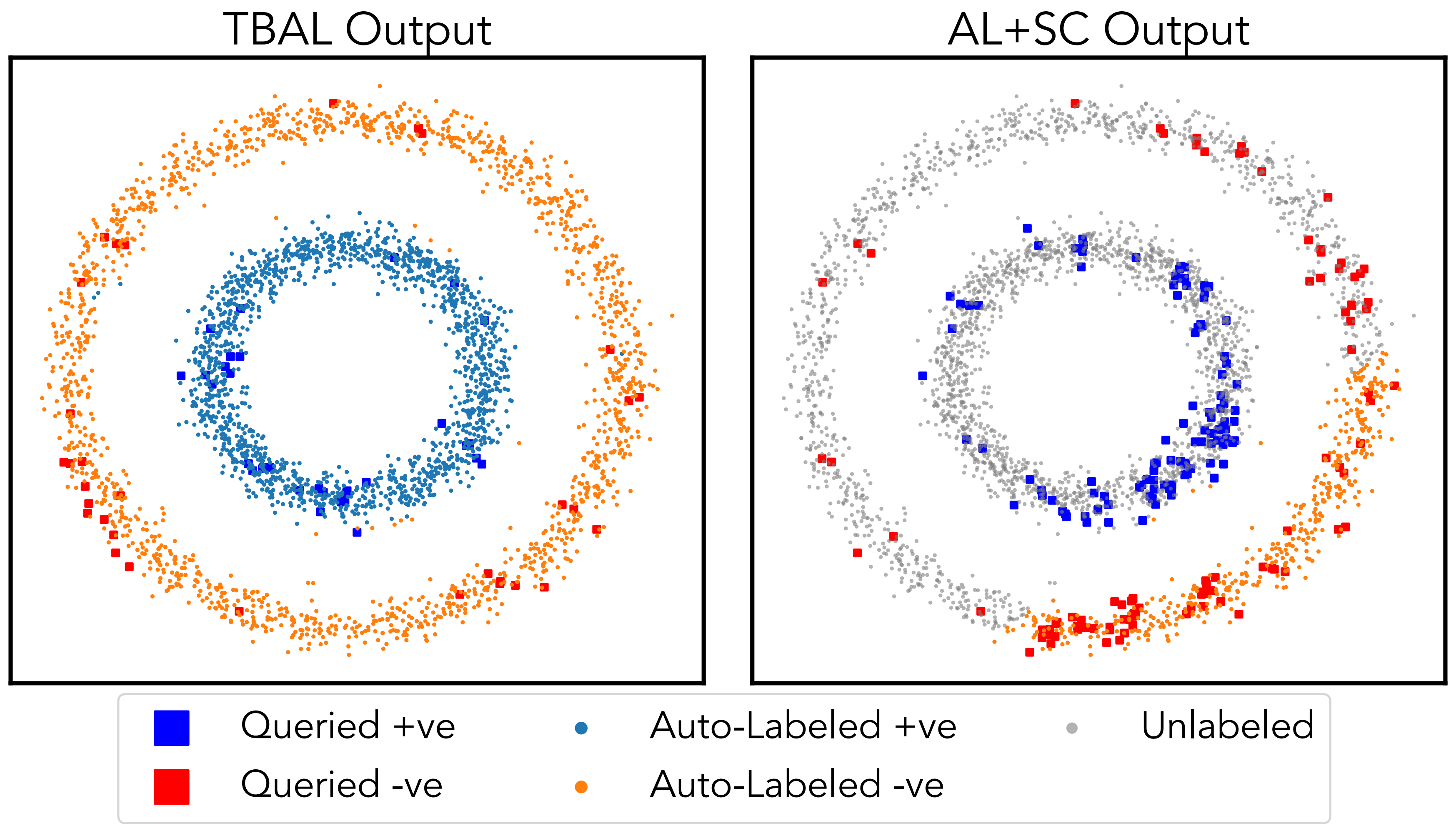
El-Yaniv & Weiner, 2010; Cortes, Desalvo, Mohri 2016;
Gelbhart & El-Yaniv 2019; Fisch, Jakkola et al. 2022;

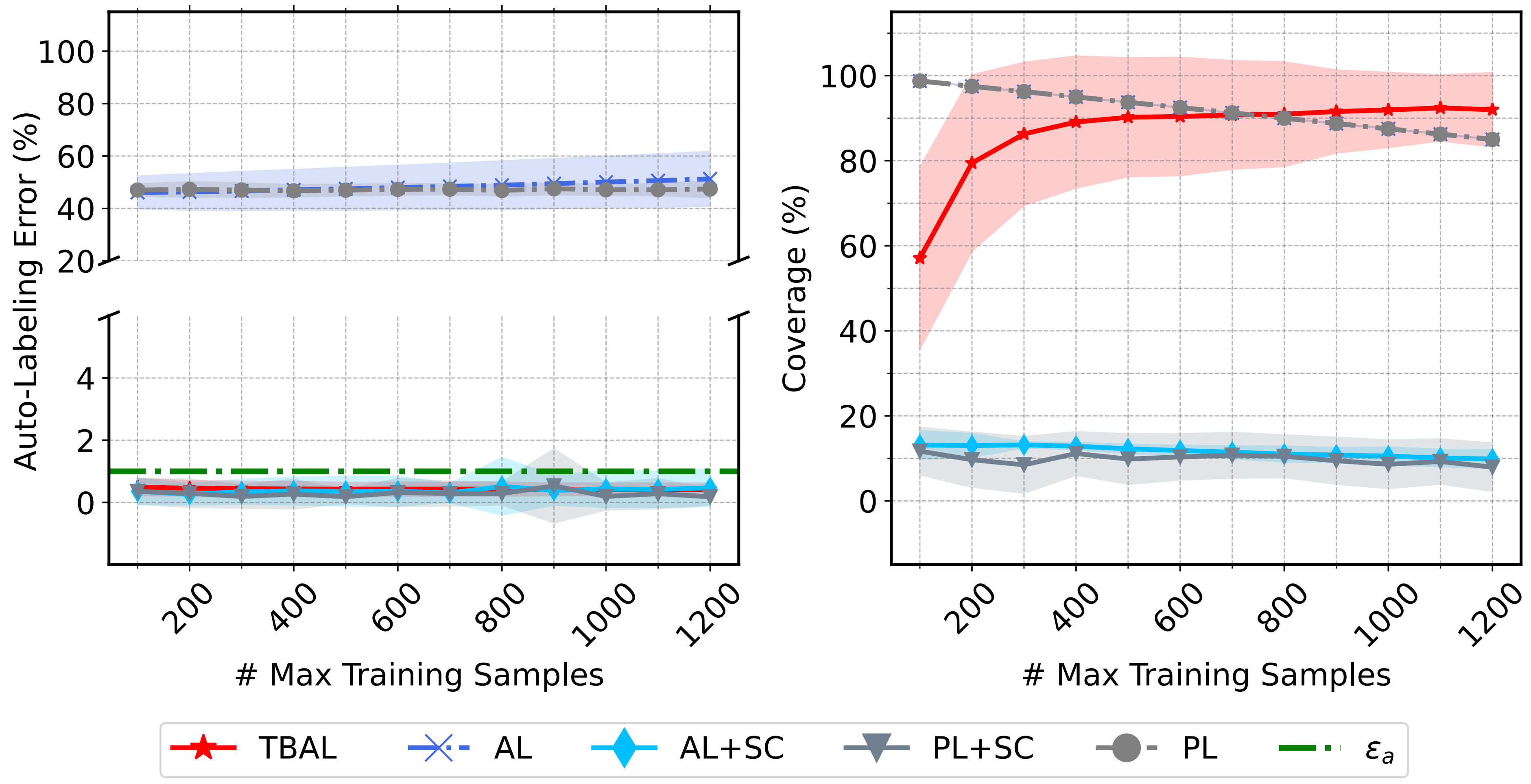
A natural auto-labeling strategy (AL+SC):
First learn the best classifier using Active Learning,
then auto-label using selective classification.

The methods work as expected on the circles example

Misspecified setting: Using incorrect model class, (in practice the correct class is not known)







We validate the results empirically

Fix the auto-labeling error tolerance and the max number of training points algorithm can use.

Vary the number of validation points

Increasing Validation data

Unit ball (Synthetic)

N_v	Error (%) TBAL	Error (%) AL+SC	Coverage (%) TBAL	Coverage (%) AL+SC
100	3.10 $\pm$1.80	0.68 $\pm$0.81	71.43 $\pm$8.86	96.95 $\pm$1.01
400	1.65 $\pm$0.65	0.32 $\pm$0.15	93.27 $\pm$2.50	96.91 $\pm$0.99
800	1.08 $\pm$0.47	0.24 $\pm$0.16	96.01 $\pm$1.16	96.31 $\pm$1.36
1200	0.78 $\pm$0.27	0.17 $\pm$0.11	96.82 $\pm$0.84	95.96 $\pm$1.40
1600	0.65 $\pm$0.20	0.13 $\pm$0.08	96.93 $\pm$0.57	95.70 $\pm$1.38
2000	0.54 $\pm$0.16	0.21 $\pm$0.11	97.23 $\pm$0.42	96.36 $\pm$1.13

Classes = 2 ϵ_a = 1%

Max # training points = 500

IMDB

N_v	Error (%) TBAL	Error (%) AL+SC	Coverage (%) TBAL	Coverage (%) AL+SC
200	2.28 $\pm$0.21	3.11 $\pm$0.86	68.24 $\pm$6.20	57.77 $\pm$13.09
400	1.29 $\pm$0.10	1.98 $\pm$0.40	63.81 $\pm$4.86	63.06 $\pm$10.70
600	1.41 $\pm$0.20	1.81 $\pm$0.22	69.64 $\pm$3.98	62.92 $\pm$9.20
800	1.62 $\pm$0.30	2.04 $\pm$0.35	67.45 $\pm$3.72	63.22 $\pm$7.89
1000	1.64 $\pm$0.23	1.97 $\pm$0.26	70.28 $\pm$2.82	66.11 $\pm$8.00

Classes = 2 ϵ_a = 5%

Max # training points = 500

Tiny Imagenet

N_v	Error (%) TBAL	Error (%) AL+SC	Coverage (%) TBAL	Coverage (%) AL+SC
2000	0.0 $\pm$0.0	0.0 $\pm$0.0	0.0 $\pm$0.0	0.0 $\pm$0.0
4000	10.50 $\pm$6.01	7.37 $\pm$4.57	0.47 $\pm$0.05	0.48 $\pm$0.06
6000	10.61 $\pm$0.62	7.71 $\pm$1.03	10.16 $\pm$1.10	4.31 $\pm$1.10
8000	9.90 $\pm$0.63	6.80 $\pm$0.77	25.84 $\pm$1.57	14.43 $\pm$2.01
10000	8.97 $\pm$0.36	6.87 $\pm$0.48	32.19 $\pm$1.34	21.96 $\pm$1.35

Classes = 200 ϵ_a = 10%

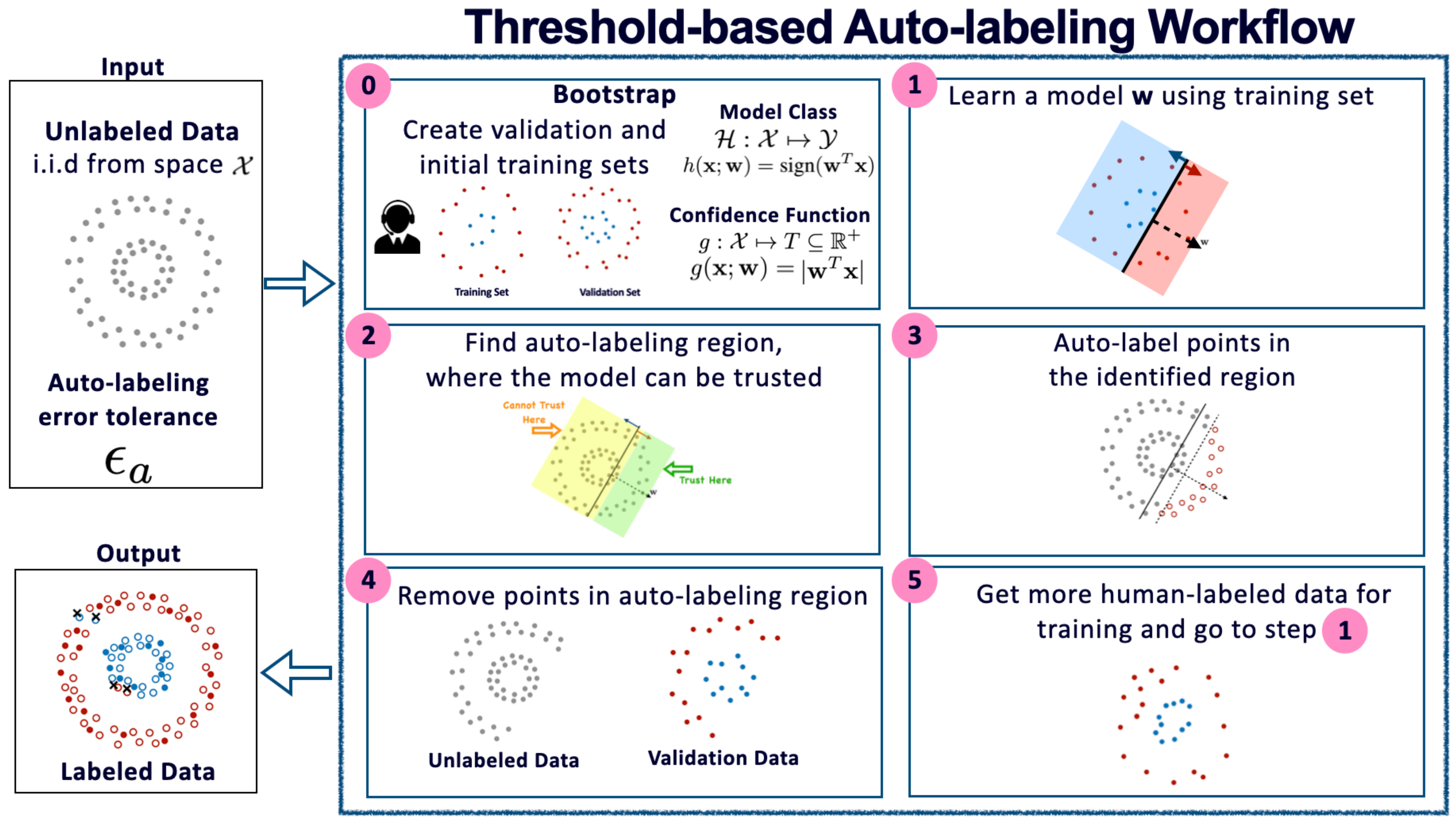
Max # training points = 10000

As expected, we observe

Less validation data ⇨ high auto-labeling errors and high variance in coverage

Suff. Large validation data ⇨ less auto-labeling errors and less variance in coverage

Summary and Takeaways



1. Auto labeling is a promising solution to obtain labeled data.

2. Our work develops a theoretical understanding of auto-labeling systems.

3. **The promise —** Seemingly bad models can auto-label significant portion of data with good accuracy.

4. **The pitfall —** Hidden downside is it may need large amount validation data to ensure good accuracy.

Thank You

Checkout our paper and code!

Code

Paper

Paper https://openreview.net/pdf?id=RUCFAKNDb2

Code https://github.com/harit7/TBAL-NeurIPS-23

Come to our poster @ NeurIPS

Hall B1 + B2 #1103

Wed 13 Dec

3 p.m. - 5 p.m. PST

Contact us

Harit Vishwakarma
hvishwakarma@cs.wisc.edu

Heguang Lin
hglin@seas.upenn.edu

Frederic Sala
fredsala@cs.wisc.edu

Ramya Korlakai Vinayak
ramya@ece.wisc.edu