Learning from positive and unlabeled data

3. Assumptions to enable PU Learning

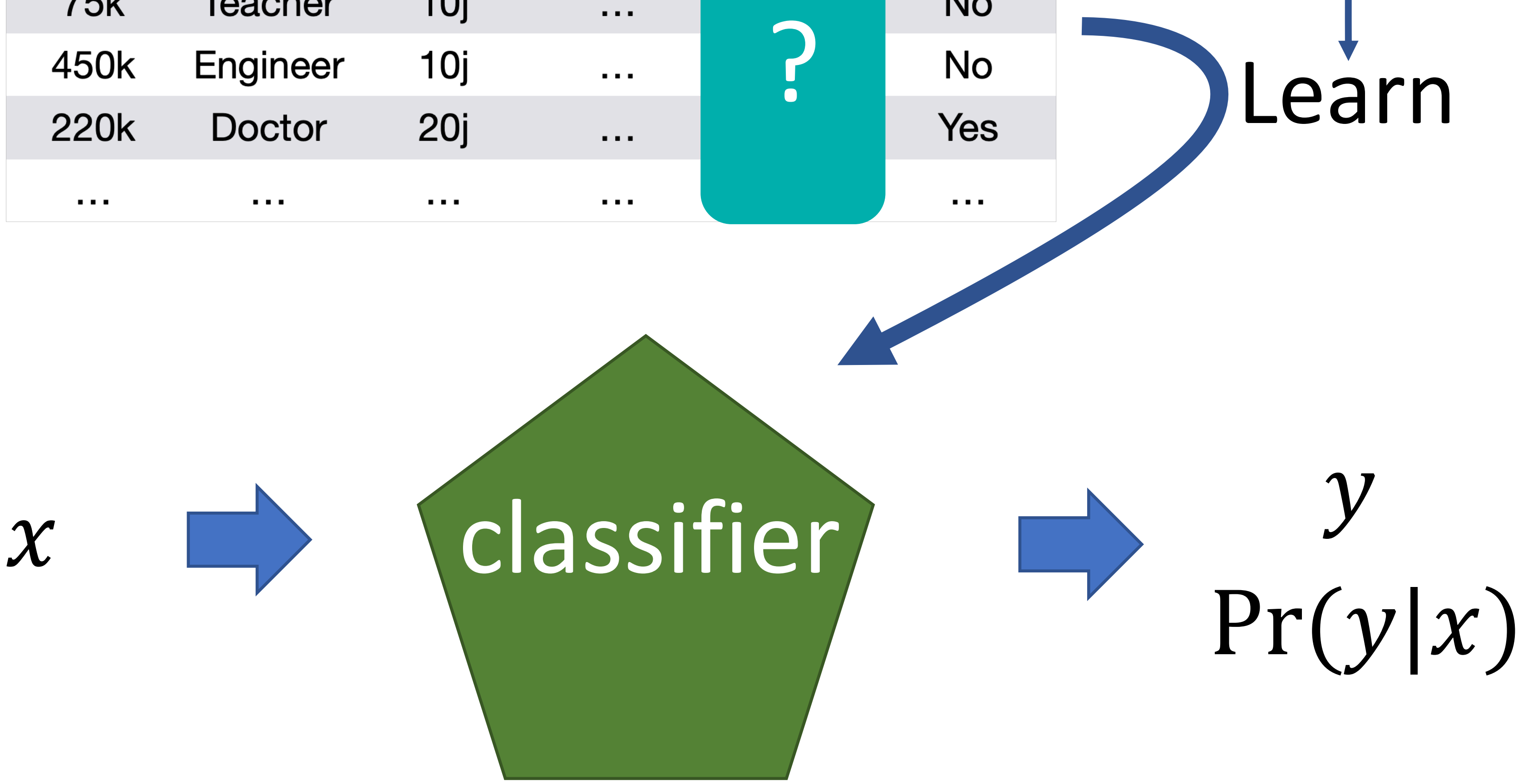
Section 3 in the survey paper

Assumptions in PU Learning

Labeling mechanism

Features x				Class y	Label s
Amount	**Job**	**Repayment period**	**…**	**Qualify**	**Approval**
300k	Lawyer	15j	…	?	No
75k	Teacher	10j	…		No
450k	Engineer	10j	…		No
220k	Doctor	20j	…		Yes
…	…	…	…		…

enable



Assumptions

- Enough data
- Training data is iid sample of population

  $$\boldsymbol{x} \sim f(x)$$
  $$\sim \alpha f_{+}(x) + (1-\alpha) f_{-}(x)$$

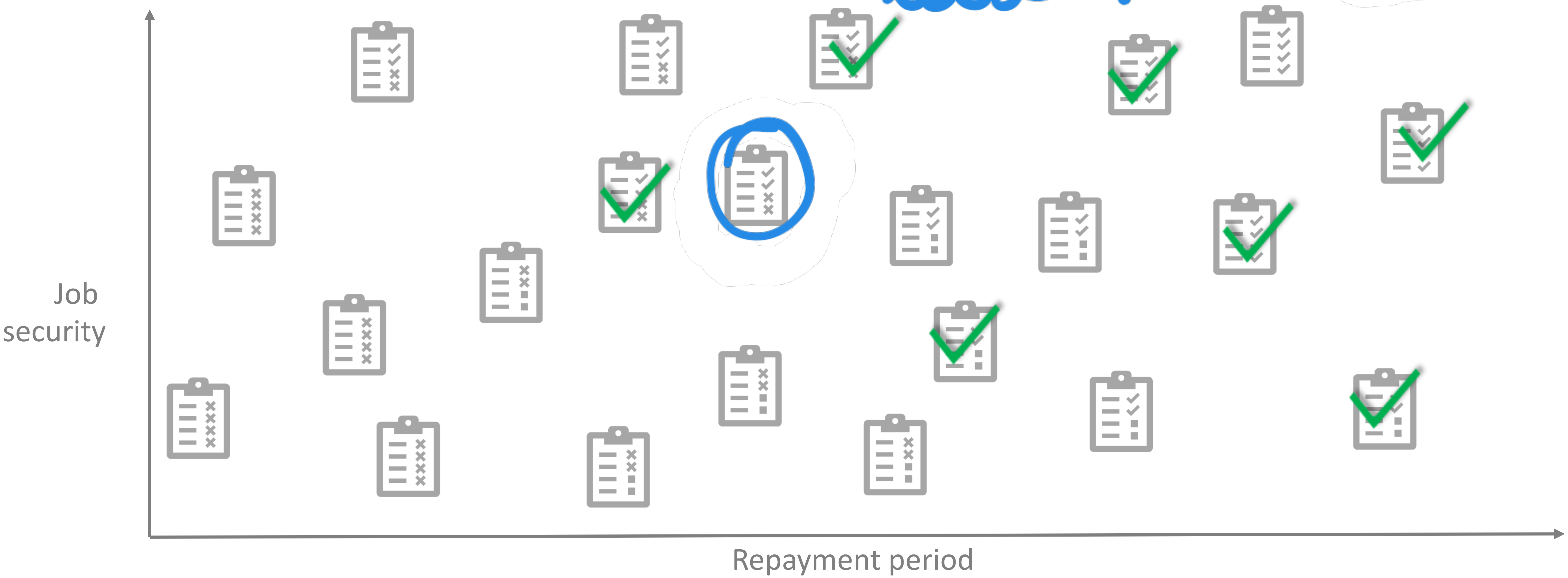
- Class distribution assumptions (classifier bias)
- Labeling mechanism assumptions

Not all the assumptions need to be strong individually, but together they need to be strong enough to enable learning.

Assumptions in PU Learning

- no?
- label?

Assume enough data and iid sample



Assumptions in PU Learning

Additional assumption are necessary to enable learning

1. Labeling mechanism assumptions
2. Data (class distribution) assumptions

Label mechanism assumptions

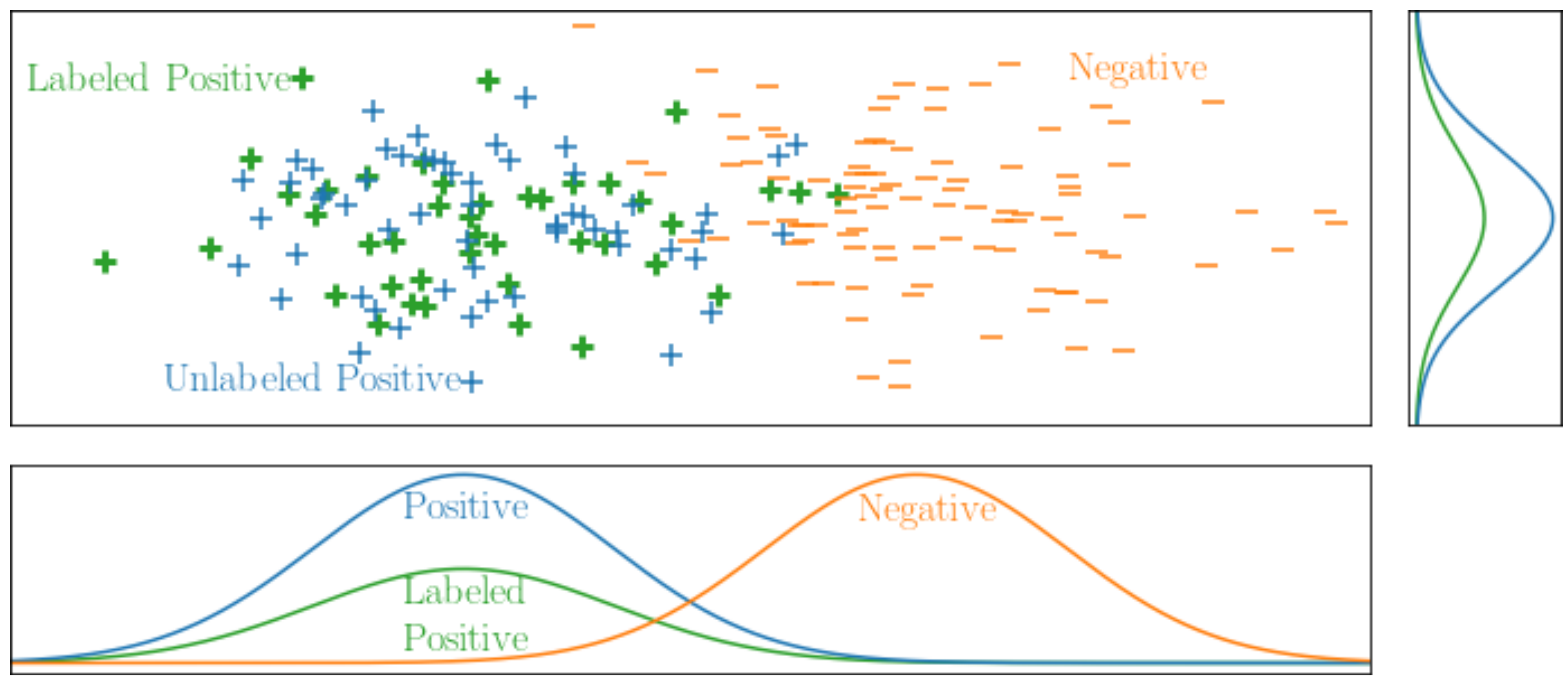
Section 3.1 in the survey paper

Selected Completely At Random (SCAR)

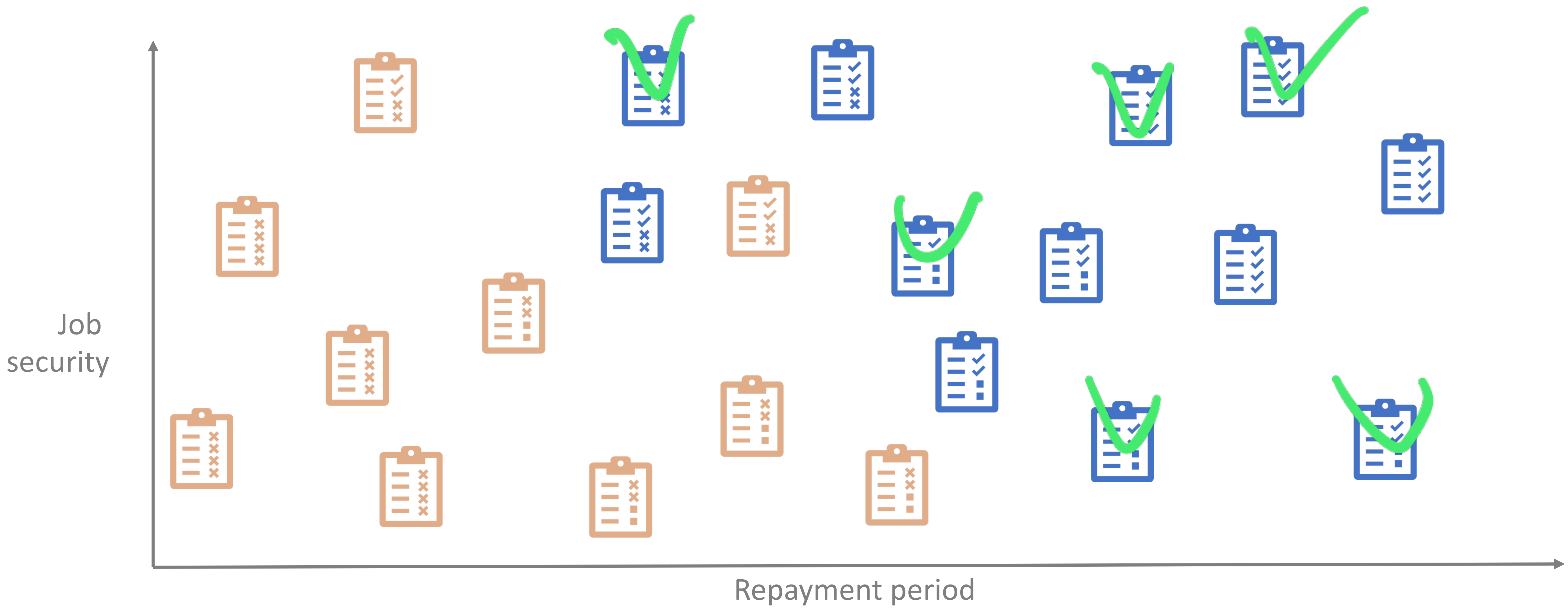
Every positive example has the same probability c to be labeled

$$e(x) = \Pr(s = 1|x, y = 1)$$
$$= \Pr(s = 1|y = 1)$$
$$= c$$

Enables reducing PU learning to binary classification by weighting the data or minor algorithm modifications



Selected Completely At Random (SCAR)



Selected Completely At Random (SCAR)

Probability to be labeled directly proportional to positive probability

$$\Pr(s = 1|x) = c \Pr(y = 1|x)$$

Non-traditional classifier predicts $\Pr(s = 1|x)$ can be learned from PU data using traditional techniques [1].

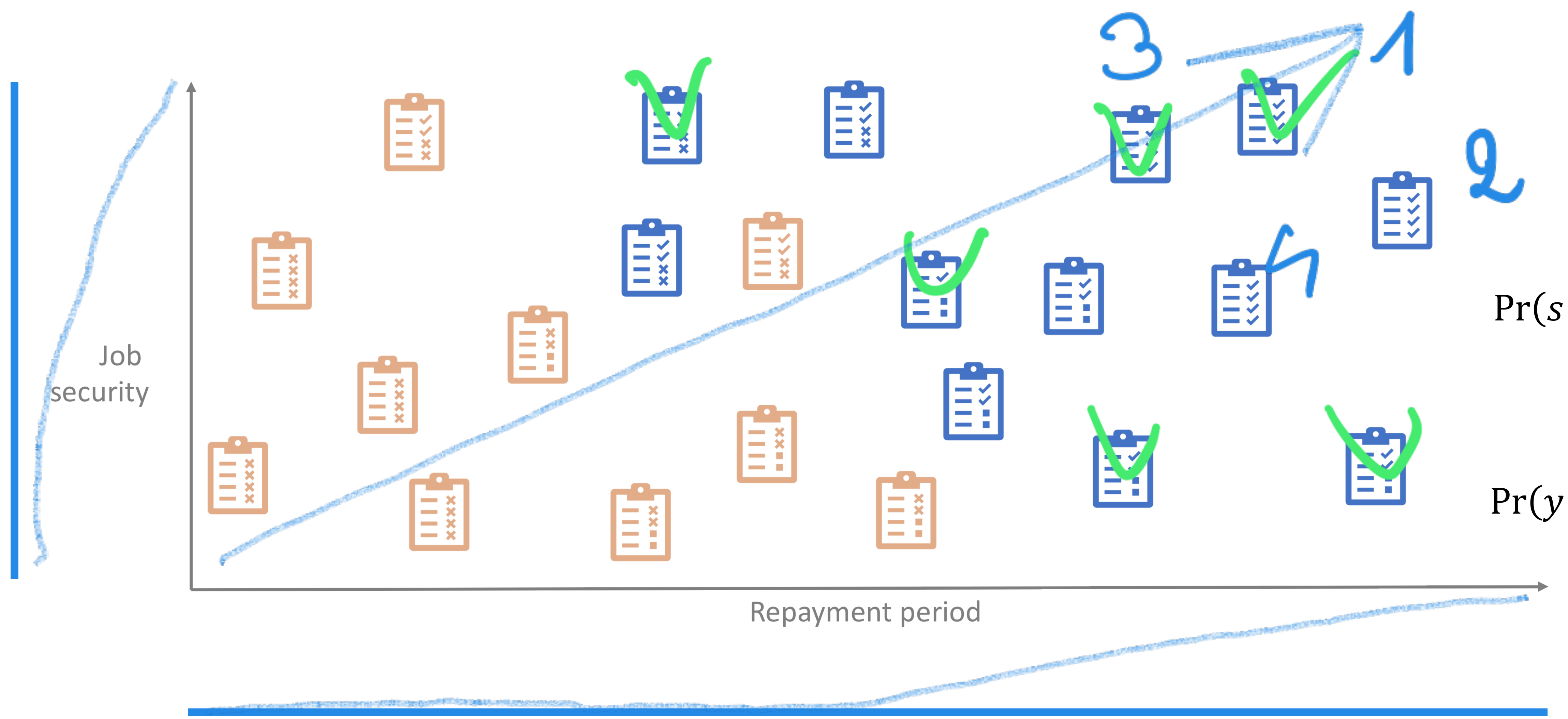
Interesting properties:

- Ranking order
- Recall
- Get class probabilities by dividing non-traditional classifier by c

[1] Elkan & Noto. Learning Classifiers from Only Positive and Unlabeled Data. KDD. 2008

SCAR: Ranking order

Non-traditional classifier predicts $\Pr(s=1|x)$



$$\Pr(s=1|x_1) > \Pr(s=1|x_2)$$

$$\Updownarrow$$

$$\Pr(y=1|x_1) > \Pr(y=1|x_2)$$

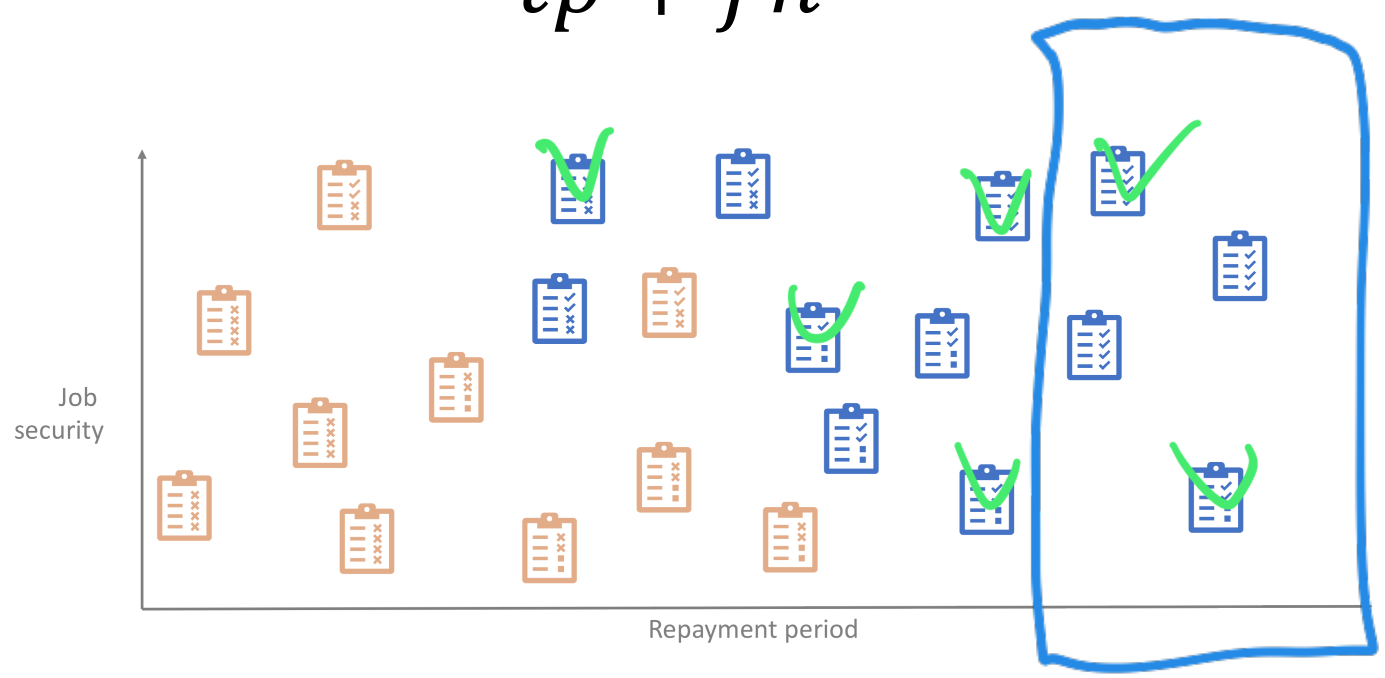
[1] Elkan & Noto. Learning Classifiers from Only Positive and Unlabeled Data. KDD. 2008

Training a non-traditional classifier subject to a recall is equivalent to training a traditional classifier subject to that same recall.

SCAR: Recall

Non-traditional classifier predicts $\Pr(s = 1|x)$

$$recall = \frac{tp}{tp + fn} = \Pr(\hat{y} = 1|y = 1)$$



$12 \quad 4 \quad \frac{1}{3}$

$6 \quad 2 \quad \frac{1}{3}$

$$\Pr(\hat{y} = 1|y = 1) = \Pr(\hat{y} = 1|s = 1)$$

[1] Liu et al. Partially supervised classification of text documents. ICML 2002
[2] Blanchard et al. Semi-supervised novelty detection. JMLR. 2010

SCAR: Class probabilities

Non-traditional classifier predicts $\Pr(s = 1|x)$

From

$$\Pr(s = 1|x) = c \Pr(y = 1|x)$$

It follows directly

$$\Pr(y = 1|x) = \frac{1}{c} \Pr(s = 1|x)$$

1) expert knowledge
2) labeled valid. data
3) estimate from PU data

[1] Elkan & Noto. Learning Classifiers from Only Positive and Unlabeled Data. KDD. 2008

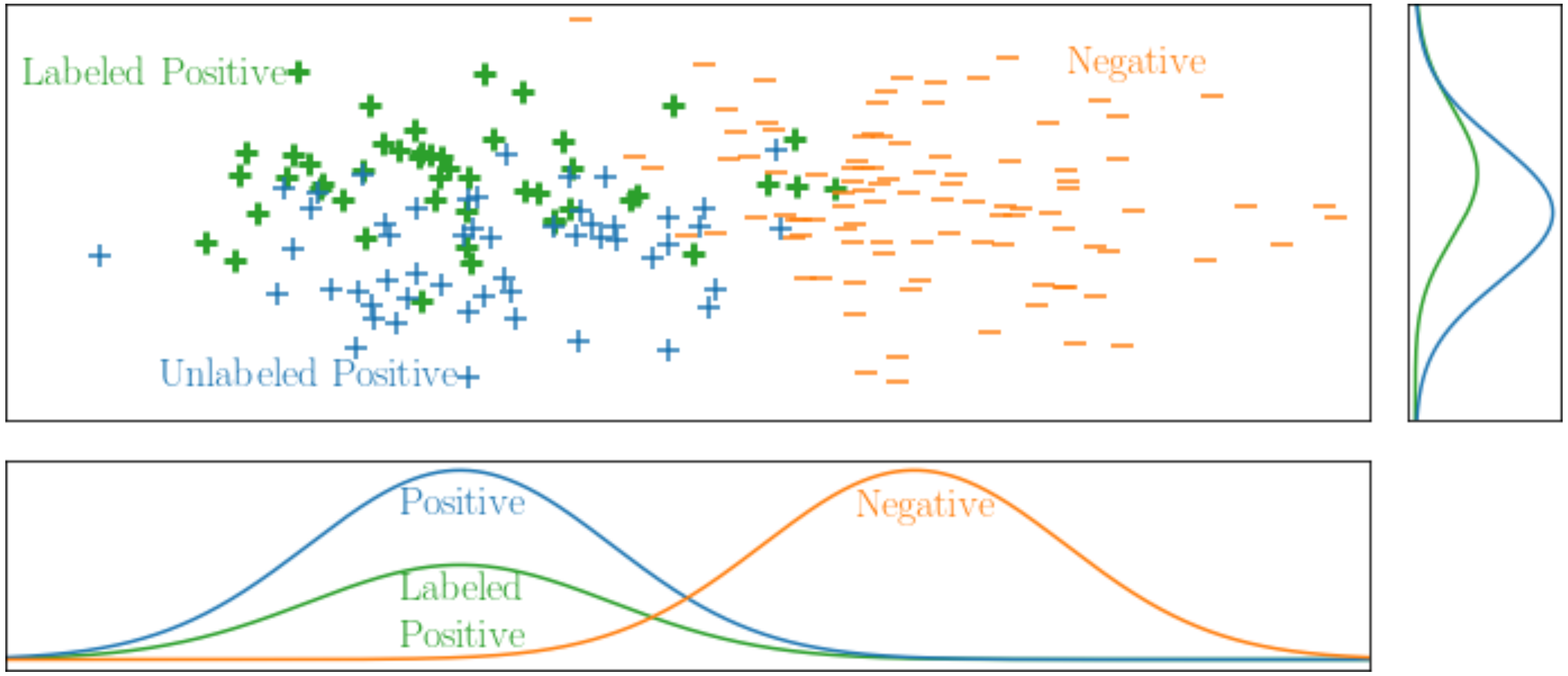
Selected Completely At Random (SCAR)

The ability to employ non-traditional classifiers is what makes the SCAR assumption so attractive and widely used

Selected At Random (SAR) (better name: selected conditionally at random)

Labeled examples are a biased sample from the positive distribution, where the bias completely depends on the attributes x [1]

$$e(x) = \Pr(s = 1 | x, y = 1)$$

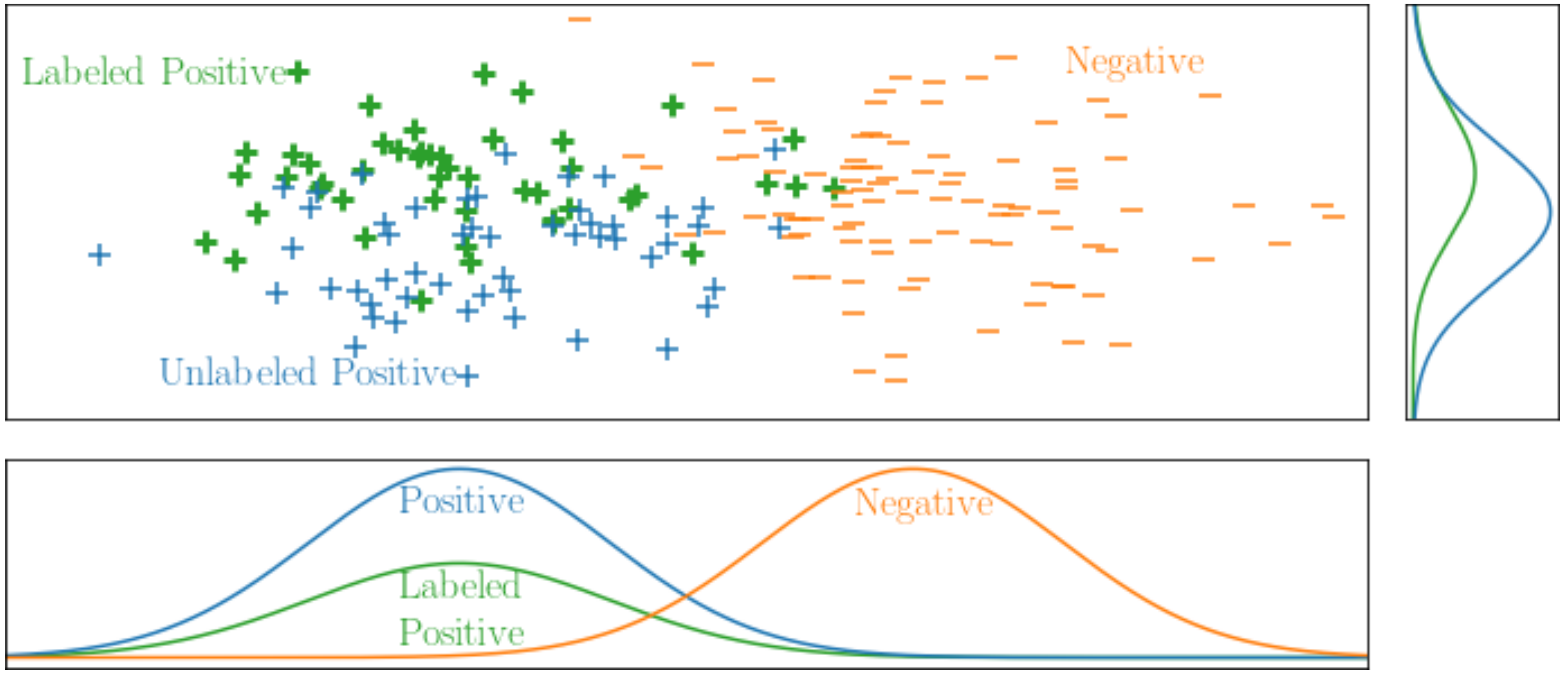


[1] Bekker & Davis. Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data. ECML-PKDD. 2019

Selected At Random (SAR) (better name: selected conditionally at random)

Labeled examples are a biased sample from the positive distribution, where the bias completely depends on the attributes x [1]

$$e(x) = \Pr(s = 1 | x, y = 1)$$

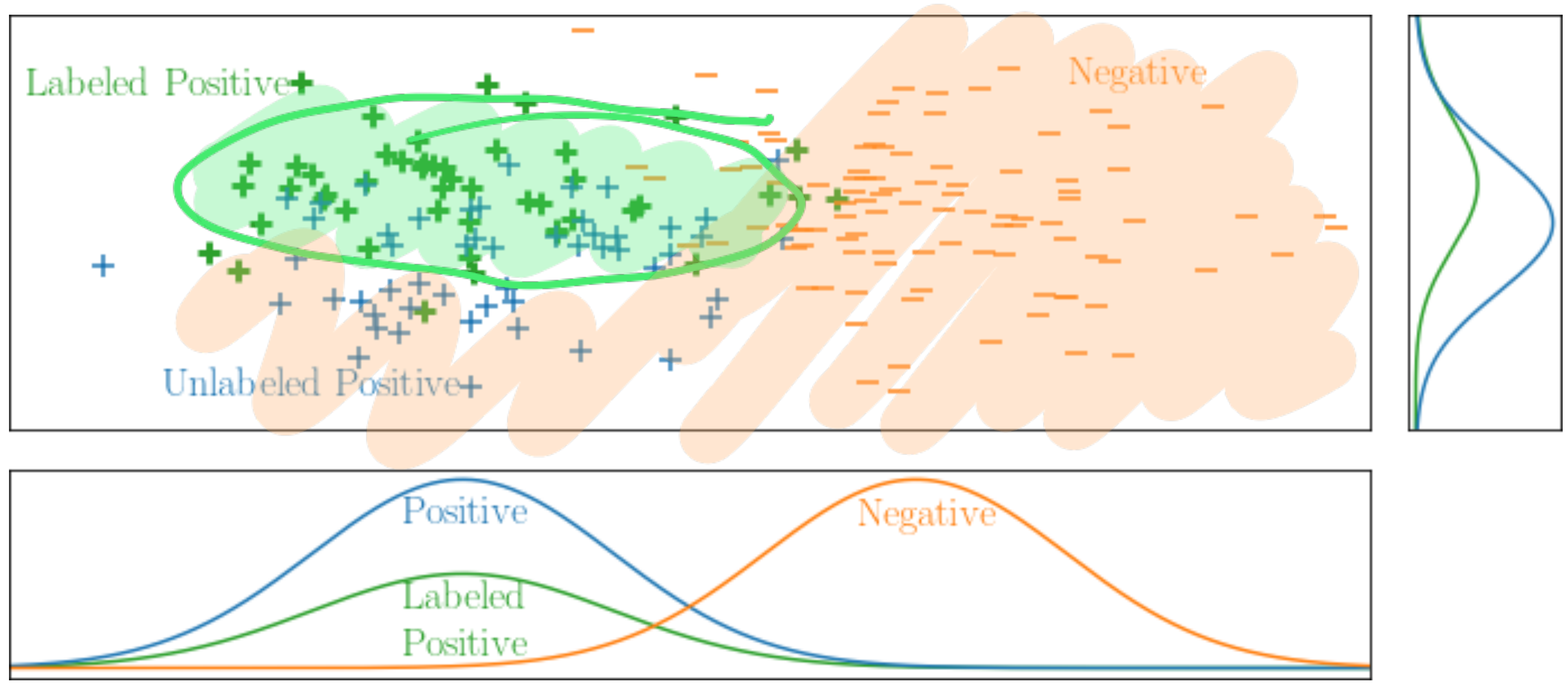


[1] Bekker & Davis. Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data. ECML-PKDD. 2019

Selected At Random (SAR) (better name: selected conditionally at random)

Labeled examples are a biased sample from the positive distribution, where the bias completely depends on the attributes x [1]

$$e(x) = \Pr(s = 1 | x, y = 1)$$

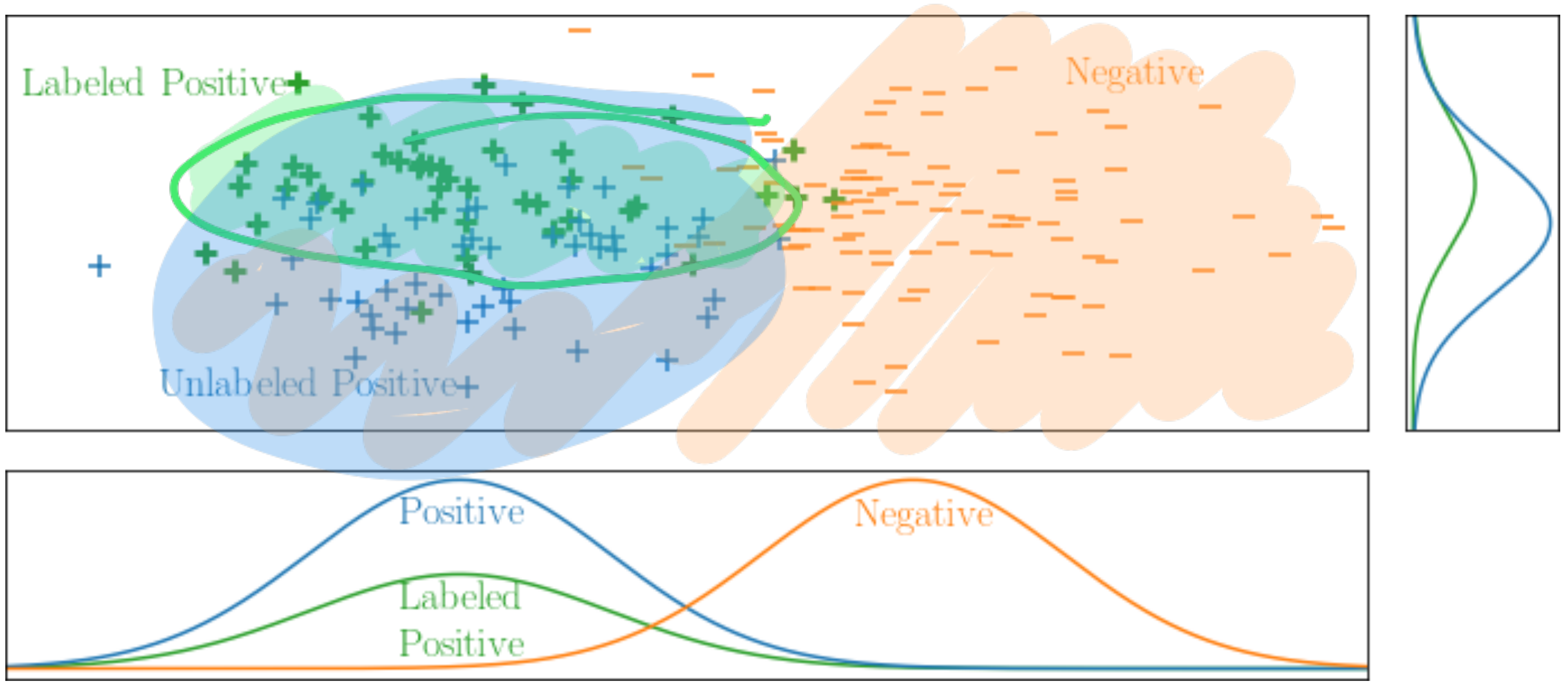


[1] Bekker & Davis. Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data. ECML-PKDD. 2019

Selected At Random (SAR) (better name: selected conditionally at random)

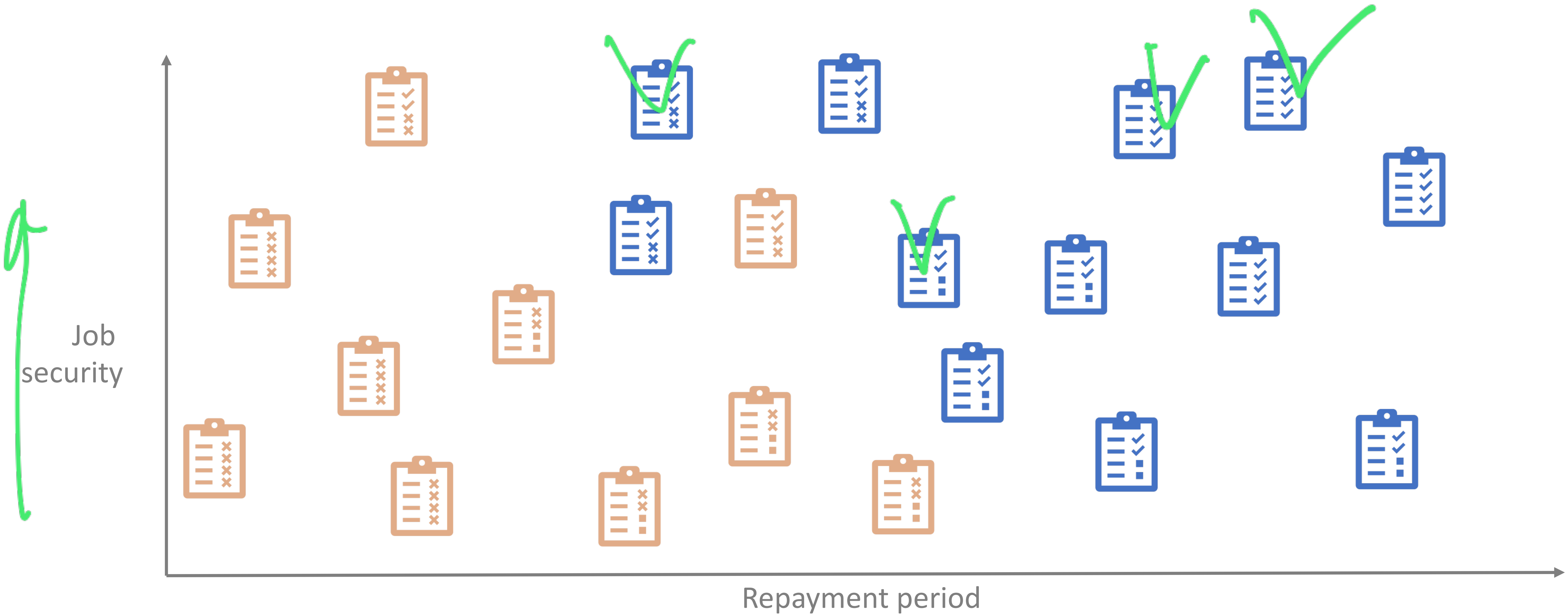
Labeled examples are a biased sample from the positive distribution, where the bias completely depends on the attributes x [1]

$$e(x) = \Pr(s = 1 | x, y = 1)$$



[1] Bekker & Davis. Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data. ECML-PKDD. 2019

Selected At Random (SAR)



Selected At Random (SAR)

If labeling mechanism is known, learning is possible.

Otherwise, additional assumptions are needed.

- class distribution
- labeling mechanism
  - depends only on subset of x

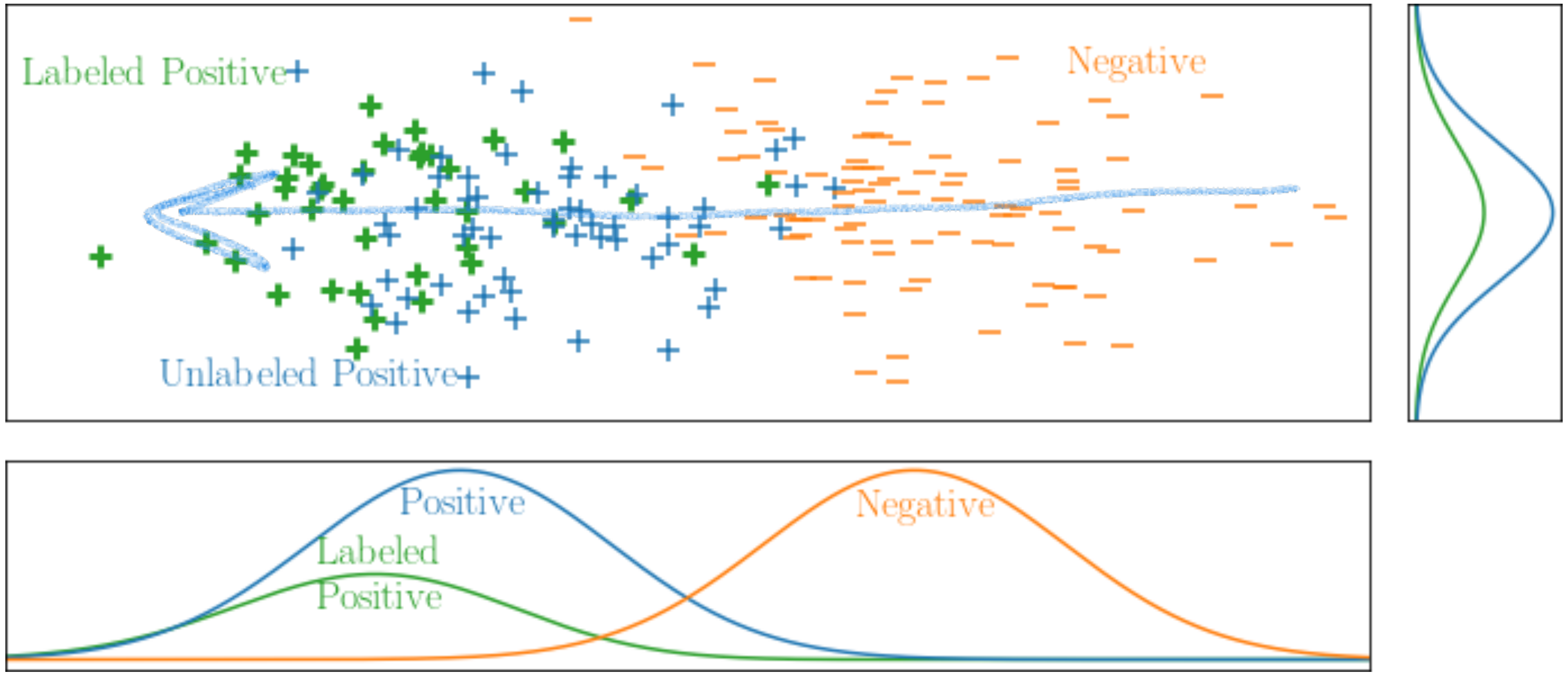
Important field for future research!

Probabilistic Gap (type of SAR)

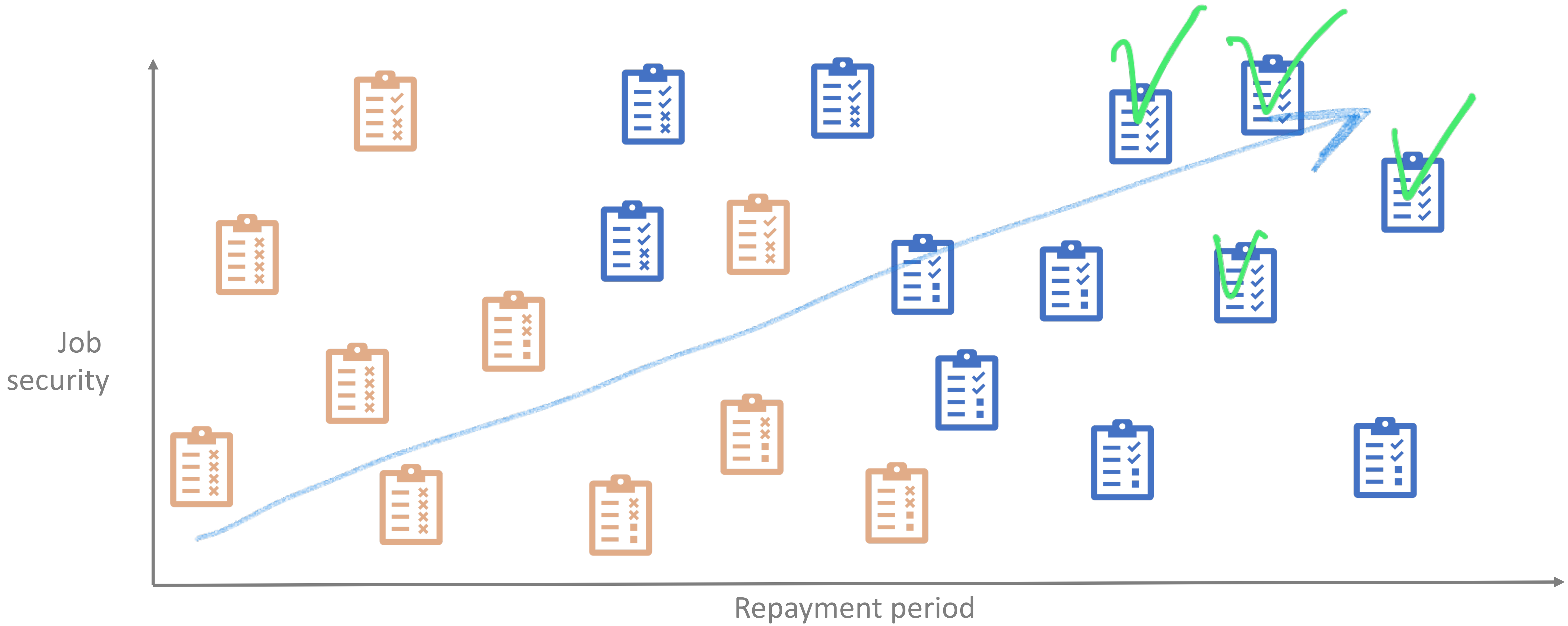
Positive examples that resemble negative examples are less likely to be labeled

$$e(x) = \Pr(s = 1|x, y = 1)$$
$$= f(\Pr(y = 1|x))$$

With $\frac{d}{dt} f(t) > 0$



Probabilistic Gap (type of SAR)



Probabilistic Gap (type of SAR)

Non-traditional classifier predicts $\Pr(s = 1|x)$ again useful!

Properties:

- Non-traditional classifier predicts positive
  → reliable positive example

$$\Pr(s = 1|x) \geq 0.5 \Rightarrow \Pr(y = 1|x) \geq 0.5$$

- Ranking order preserved

$$\Pr(s = 1|x_1) > \Pr(s = 1|x_2) \Leftrightarrow \Pr(y = 1|x_1) > \Pr(y = 1|x_2)$$

[1] He et al. Instance-dependent pu learning by Bayesian optimal relabeling. 2020
[2] Bekker & Davis. Learning from positive and unlabeled data: a survey. MLj. 2020

Data assumptions

Section 3.2 in the survey paper

Negativity

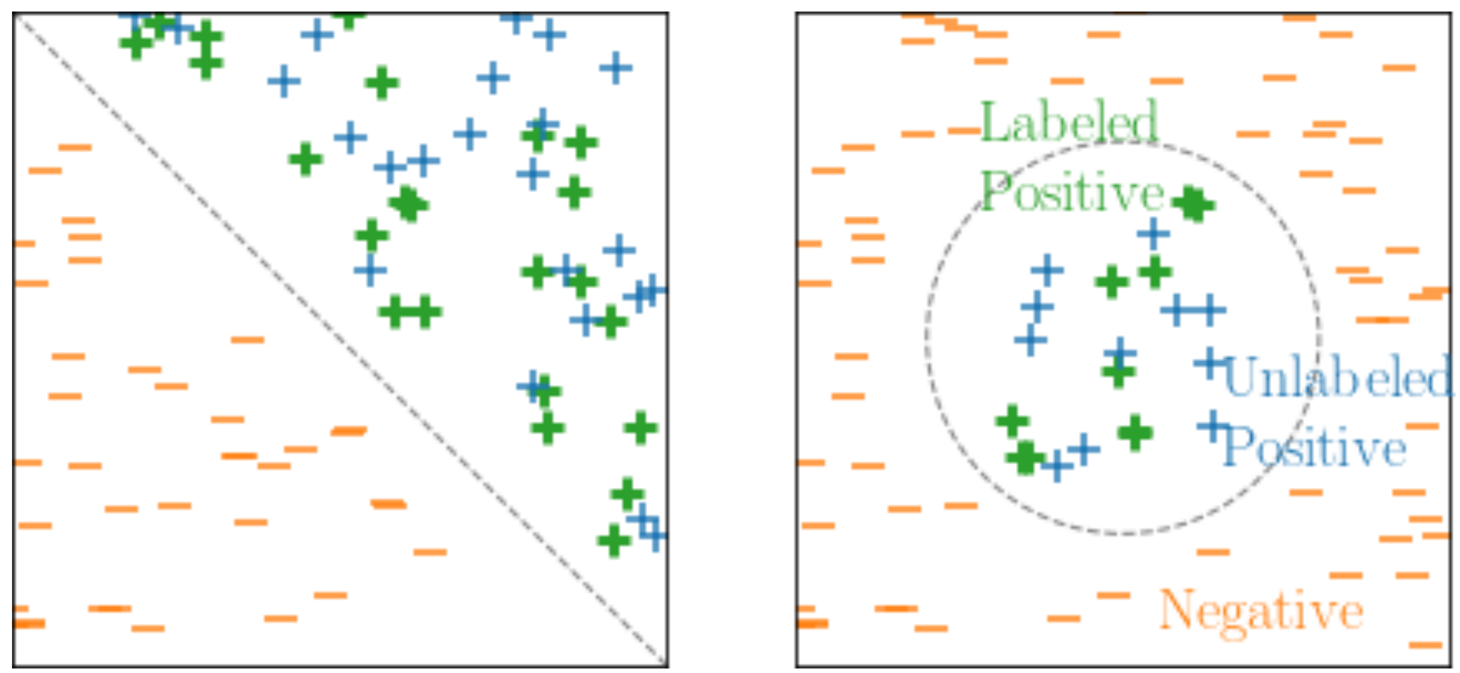
Just ignore that it is PU data….

In combination with SCAR, not necessarily so bad

Only interested in most confident predictions?
→ Ranking order property

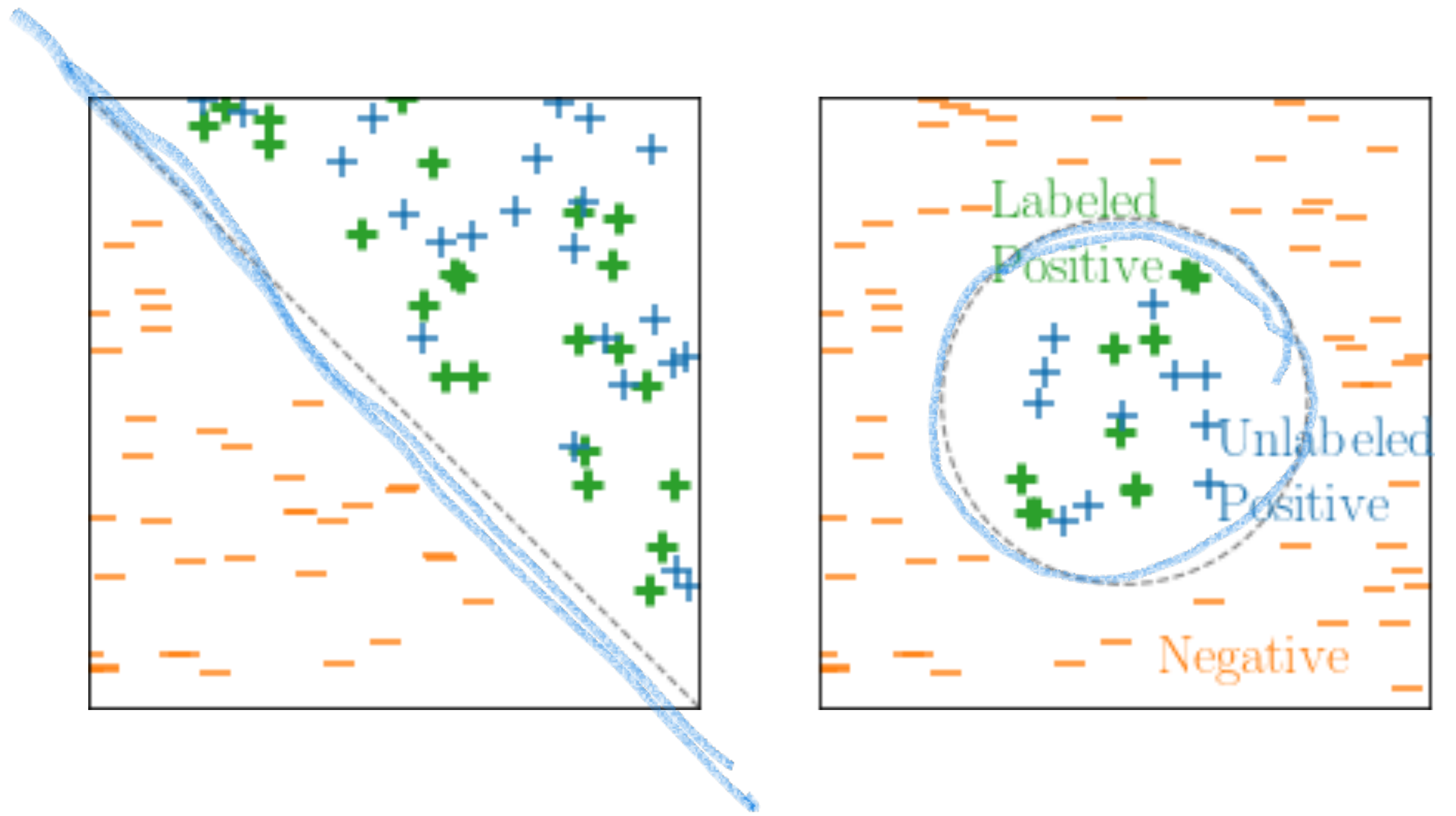
Interested in certain recall?
→ Recall property

Separability



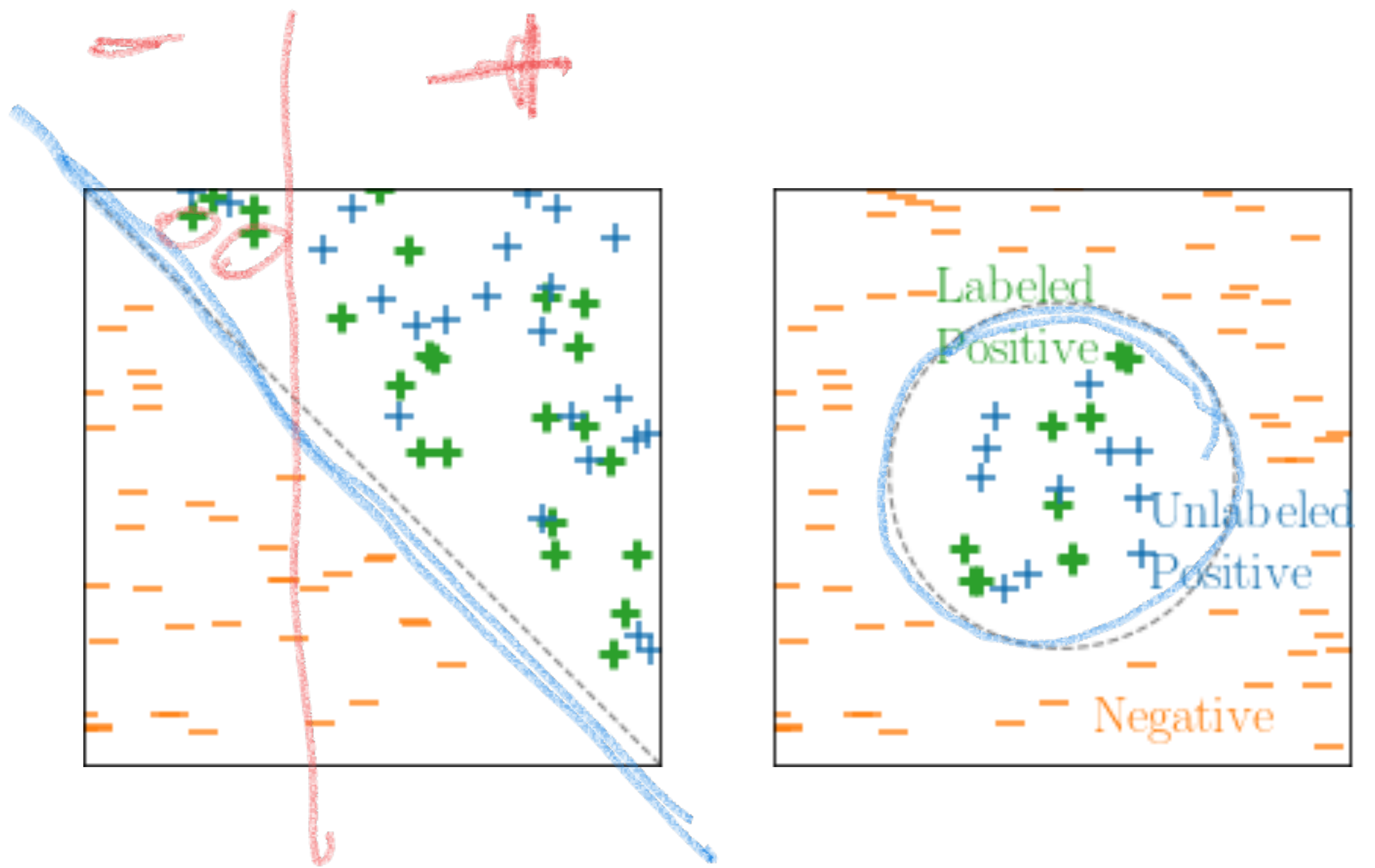
[1] Liu et al. Partially supervised classification of text documents. ICML 2002
[2] Blanchard et al. Semi-supervised novelty detection. JMLR. 2010

Separability



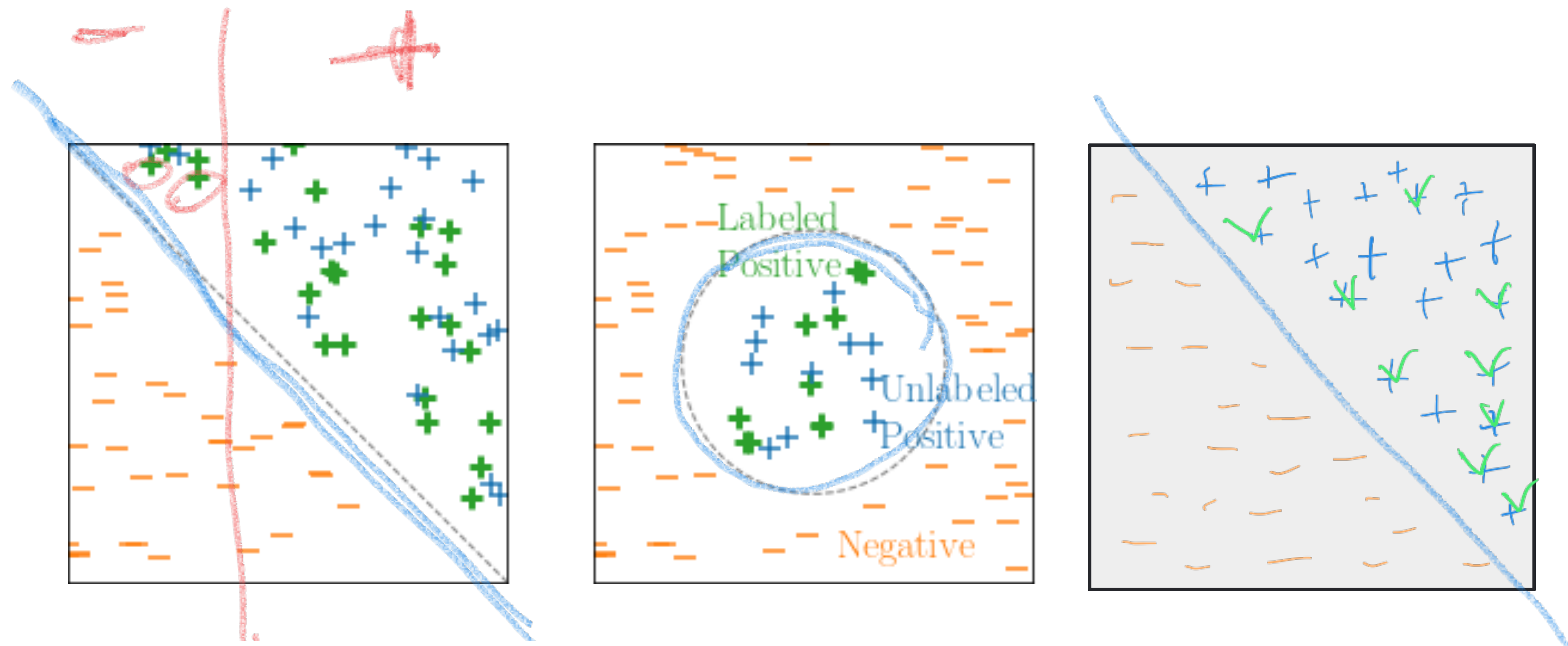
[1] Liu et al. Partially supervised classification of text documents. ICML 2002
[2] Blanchard et al. Semi-supervised novelty detection. JMLR. 2010

Separability



[1] Liu et al. Partially supervised classification of text documents. ICML 2002
[2] Blanchard et al. Semi-supervised novelty detection. JMLR. 2010

Separability



Good classifier [1,2]:

1. Classifies all labeled examples as positive
2. Classifies as few as possible examples as positive

[1] Liu et al. Partially supervised classification of text documents. ICML 2002
[2] Blanchard et al. Semi-supervised novelty detection. JMLR. 2010

Smoothness

Examples that are similar to each other
are likely to have the same label

Useful for identifying reliable negative examples:

examples that are not similar to any of the labeled positives

Up next...