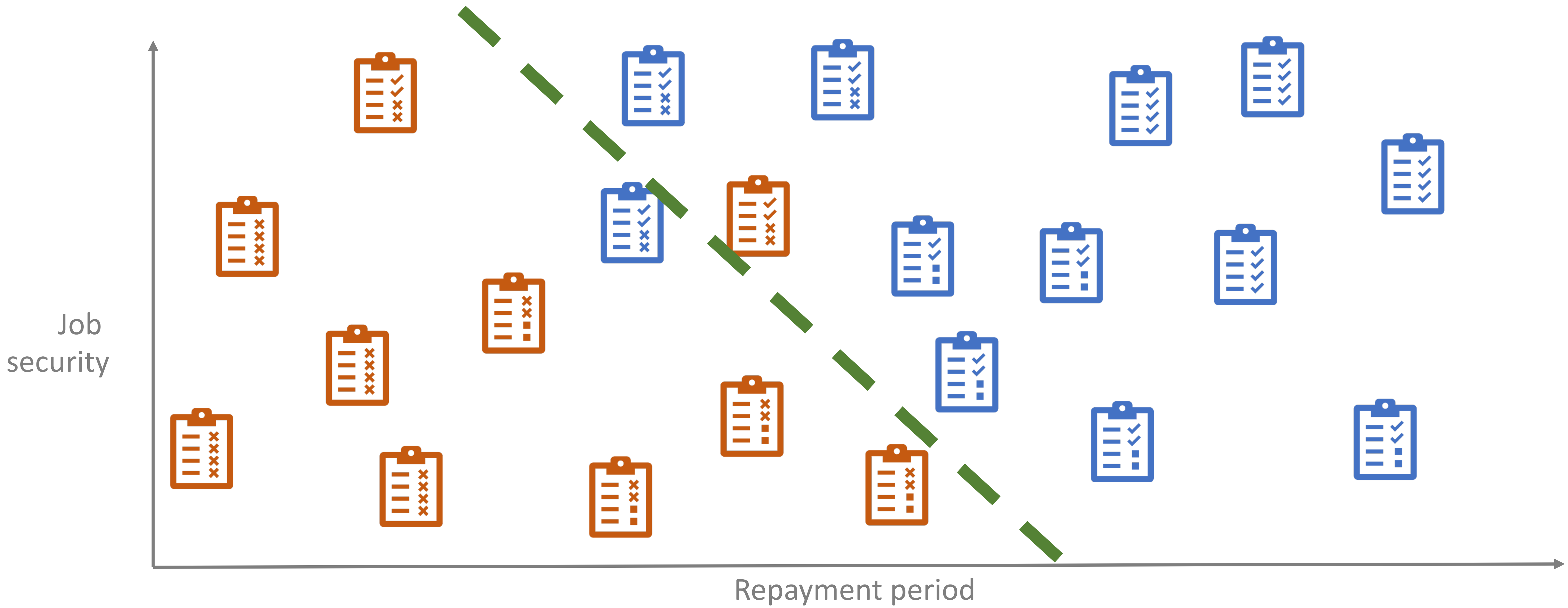
Learning from positive and unlabeled data

2. PU Learning definitions

Section 2 in the survey paper

Binary Classification



Binary Classification

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

enable

Learn

Assumptions

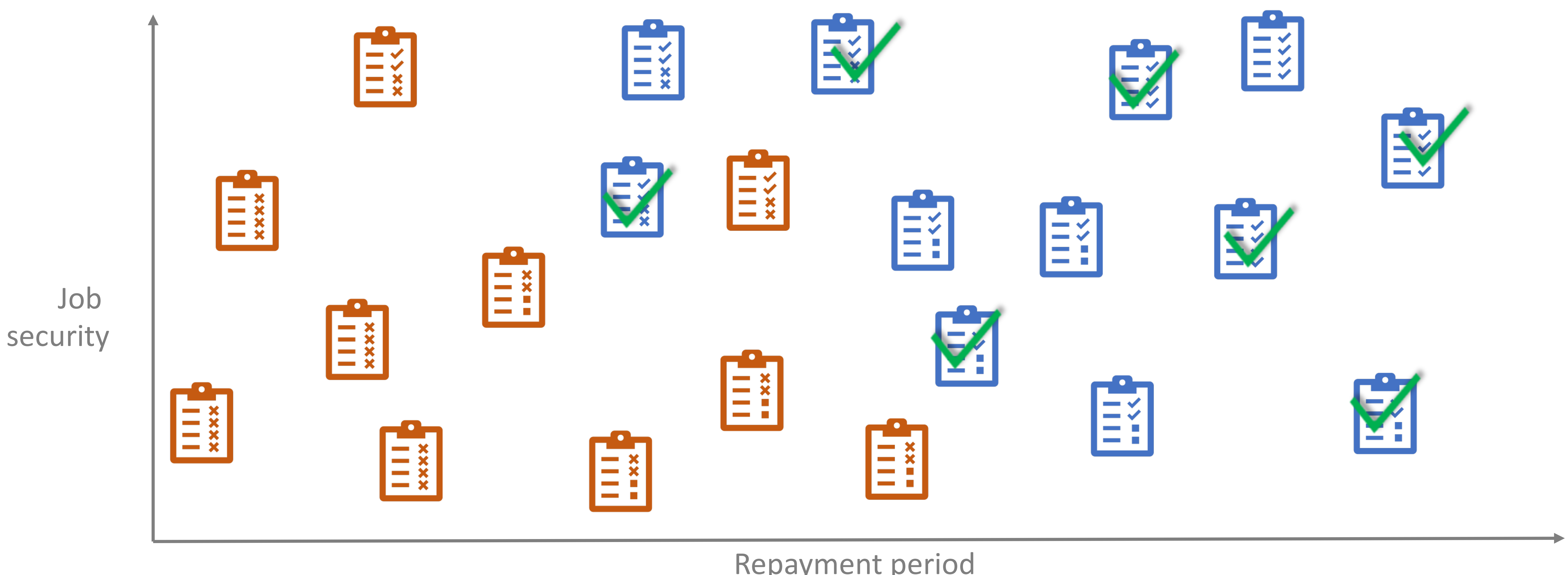
- Enough data
- Training data is iid sample of population

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

  class prior ↳ class prior of negative class
- Class distribution assumptions (classifier bias)

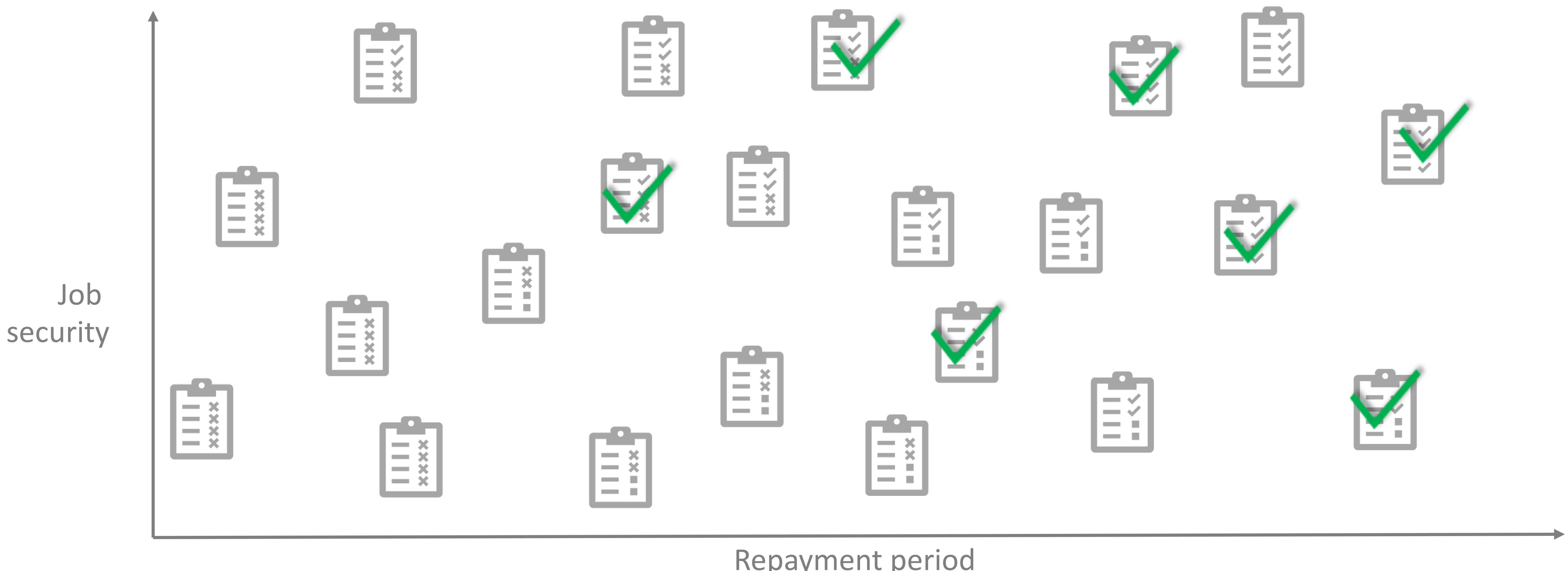
Not all the assumptions need to be strong individually, but together they need to be strong enough to enable learning.

Learning from positive and unlabeled data



In PU data: only a subset of the positive examples are labeled

Learning from positive and unlabeled data



Now it is less clear what the decision boundary should be

Learning from positive and unlabeled data

$y=0 \rightarrow s=0$

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

Learn

enable

Assumptions

- Enough data
- Training data is iid sample of population

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

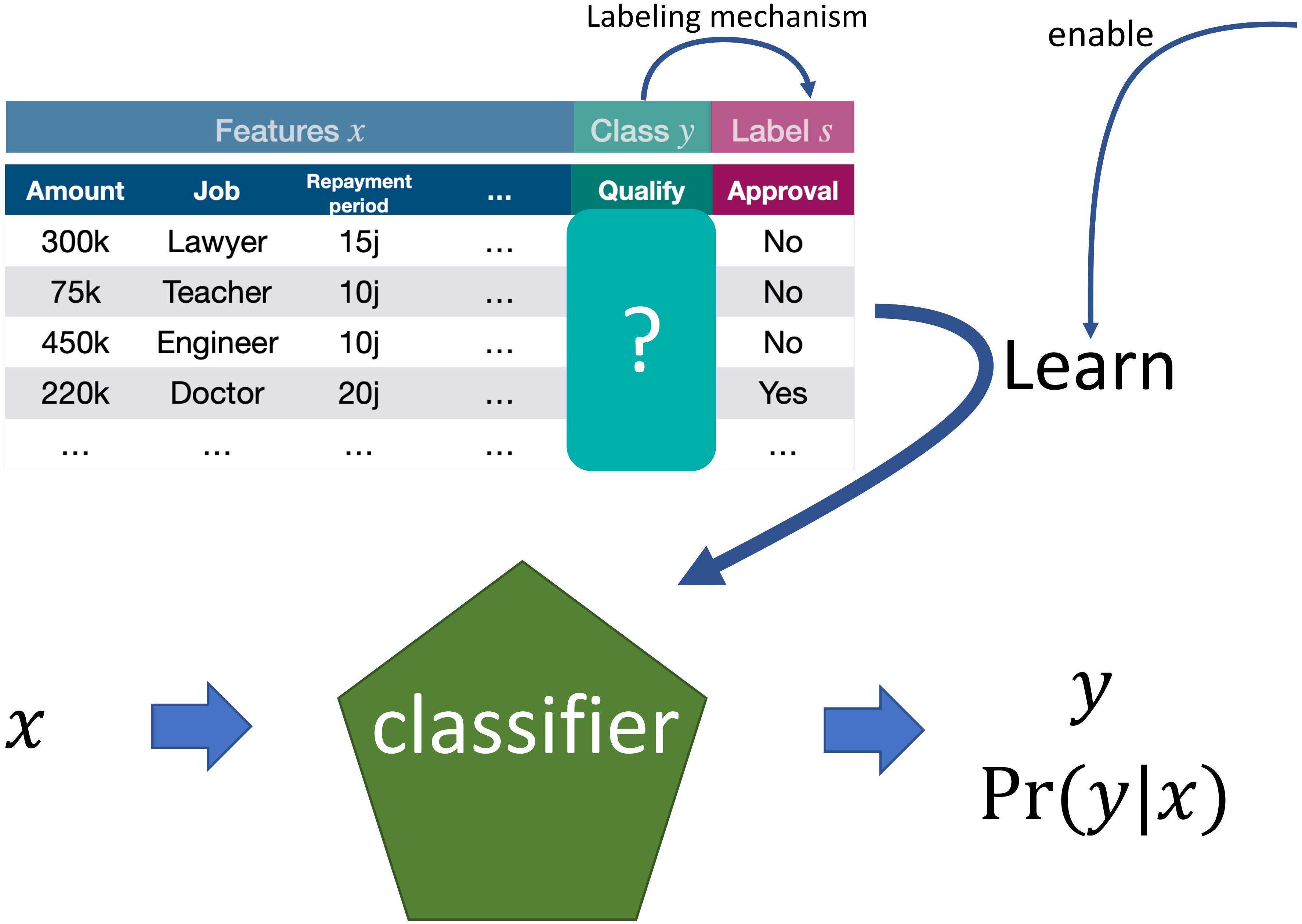
- Class distribution assumptions (classifier bias)

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

Learning from positive and unlabeled data

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
...	...	...	...		...



Assumptions

- Enough data
- Training data is iid sample of population

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

- 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.

Labeling mechanism

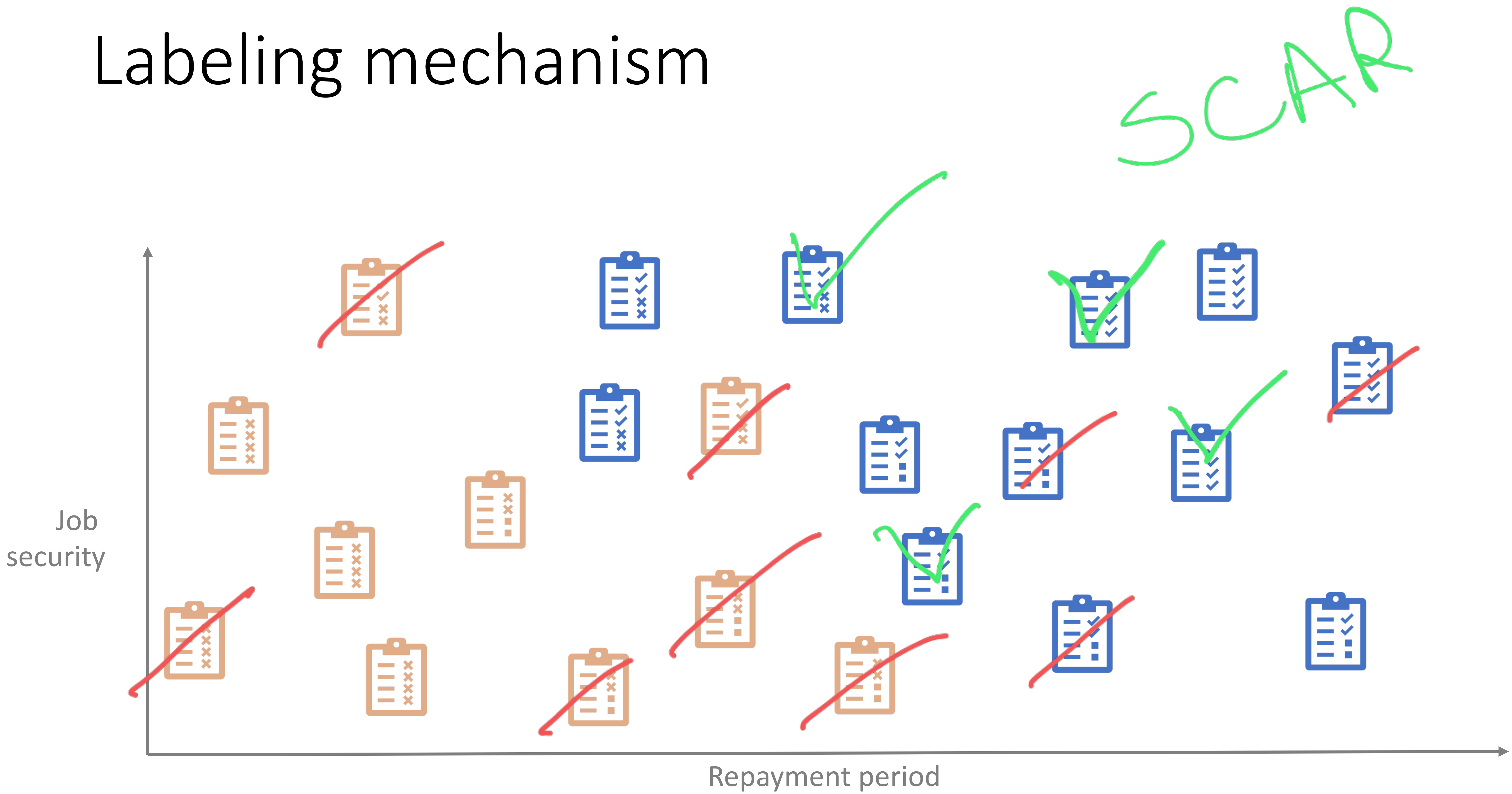
How are positive examples selected to be labeled?

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

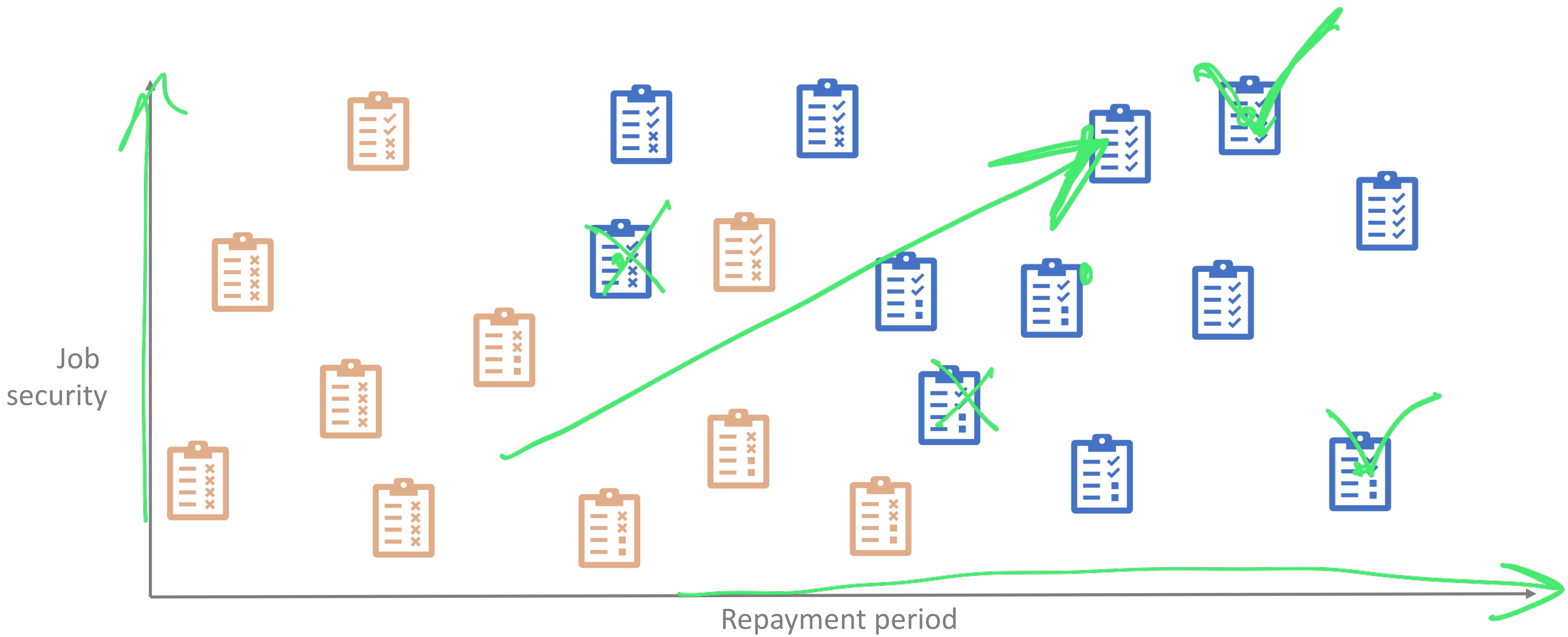
Propensity score

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

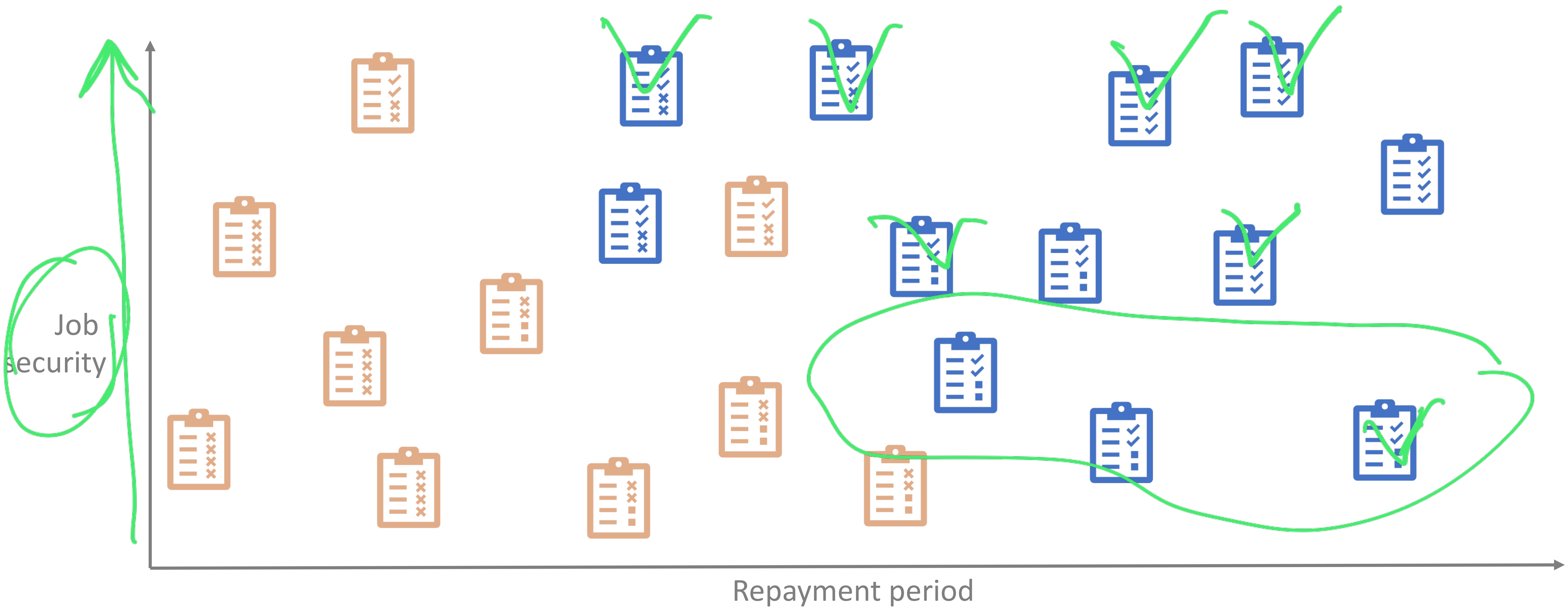
Labeling mechanism



Labeling mechanism



Labeling mechanism



Labeling mechanism

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

The labeled distribution is a biased version of the positive distribution

$$f_l(x) = \Pr(x | s = 1) \sim e(x) f_+(x)$$

→ $P(x | y = 1)$

$$f_l(x) = \frac{e(x)}{c} f_+(x)$$

Normalization constant c is the *label frequency*

$$c = \mathbb{E}_{f_+}[e(x)] = \Pr(s = 1 | y = 1)$$

Labeling mechanism

$e(x) = \Pr(s = 1|y = 1, x)$
$c = \mathbb{E}_{f_+} e(x) = \Pr(s = 1|y = 1$

Important special case: $e(x) = c$

$$f_l(x) = f_+(x)$$

This is called the *Selected Completely At Random (SCAR)* assumption [1]

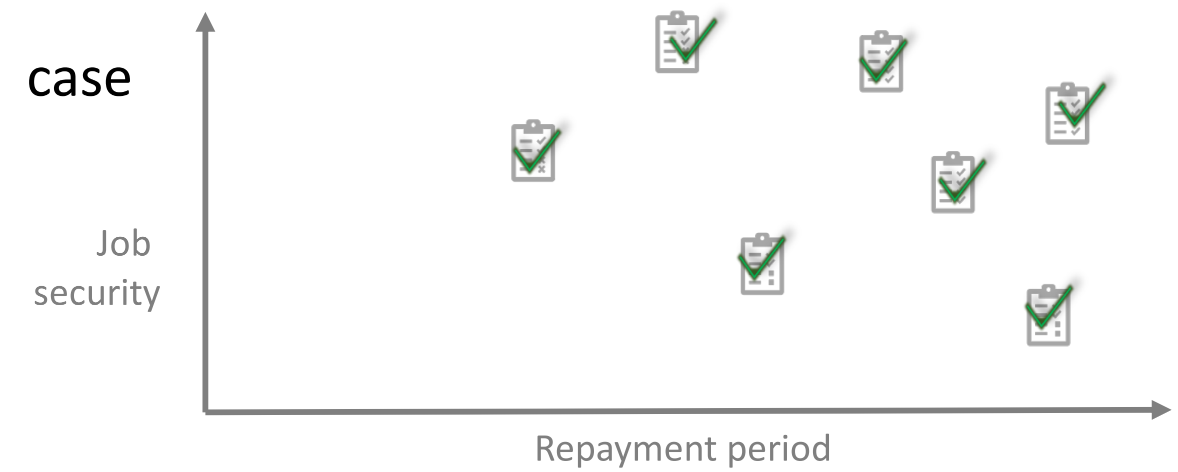
[1] Elkan & Noto. Learning Classifiers from Only Positive and Unlabeled Data. KDD. 2008

Single-training-set vs case-control scenario

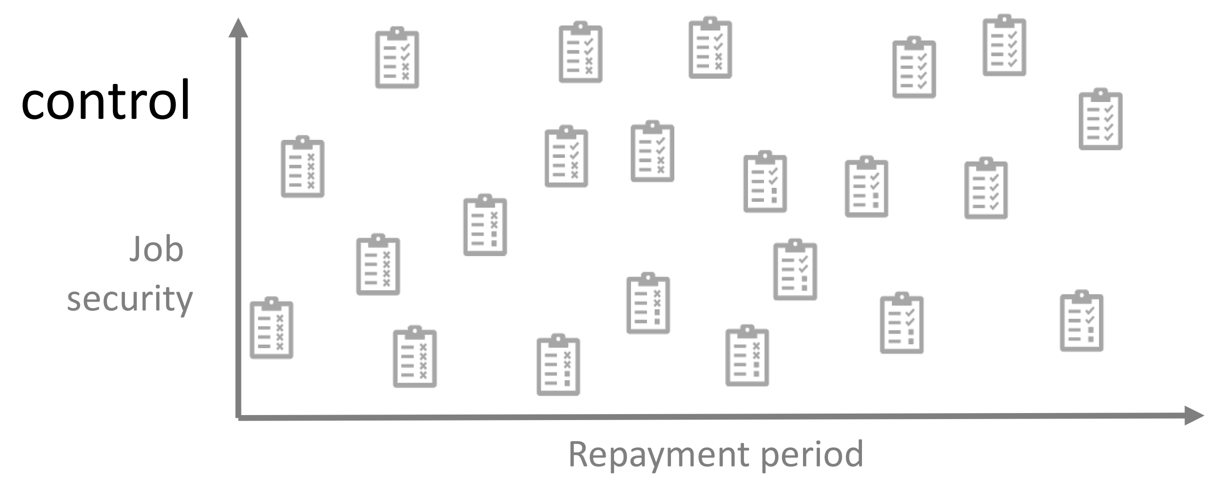
Single-training-set

$$\boldsymbol{x} \sim f(x)$$
$$\sim \alpha c f_l(x) + (1 - \alpha c) f_u(x)$$

Case-control



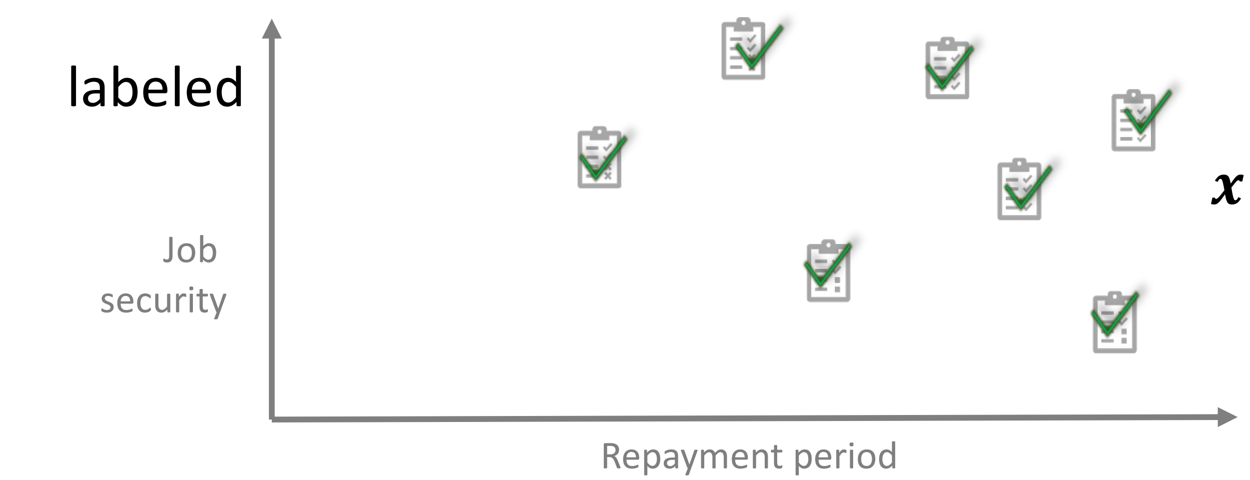
$$\boldsymbol{x} \sim f_l(x)$$



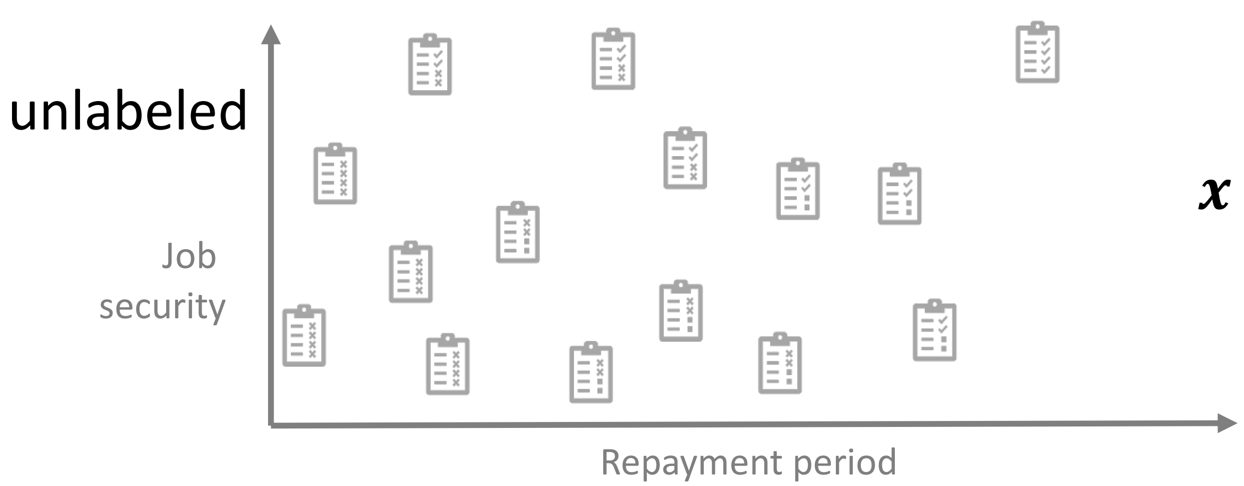
$$\boldsymbol{x} \sim f_u(x)$$
$$\sim f(x)$$

Single-training-set vs case-control scenario

Single-training-set

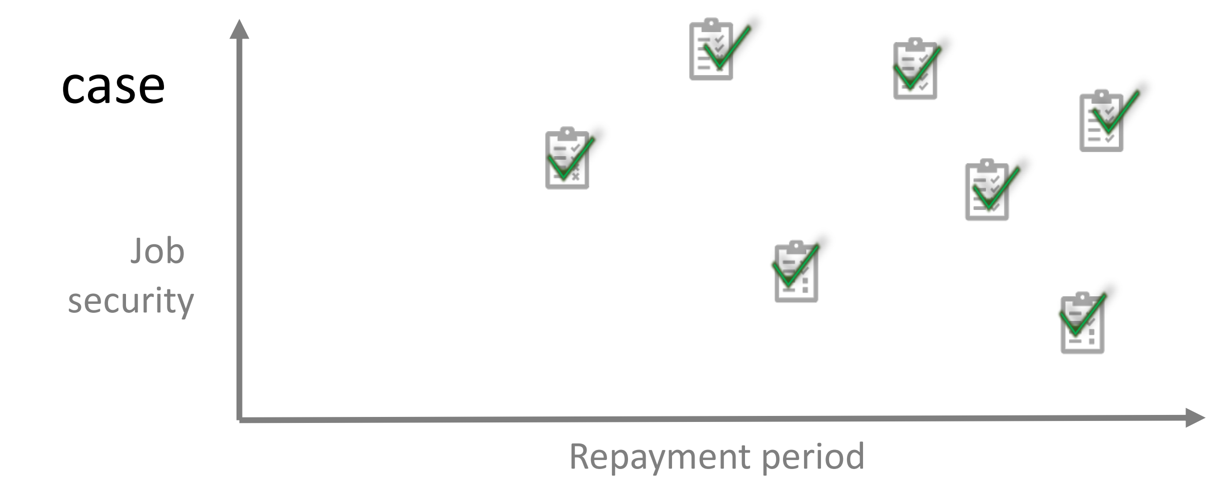


$\boldsymbol{x} \sim f_l(x)$

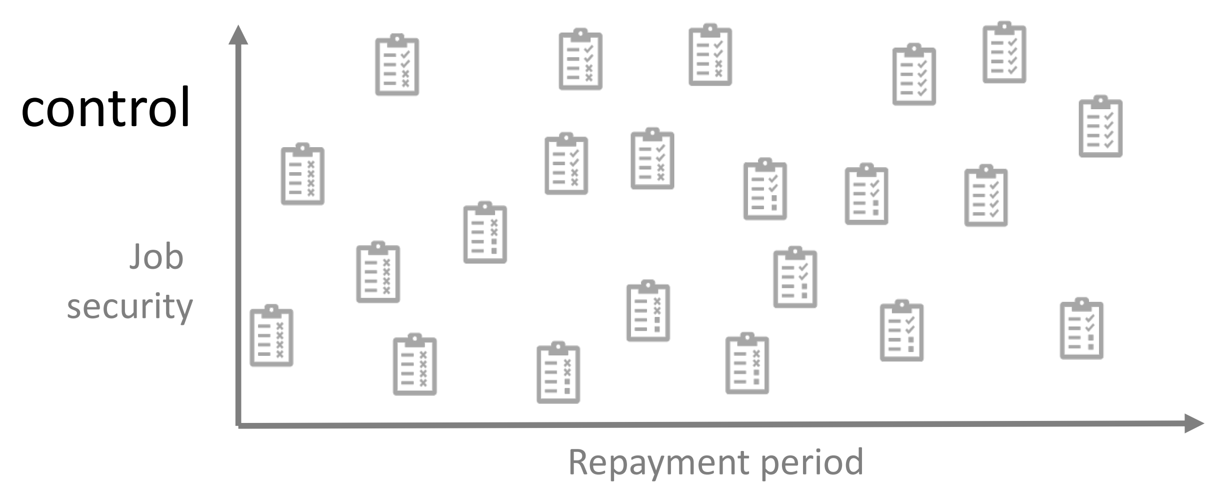


$\boldsymbol{x} \sim f_u(x)$

Case-control



$\boldsymbol{x} \sim f_l(x)$



$\boldsymbol{x} \sim f_u(x)$

$\sim f(x)$

Single-training-set vs case-control scenario

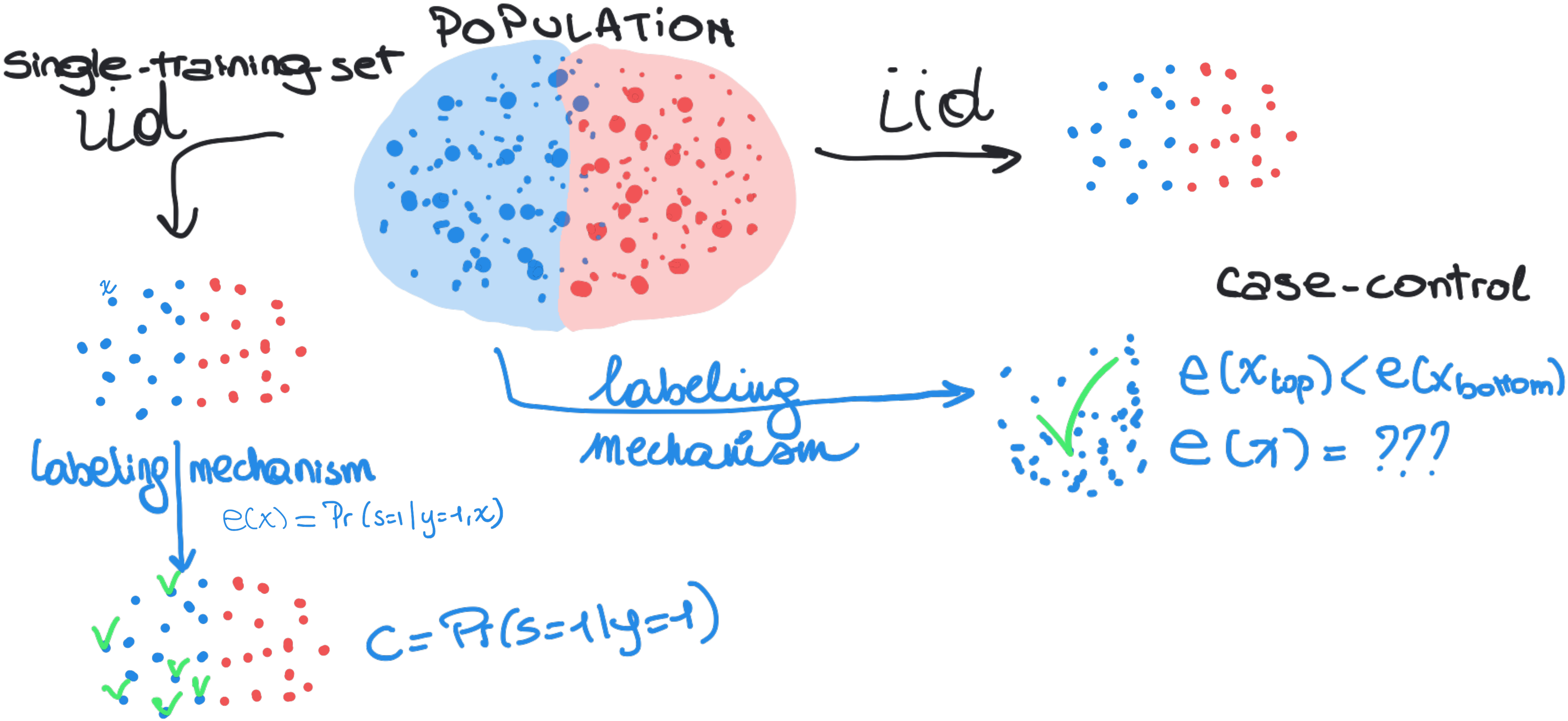
In both scenarios: learner has access to

1. i.i.d. sample from true distribution
2. sample from positive distribution, according to labeling mechansim

➔ Most PU learning methods can handle both scenarios, but some conversion is necessary.

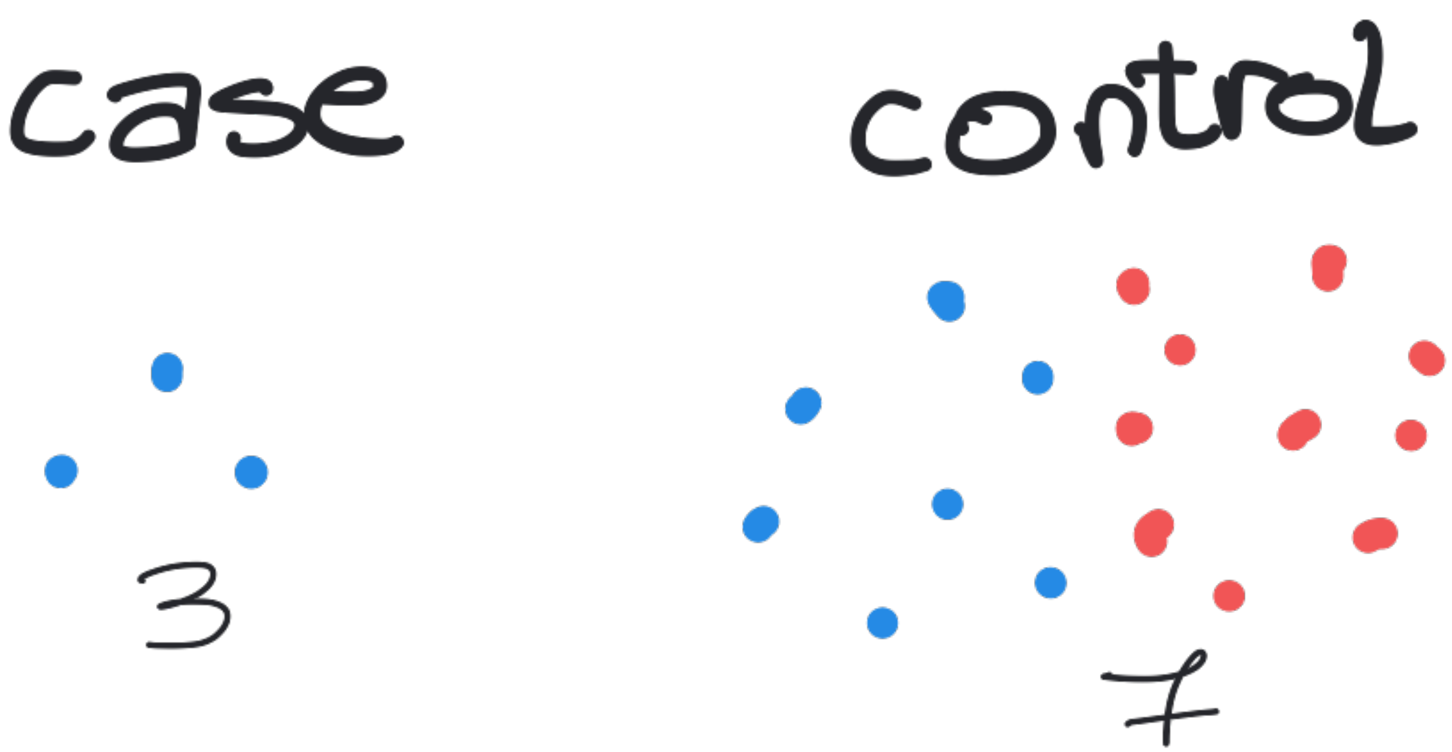
➔ Pay attention to the scenario when using methods/implementations

Single-training-set vs case-control scenario



Label frequency in case-control scenario

$$c = \Pr_{l+u}(s=1 \mid y=1)$$



$$c = \frac{3}{3+7} = 3/10$$

Class prior and label frequency

In single-training-set scenario

Class prior $\alpha = \Pr(y = 1)$

Label frequency $c = \Pr(s = 1 | y = 1)$

Conversion between the two is possible, given label prior $\Pr(s = 1)$, which can be estimated by counting the labeled examples in the data:

$$\Pr(s = 1) = \alpha c$$

$c = 0.5$

$\alpha = 0.5$

$\Pr(s=1) = 0.5 \cdot 0.5$

$= 0.25$

Class prior and label frequency

In case-control scenario

Class prior $\alpha = \Pr_{u}(y = 1)$

Label frequency $c = \Pr_{u+l}(s = 1 | y = 1) = \frac{\Pr_{u+l}(s=1)}{\Pr_{u+l}(y=1)}$

Label prior: $\Pr_{u+l}(s = 1)$

$$\Pr_{u+l}(y=1) = \Pr_{u+l}(s=1)\underbrace{\Pr_{l}(y=1)}_{=1} + (1-\Pr_{u+l}(s=1))\underbrace{\Pr_{u}(y=1)}_{=\alpha}$$

$$c = \frac{\Pr(s=1)}{\alpha(1-\Pr(s=1)) + \Pr(s=1)}$$

$$\alpha = \frac{1-c}{c}\,\frac{\Pr(s=1)}{1-\Pr(s=1)}$$

Single-training-set vs case-control scenario

In this tutorial, we assume the single-training-set scenario,
unless explicitly said otherwise

Up next...