Learning from positive and unlabeled data

An introductory tutorial

4. Two-step techniques

Section 5.1 in the survey paper

General idea

1. Identify reliable negative (and positive) examples
2. Train a classifier using (semi-)supervised techniques
3. Select the best classifier

Combinations of steps in literature

Table 4 Two-step techniques

Method	Step 1	Step 2	Step 3
S-EM Liu et al. (2002)	Spy	EM NB	ΔE
Roc-SVM Li and Liu (2003)	Rocchio	Iterative SVM	$FNR > 5\%$
Roc-Clu-SVM Li and Liu (2003)	Rocchio*	Iterative SVM	$FNR > 5\%$
PEBL Yu et al. (2002); Yu et al. (2004)	1-DNF	Iterative SVM	Last
A-EM Li and Liu (2005)	Augmented Negatives	EM NB	ΔF
LGN Li et al. (2007)	Single Negative	BN	/
PE_PUC Yu and Li (2007)	PE	(EM) NB	Unspecified
WVC/PSOC Peng et al. (2007)	1-DNF*	Iterative SVM	Vote
CR-SVM Li et al. (2010)	Rocchio*	SVM	/
MCLS Chaudhari and Shevade (2012)	k-means	Iterative LS-SVM	Last
C-CRNE Liu and Peng (2014)	C-CRNE	TFIPNDF	/
Pulce Ienco and Pensa (2016)	DILCA	DILCA-KNN	/
PGPU He et al. (2018)	PGPU	Biased SVM	/

Step 1: identifying reliable negative examples

Based on smoothness assumption

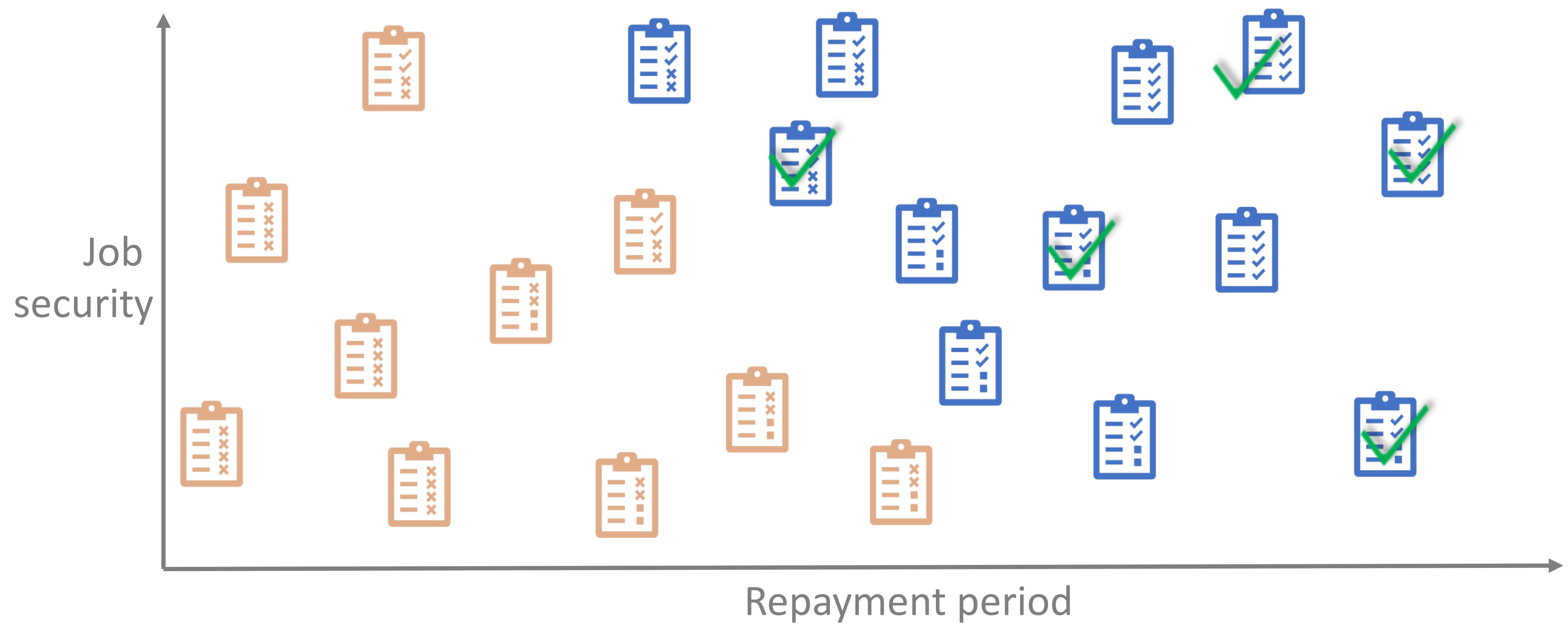
- Use distance metric directly, or TF-IDF
- Train non-traditional classifier and use those probabilities for the distance $Pr(s=1|x)$

- Additional problem: what is far enough?

Step 1: 1-DNF

1. Find strong positive features

2. reliable negative examples
   =
   examples without strong positive features



[1] Yu et al. PEBL: Positive Example Based Learning for Web Page Classification Using SVM. KDD. 2002

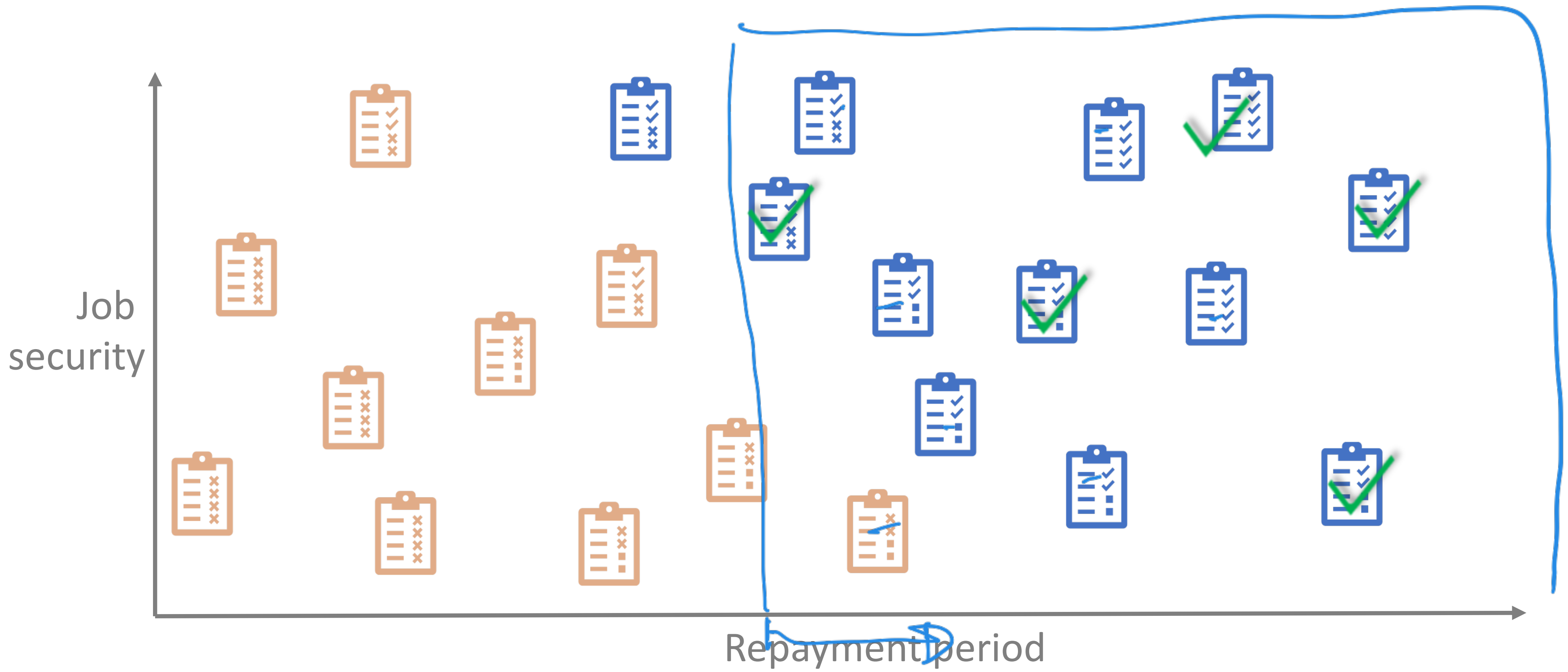
5/5 L

7/18 U-

Step 1: 1-DNF

1. Find strong positive features

2. reliable negative examples
=
examples without strong positive features



[1] Yu et al. PEBL: Positive Example Based Learning for Web Page Classification Using SVM. KDD. 2002

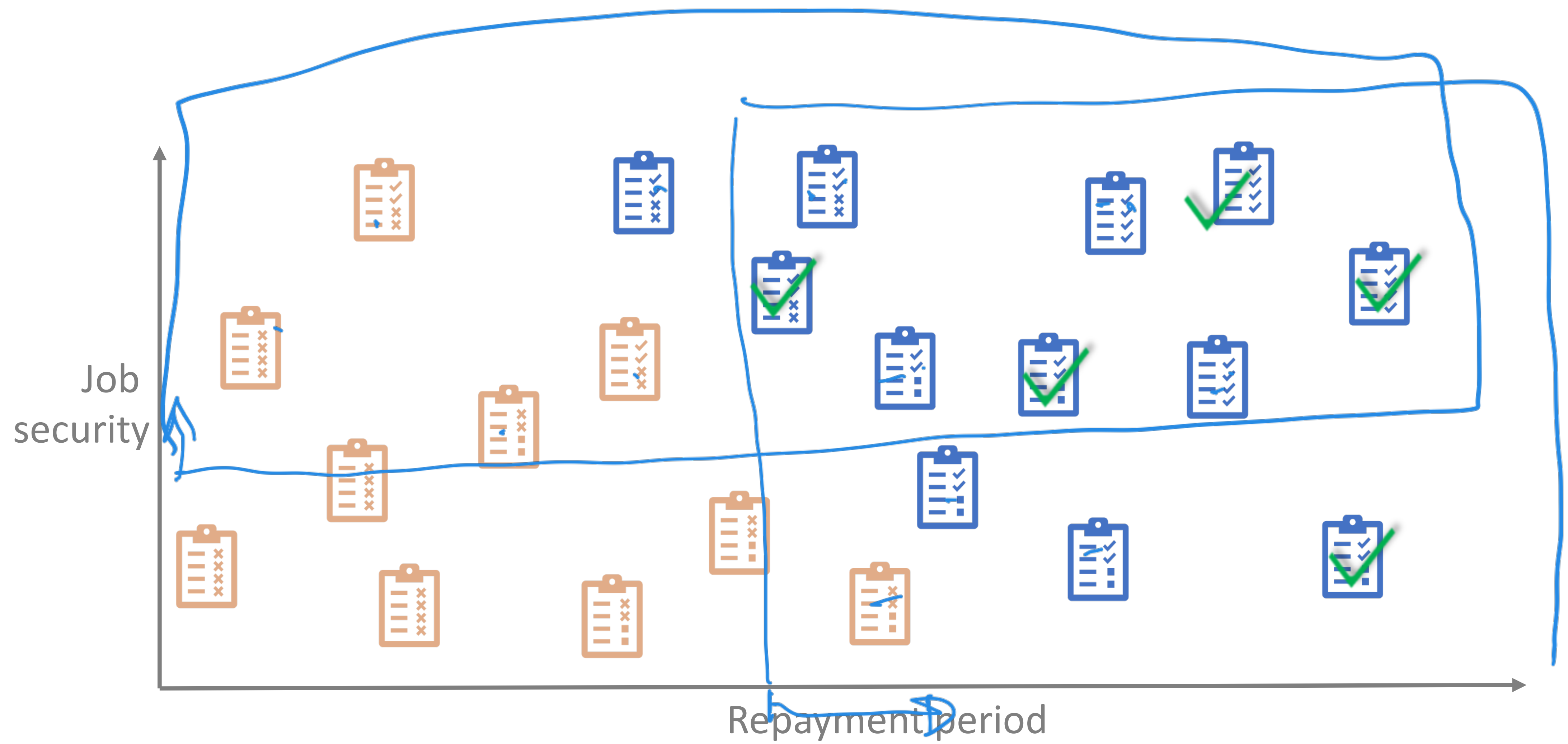
4/5 L   5/5 L

9/18 U   7/18 U-

Step 1: 1-DNF

1. Find strong positive features

2. reliable negative examples
   =
   examples without strong positive features



[1] Yu et al. PEBL: Positive Example Based Learning for Web Page Classification Using SVM. KDD. 2002

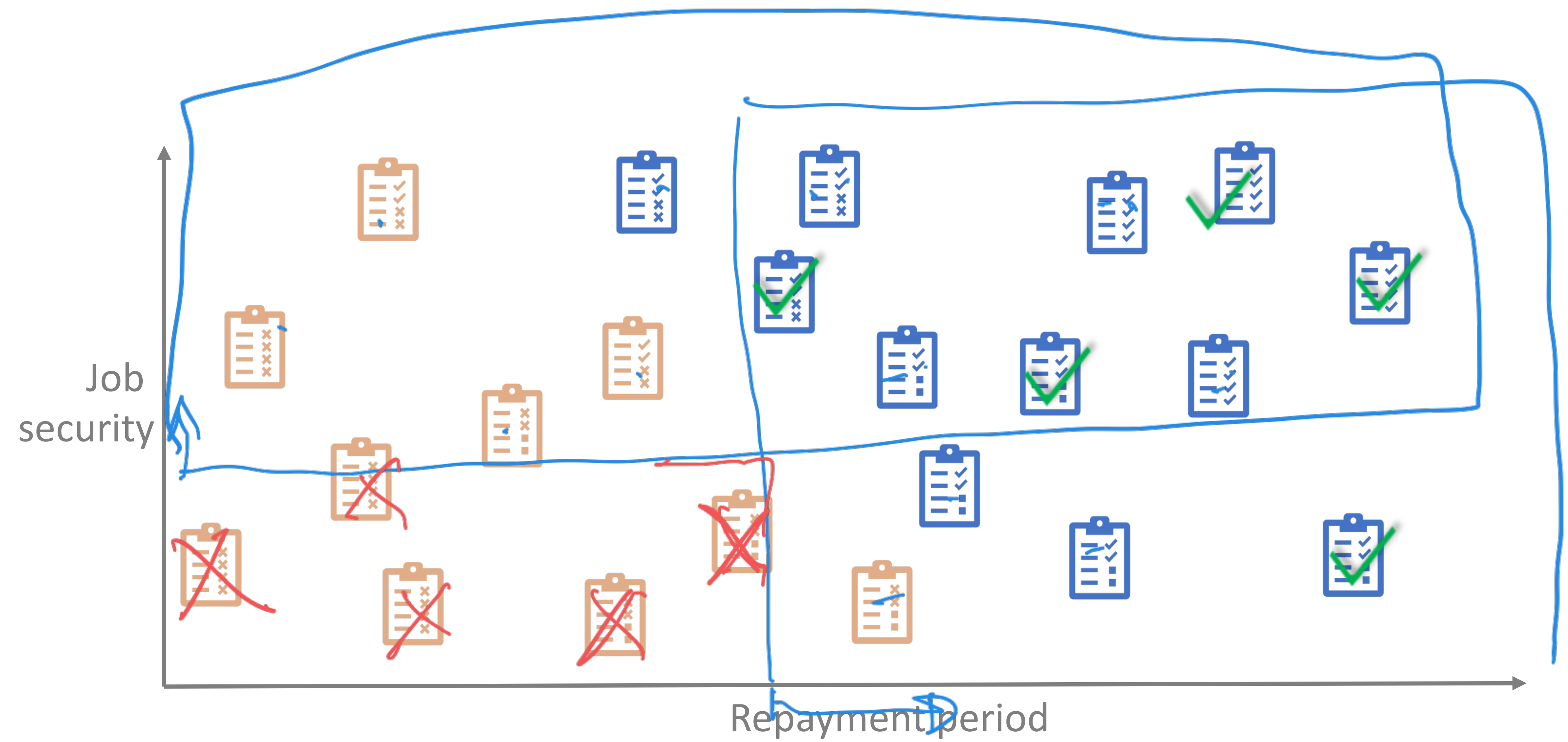
Step 1: 1-DNF

4/5 L
9/18 U

5/5 L
7/18 U-

1. Find strong positive features

2. reliable negative examples
=
examples without strong positive features



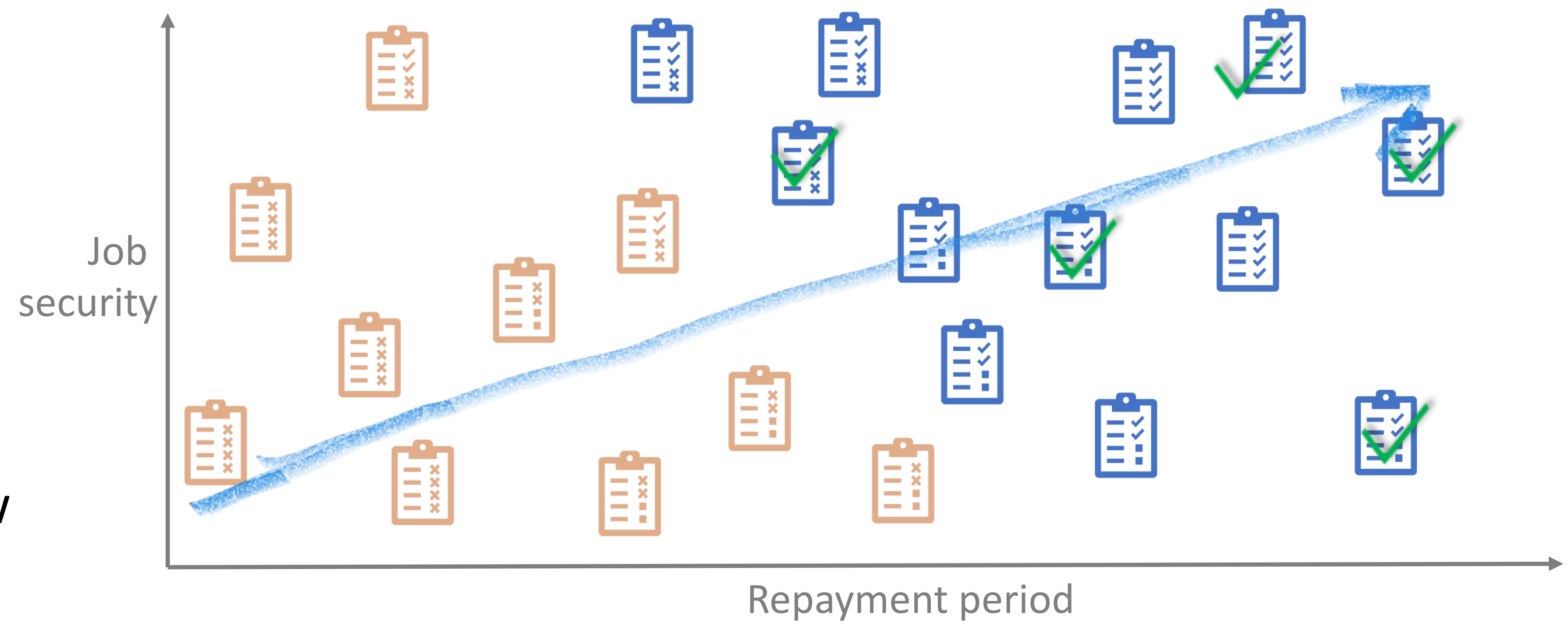
[1] Yu et al. PEBL: Positive Example Based Learning for Web Page Classification Using SVM. KDD. 2002

Step 1: Non-traditional classifier

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

1. Train non-traditional classifier

2. reliable negative examples = examples with low probabilities $\Pr(s = 1|x)$

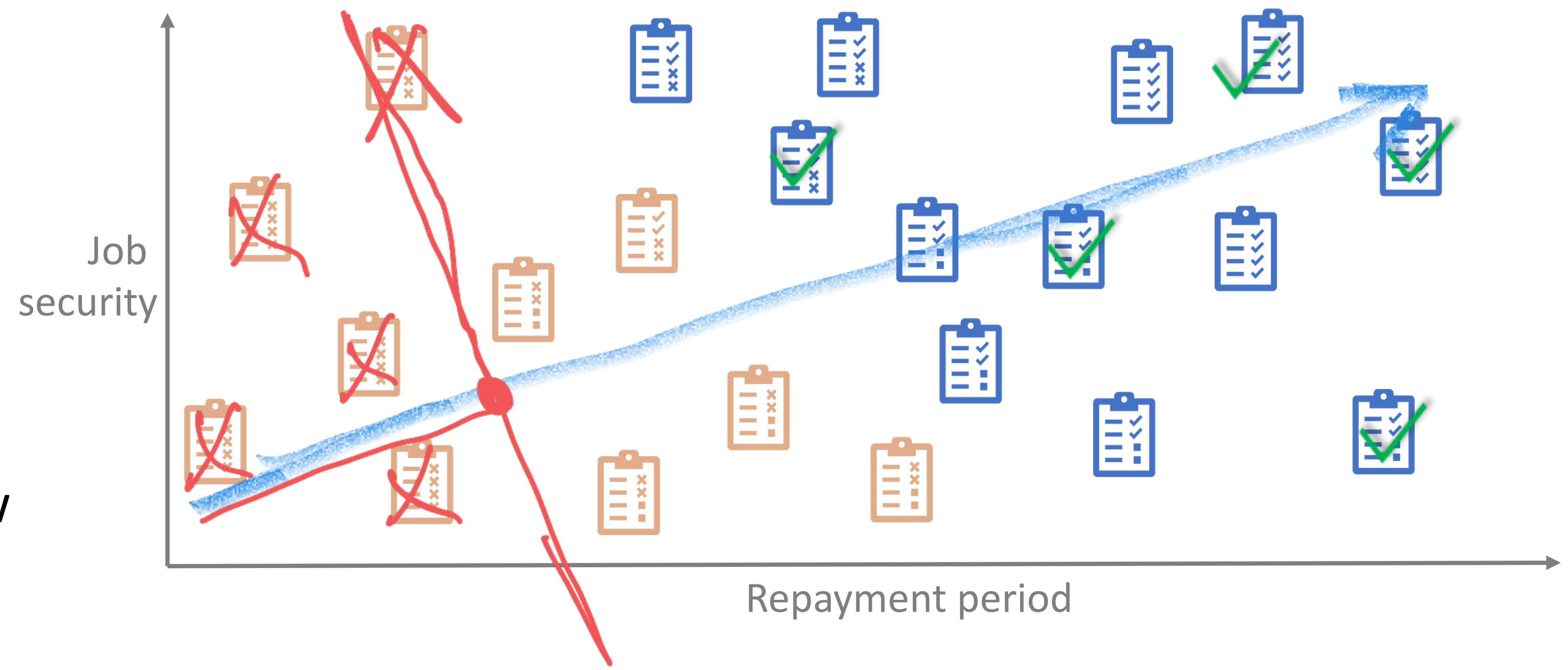


[1] Liu et al. Partially supervised classification of text documents. ICML. 2002
[2] Liu et al. Building text classifiers using positive and unlabeled examples. ICDM. 2003

Step 1: Non-traditional classifier

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

1. Train non-traditional classifier
2. reliable negative examples = examples with low probabilities $\Pr(s = 1|x)$



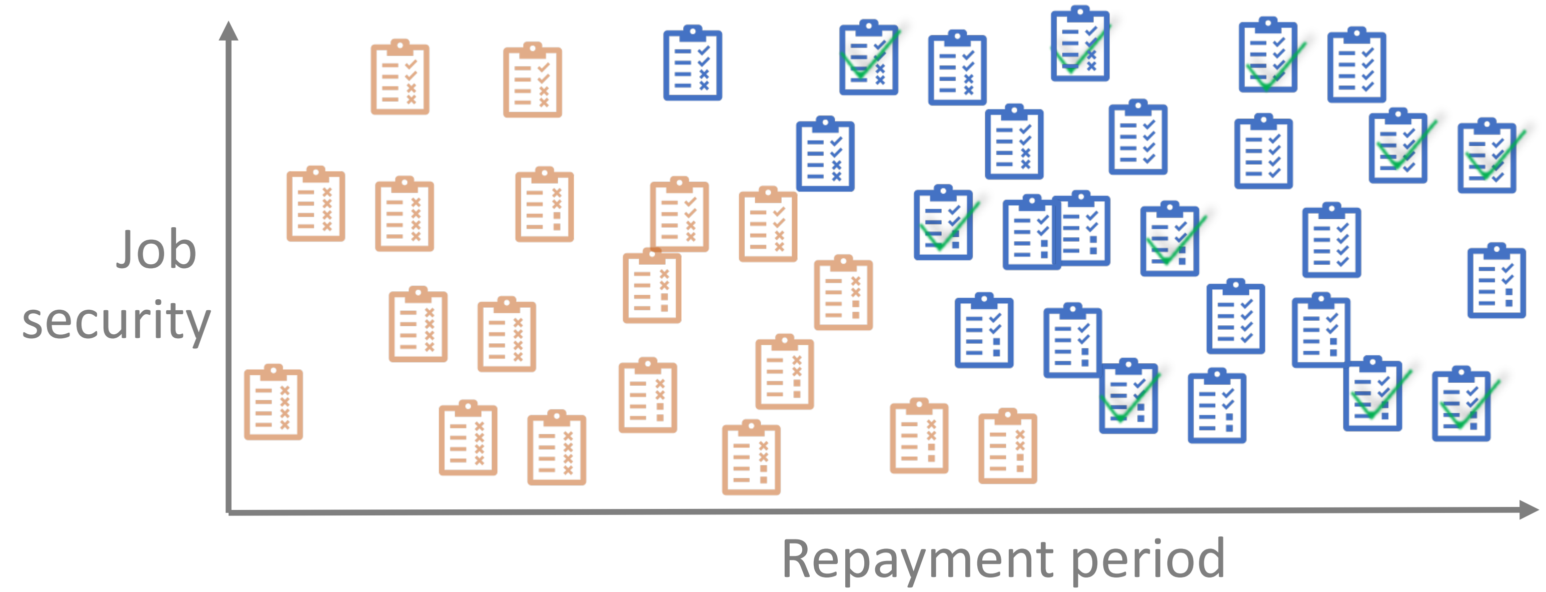
[1] Liu et al. Partially supervised classification of text documents. ICML. 2002
[2] Liu et al. Building text classifiers using positive and unlabeled examples. ICDM. 2003

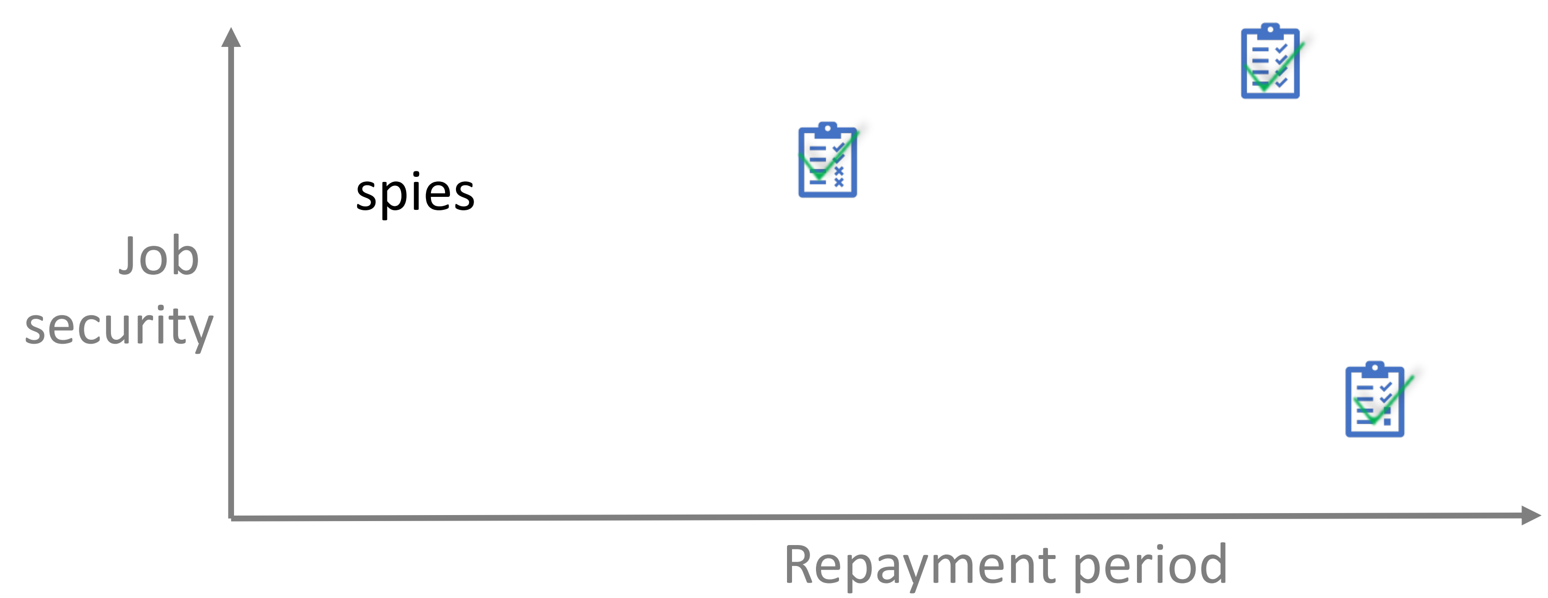
Step 1: Spy

1. Keep set of labeled “spies” behind when training non-traditional classifier

2. reliable negative examples
   =
   examples with probabilities lower than spy probabilities

   $\Pr(s = 1 | x_{neg}) < \Pr(s = 1 | x_{spy})$





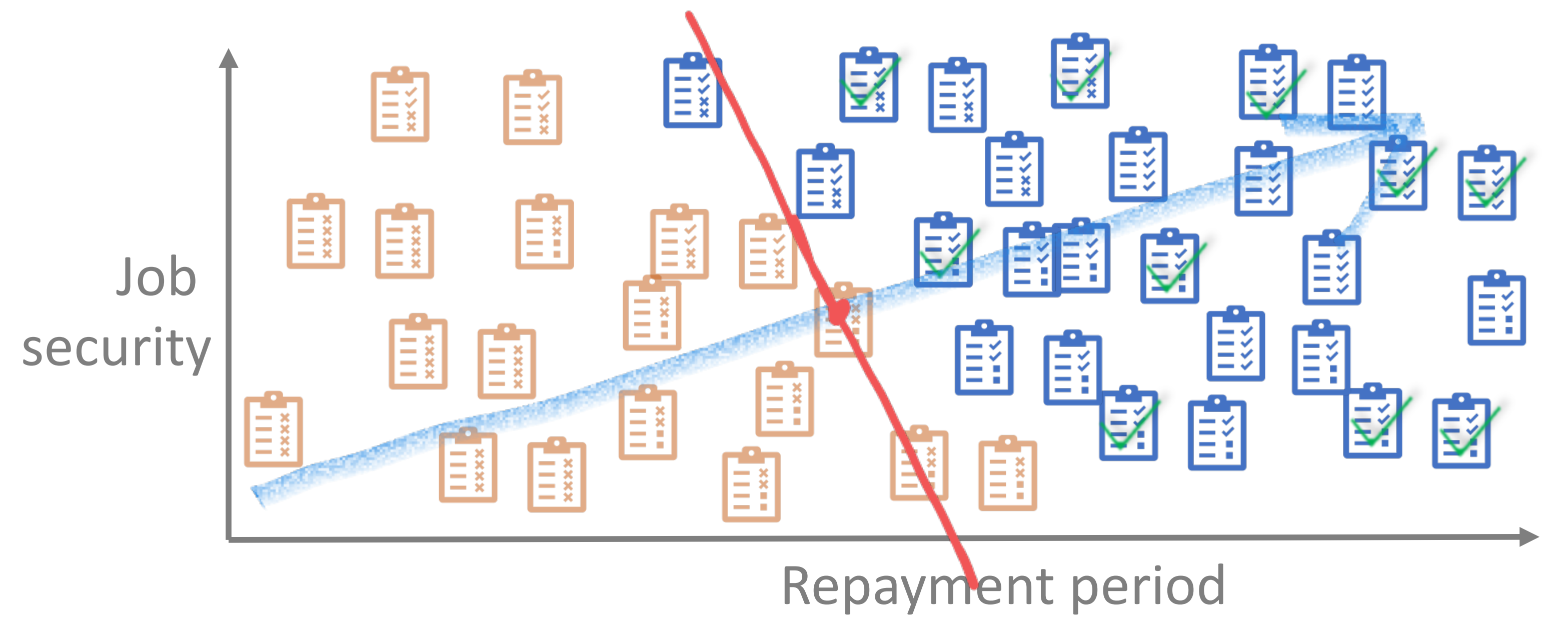
[1] Liu et al. Partially supervised classification of text documents. ICML. 2002

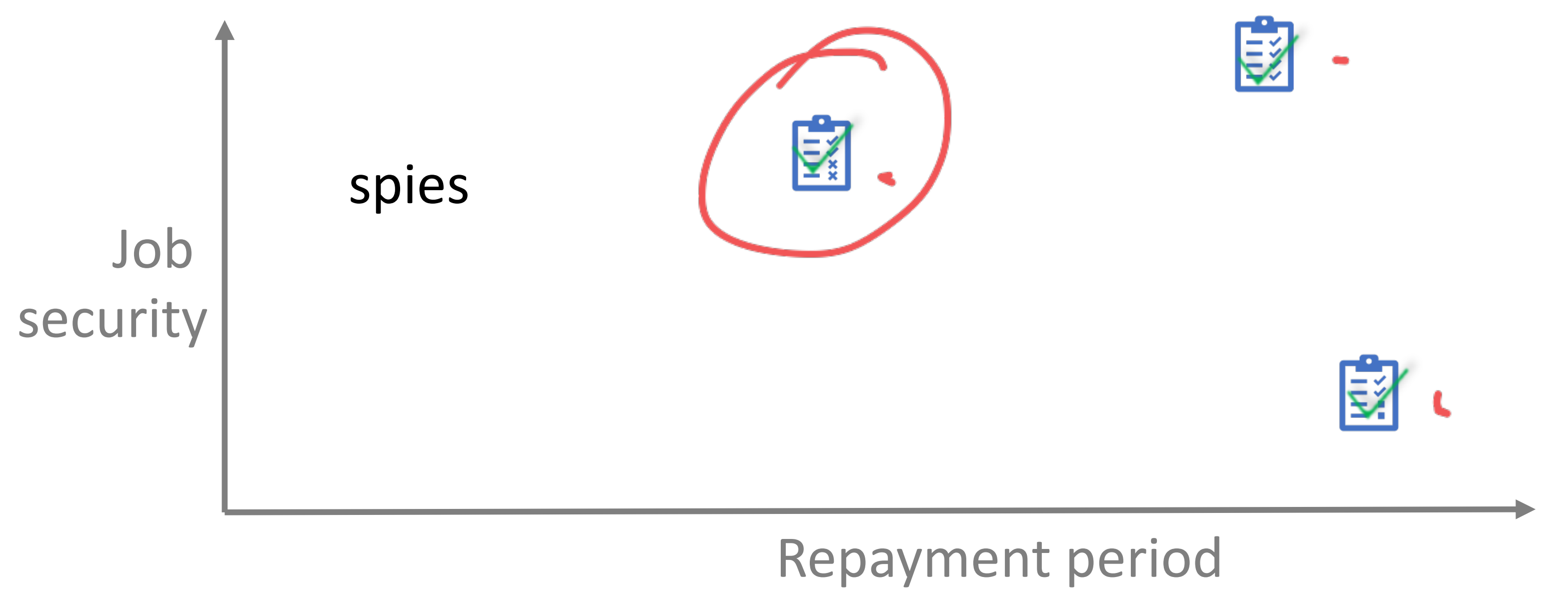
Step 1: Spy

1. Keep set of labeled "spies" behind when training non-traditional classifier

2. reliable negative examples
   =
   examples with probabilities lower than spy probabilities

   $$\Pr(s = 1 | x_{neg}) < \Pr(s = 1 | x_{spy})$$





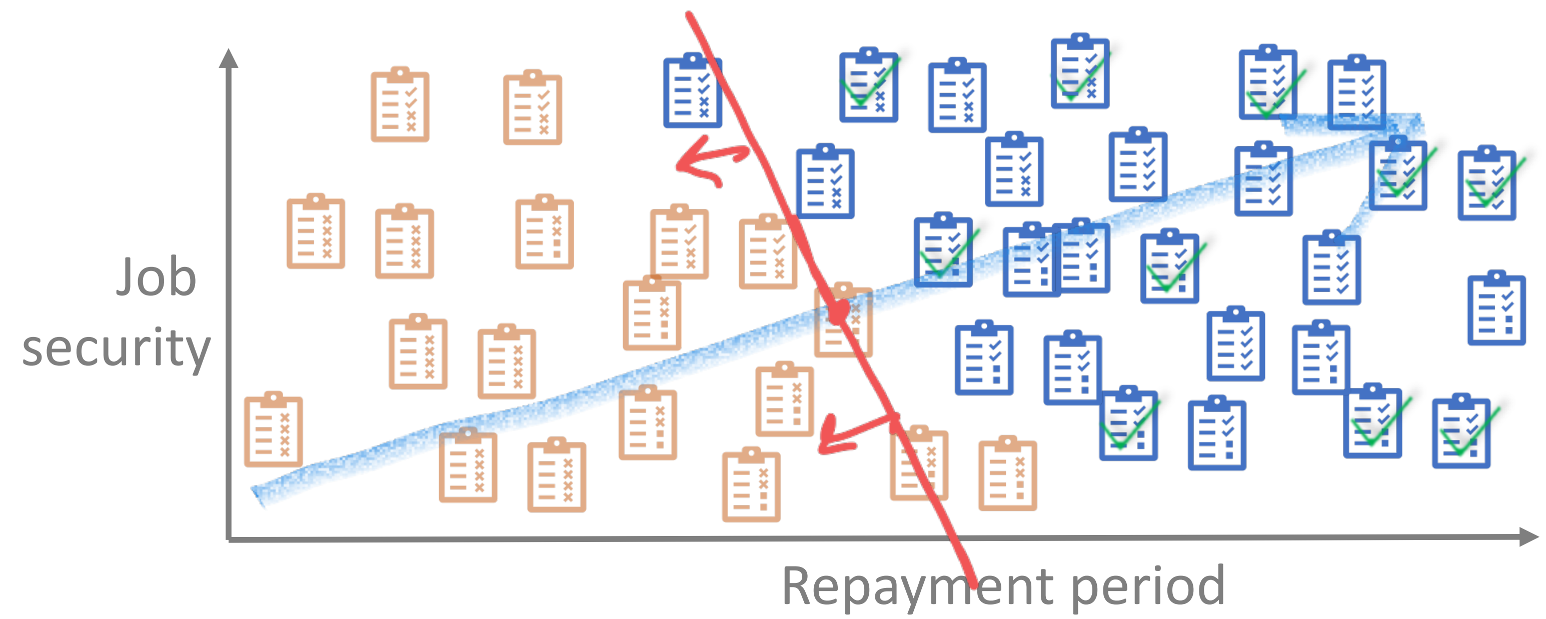
[1] Liu et al. Partially supervised classification of text documents. ICML. 2002

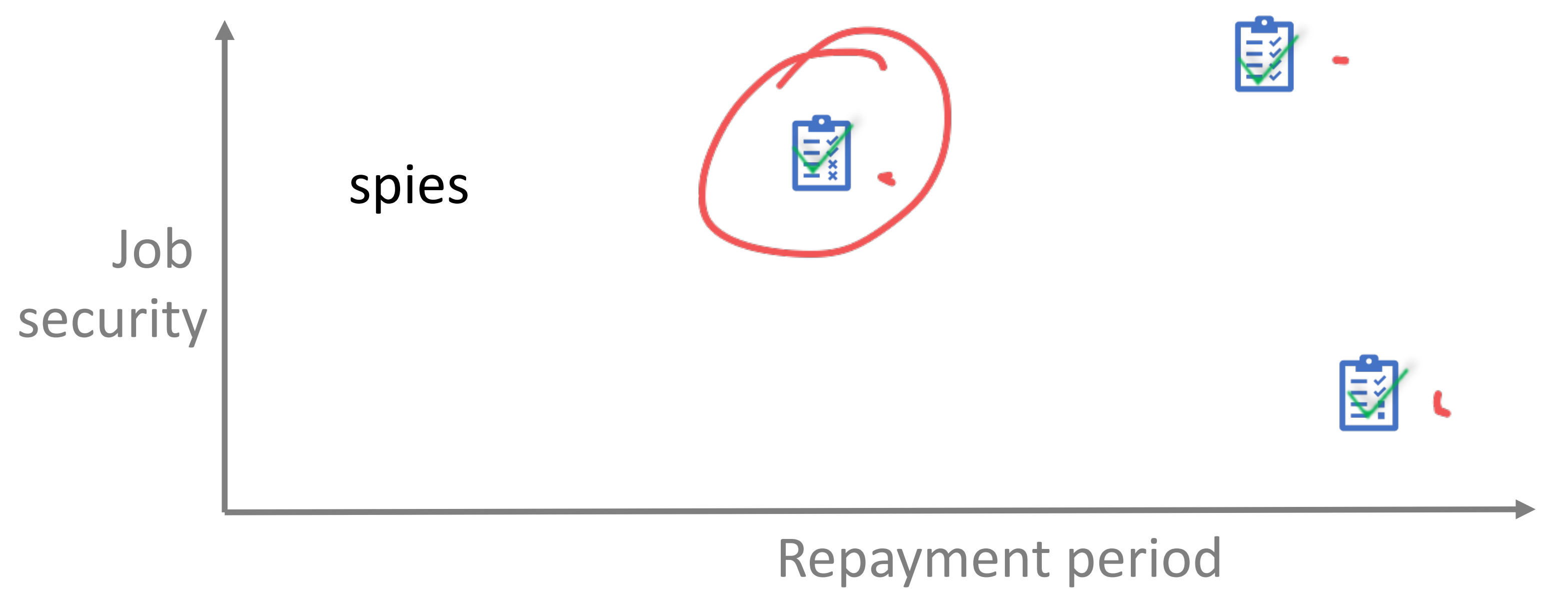
Step 1: Spy

1. Keep set of labeled “spies” behind when training non-traditional classifier

2. reliable negative examples
   =
   examples with probabilities lower than spy probabilities

   $\Pr(s = 1 | x_{neg}) < \Pr(s = 1 | x_{spy})$





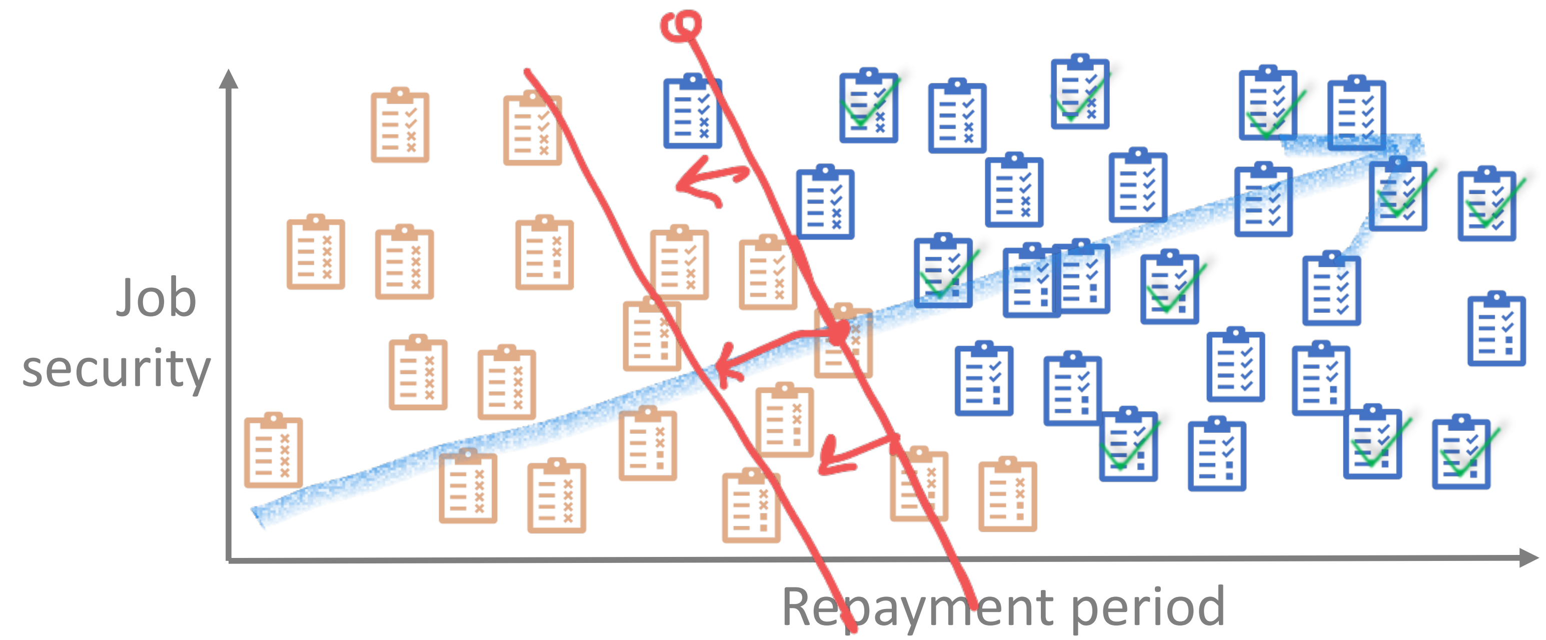
[1] Liu et al. Partially supervised classification of text documents. ICML. 2002

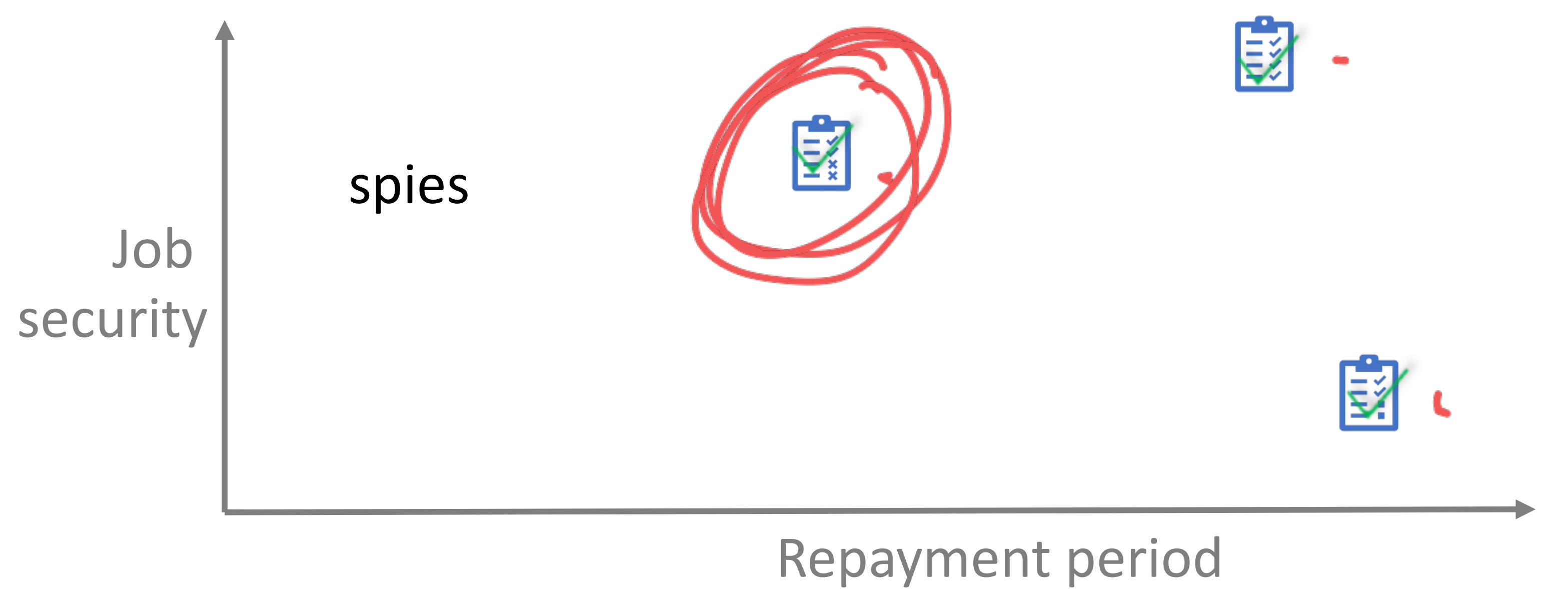
Step 1: Spy

1. Keep set of labeled "spies" behind when training non-traditional classifier

2. reliable negative examples
   =
   examples with probabilities lower than spy probabilities

   $$\Pr(s = 1 | x_{neg}) < \Pr(s = 1 | x_{spy})$$





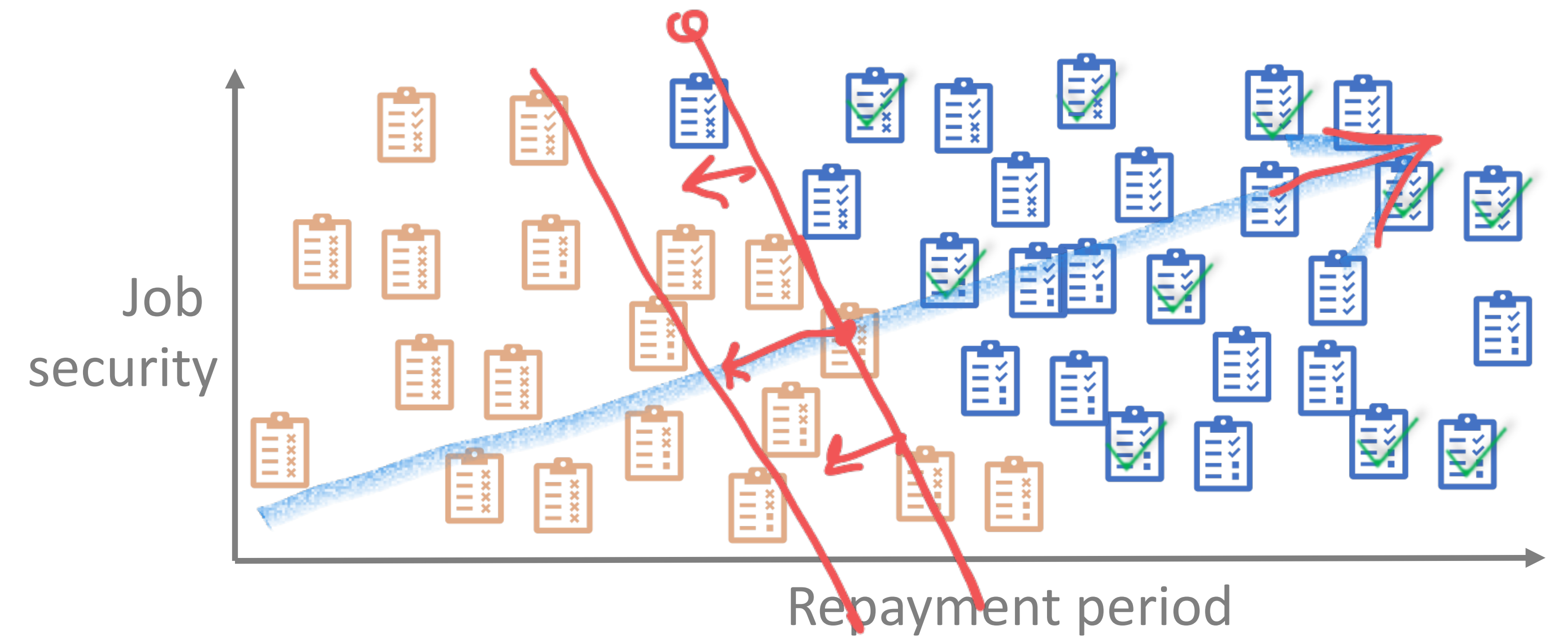
[1] Liu et al. Partially supervised classification of text documents. ICML. 2002

Step 1: Spy

1. Keep set of labeled “spies” behind when training non-traditional classifier

2. reliable negative examples
=
examples with probabilities lower than spy probabilities

$$\Pr(s = 1 | x_{neg}) < \Pr(s = 1 | x_{spy})$$



[1] Liu et al. Partially supervised classification of text documents. ICML. 2002

Step 2: training a classifier

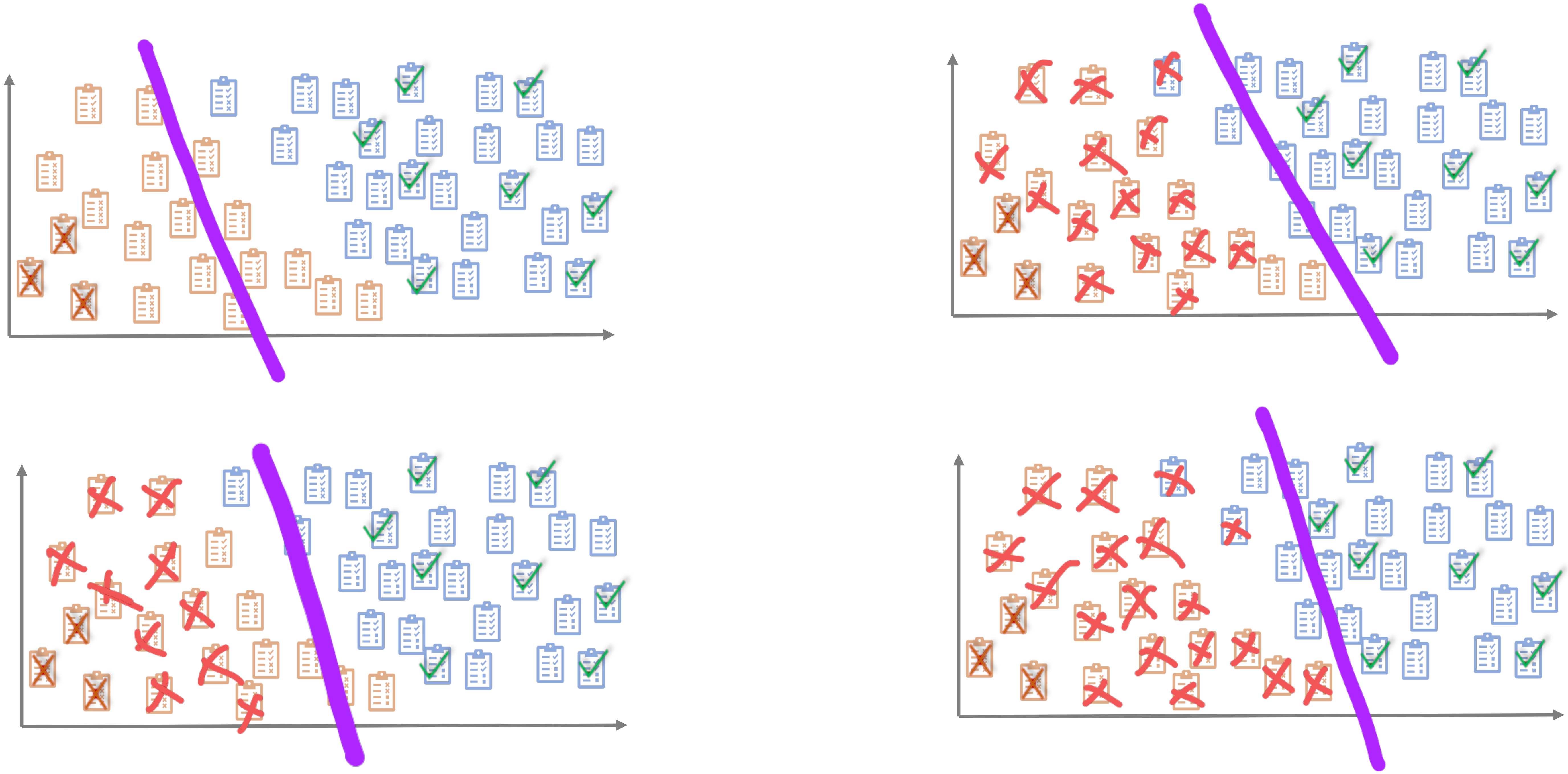
✓ labeled positive
✗ reliable negative
? unlabeled (leftover)

Any (semi-)supervised method can be used

SVM
application-specific classifier
EM

iterative SVM

Step 2: Iterative SVM



[1] Liu et al. Partially supervised classification of text documents. ICML. 2002

Step 3: selecting a classifier

- Last iteration

[1] Li & Liu. Learning to classify texts using positive and unlabeled data. IJCAI. 2003
[2] Li & Liu. Learning from positive and unlabeled examples with different data distributions. ECML. 2005

Step 3: selecting a classifie

- Last iteration
- Recall [1] SCAR

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

[1] Li & Liu. Learning to classify texts using positive and unlabeled data. IJCAI. 2003
[2] Li & Liu. Learning from positive and unlabeled examples with different data distributions. ECML. 2005

Step 3: selecting a classifie

- Last iteration
- Recall [1] SCAR

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

- $$F_1 = \frac{2pr}{p + r} \qquad p = \Pr(y = 1 | \hat{y} = 1)$$

F_1 high when p and r are high → same goal for F_1'

$$F_1' = \frac{pr}{\Pr(y = 1)} = \frac{r^2}{\Pr(\hat{y} = 1)}$$

[1] Li & Liu. Learning to classify texts using positive and unlabeled data. IJCAI. 2003
[2] Li & Liu. Learning from positive and unlabeled examples with different data distributions. ECML. 2005

Up next...