robROSE

An approach for dealing with imbalanced data in fraud detection

Sebastiaan Höppner

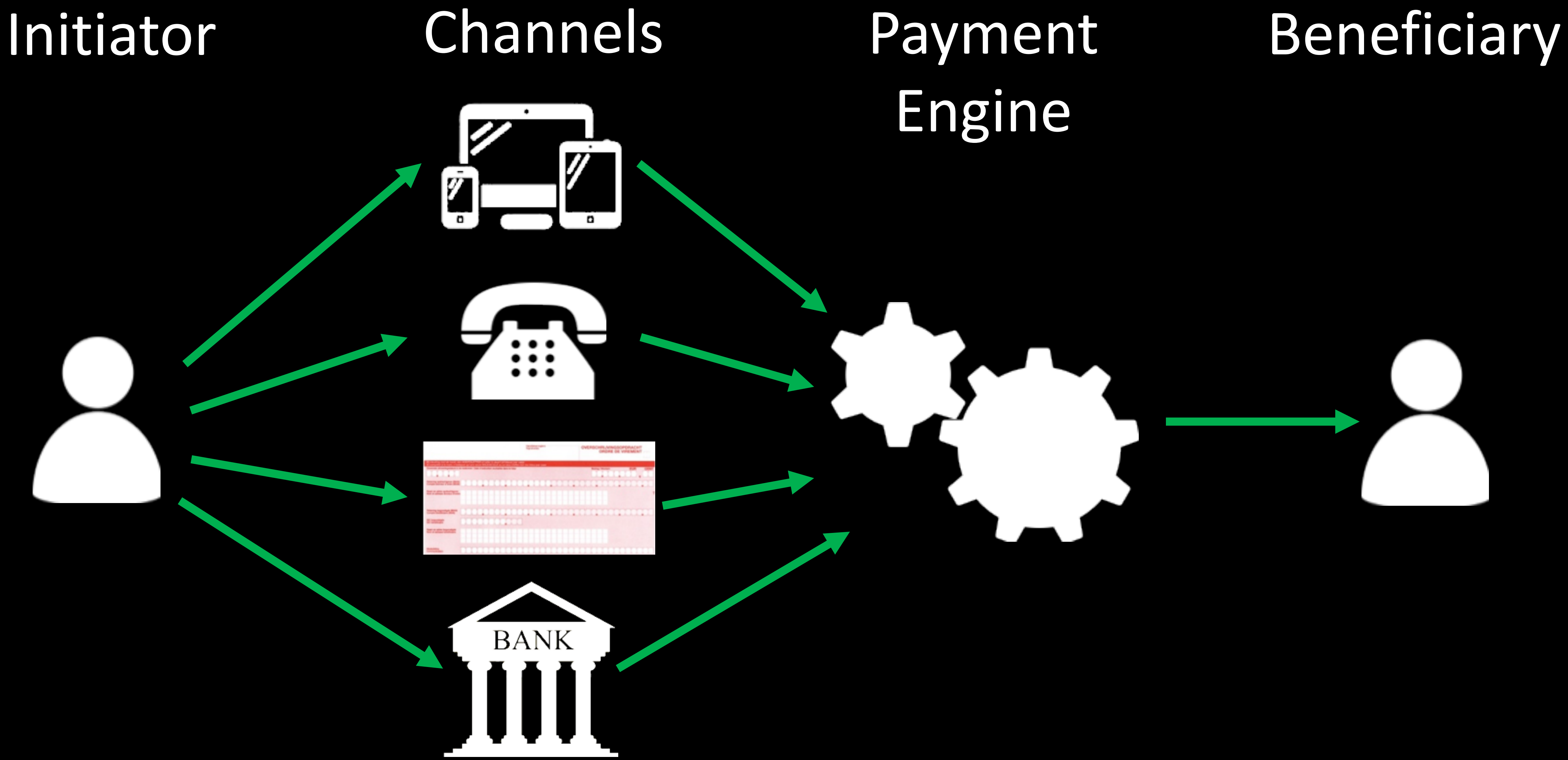
joint work with
Iréne Ortner, Bart Baesens & Tim Verdonck

11 July 2019

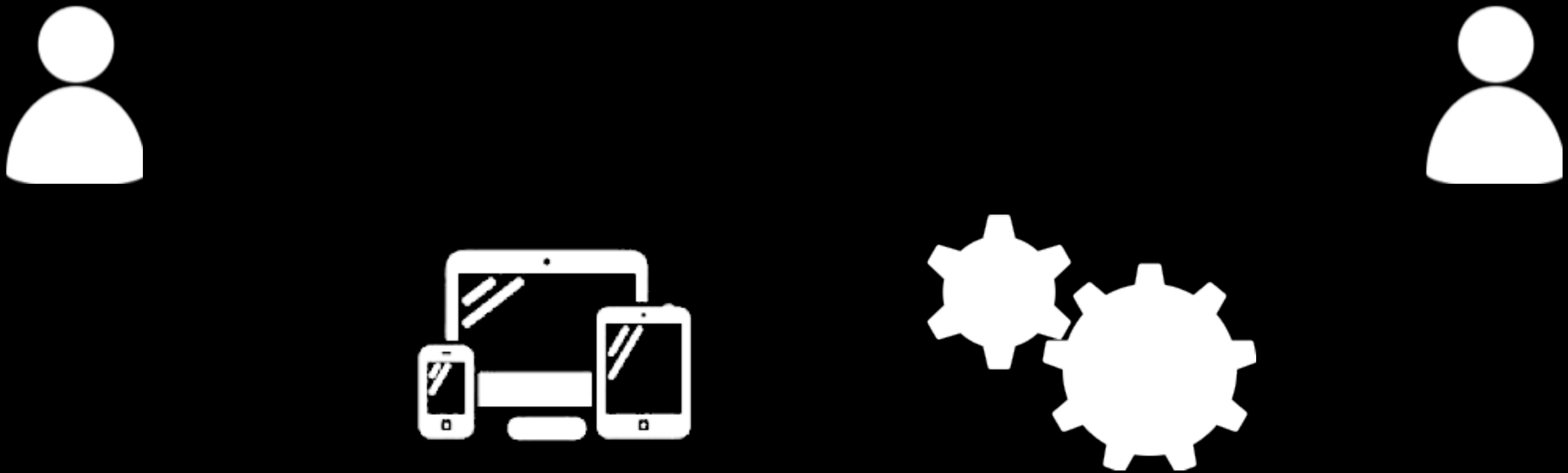
KU LEUVEN

Benford's Law Conference
10-12 July 2019 - Stresa, Italy

Credit transfers

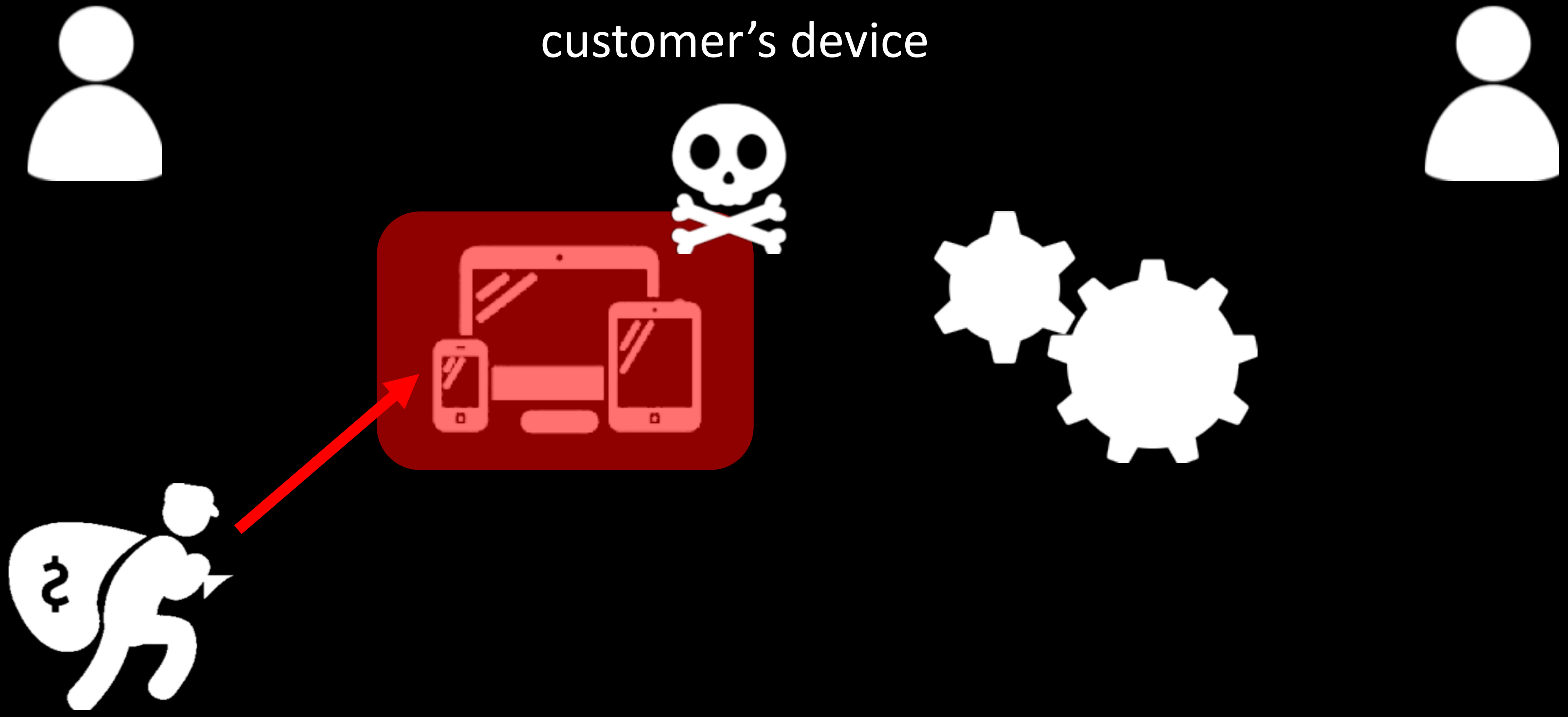


Hacking



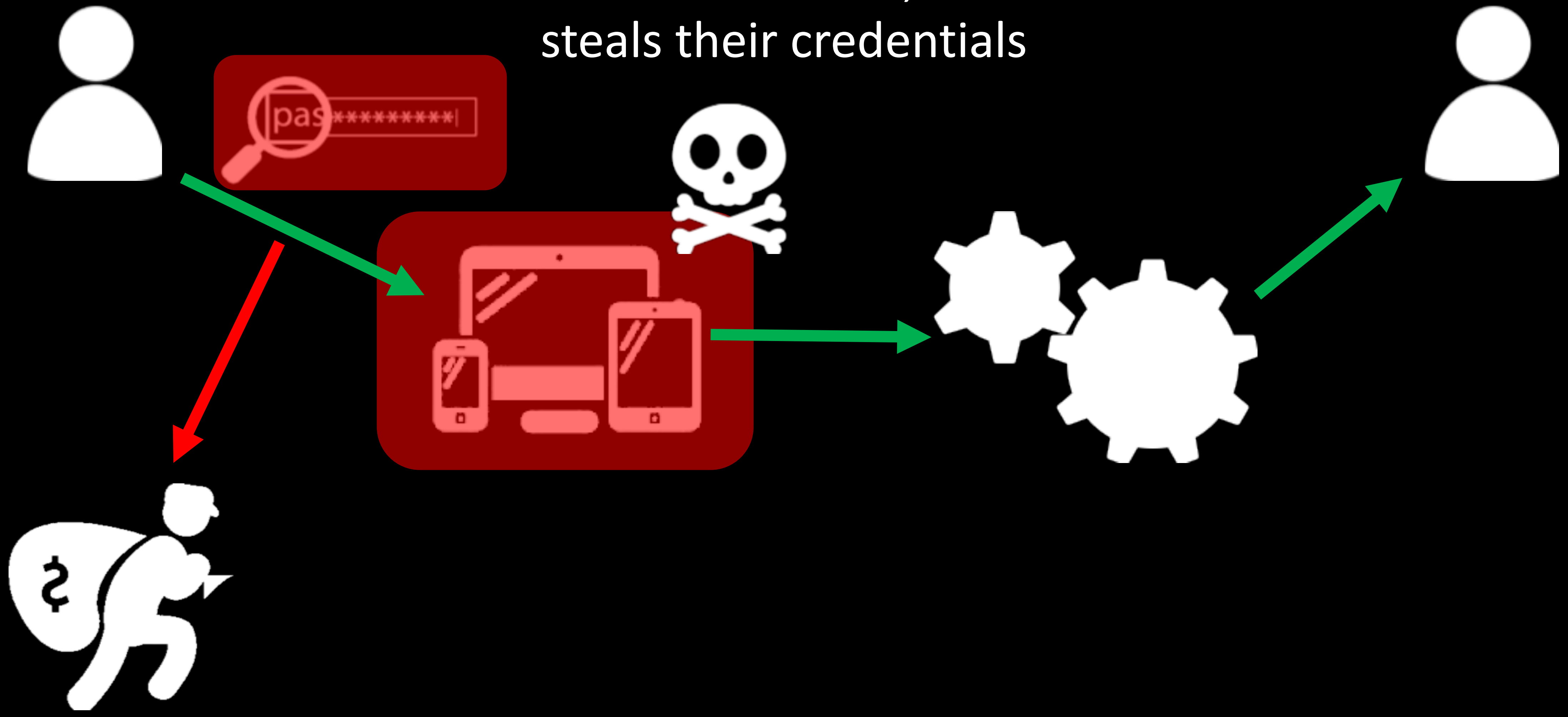
Hacking

Step 1: the fraudster installs malware on the customer's device



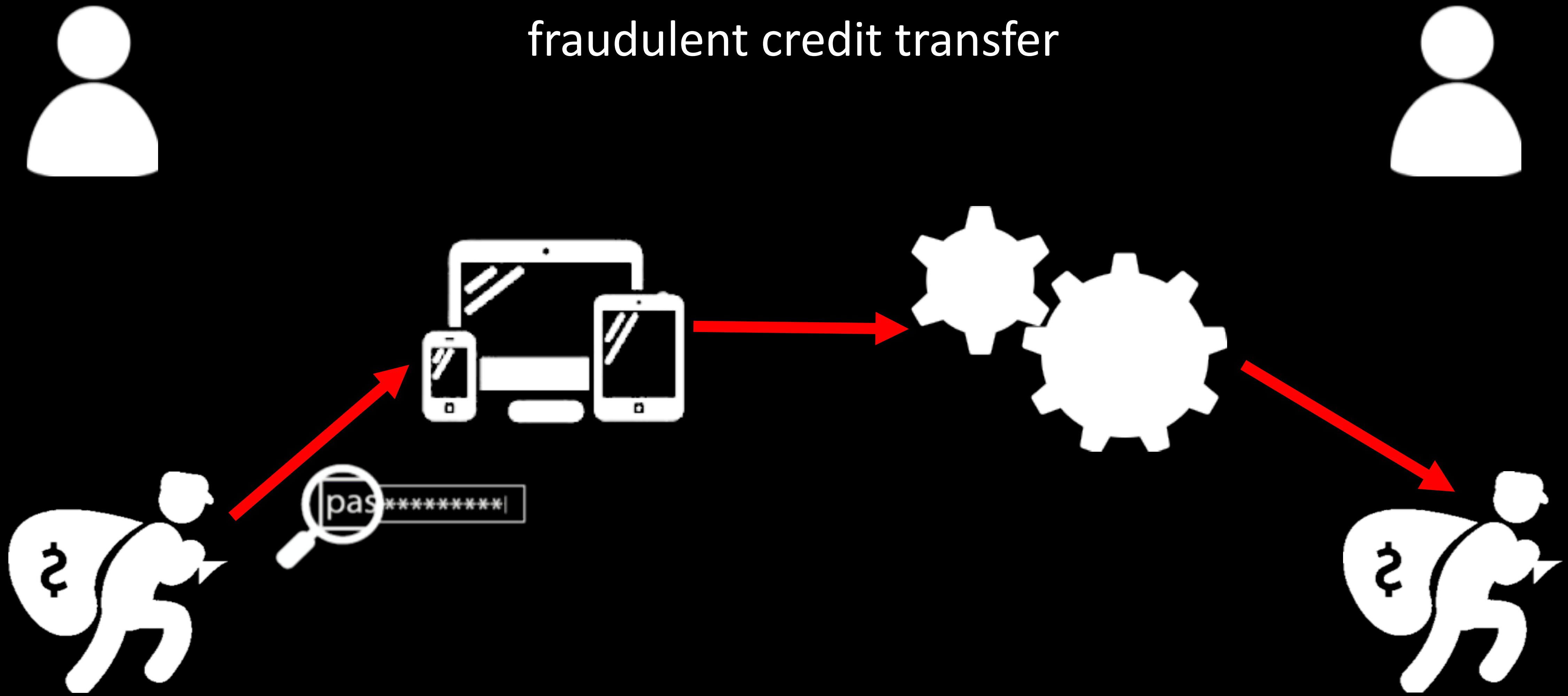
Hacking

Step 2: when the customer uses their device, the fraudster steals their credentials

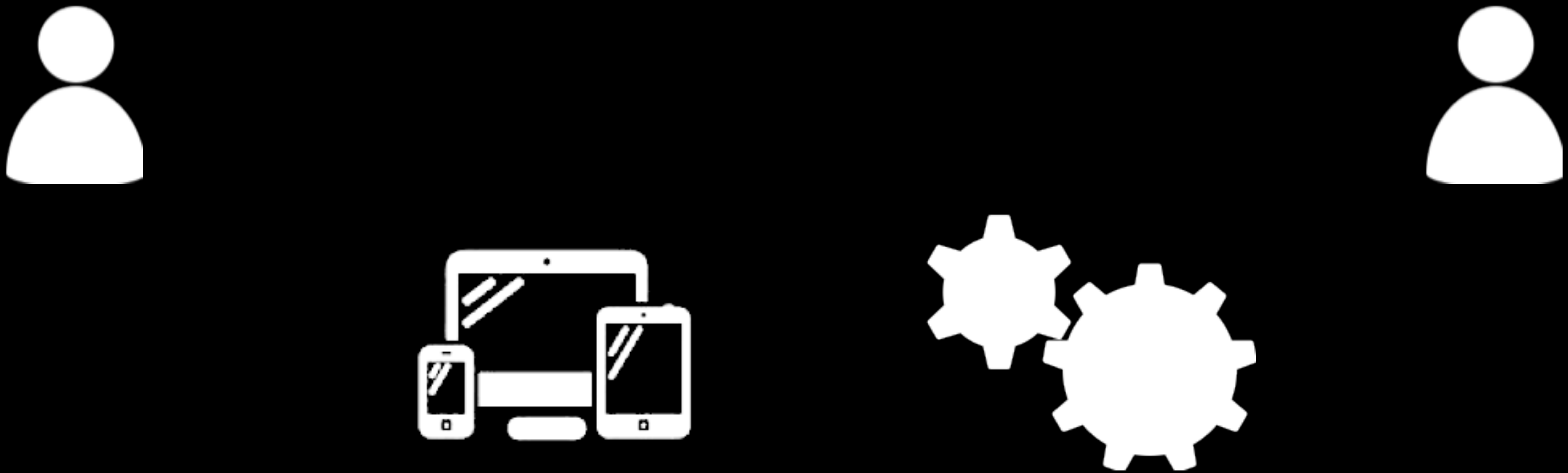


Hacking

Step 3: the fraudster uses the stolen credentials to book a fraudulent credit transfer

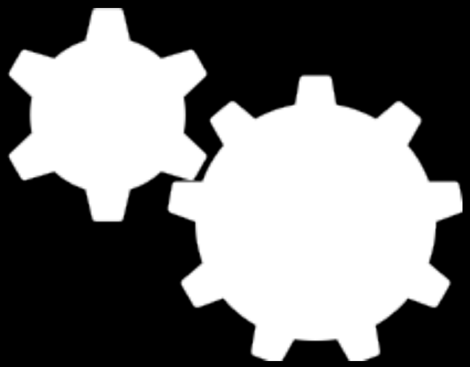


Phishing / vishing



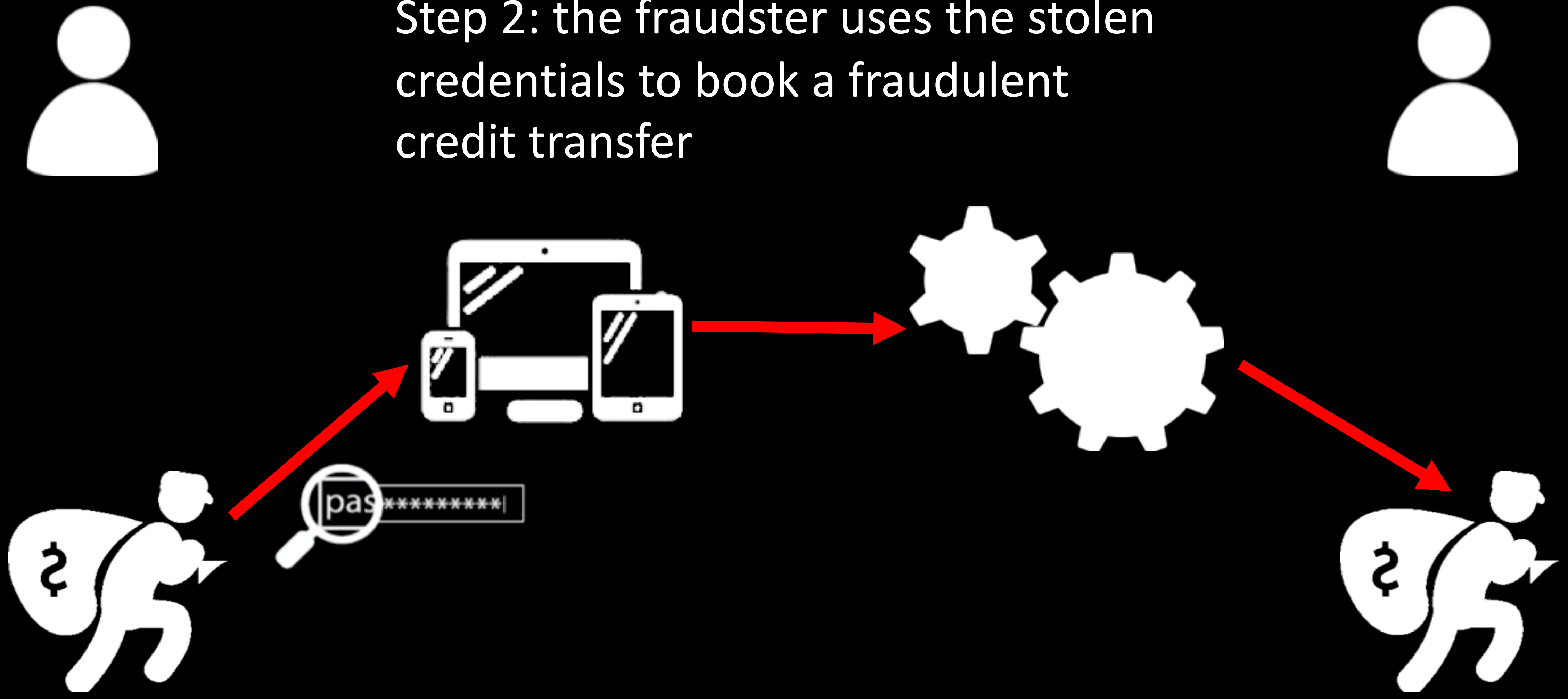
Phishing / vishing

Step 1: a fraudster tricks a customer into sharing their credentials

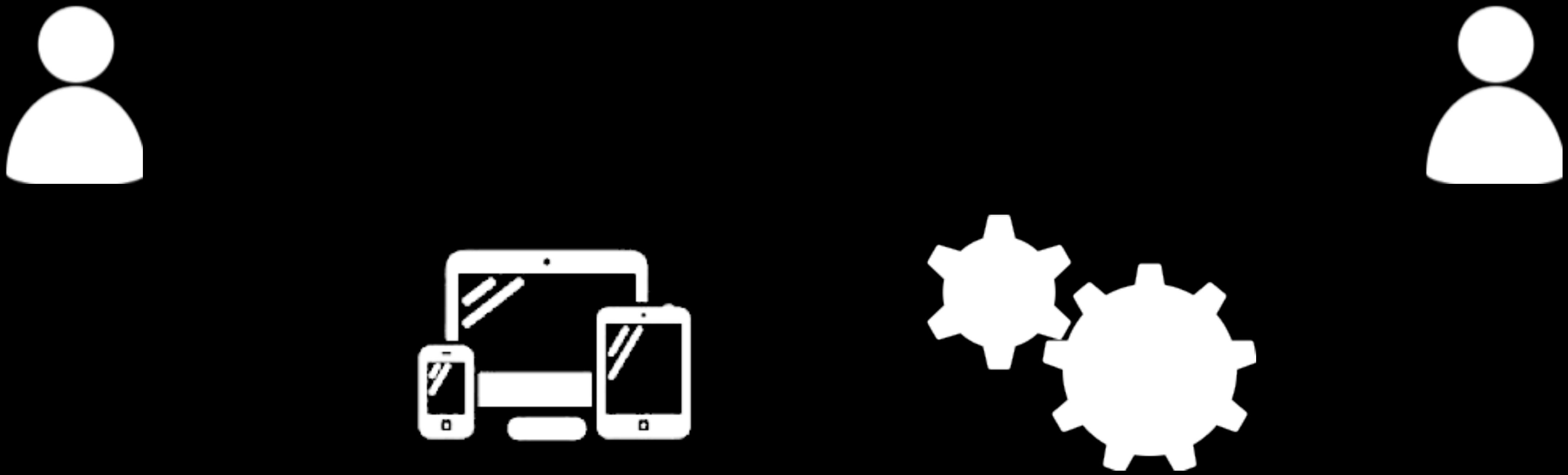


Phishing / vishing

Step 2: the fraudster uses the stolen credentials to book a fraudulent credit transfer

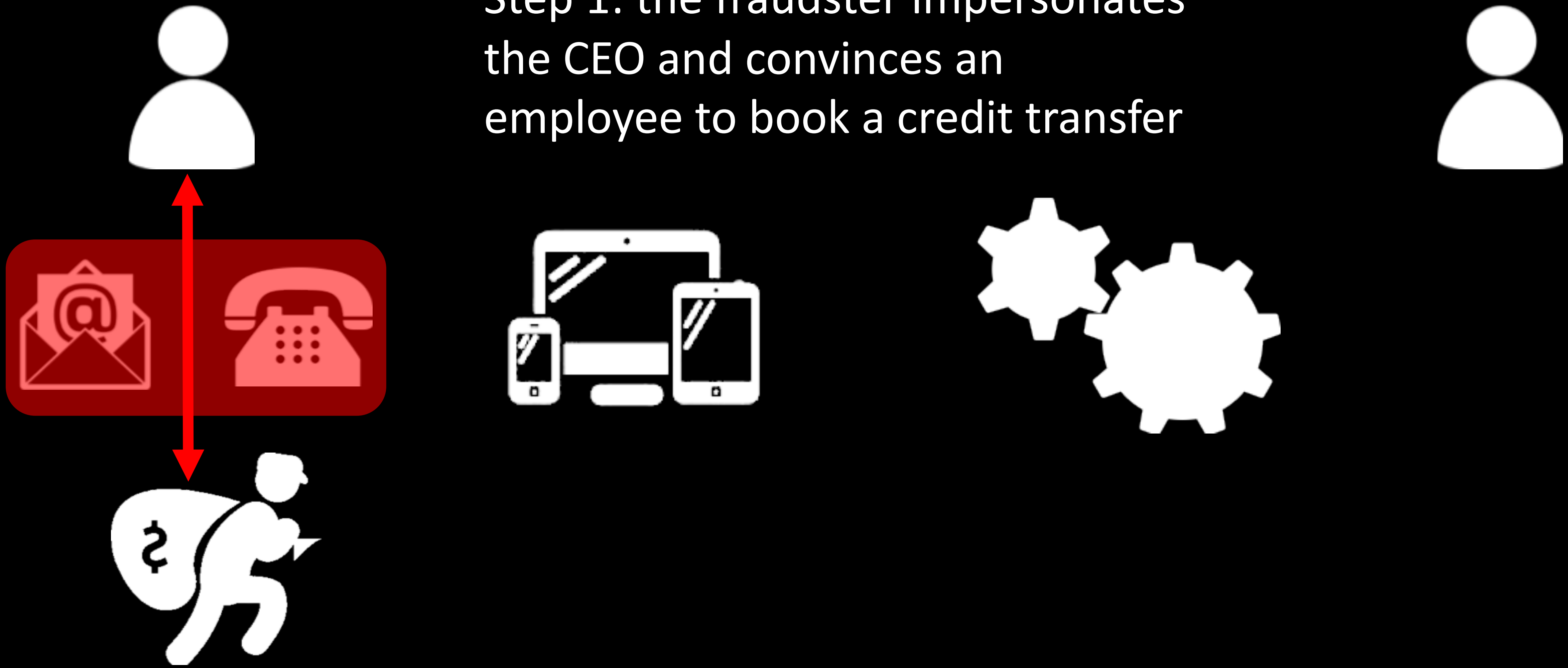


CEO fraud



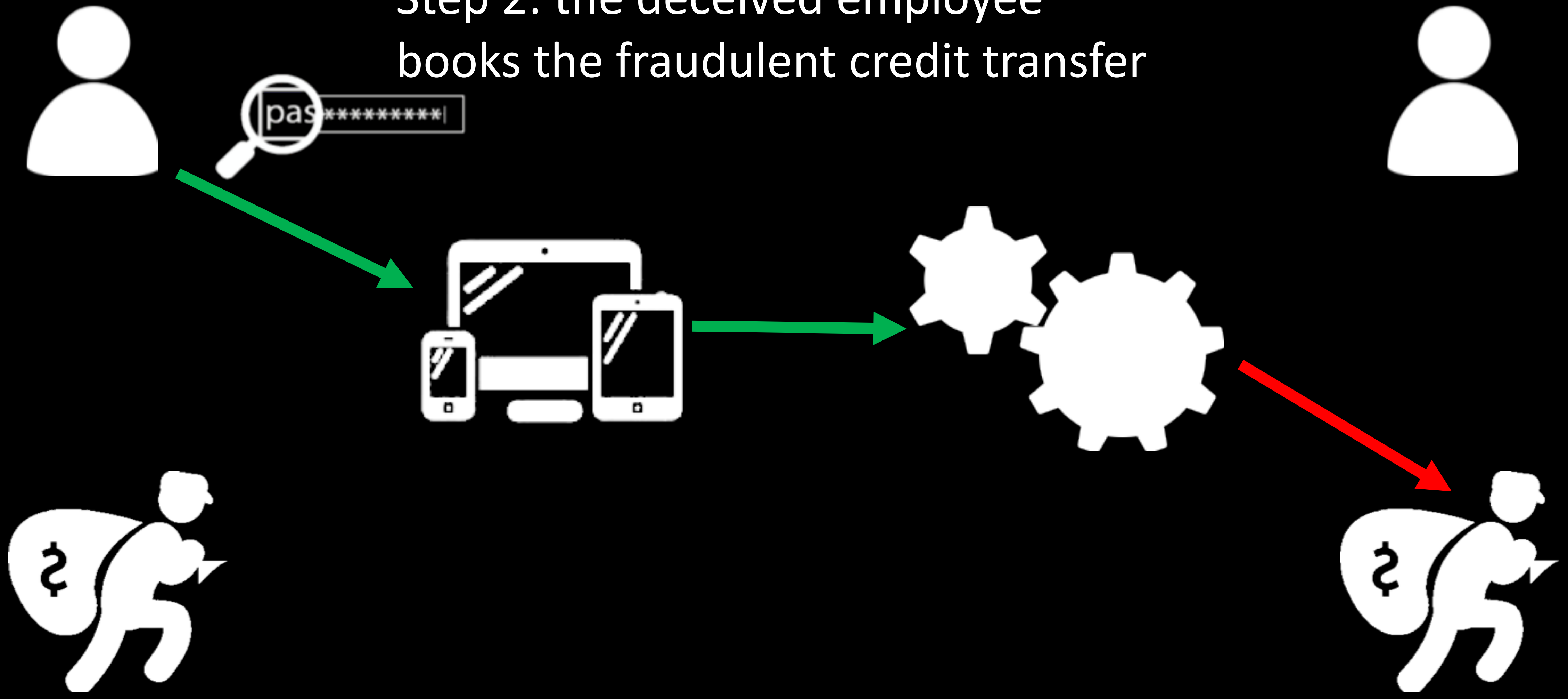
CEO fraud

Step 1: the fraudster impersonates the CEO and convinces an employee to book a credit transfer



CEO fraud

Step 2: the deceived employee books the fraudulent credit transfer



Problem: imbalanced data

- ***Binary classification***
  legitimate vs fraud

Problem: imbalanced data

- ***Binary classification***
  legitimate vs fraud

- ***Imbalanced data***
  (very) large difference in number of observations of both groups

Problem: imbalanced data

- ***Binary classification***
  legitimate vs fraud
- ***Imbalanced data***
  (very) large difference in number of observations of both groups
- ***Credit card fraud***
  less than 1 out 10m transactions
  (< 0.00001%)

Problem: imbalanced data

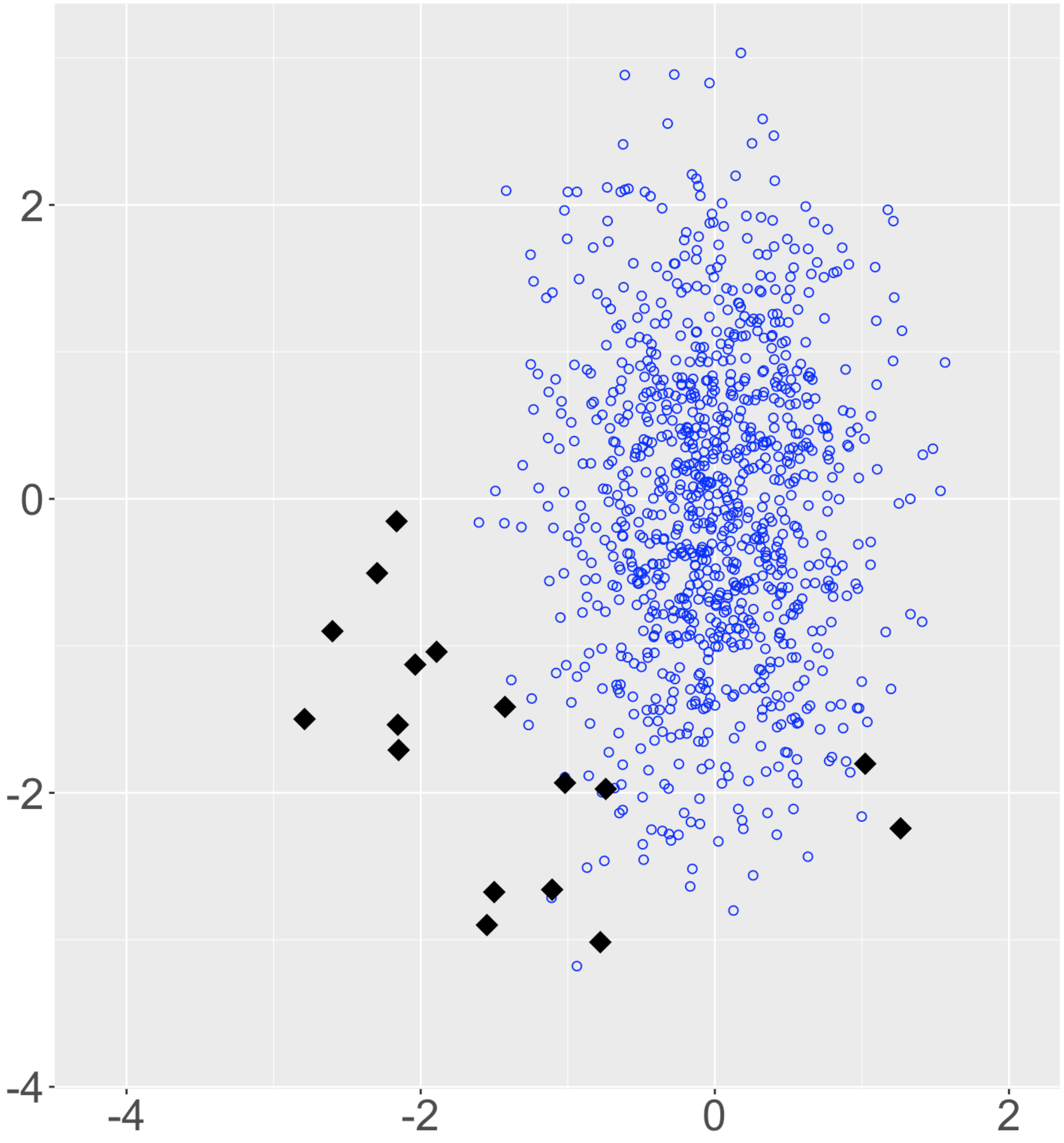
- ***Binary classification***
  legitimate vs fraud

- ***Imbalanced data***
  (very) large difference in number of observations of both groups

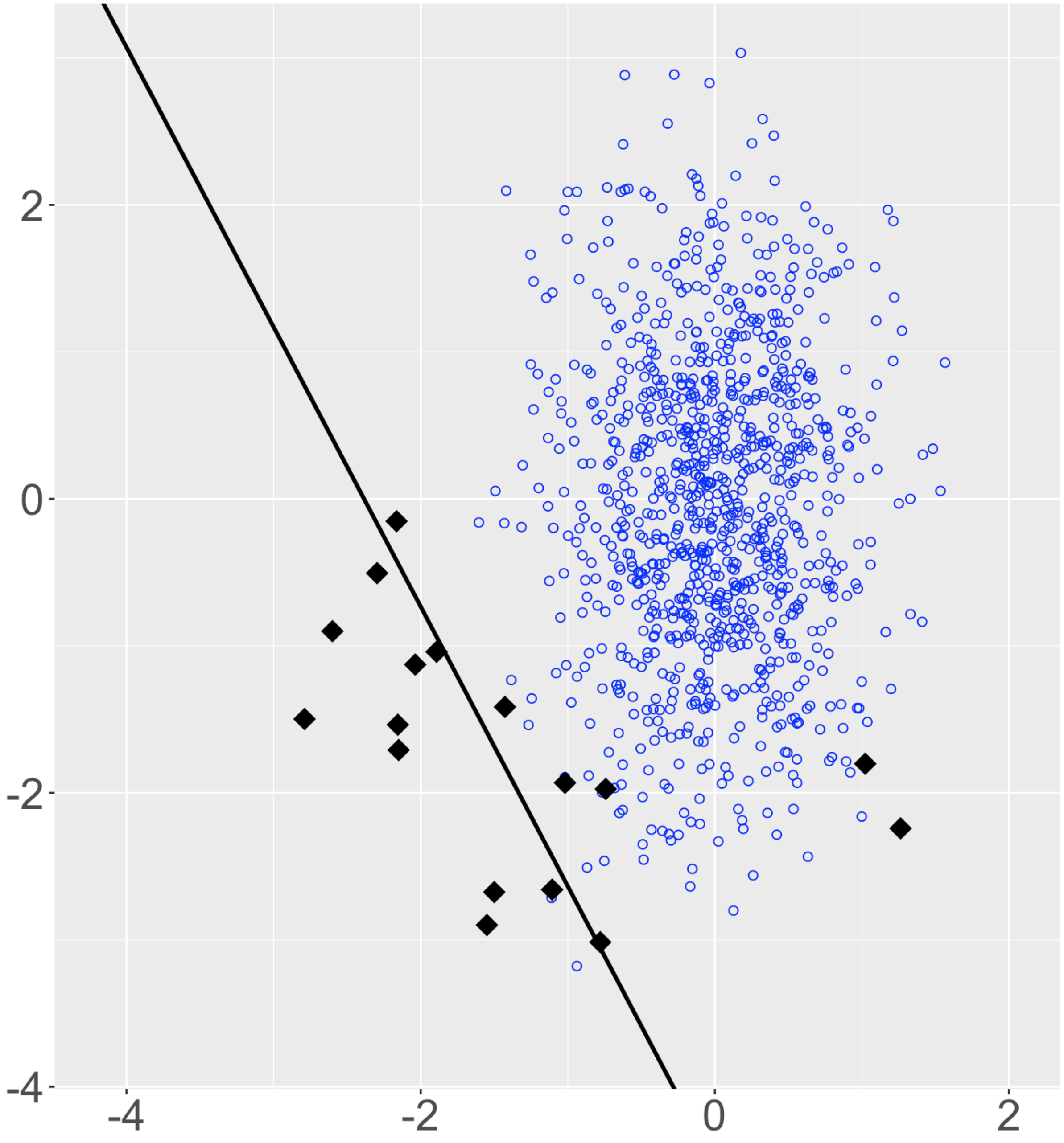
- ***Credit card fraud***
  less than 1 out 10m transactions
  (< 0.00001%)

- Typically very few “cases of interest” compared to legitimate observations
  (20% - 0.01%)

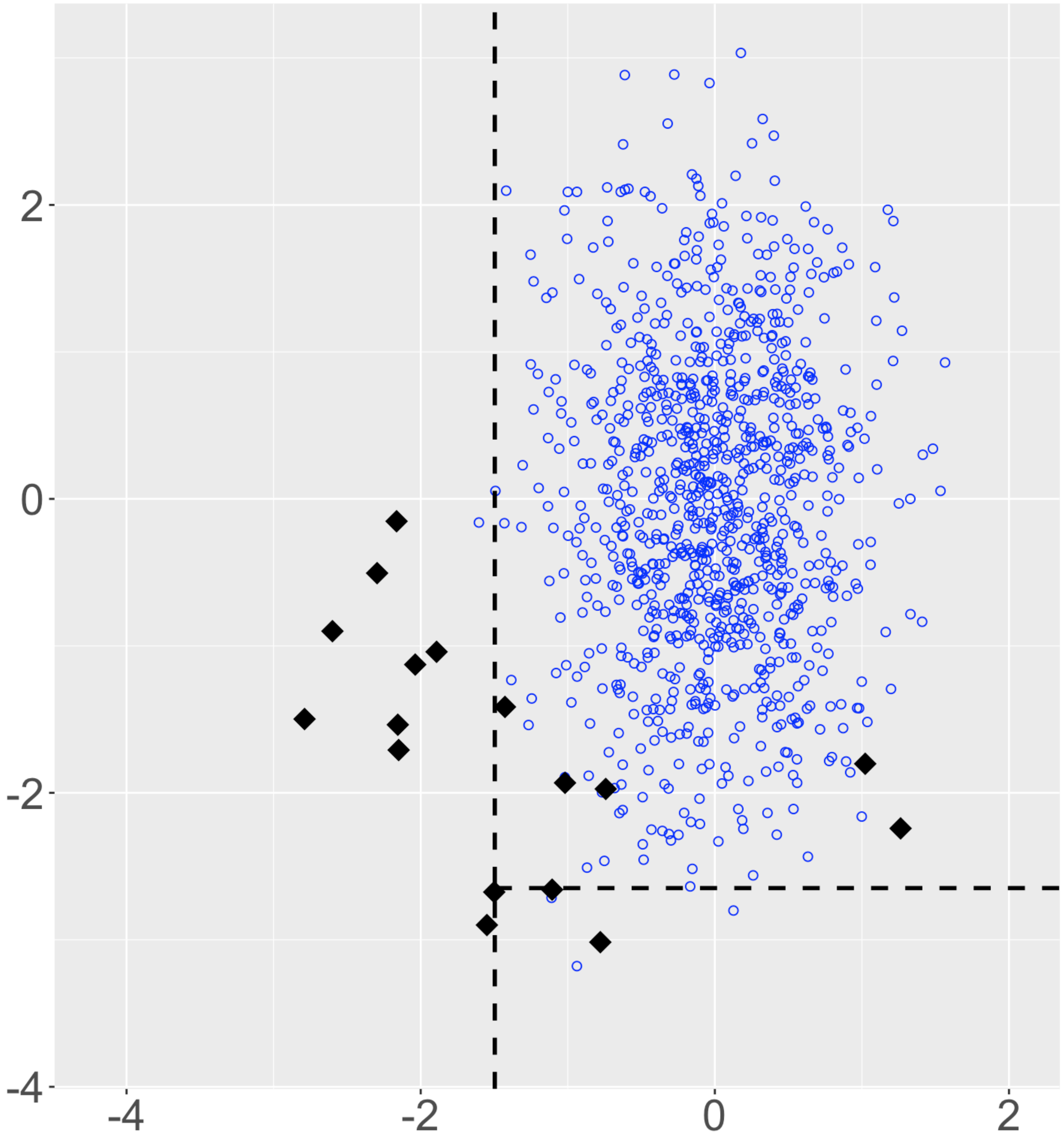
Original data



Logistic regression



Classification tree (CART)



"How to reduce imbalance in training data?"

"How to reduce imbalance in training data?"

Reduce # legitimate cases	Increase # fraud cases

“How to reduce imbalance in training data?”

Reduce # legitimate cases

- Random under-sampling: randomly sub-sample legitimate cases

Increase # fraud cases

- Random over-sampling: sampling with replacement of fraud samples
- ***Generate synthetic minority/fraud cases***

“How to reduce imbalance in training data?”

Reduce # legitimate cases

- Random under-sampling: randomly sub-sample legitimate cases

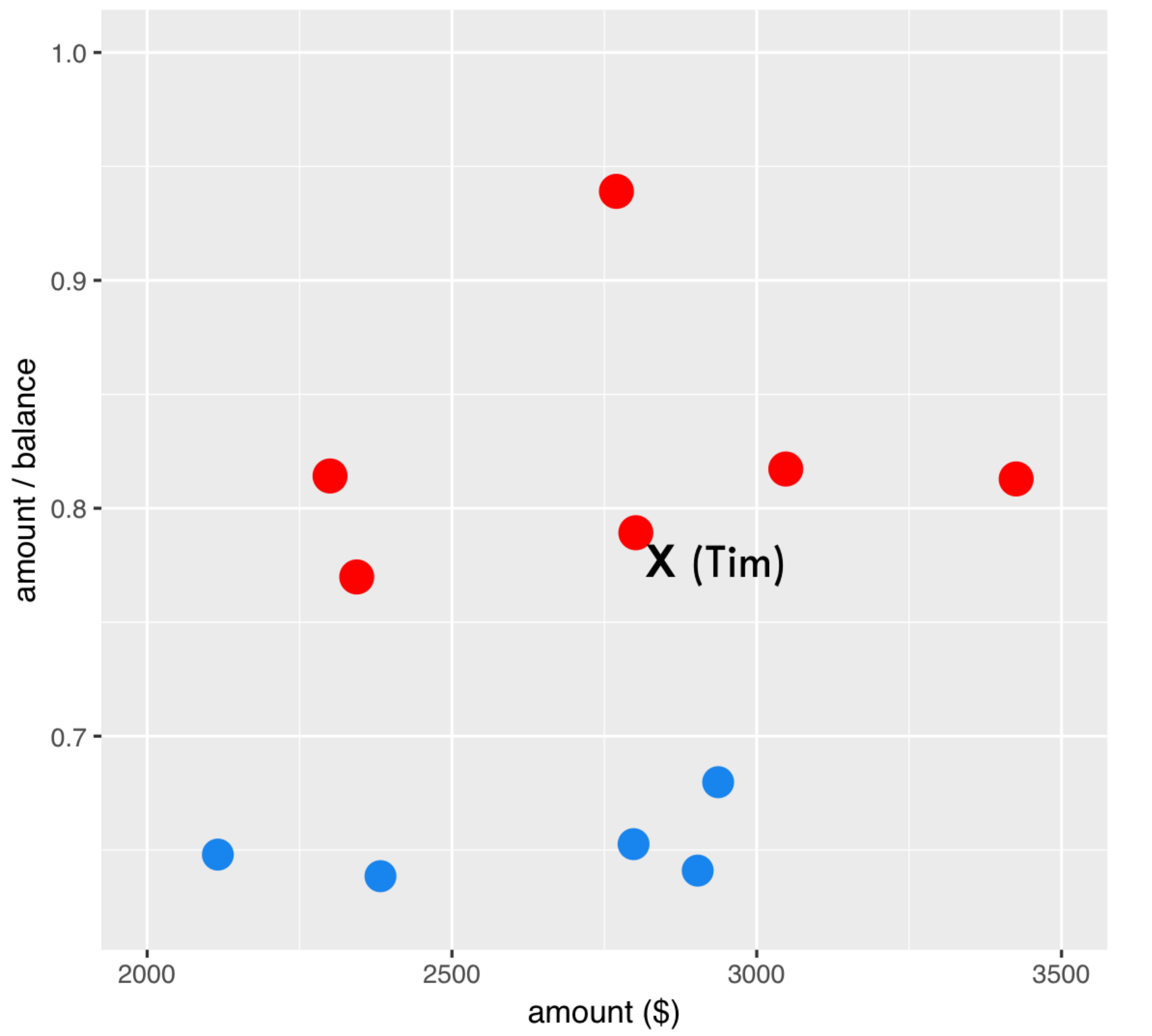
Increase # fraud cases

- Random over-sampling: sampling with replacement of fraud samples
- ***Generate synthetic minority/fraud cases***

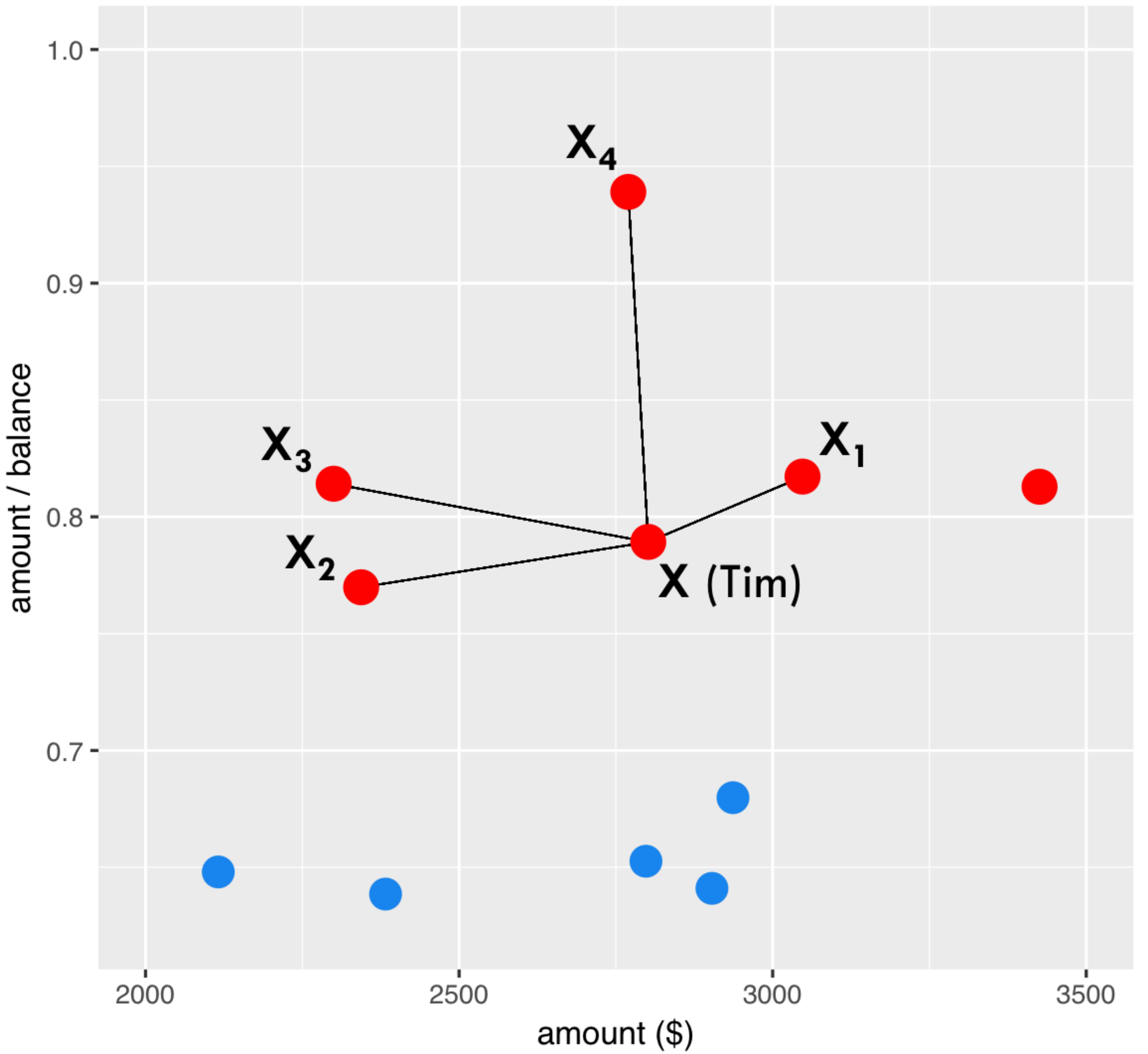
Evaluate model on imbalanced “original” test data !!!

SMOTE -
Synthetic Minority Over-sampling Technique

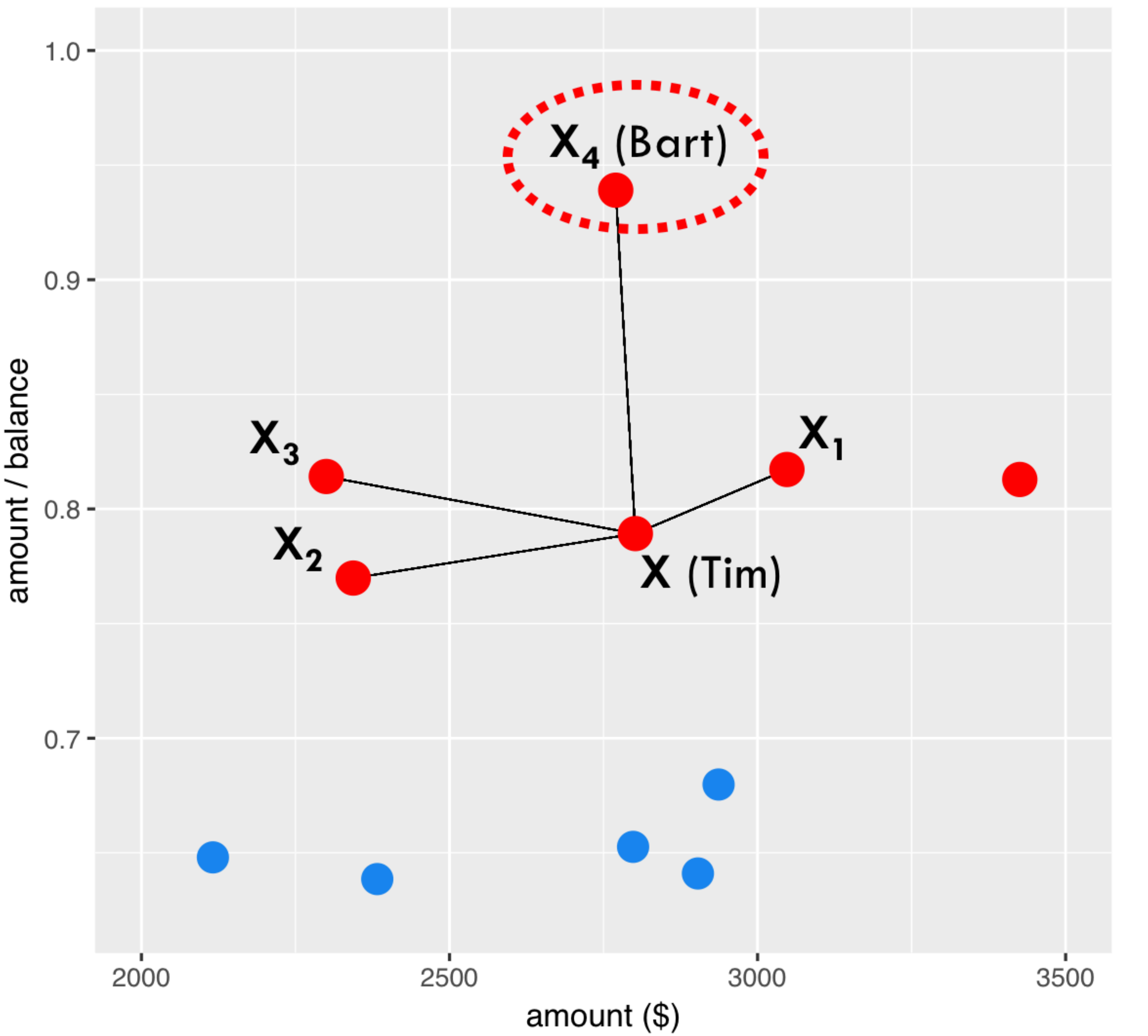
Chawla, Bowyer, Hall & Kegelmeyer (2002)

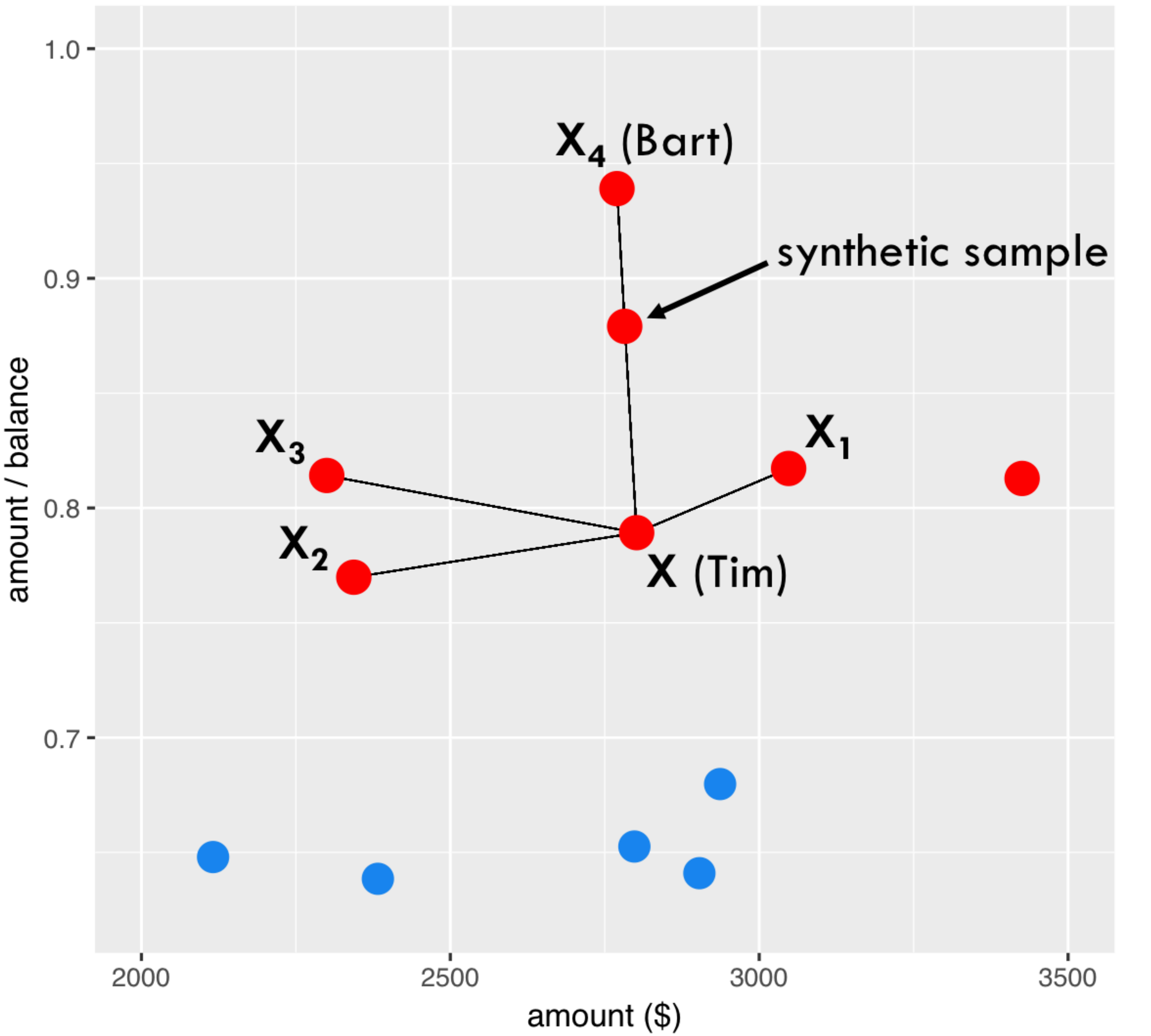


1.0
0.9
0.8
0.7
amount / balance
X (Tim)
2000
2500
3000
3500
amount ($)



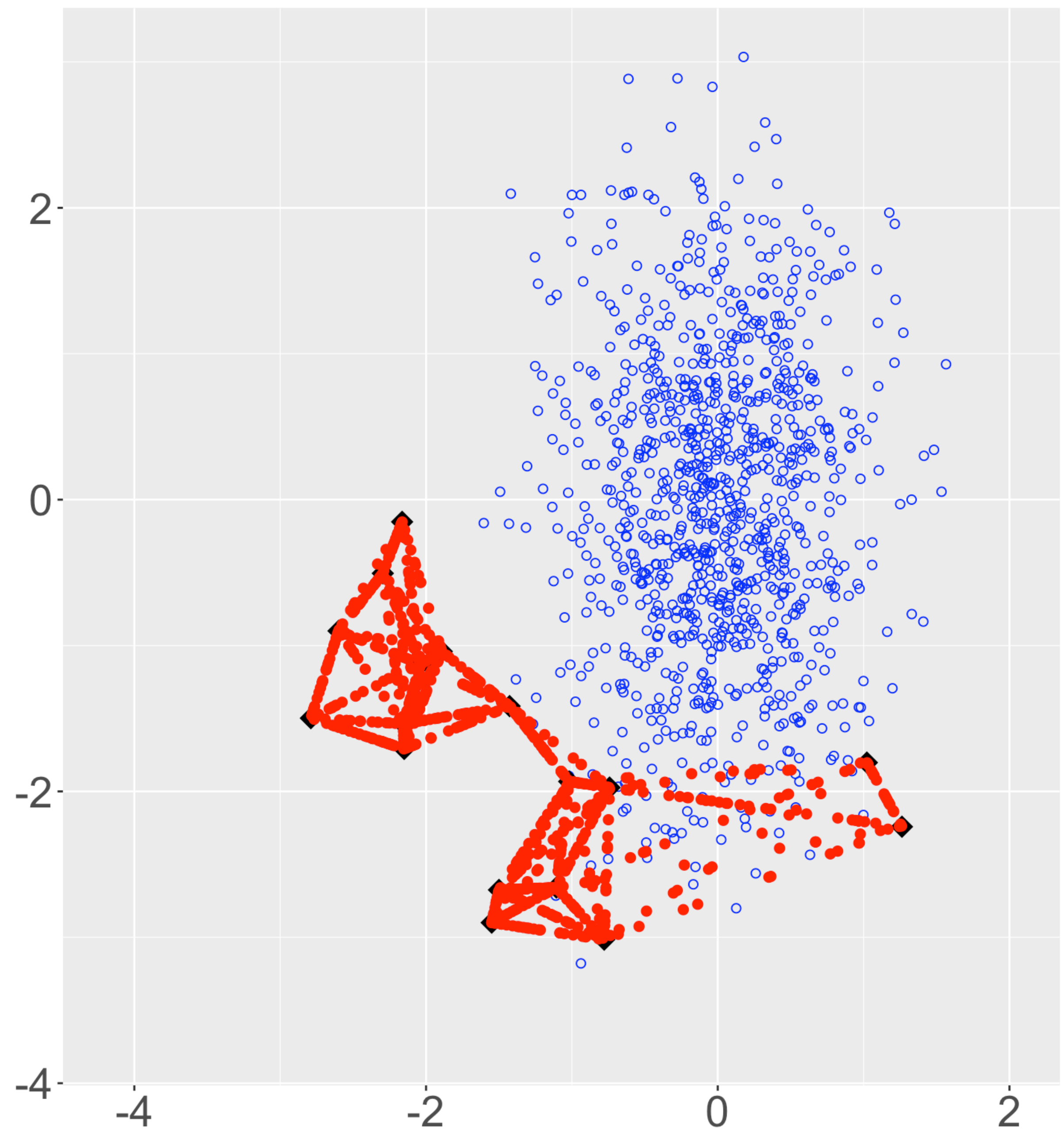
1.0
0.9
0.8
0.7
amount / balance
2000
2500
3000
3500
amount ($)
X_4
X_3
X_1
X_2
X (Tim)


X4 (Bart)
X3
X1
X2
X (Tim)
1.0
0.9
0.8
0.7
amount / balance
2000
2500
3000
3500
amount ($)



X4 (Bart)
synthetic sample
X3
X1
X2
X (Tim)
1.0
0.9
0.8
0.7
amount / balance
2000
2500
3000
3500
amount ($)

SMOTE



ROSE - Random Over-Sampling Examples

Menardi & Torelli (2014)

ROSE - Random Over-Sampling Examples

1. Randomly select an observation $\boldsymbol{x}_i$ from the minority group

Menardi & Torelli (2014)

ROSE - Random Over-Sampling Examples

1. Randomly select an observation $\boldsymbol{x}_i$ from the minority group
2. Estimate normal density distribution $\mathcal{N}(\boldsymbol{x}_i, H)$

Menardi & Torelli (2014)

ROSE - Random Over-Sampling Examples

1. Randomly select an observation $\boldsymbol{x}_i$ from the minority group
2. Estimate normal density distribution $\mathcal{N}(\boldsymbol{x}_i, H)$
    - selected observation $\boldsymbol{x}_i$ as center

Menardi & Torelli (2014)

ROSE - Random Over-Sampling Examples

1. Randomly select an observation $\boldsymbol{x}_i$ from the minority group

2. Estimate normal density distribution $\mathcal{N}(\boldsymbol{x}_i, H)$
   - selected observation $\boldsymbol{x}_i$ as center
   - *smoothing* matrix $H = diag(h_1, \ldots, h_d)$

$$h_q = \left(\frac{4}{(d+2)n}\right)^{1/(d+4)} \hat{\sigma}_q \quad (q = 1, \ldots, d)$$

$\hat{\sigma}_q$ = sample standard deviation of q-th variable of minority group

Menardi & Torelli (2014)

ROSE - Random Over-Sampling Examples

1. Randomly select an observation $\boldsymbol{x}_i$ from the minority group

2. Estimate normal density distribution $\mathcal{N}(\boldsymbol{x}_i, H)$
   - selected observation $\boldsymbol{x}_i$ as center
   - *smoothing* matrix $H = \boxed{diag}(h_1, \ldots, h_d)$

$$h_q = \left(\frac{4}{(d+2)n}\right)^{1/(d+4)} \boxed{\hat{\sigma}_q} \quad (q = 1, \ldots, d)$$

$\hat{\sigma}_q$ = sample standard deviation of q-th variable of minority group

Menardi & Torelli (2014)

ROSE - Random Over-Sampling Examples

1. Randomly select an observation $\boldsymbol{x}_i$ from the minority group

2. Estimate normal density distribution $\mathcal{N}(\boldsymbol{x}_i, H)$
   - selected observation $\boldsymbol{x}_i$ as center
   - *smoothing* matrix $H = diag(h_1, \ldots, h_d)$

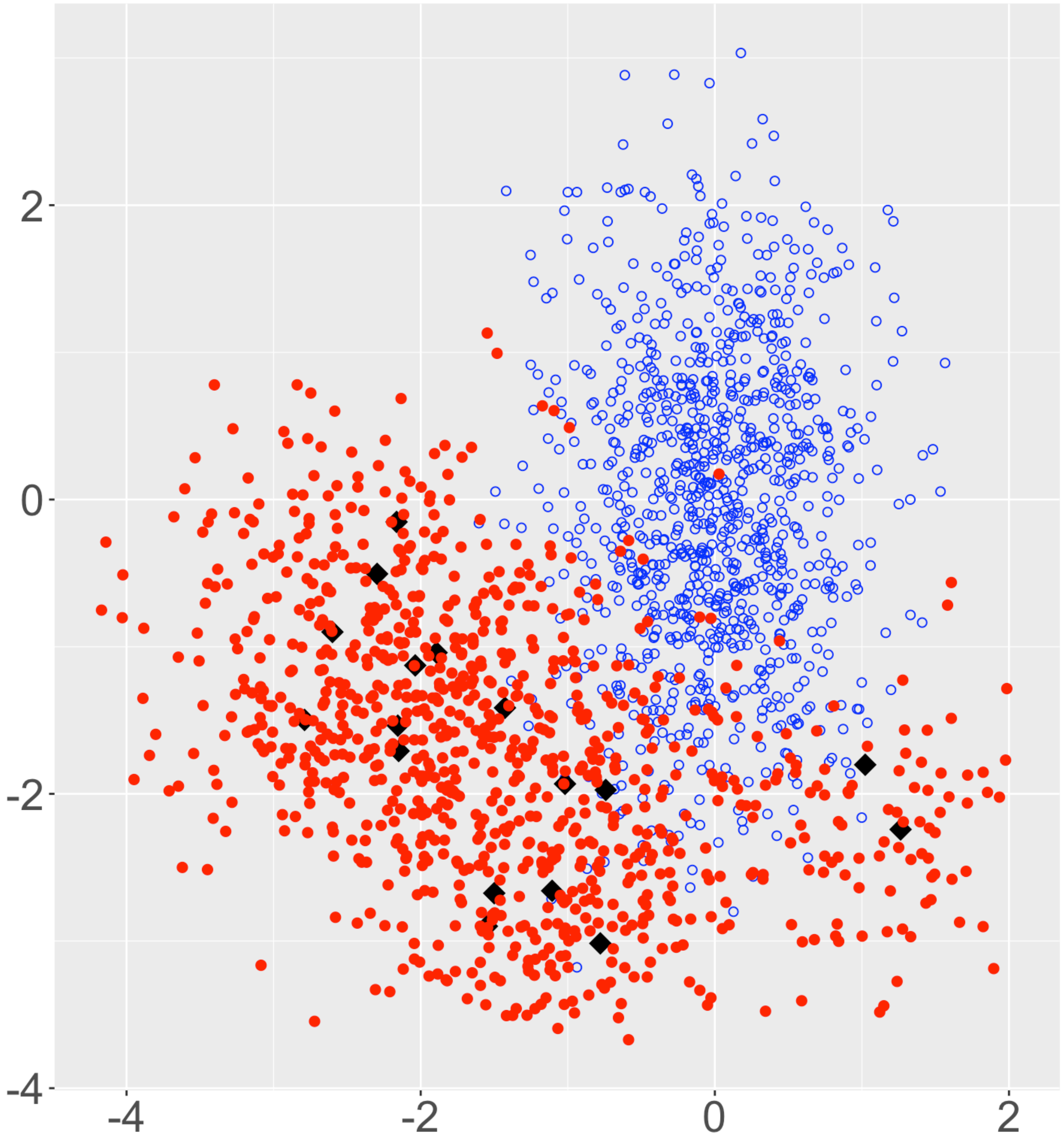
$$h_q = \left(\frac{4}{(d+2)n}\right)^{1/(d+4)} \hat{\sigma}_q \quad (q = 1, \ldots, d)$$

$\hat{\sigma}_q$ = sample standard deviation of q-th variable of minority group

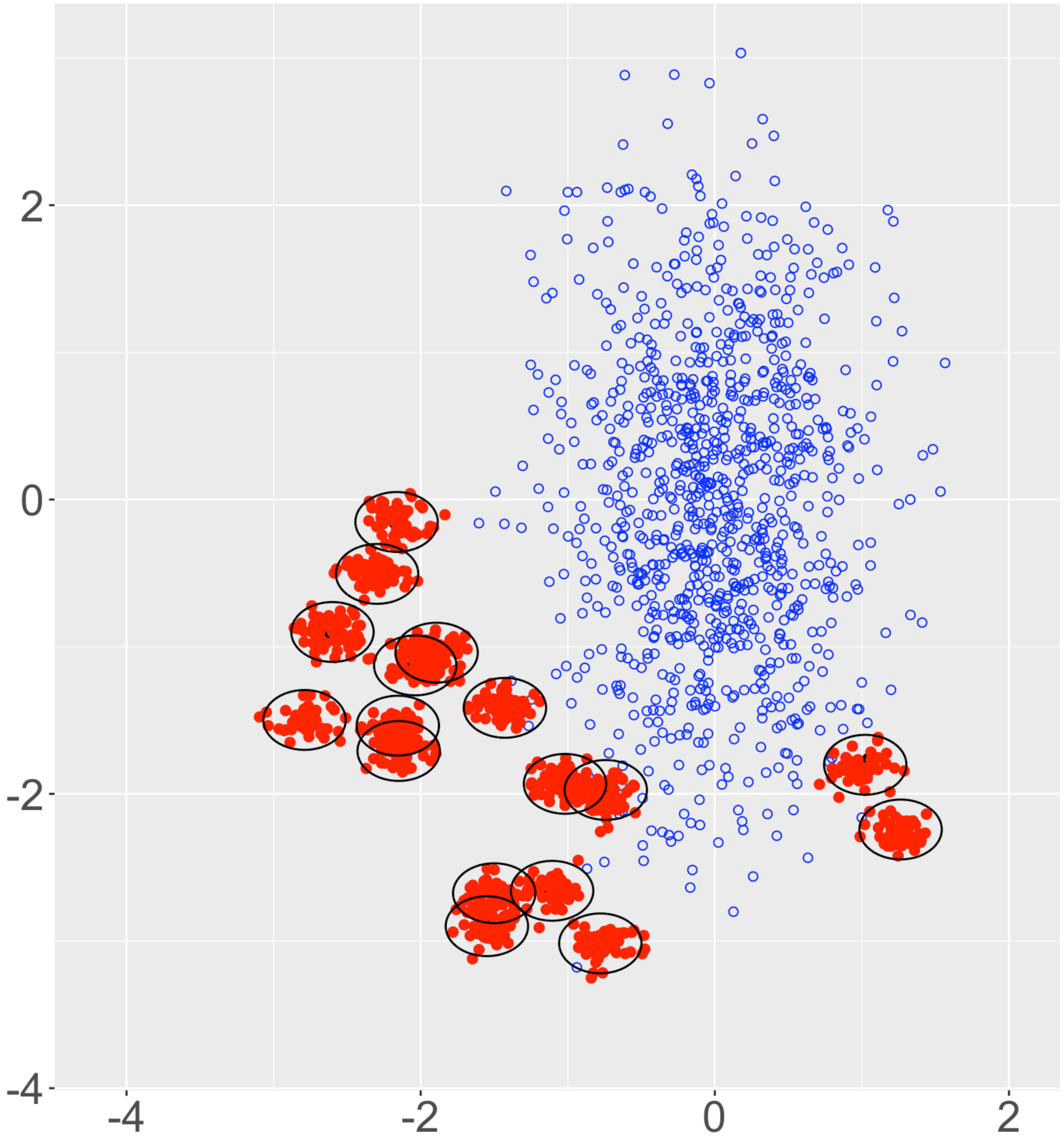
3. Generate a new observation from this normal density estimate

Menardi & Torelli (2014)

ROSE (h.minor = 1) (default)



ROSE (h.minor = 0.15)



Robust ROSE

- Identify “outlying” minority cases based on Mahalanobis distance (MD) using the robust MCD[1] estimator

[1] Rousseeuw and Van Driessen (1999)

Robust ROSE

- Identify "outlying" minority cases based on Mahalanobis distance (MD) using the robust MCD[1] estimator
- Synthetic cases generated only for minority cases x_i with $MD^2(x_i) < \chi_d^2(1-\alpha)$, e.g. $\alpha = 1\%$ ➔ "non-outlying" minority cases

[1] Rousseeuw and Van Driessen (1999)

Robust ROSE

1. Select $\boldsymbol{x}_i$ from the “non-outlying” minority group with probability ∝ 1 / (shortest distance from $\boldsymbol{x}_i$ to majority case)

Robust ROSE

1. Select $\boldsymbol{x}_i$ from the "non-outlying" minority group with probability $\propto$ 1 / (shortest distance from $\boldsymbol{x}_i$ to majority case)

2. Estimate normal density distribution $\mathcal{N}(\boldsymbol{x}_i, H)$
   - selected observation $\boldsymbol{x}_i$ as center
   - *smoothing* matrix $H = \mathrm{MCD}$ covariance matrix estimate on "non-outlying" minority group

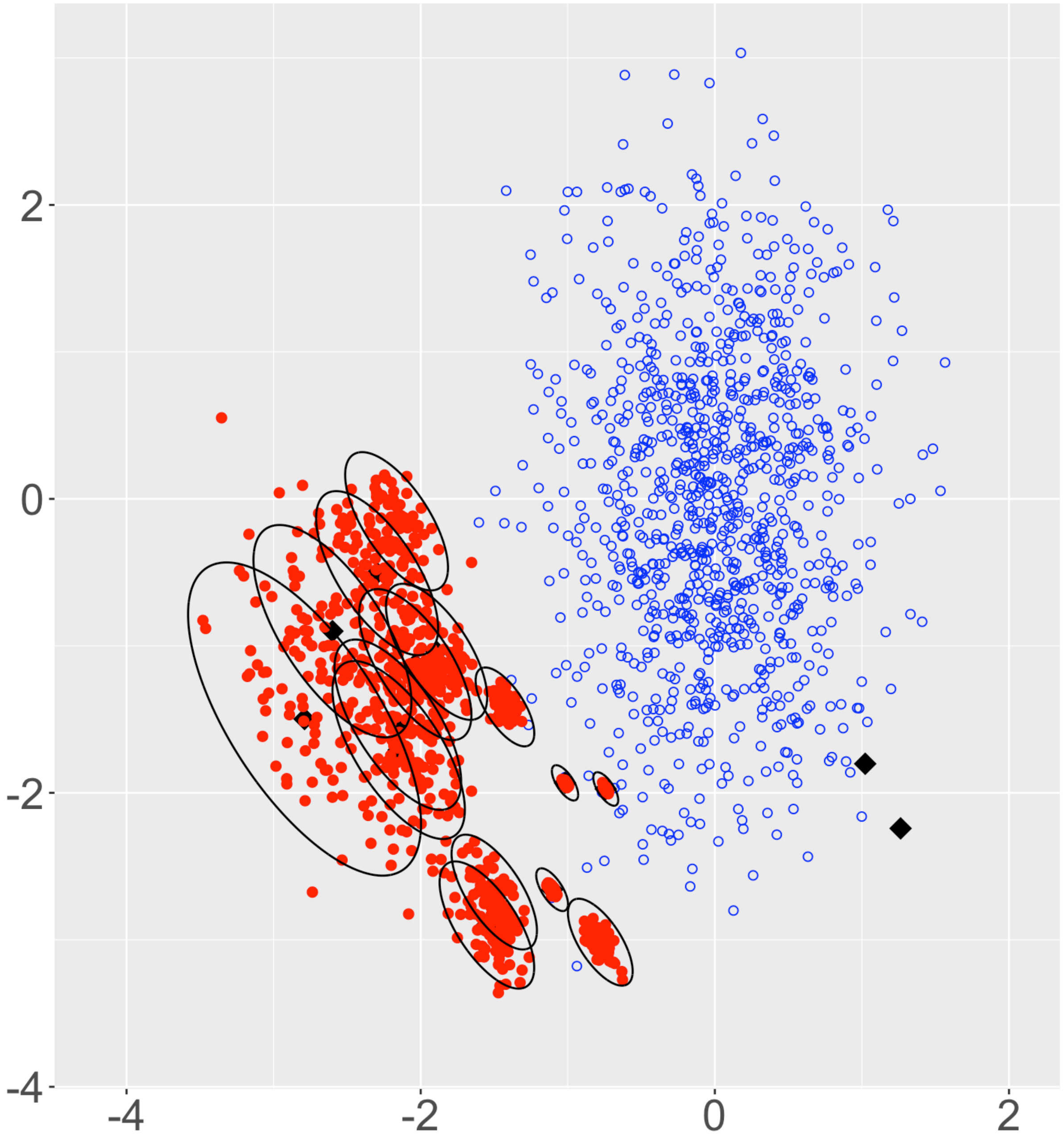
Robust ROSE

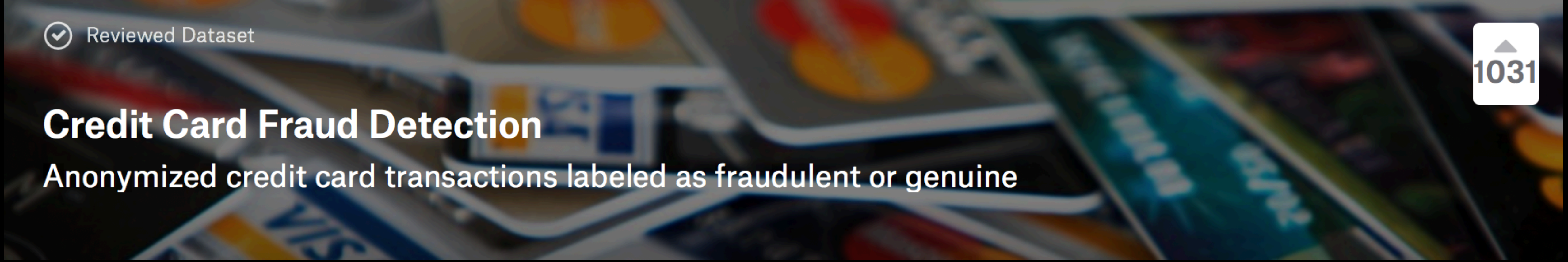
1. Select $\boldsymbol{x}_i$ from the "non-outlying" minority group with probability $\propto$ 1 / (shortest distance from $\boldsymbol{x}_i$ to majority case)

2. Estimate normal density distribution $\mathcal{N}(\boldsymbol{x}_i, H)$
   - selected observation $\boldsymbol{x}_i$ as center
   - *smoothing* matrix $H = \mathrm{MCD}$ covariance matrix estimate on "non-outlying" minority group

3. Generate a new observation from this multivariate normal density

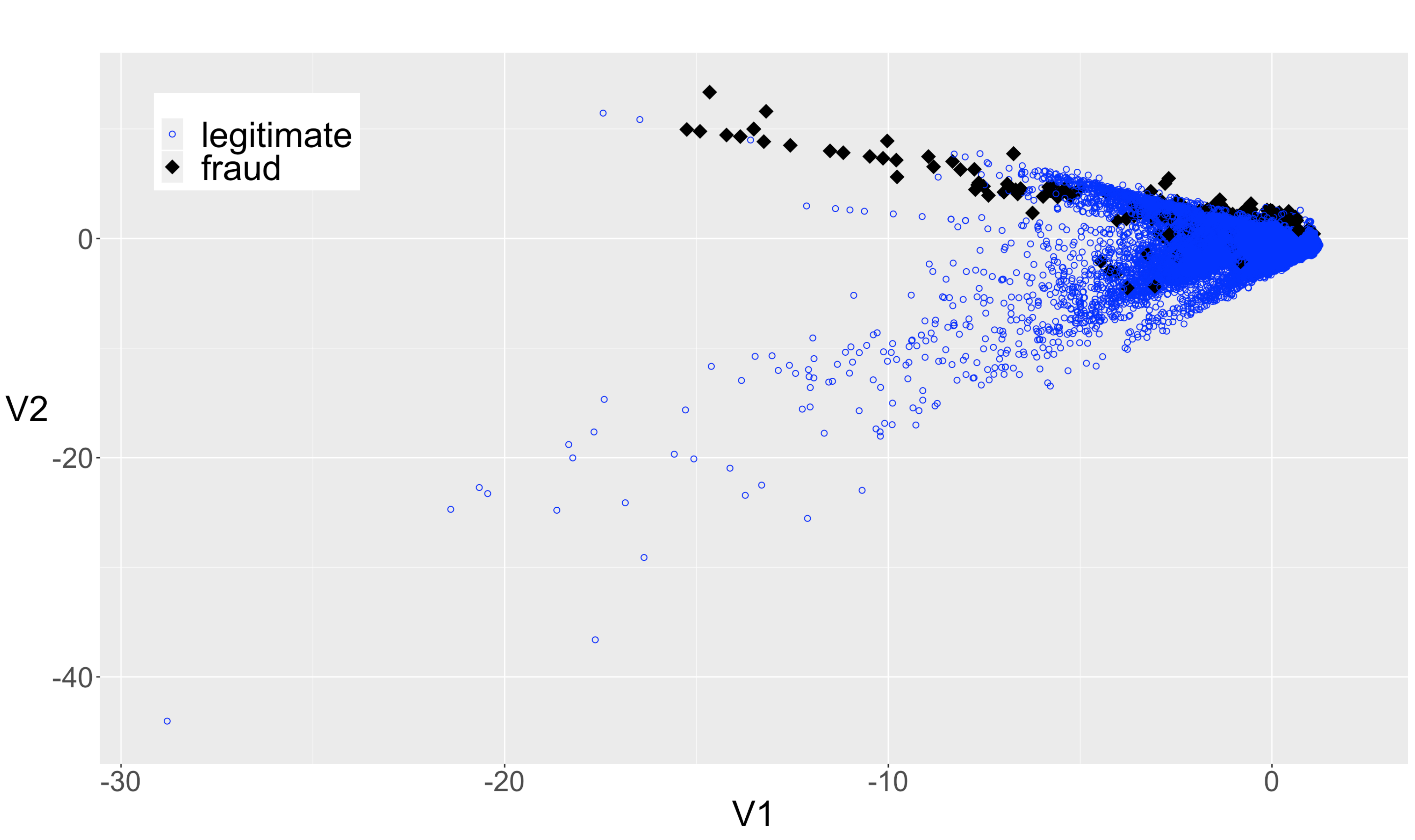
robROSE





source: kaggle.com, made available by Andrea Dal Pozzolo et al., Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

- Transactions made in two days by credit cards in September 2013 by European cardholders
- 497 frauds out of 284,807 transactions $\Rightarrow$ 0.172% is fraud


legitimate
fraud
V2
0
-20
-40
-30
-20
-10
0
V1

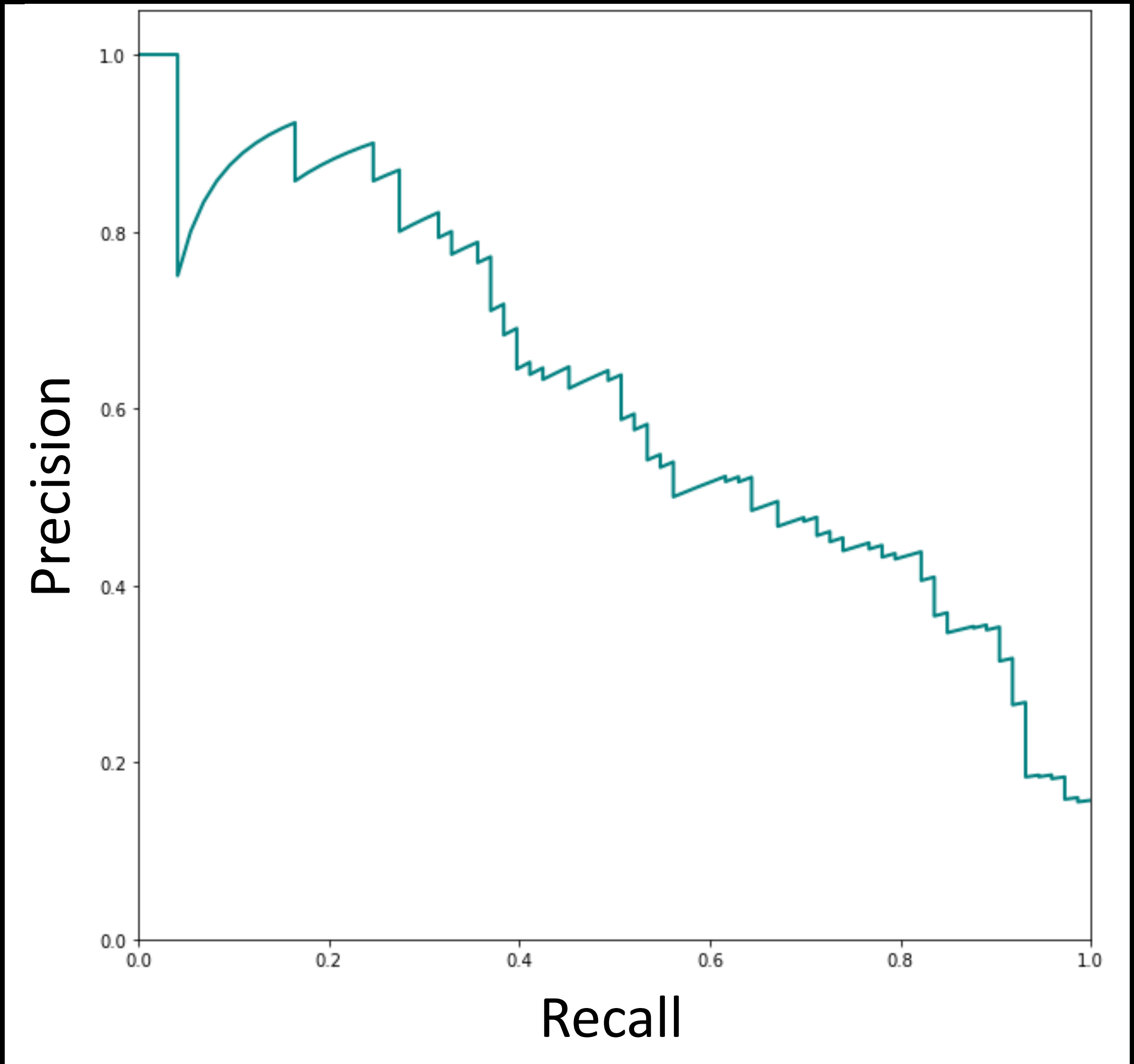
- Precision = $\frac{TP}{TP+FP}$

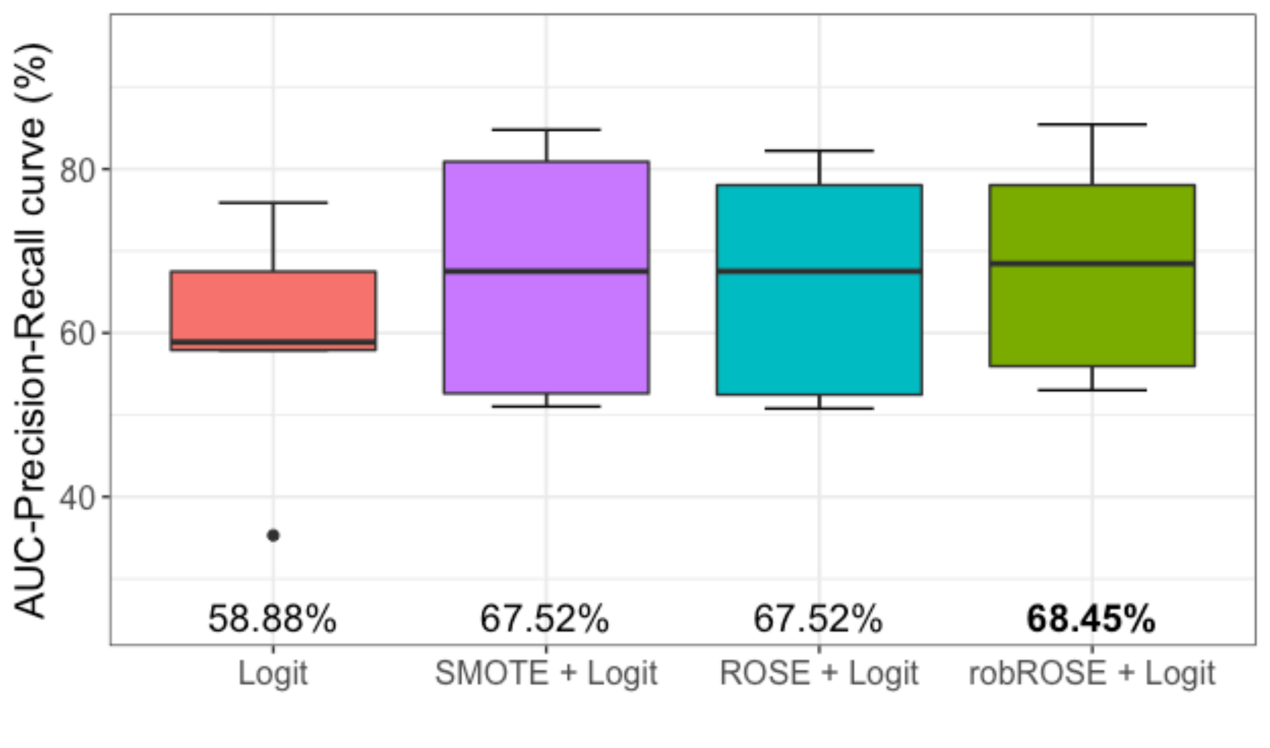
  Out of all cases classified as fraud, how many are actually fraud?

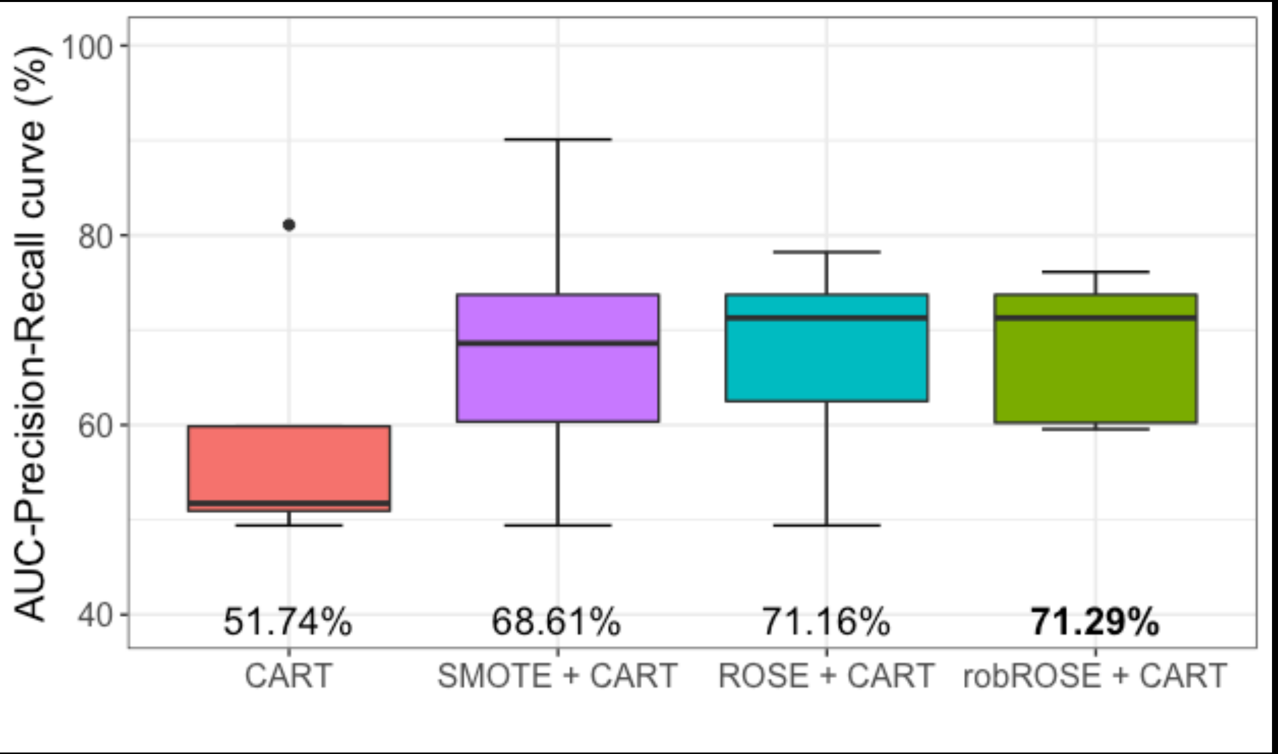
- Recall = $\frac{TP}{TP+FN}$

  Out of all fraud cases, how many are detected?

- Evaluation measure:
  area under precision-recall curve




AUC-Precision-Recall curve (%)
80
60
40
58.88%
67.52%
67.52%
68.45%
Logit
SMOTE + Logit
ROSE + Logit
robROSE + Logit


AUC-Precision-Recall curve (%)
100
80
60
40
51.74%
68.61%
71.16%
71.29%
CART
SMOTE + CART
ROSE + CART
robROSE + CART

Thank you

SMOTE

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

ROSE

Giovanna Menardi and Nicola Torelli. Rose: random over-sampling examples. *Data Mining and Knowledge Discovery*, 28(1):92–122, 2014.