# Model Selection

# Data Set

# Data Set

## Train/Test Split


70%

30%

# Data Set

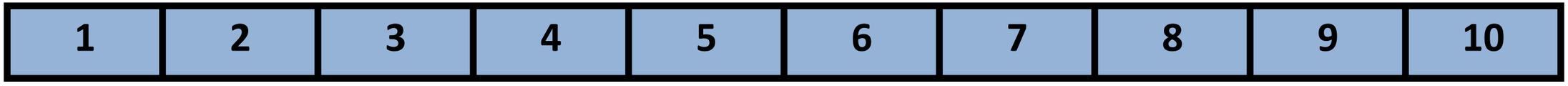
**Training Set**

**70%**

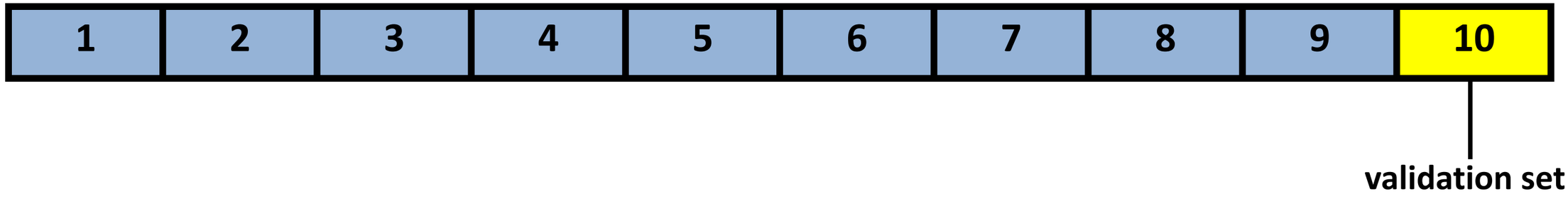
# K-fold Cross Validation

**Training Set**

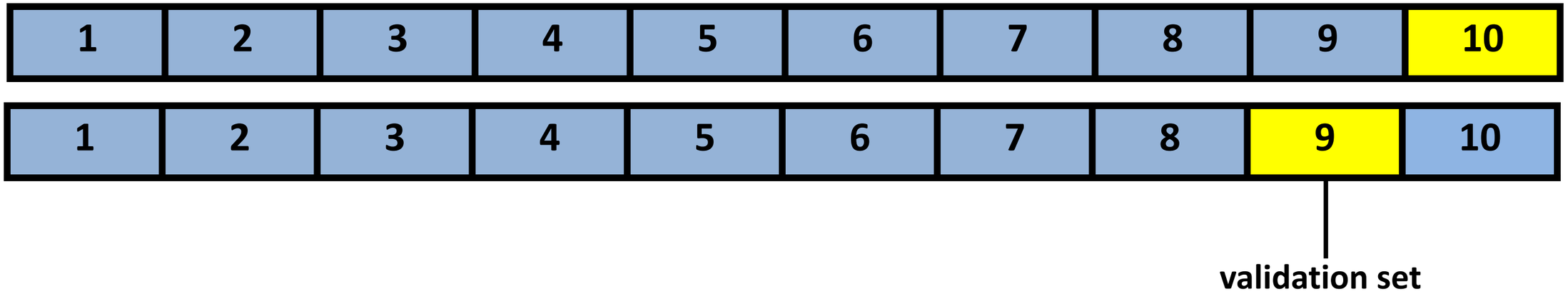
# 10-fold Cross Validation



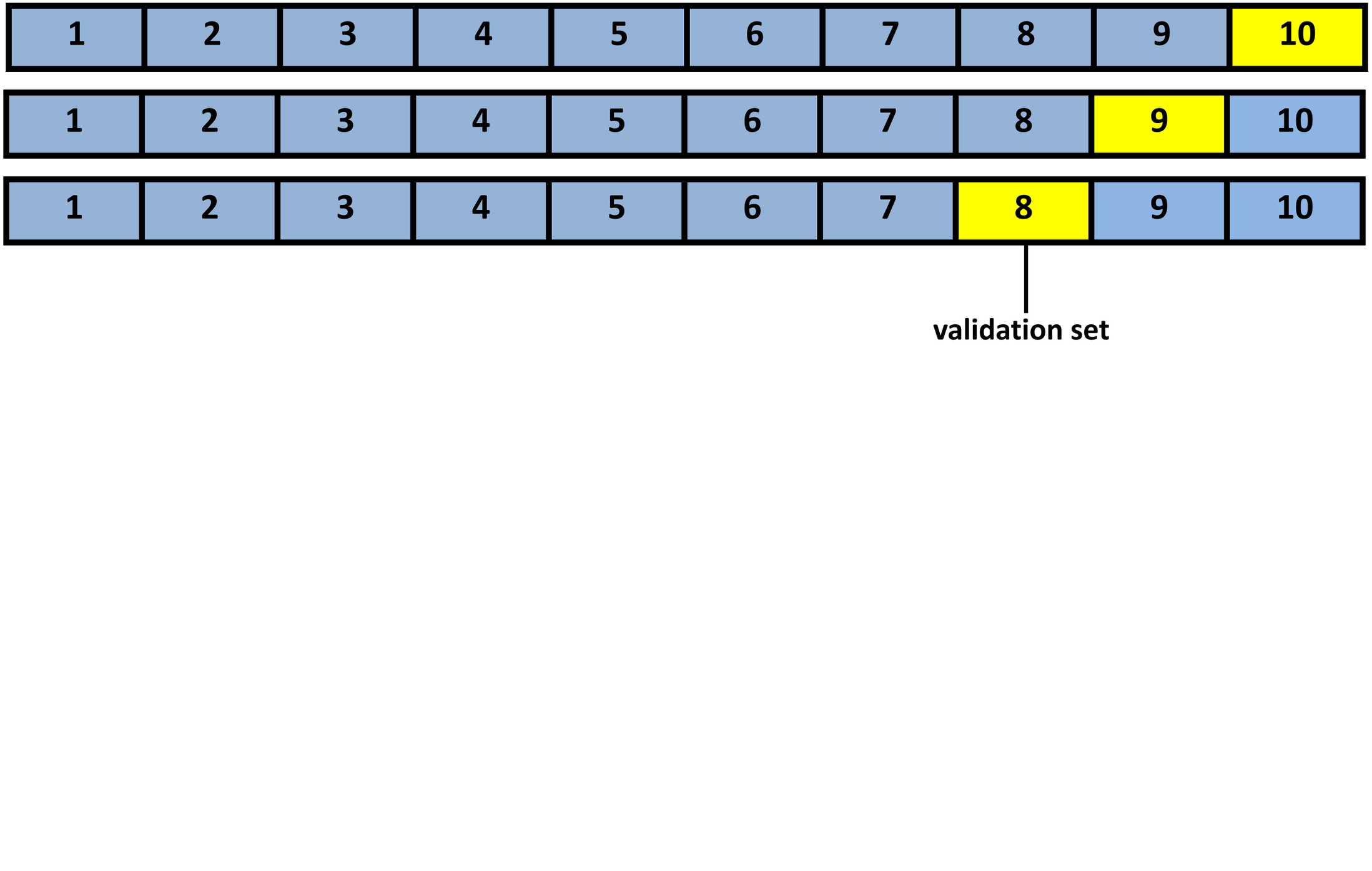
# 10-fold Cross Validation



# 10-fold Cross Validation



# 10-fold Cross Validation



# 10-fold Cross Validation

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **10** |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | **9** | 10 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | **8** | 9 | 10 |
| 1 | 2 | 3 | 4 | 5 | 6 | **7** | 8 | 9 | 10 |
| 1 | 2 | 3 | 4 | 5 | **6** | 7 | 8 | 9 | 10 |
| 1 | 2 | 3 | 4 | **5** | 6 | 7 | 8 | 9 | 10 |
| 1 | 2 | 3 | **4** | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 2 | **3** | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | **2** | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **1** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

# 10-fold Cross Validation

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **10** | .69 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | **9** | 10 | .64 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | **8** | 9 | 10 | .73 |
| 1 | 2 | 3 | 4 | 5 | 6 | **7** | 8 | 9 | 10 | .82 |
| 1 | 2 | 3 | 4 | 5 | **6** | 7 | 8 | 9 | 10 | .64 |
| 1 | 2 | 3 | 4 | **5** | 6 | 7 | 8 | 9 | 10 | .70 |
| 1 | 2 | 3 | **4** | 5 | 6 | 7 | 8 | 9 | 10 | .68 |
| 1 | 2 | **3** | 4 | 5 | 6 | 7 | 8 | 9 | 10 | .71 |
| 1 | **2** | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | .70 |
| **1** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | .69 |

# 10-fold Cross Validation

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | **10** | .69 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | **9** | 10 | .64 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | **8** | 9 | 10 | .73 |
| 1 | 2 | 3 | 4 | 5 | 6 | **7** | 8 | 9 | 10 | .82 |
| 1 | 2 | 3 | 4 | 5 | **6** | 7 | 8 | 9 | 10 | .64 |
| 1 | 2 | 3 | 4 | **5** | 6 | 7 | 8 | 9 | 10 | .70 |
| 1 | 2 | 3 | **4** | 5 | 6 | 7 | 8 | 9 | 10 | .68 |
| 1 | 2 | **3** | 4 | 5 | 6 | 7 | 8 | 9 | 10 | .71 |
| 1 | **2** | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | .70 |
| **1** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | .69 |

# Logistic Regression

| Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| .69 | .64 | .73 | .82 | .64 | .70 | .68 | .71 | .70 | .69 | **.70** |

# Logistic Regression

| Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| .69 | .64 | .73 | .82 | .64 | .70 | .68 | .71 | .70 | .69 | **.70** |

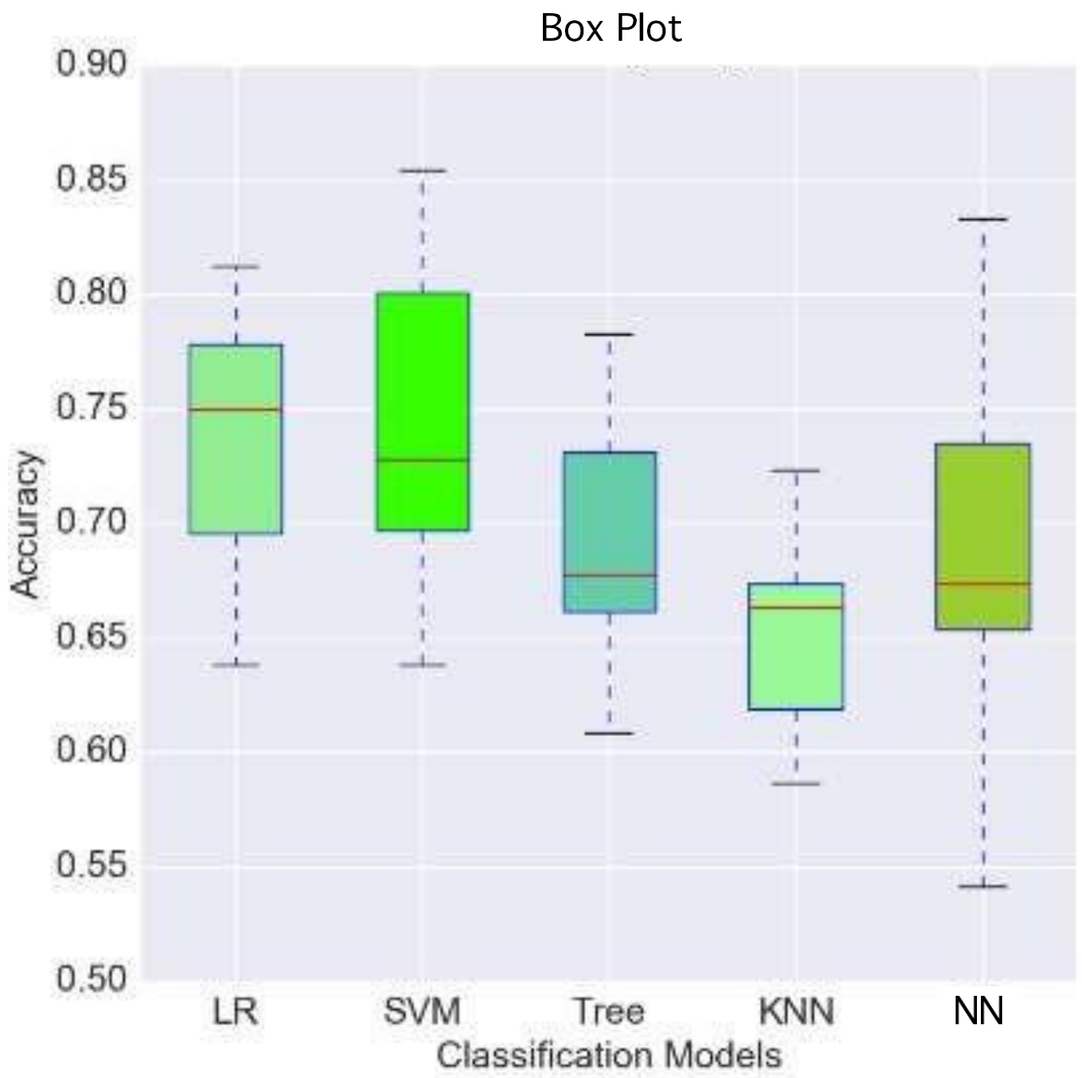
# Logistic Regression

| Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| .69 | .64 | .73 | .82 | .64 | .70 | .68 | .71 | .70 | .69 | **.70** |

# Logistic Regression

| Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| .69 | .64 | .73 | .82 | .64 | .70 | .68 | .71 | .70 | .69 | **.70** |

| Logistic Regression | Support Vector Machine | Decision Tree | K-Nearest Neighbor | Neural Network |
|---|---|---|---|---|
| 0.705 | 0.722 | 0.635 | 0.675 | 0.607 |



# Hyperparameter Tuning

# Logistic Regression

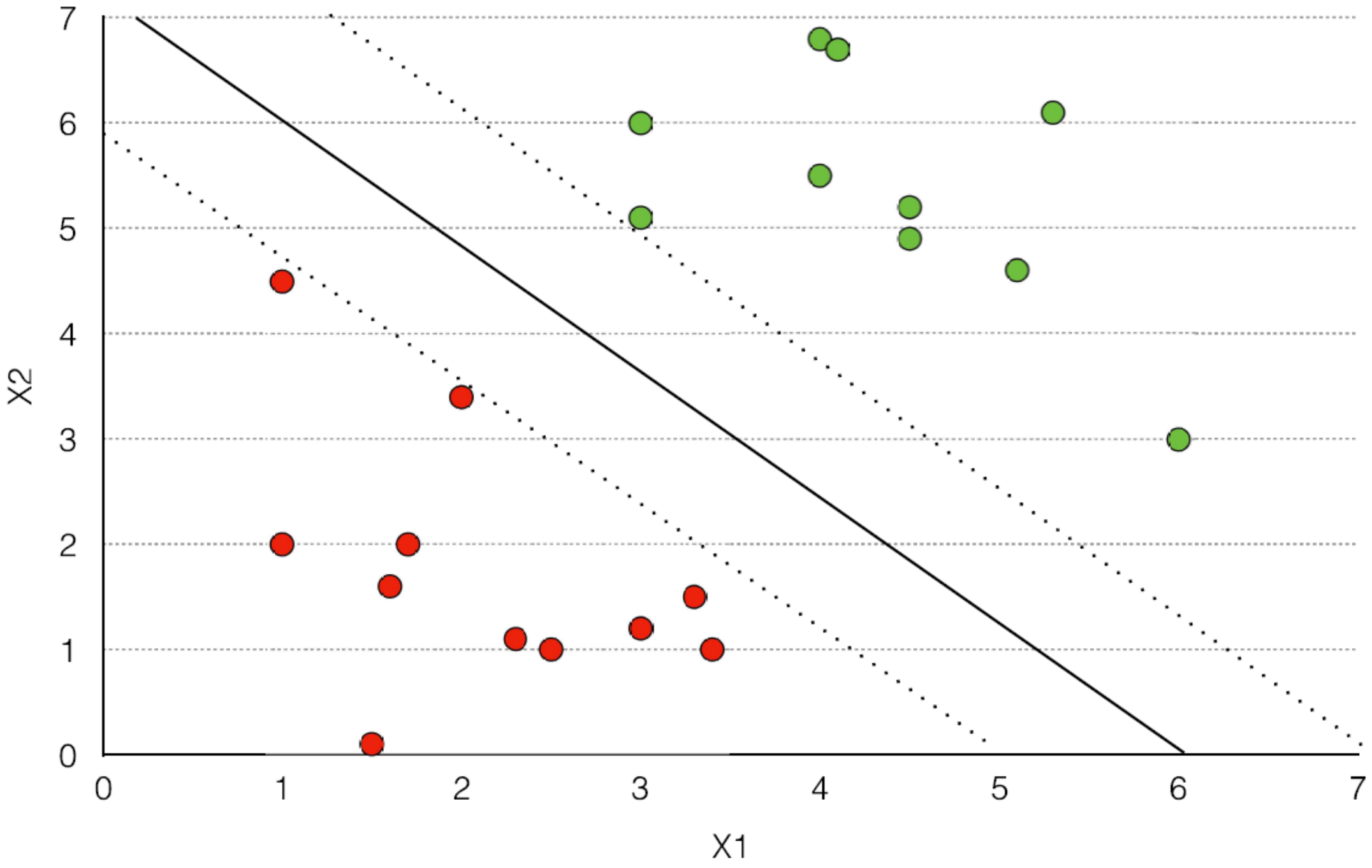
$$LOC = 227.63 + 9.51x_1 + 2.7x_2 - 7.08x_3$$

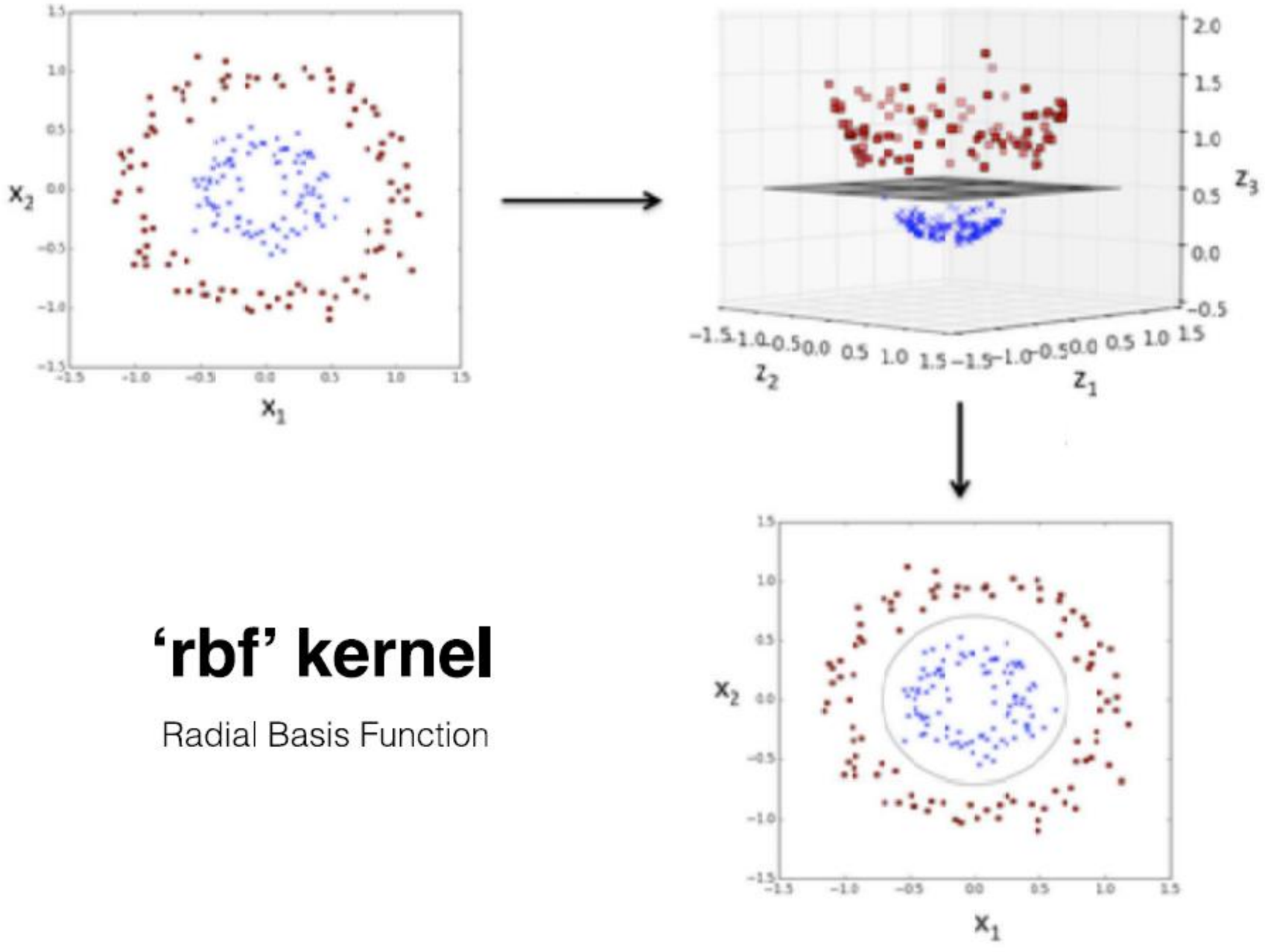
$X_1$ = hour pair programming

$X_2$ = gender (m = 0; f = 1)

$X_3$ = number of social accounts

# Support Vector Machine

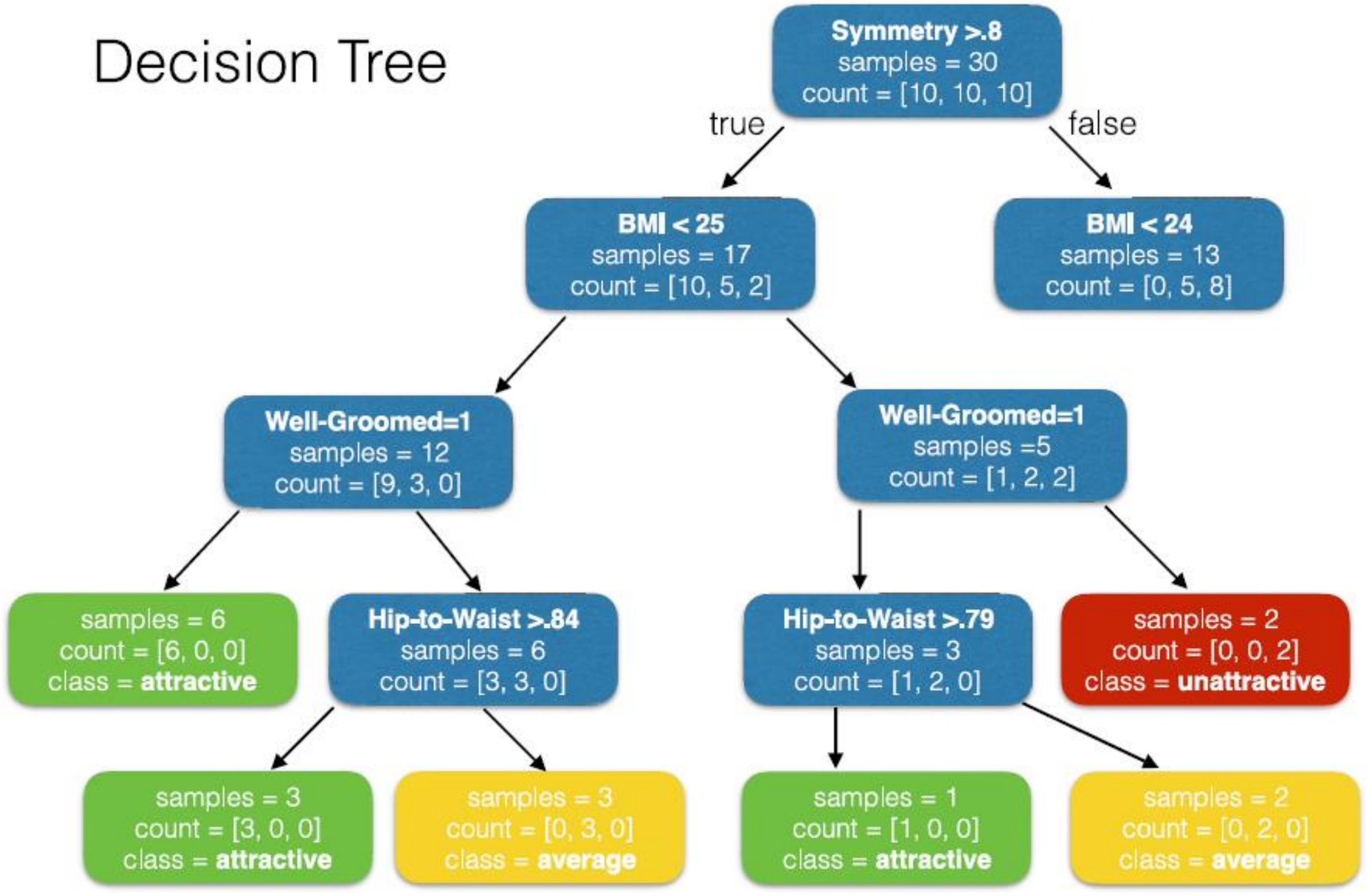





# ‘rbf’ kernel

Radial Basis Function

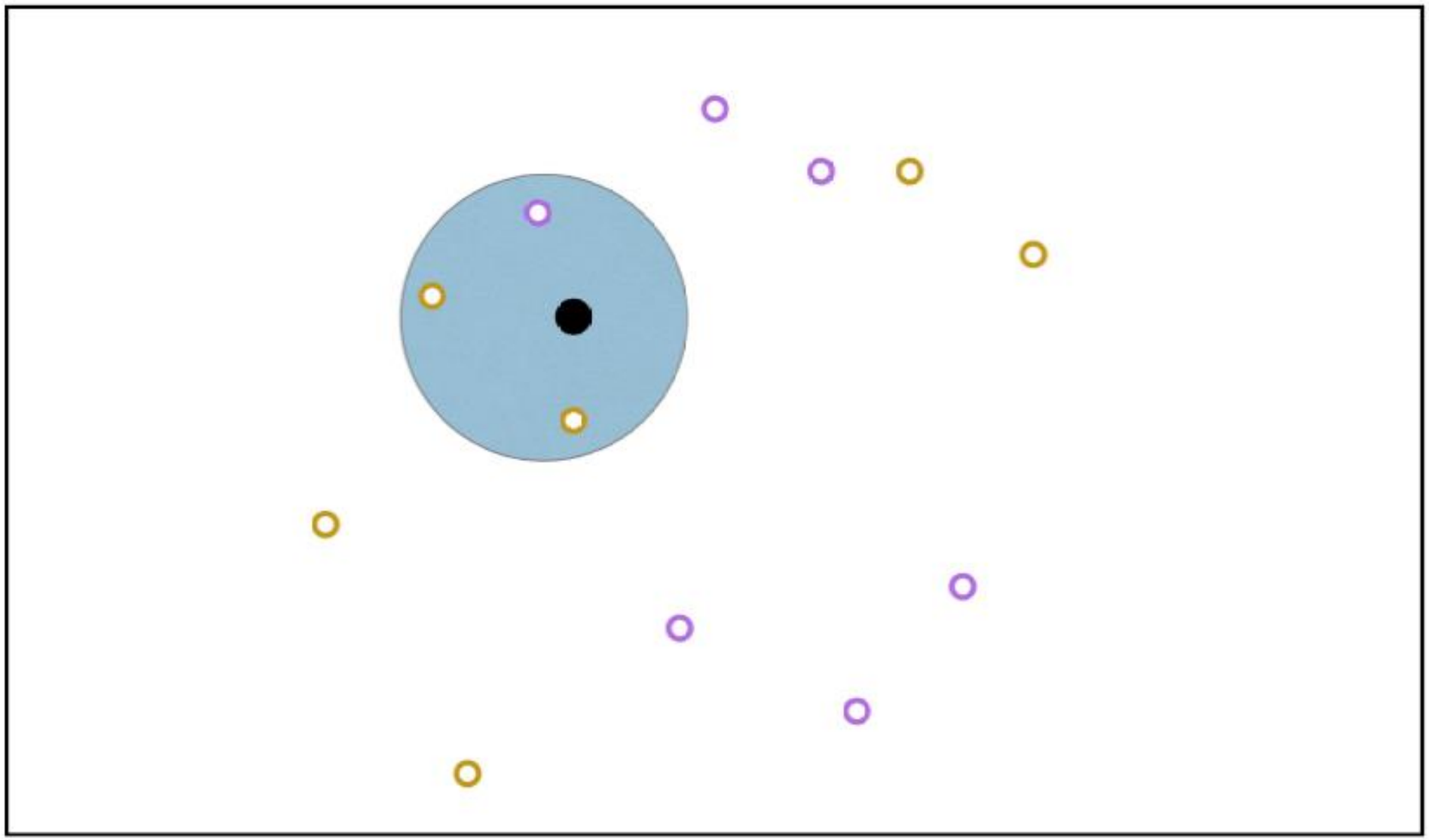
# Decision Tree



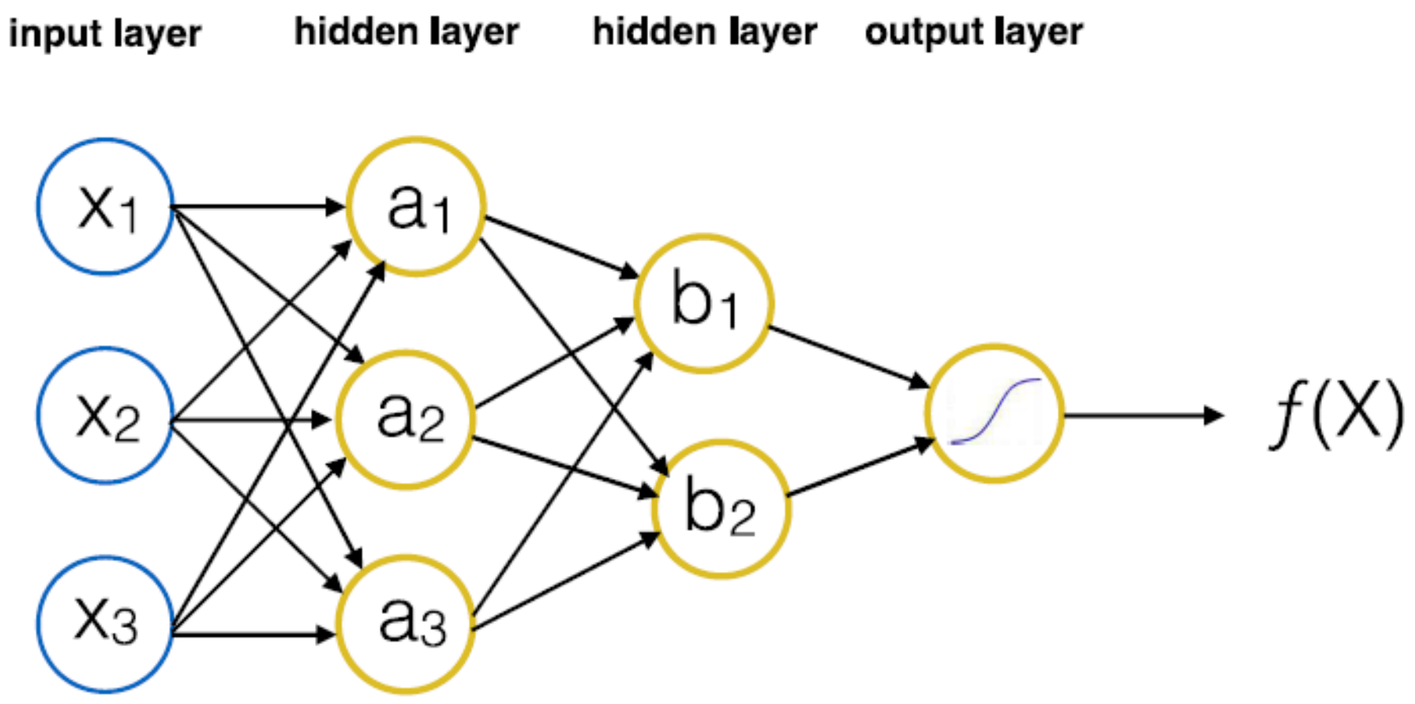

[att, ave, un]

k-Nearest Neighbor

**k = 3**



# Multi-Layer Perceptron (MLP)



# Grid Search

# Grid Search

Support Vector Machine

# Grid Search

Support Vector Machine

**param_grid=**[{‘C’: [.1, 1.0, 10], ‘kernel’: [‘linear’, ‘rbf’]}]

# Grid Search

Support Vector Machine

**param_grid**=[{**'C':** [.1, 1.0, 10], **'kernel':** ['linear', 'rbf']}]

| 'C' | 'kernel' |
|---|---|
| | |
| | |
| | |
| | |
| | |
| | |

# Grid Search

## Support Vector Machine

**param_grid**=[{‘C’: **[.1, 1.0, 10],** ‘kernel’: **[‘linear’, ‘rbf’]**}]

| ‘C’ | ‘kernel’ |
|---|---|
| 0.1 | ‘linear’ |
| 1.0 | ‘linear’ |
| 10 | ‘linear’ |
| 0.1 | ‘rbf’ |
| 1.0 | ‘rbf’ |
| 10 | ‘rbf’ |

# Grid Search

Support Vector Machine

**param_grid**=[{‘C’: **[.1, 1.0, 10],** ‘kernel’: **[‘linear’, ‘rbf’]**}]

**Best Hyperparameter Combination**

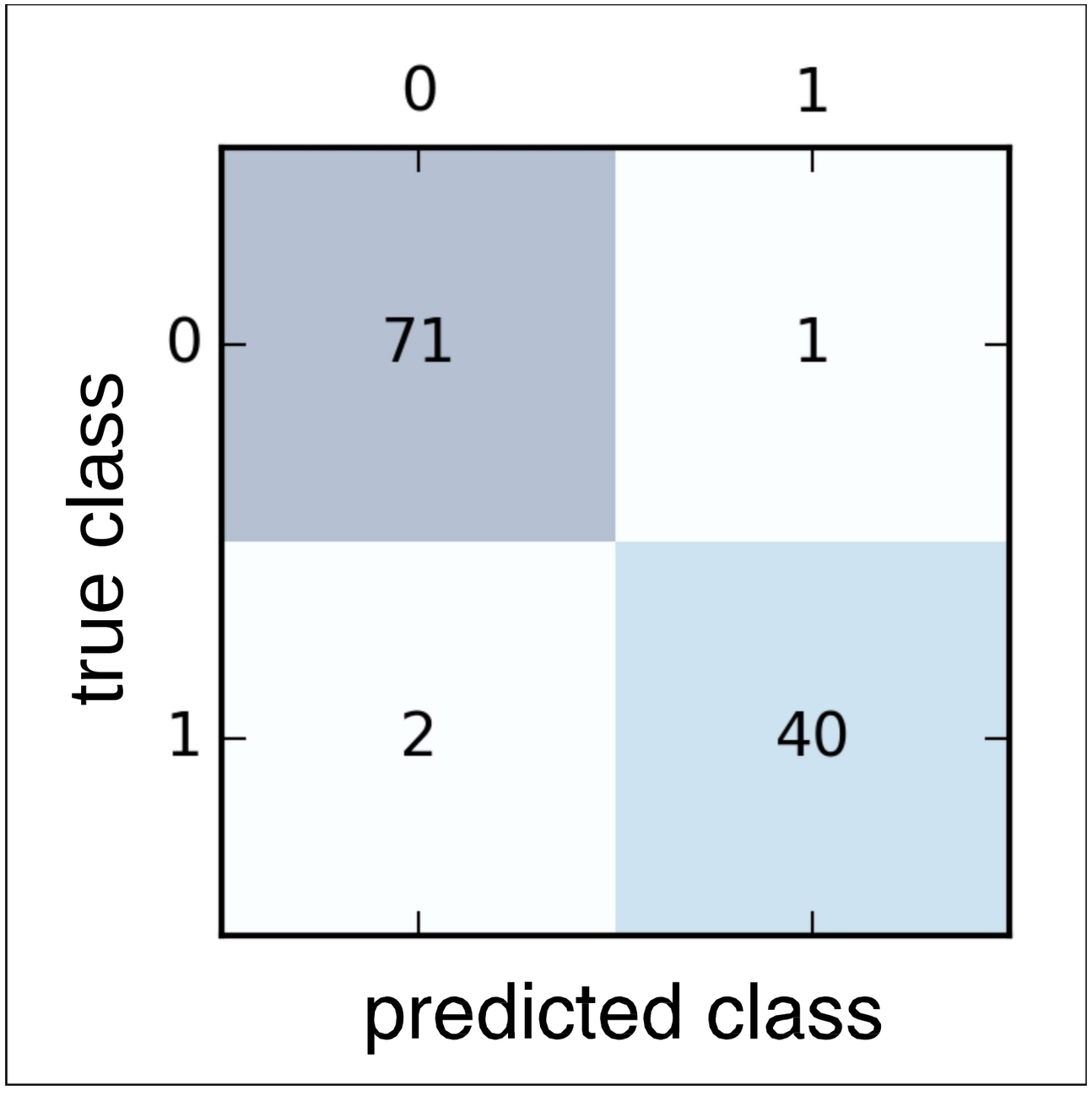
| ‘C’ | ‘kernel’ |
|---|---|
| 0.1 | ‘linear’ |
| 1.0 | ‘linear’ |
| 10 | ‘linear’ |
| 0.1 | ‘rbf’ |
| 1.0 | ‘rbf’ |
| 10 | ‘rbf’ |

# Model Evaluation Metrics

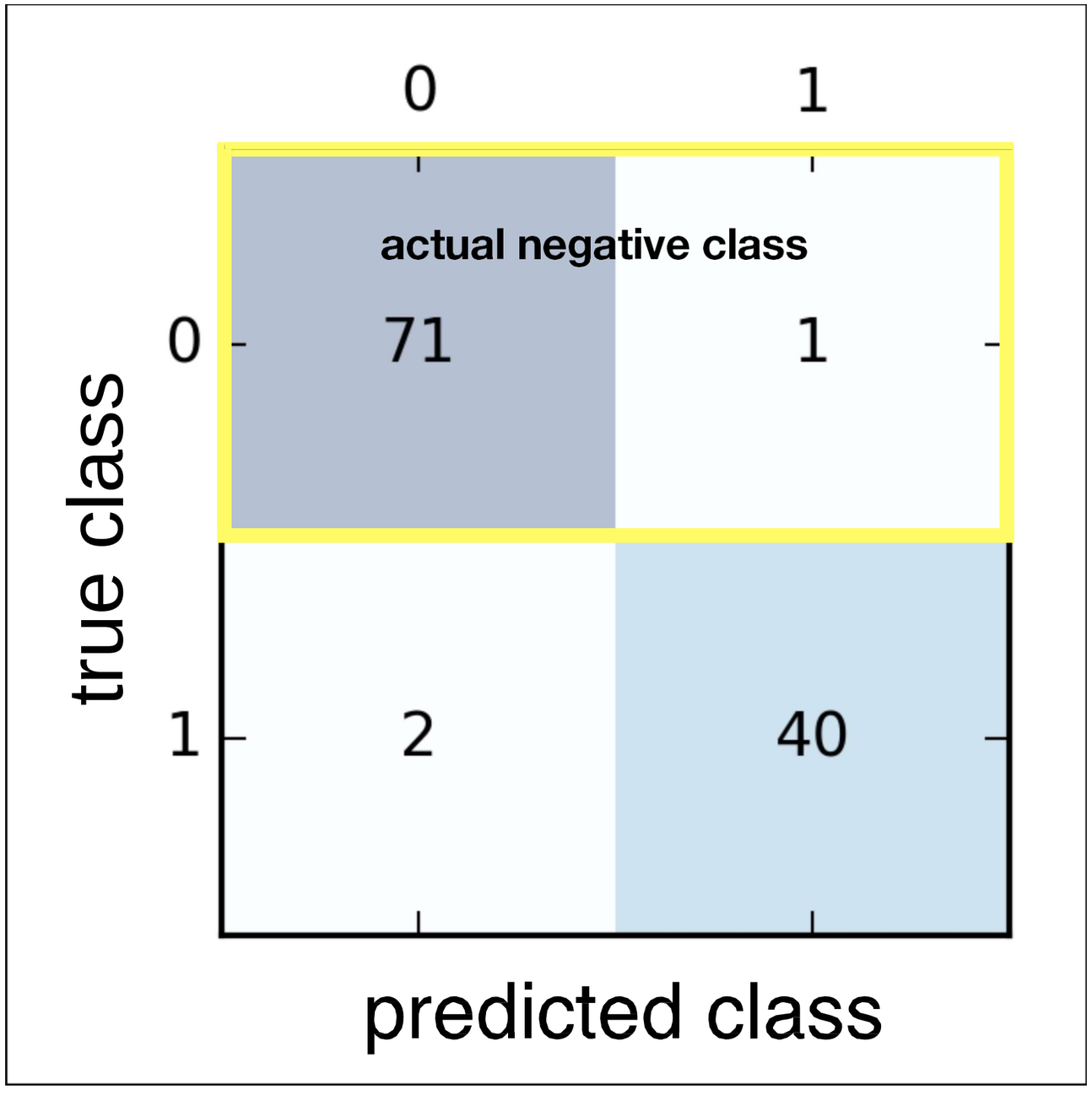
# Confusion Matrix

| | | |
|---|---|---|
| Negative Class | **TN**<br>True Negative | **FP**<br>False Positive |
| Positive Class | **FN**<br>False Negative | **TP**<br>True Positive |
| | Predicted Negative | Predicted Positive |

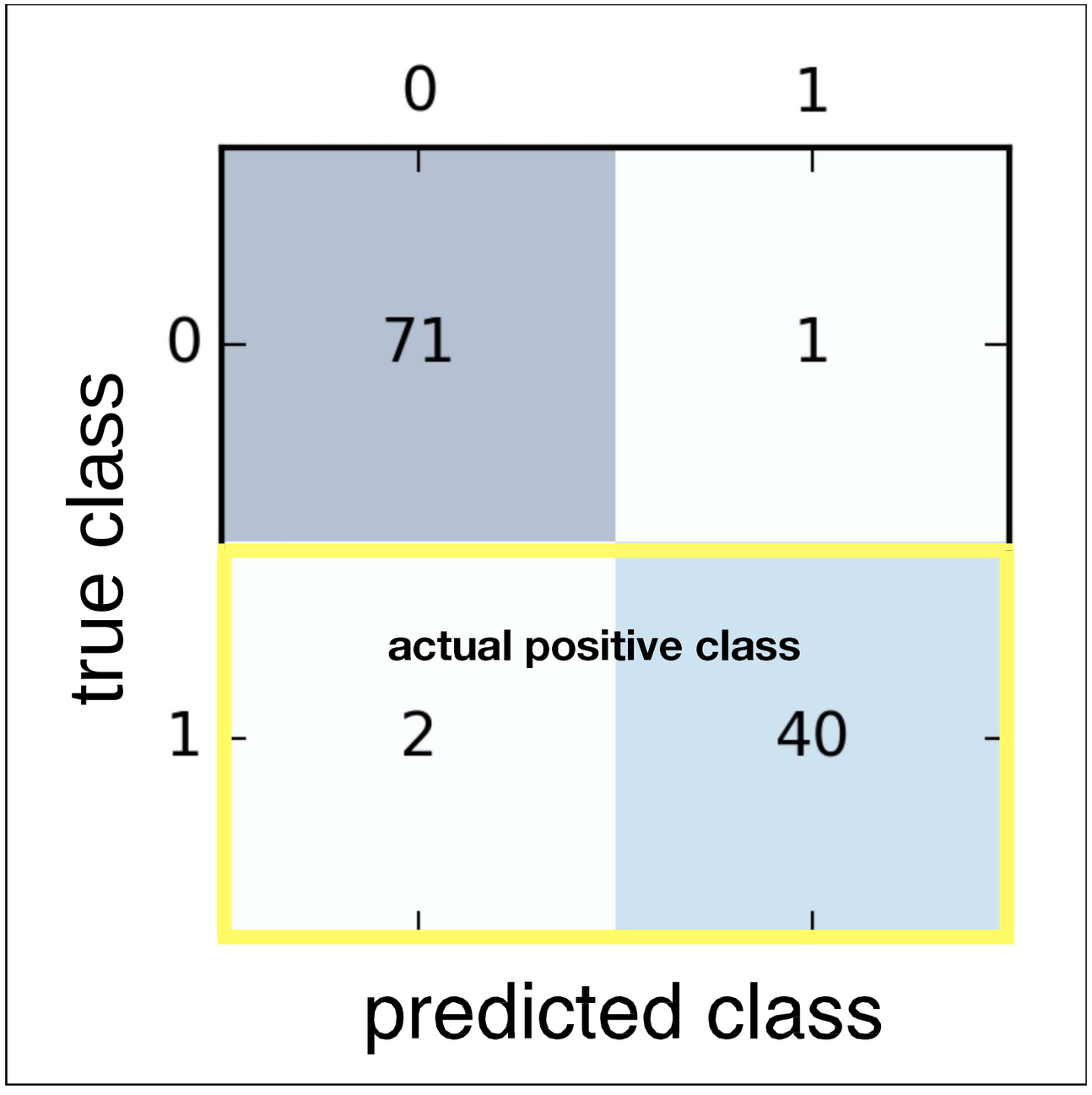
# Confusion Matrix



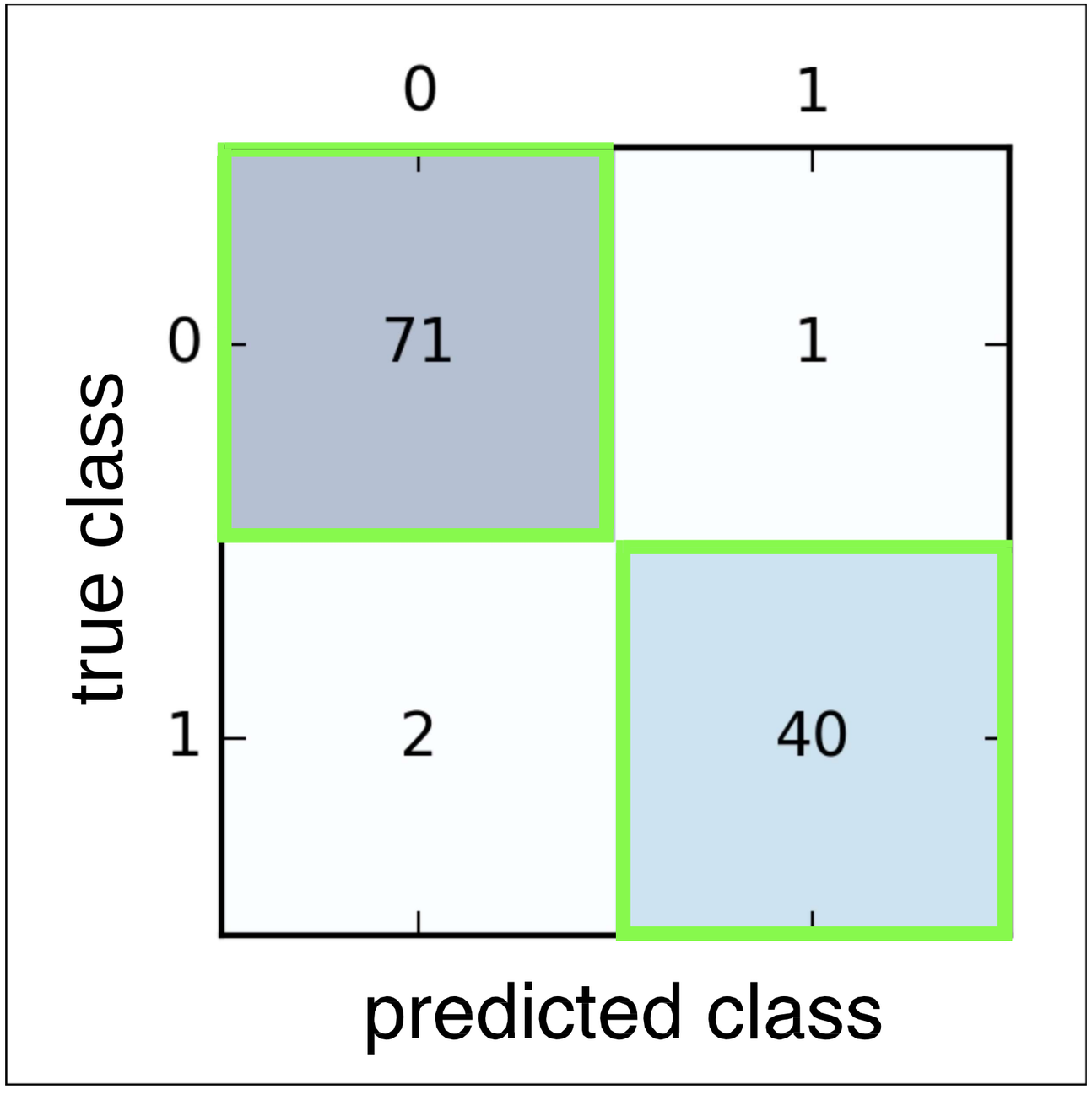
# Confusion Matrix



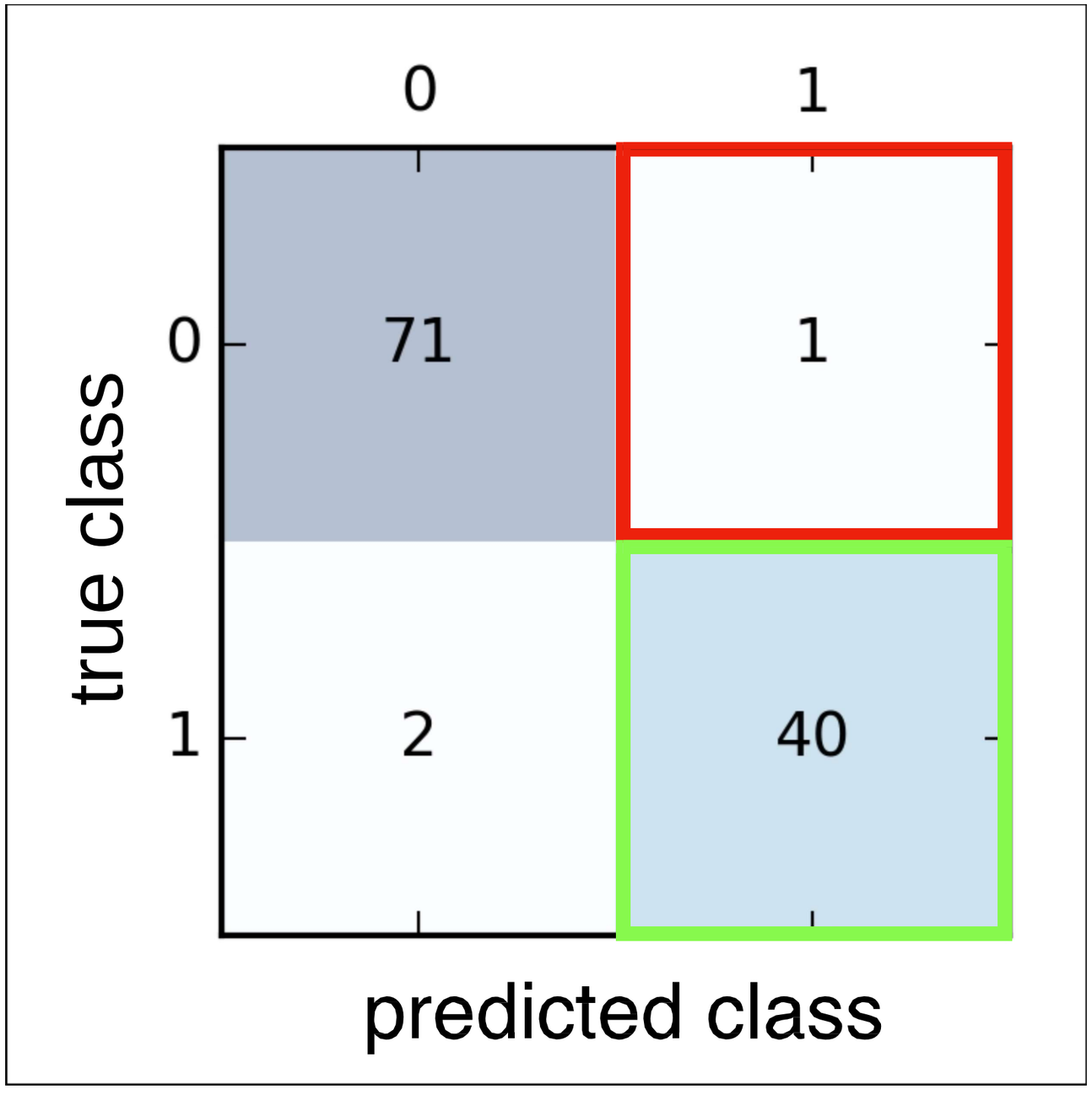
# Confusion Matrix



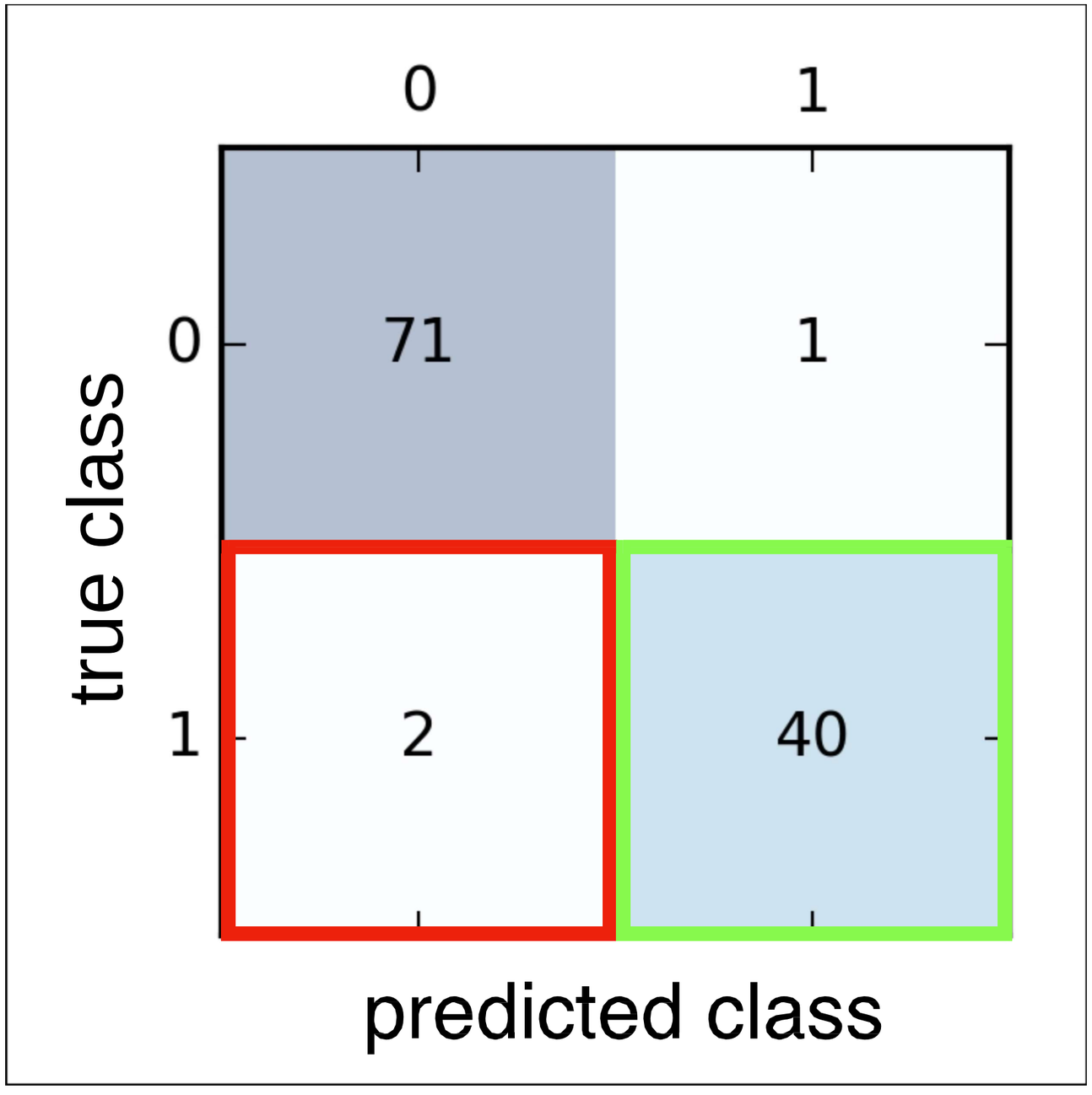
# Accuracy



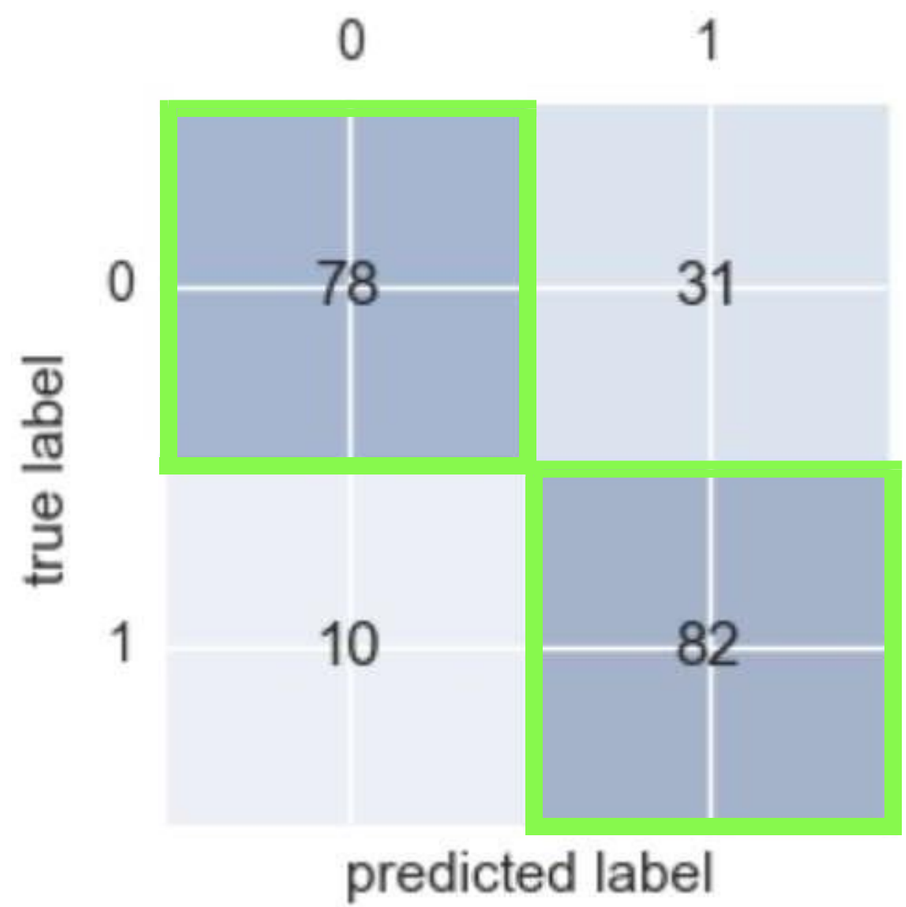
# Precision



# Recall

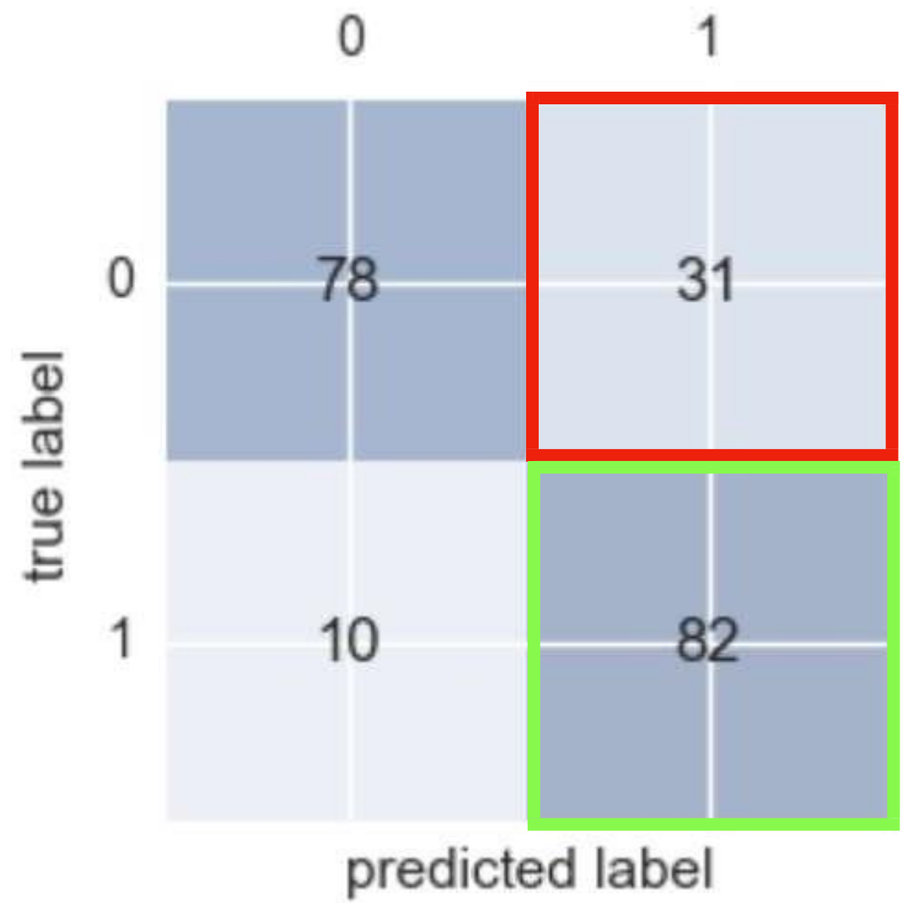


# Confusion Matrix



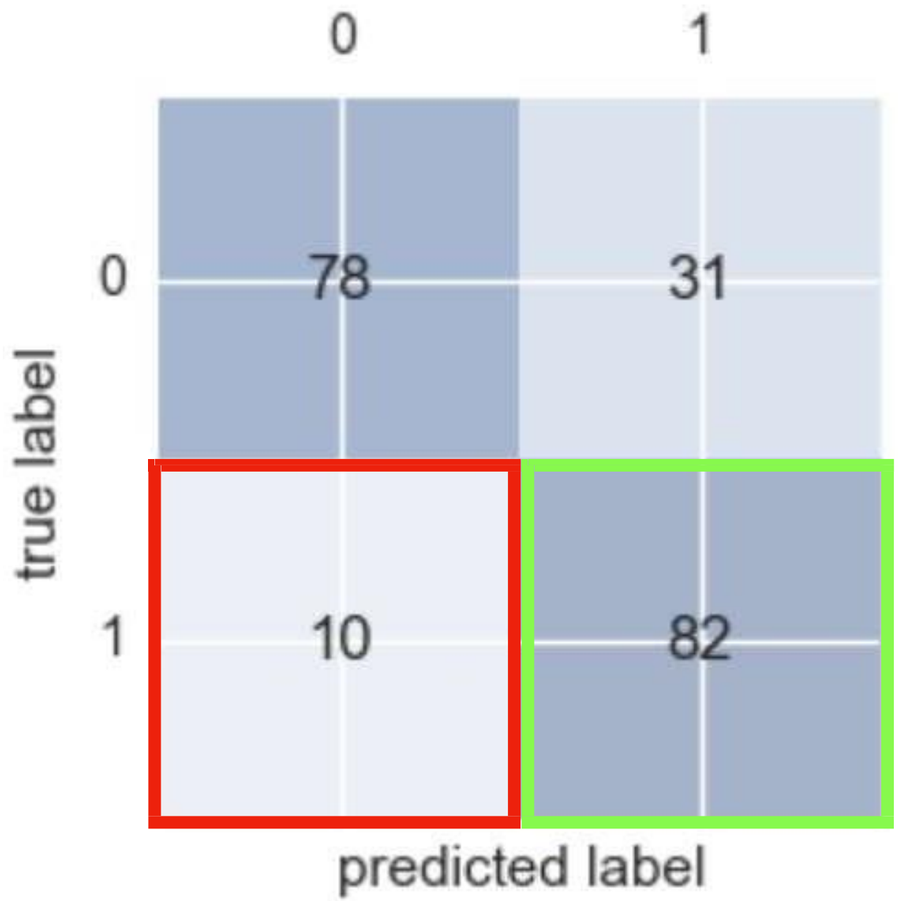

| Accuracy |
| --- |
| 0.796 |

# Confusion Matrix




| Precision |
| --- |
| 0.725 |

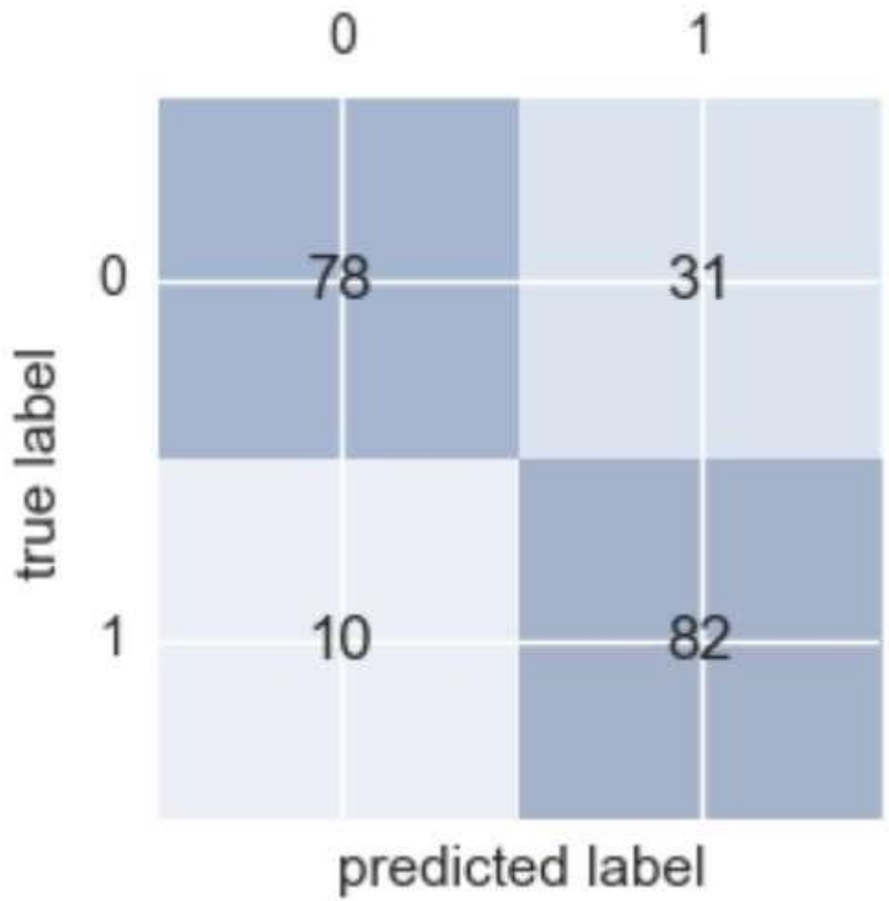
# Confusion Matrix




| Recall |
| --- |
| 0.891 |

# Model Evaluation Metrics

| | Precision | Recall | F1-score |
|---|---|---|---|
| Rejected | 0.89 | 0.72 | 0.79 |
| Liked | 0.73 | 0.89 | 0.81 |



# Fraudulent Transactions

# Fraudulent Transactions

- **The average merchant experiences 156 successful fraudulent transactions per month.**
- **The average value of a fraudulent transaction is $114**

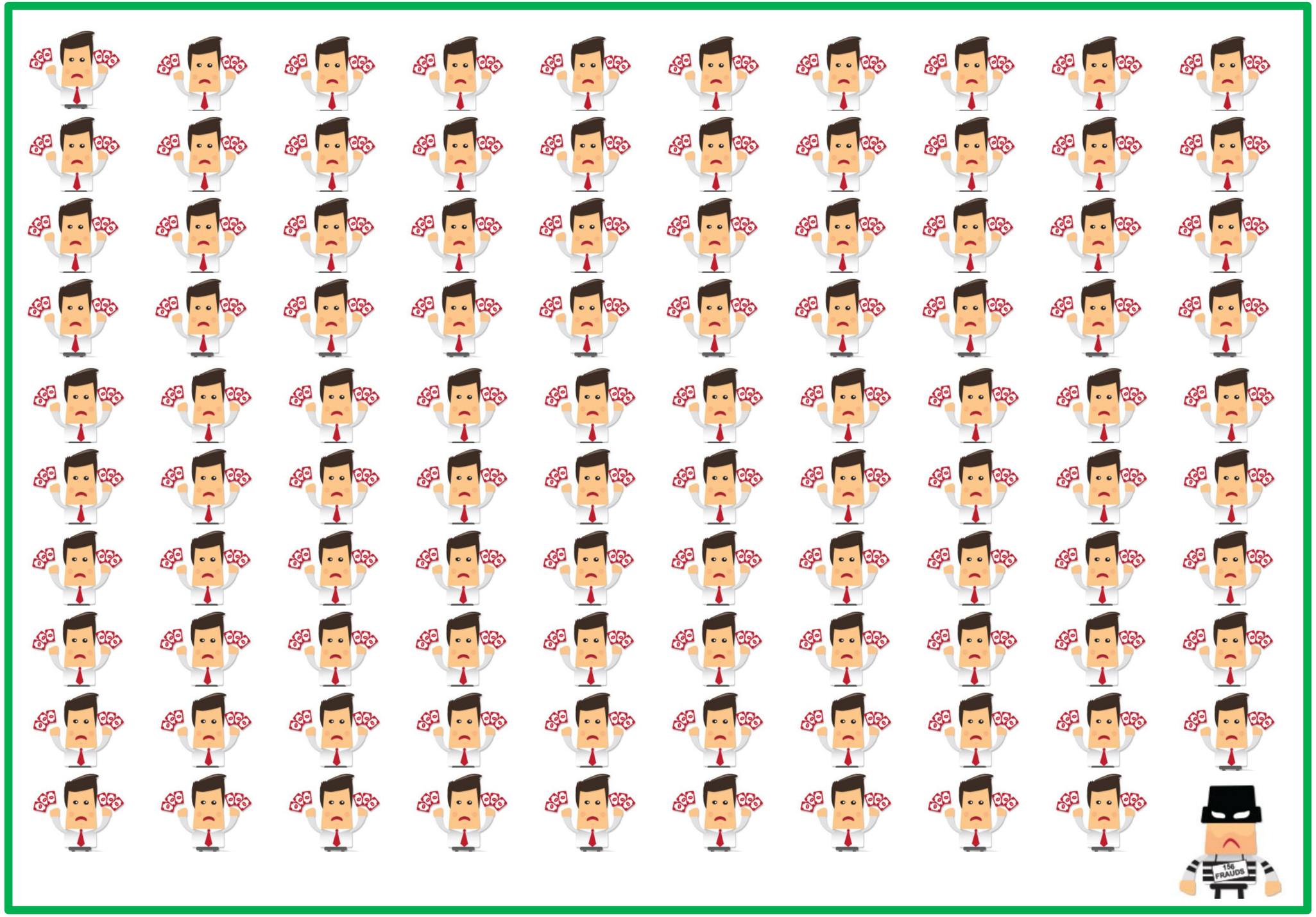
# Fraudulent Transactions

# Fraudulent Transactions

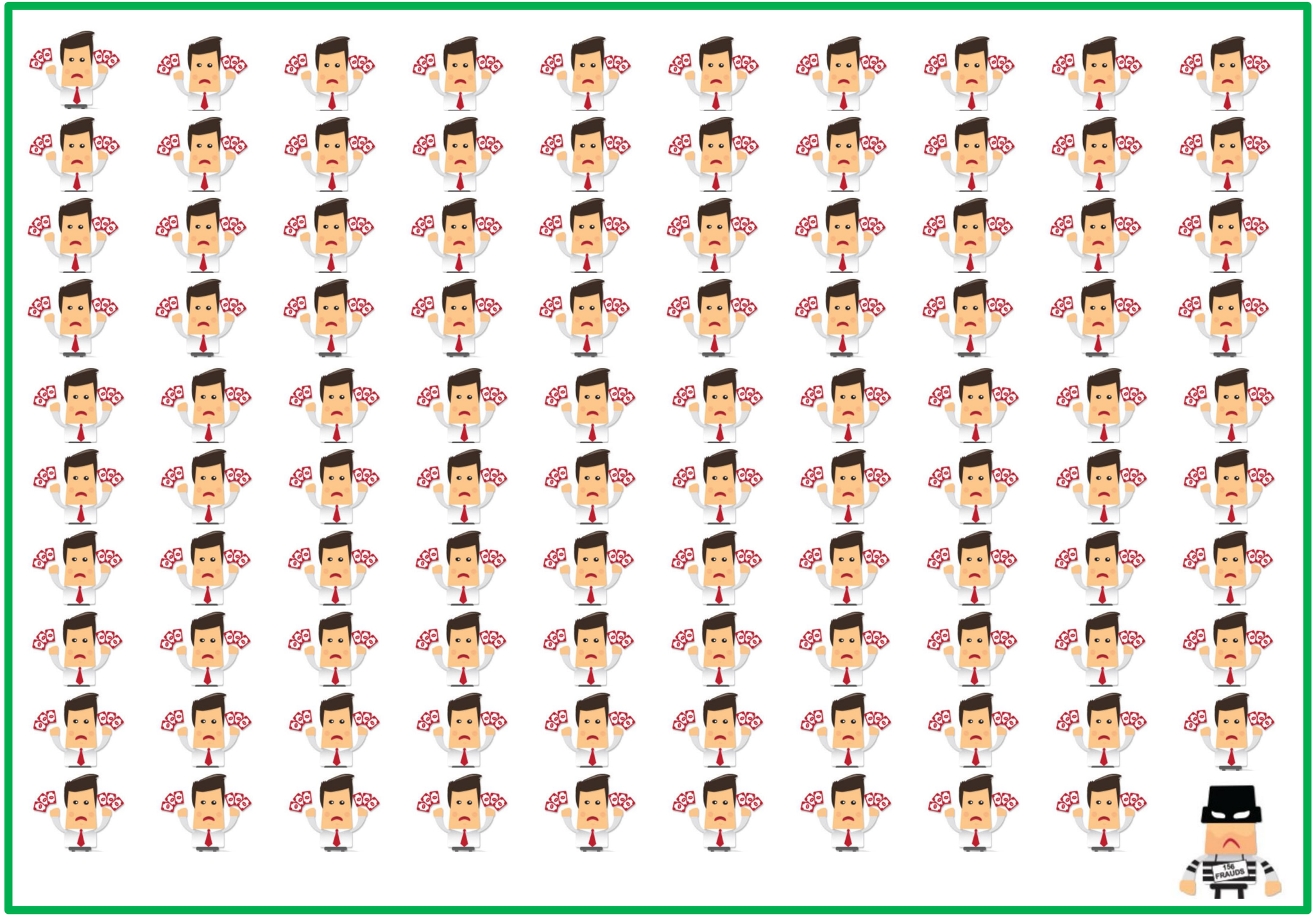

156
FRAUDS


156
FRAUDS


150
FRAUDS

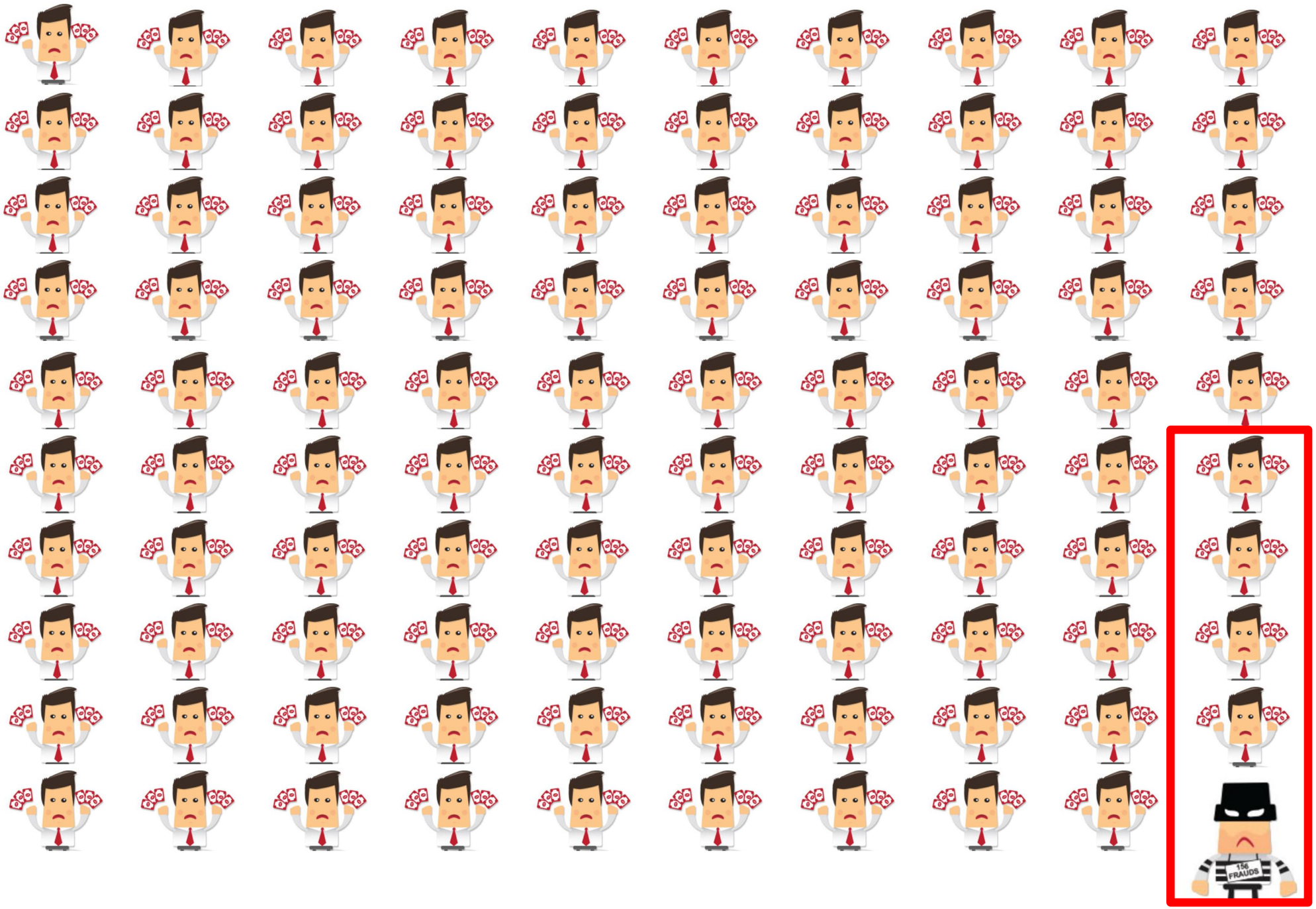


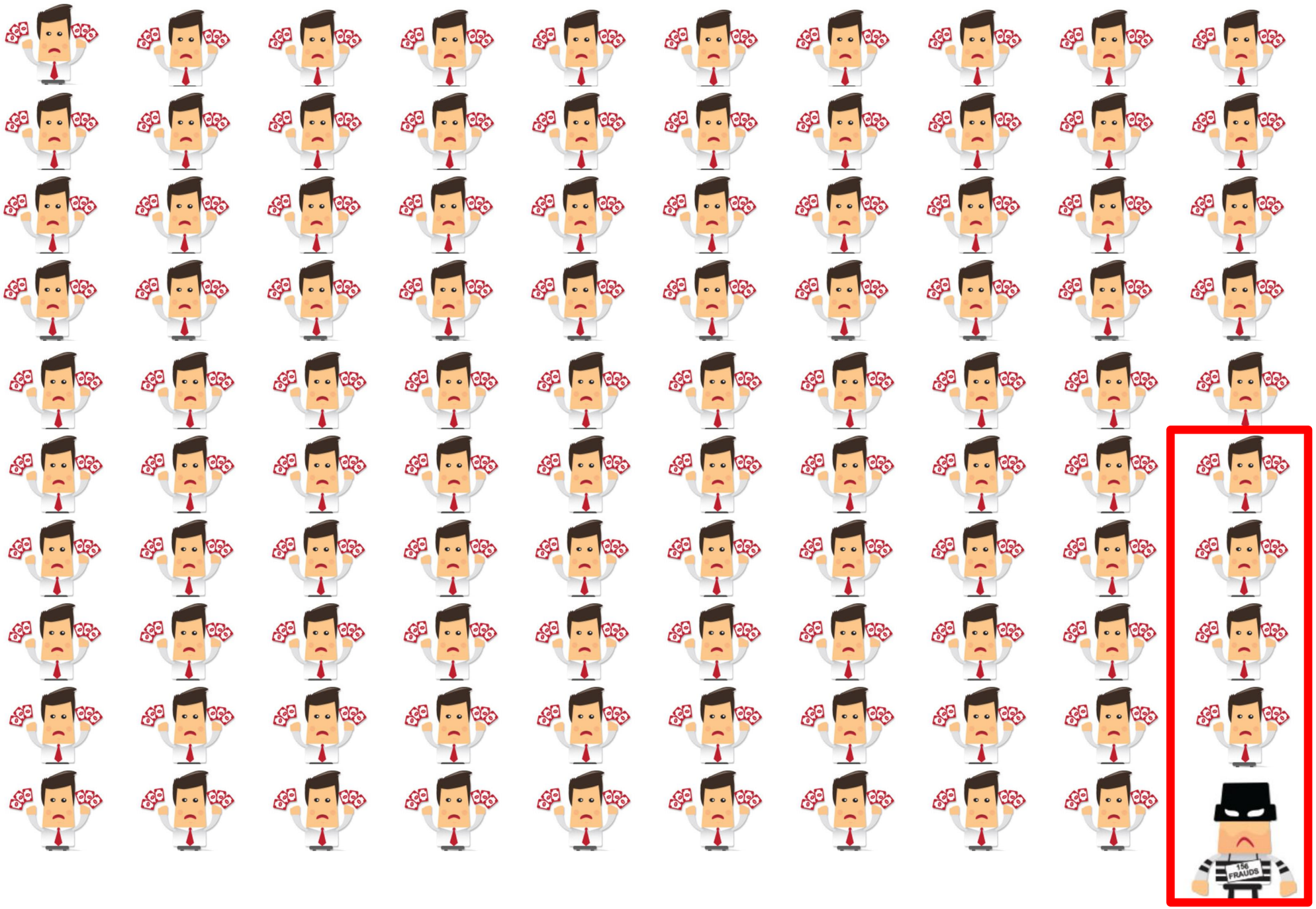

**Accuracy**



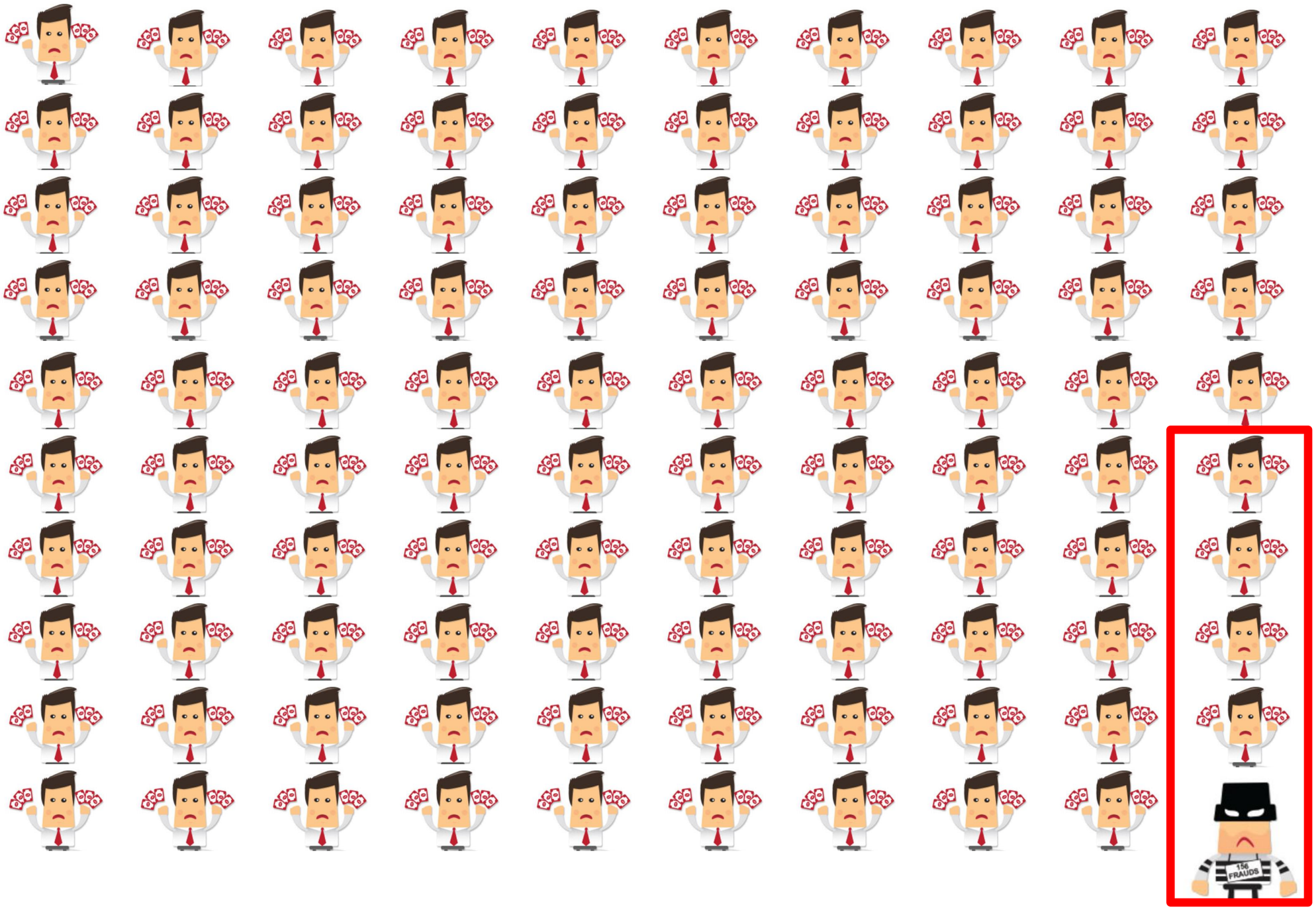

**Accuracy**

**99%**


150
FRAUDS

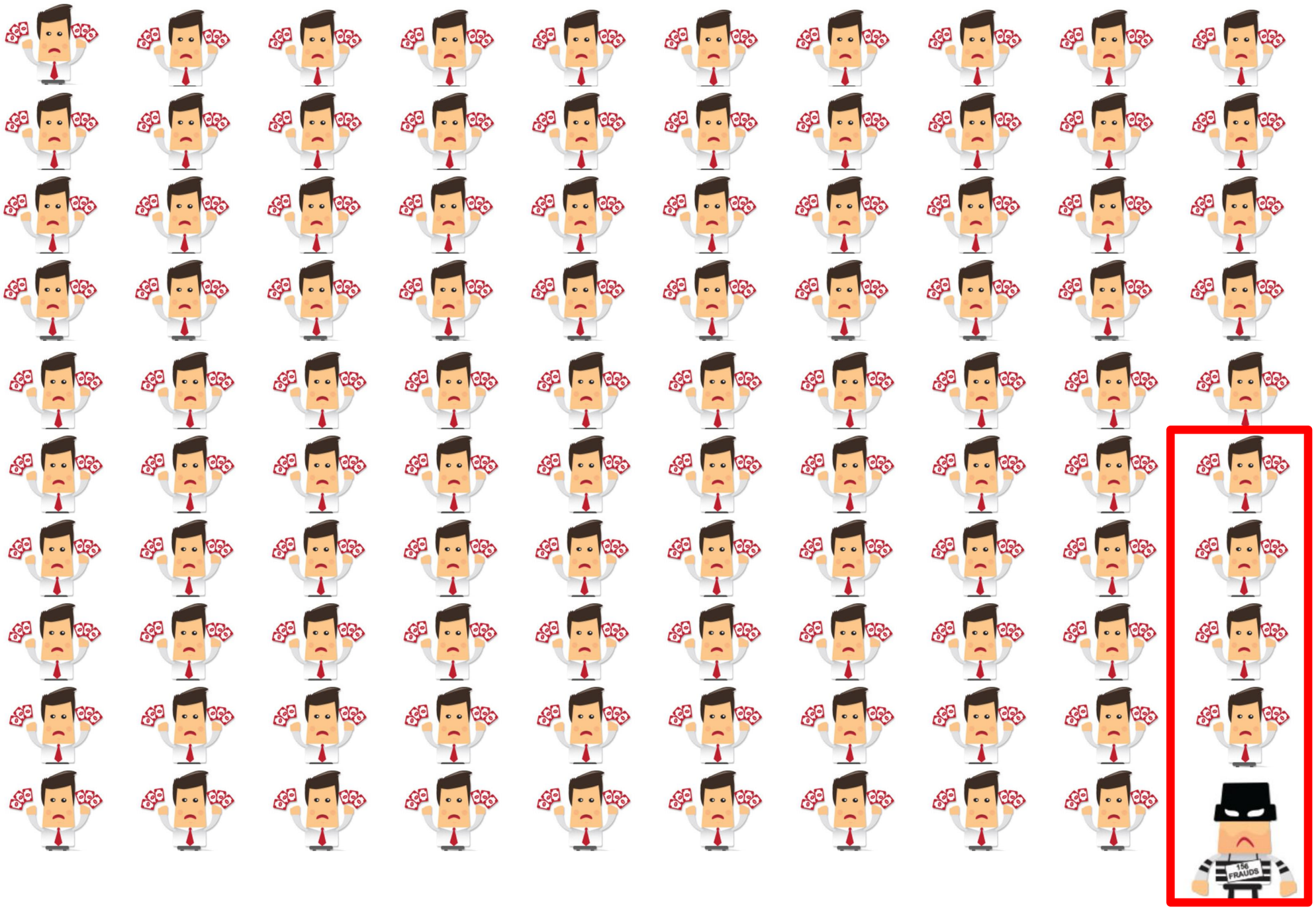



**Precision**




**Precision**

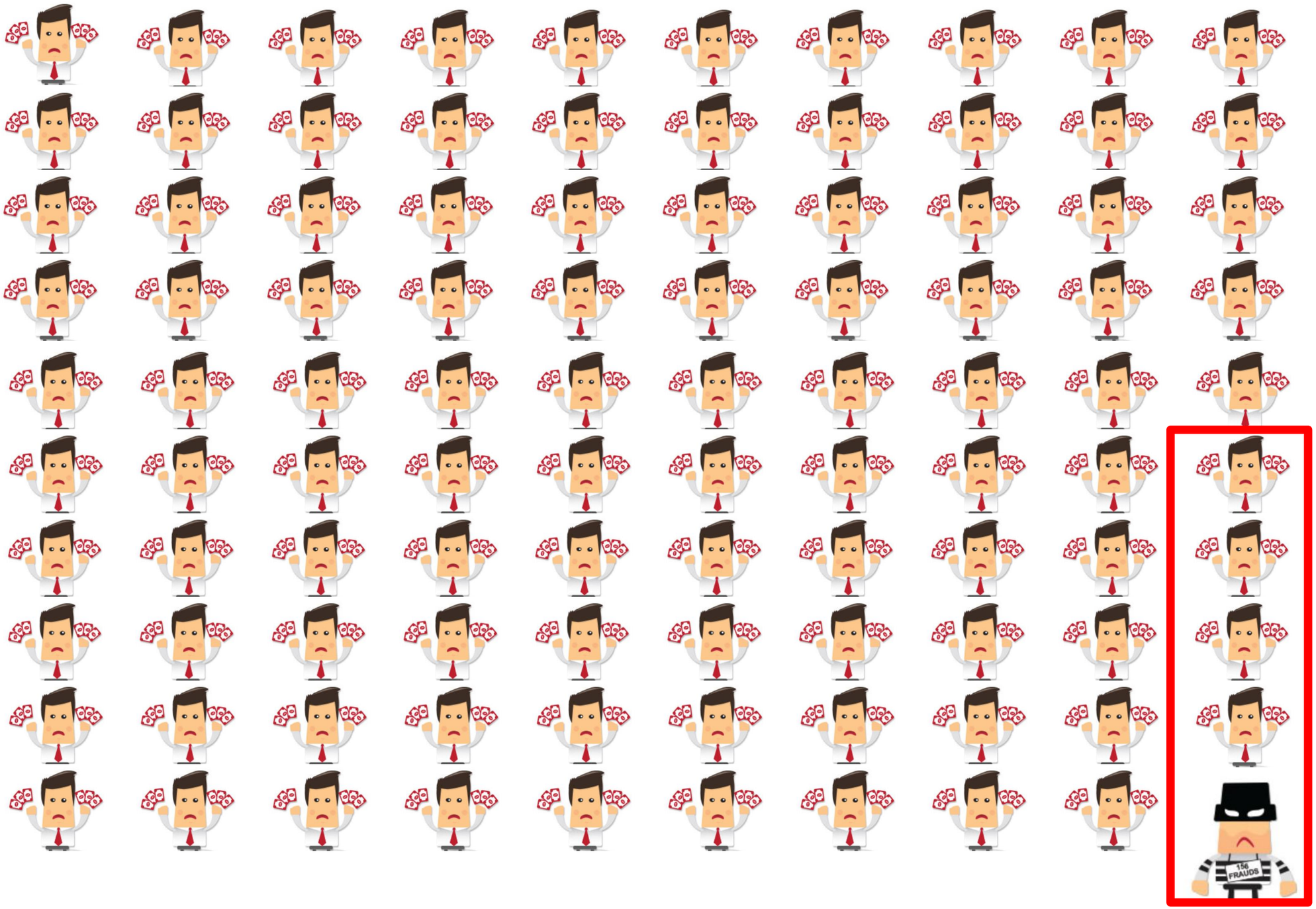
**20%**




**Precision**

**20%**

**Recall**

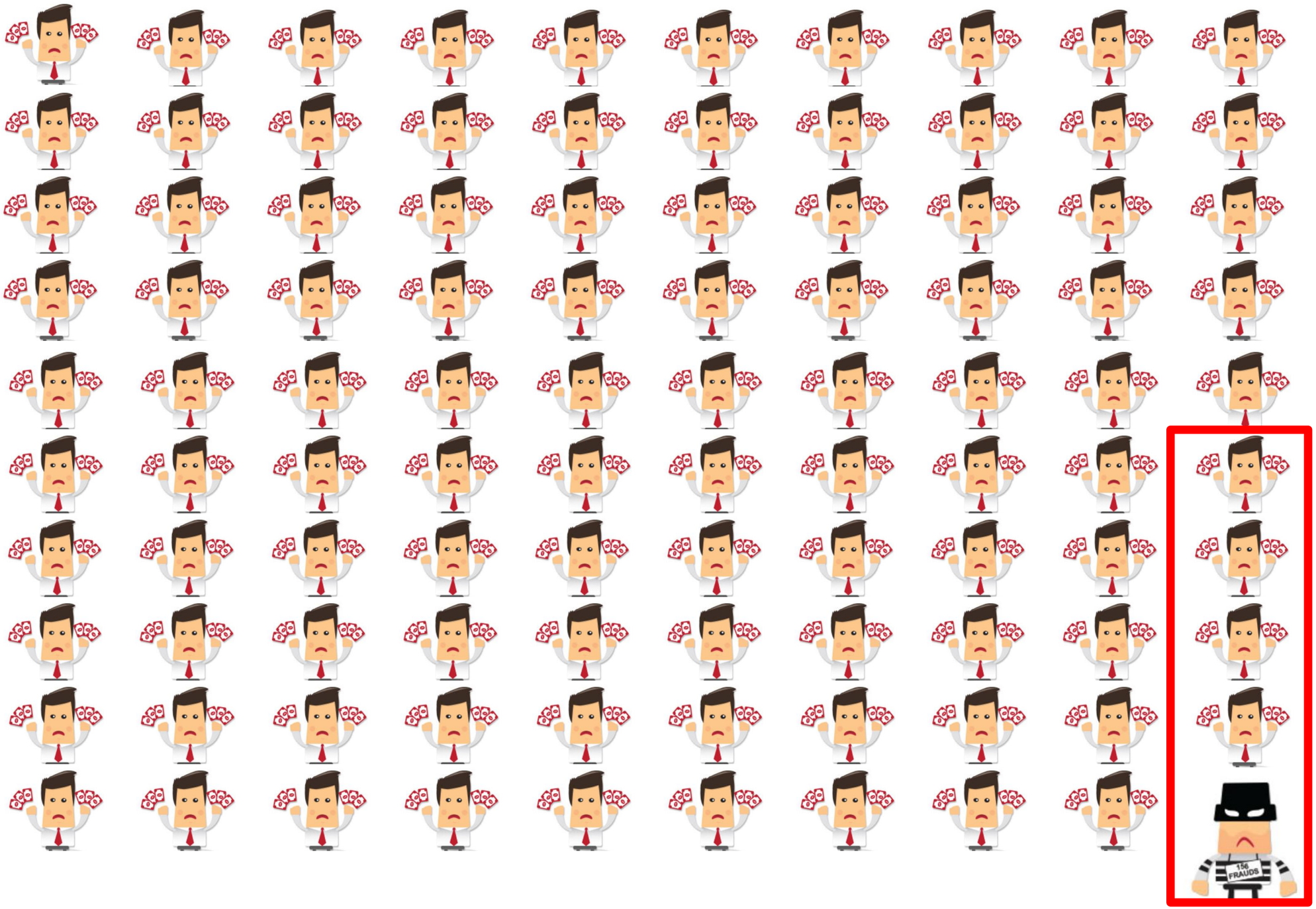



**Precision**

**20%**

**Recall**

**100%**




**Precision**

**20%**

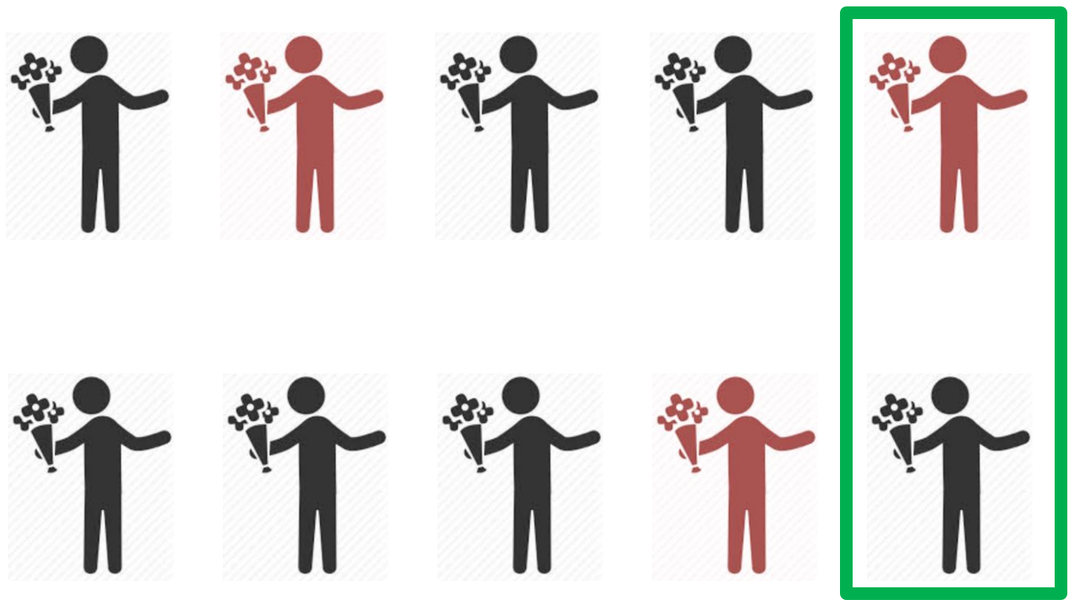
**Recall**

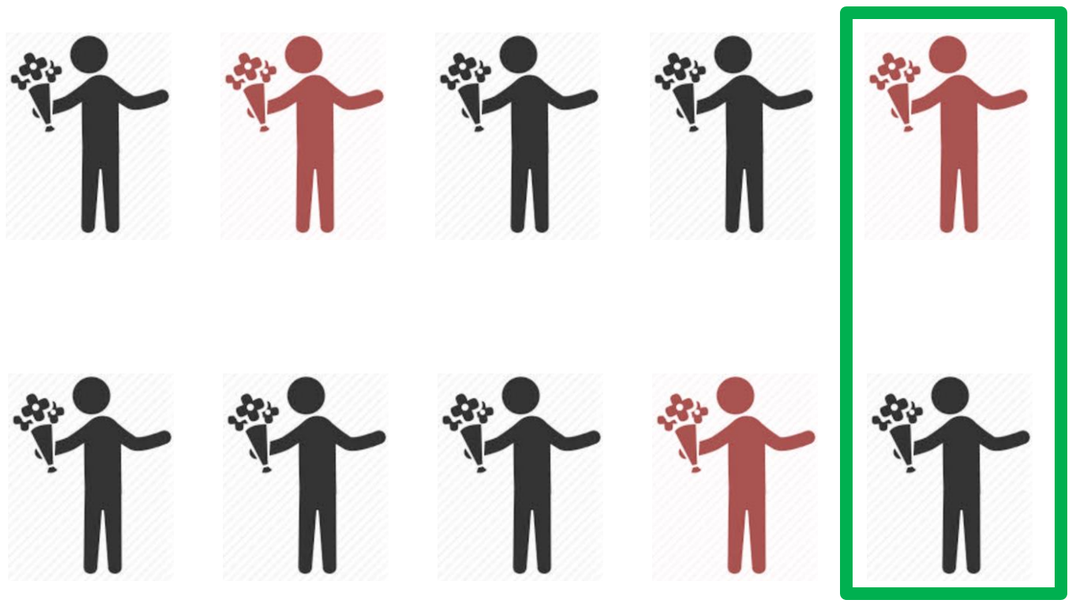
**100%**

# Online Dating

# Online Dating

- **1 in 5 relationships (20%) begins online**
- **17% of marriages in the U.S. begin online**
- **The most popular online dating site is match.com with 23.5 million users**
- **eHarmony  is responsible for 4% of all marriages in the U.S.**

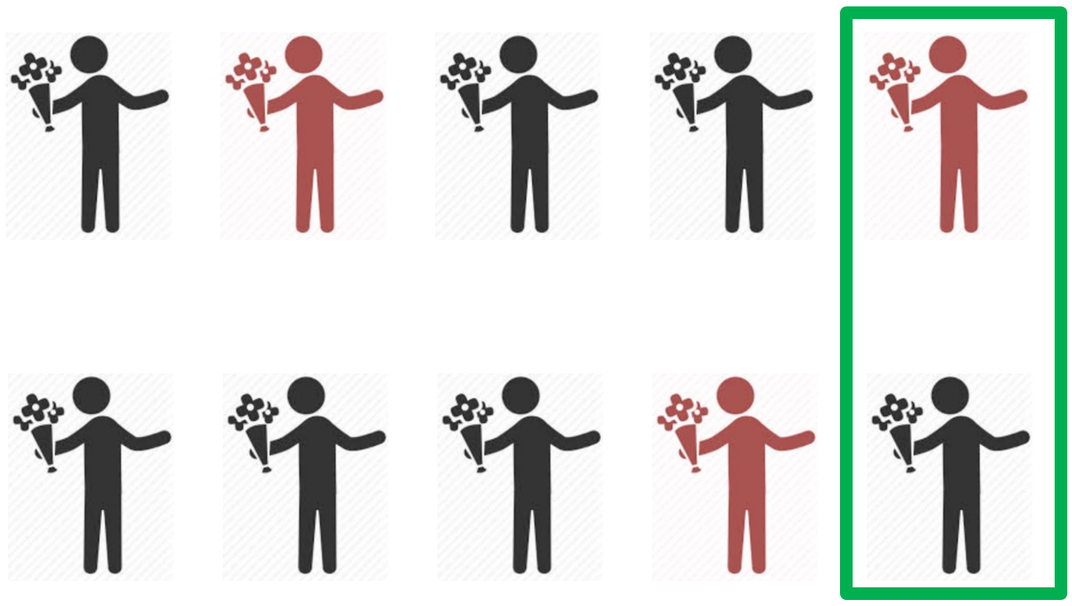




**Precision**

**Precision**

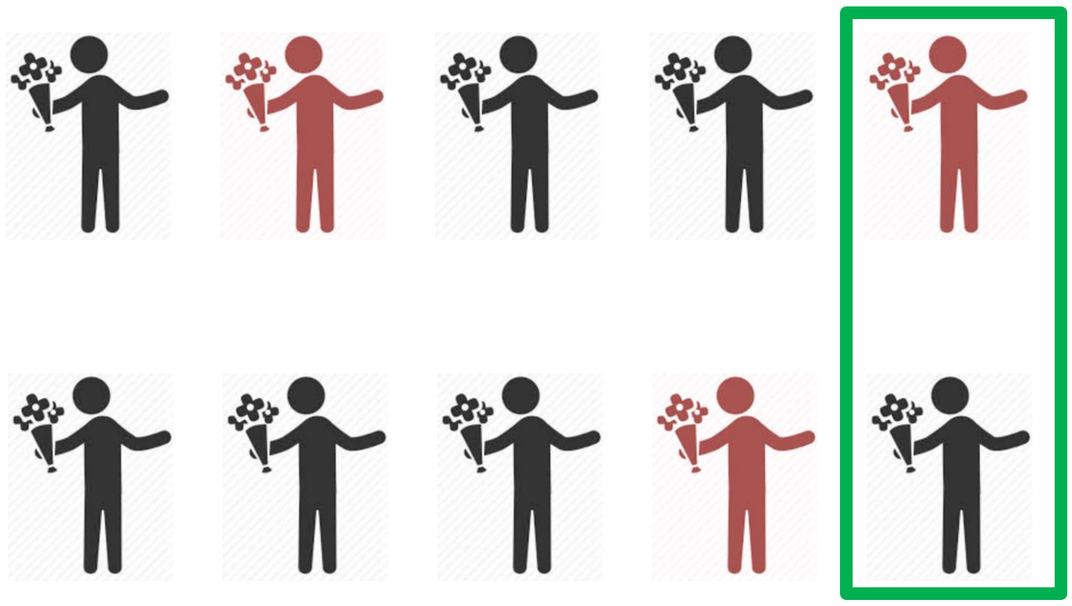
**50%**



**Precision**

**50%**

**Recall**

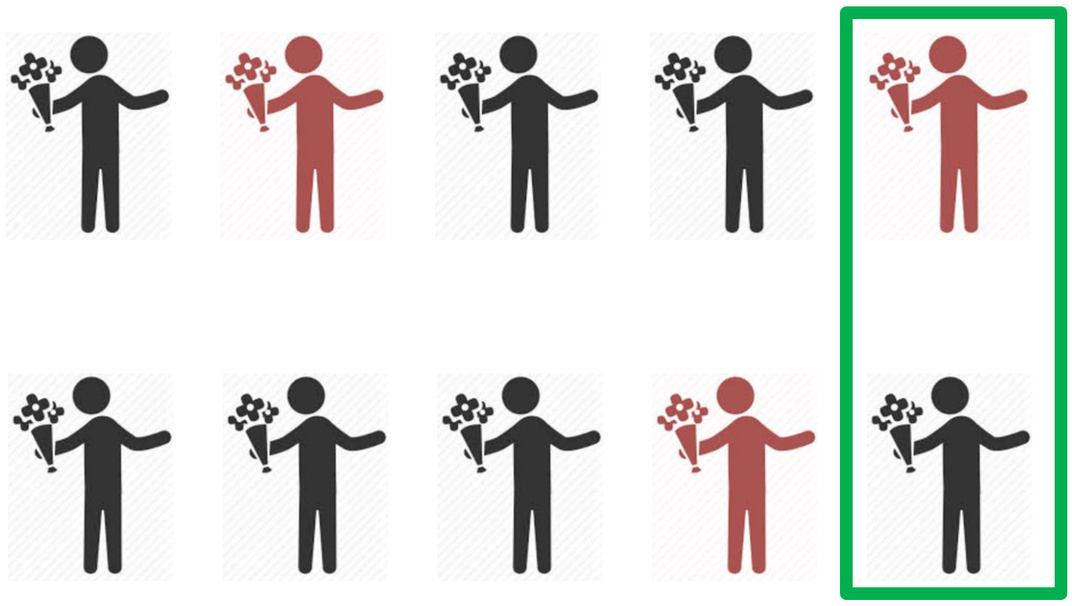


**Precision**

**50%**

**Recall**

**33%**



**Precision**

**50%**

**Recall**

**33%**