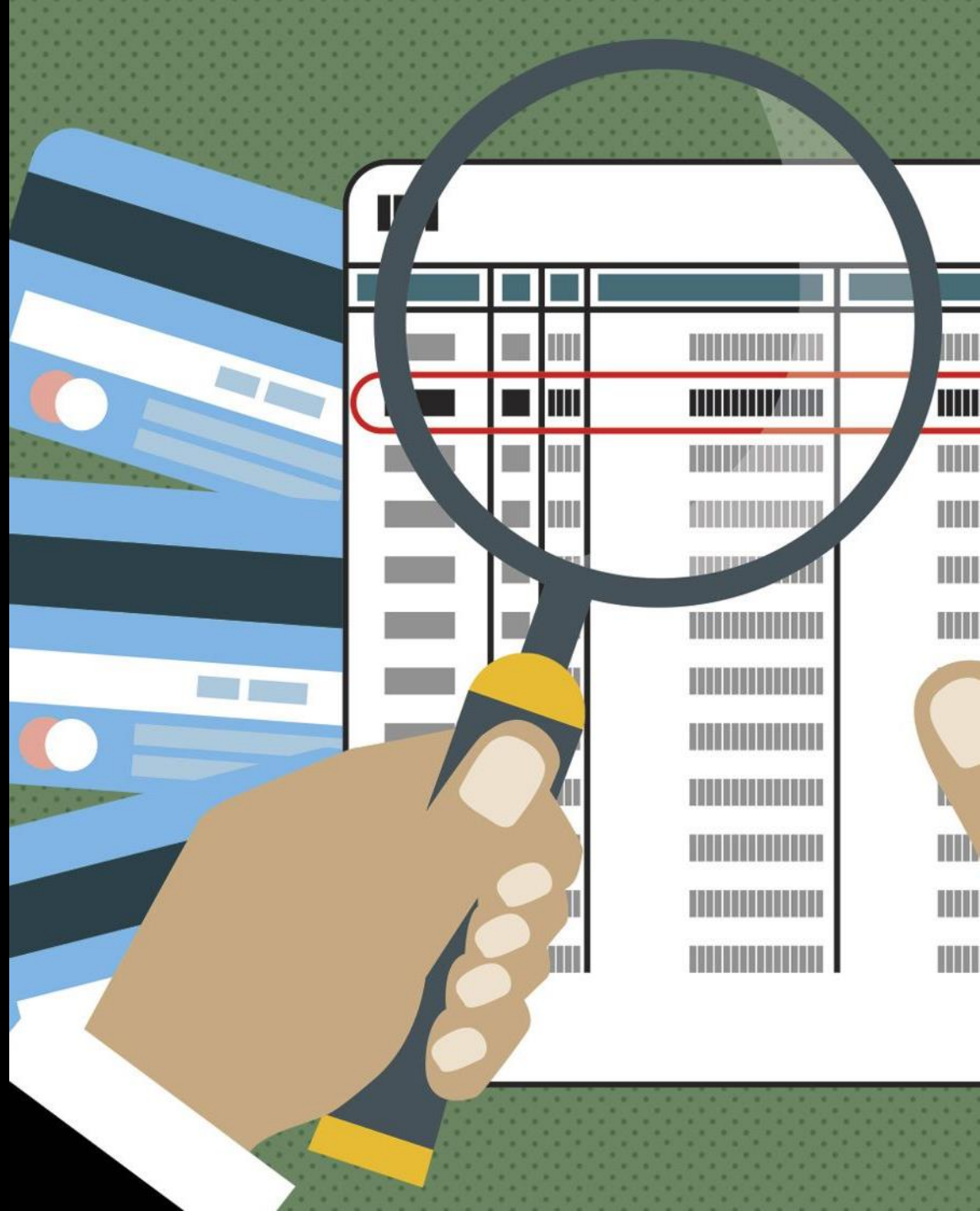
Credit Card Fraud Detection

Mai Tran

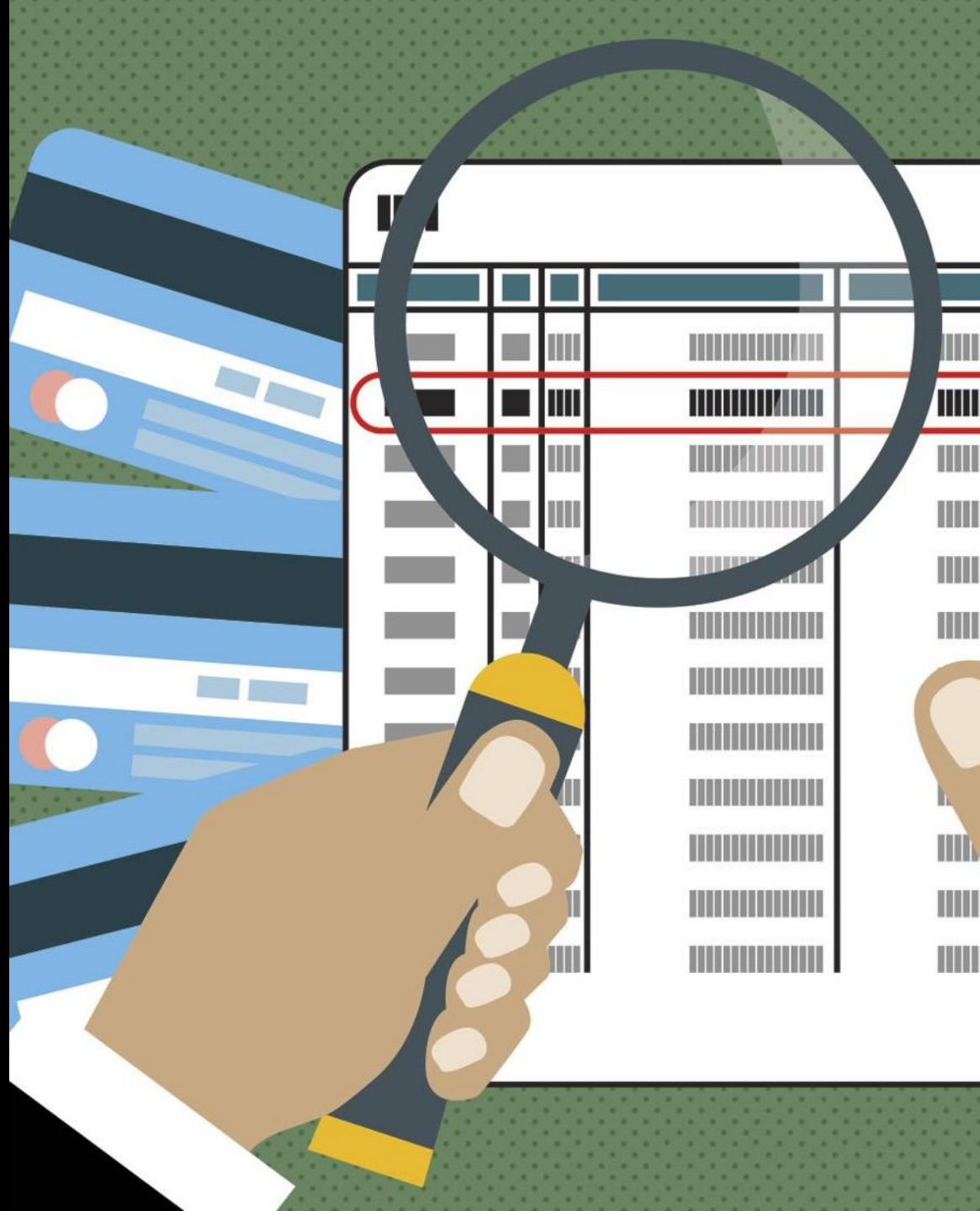
Metis

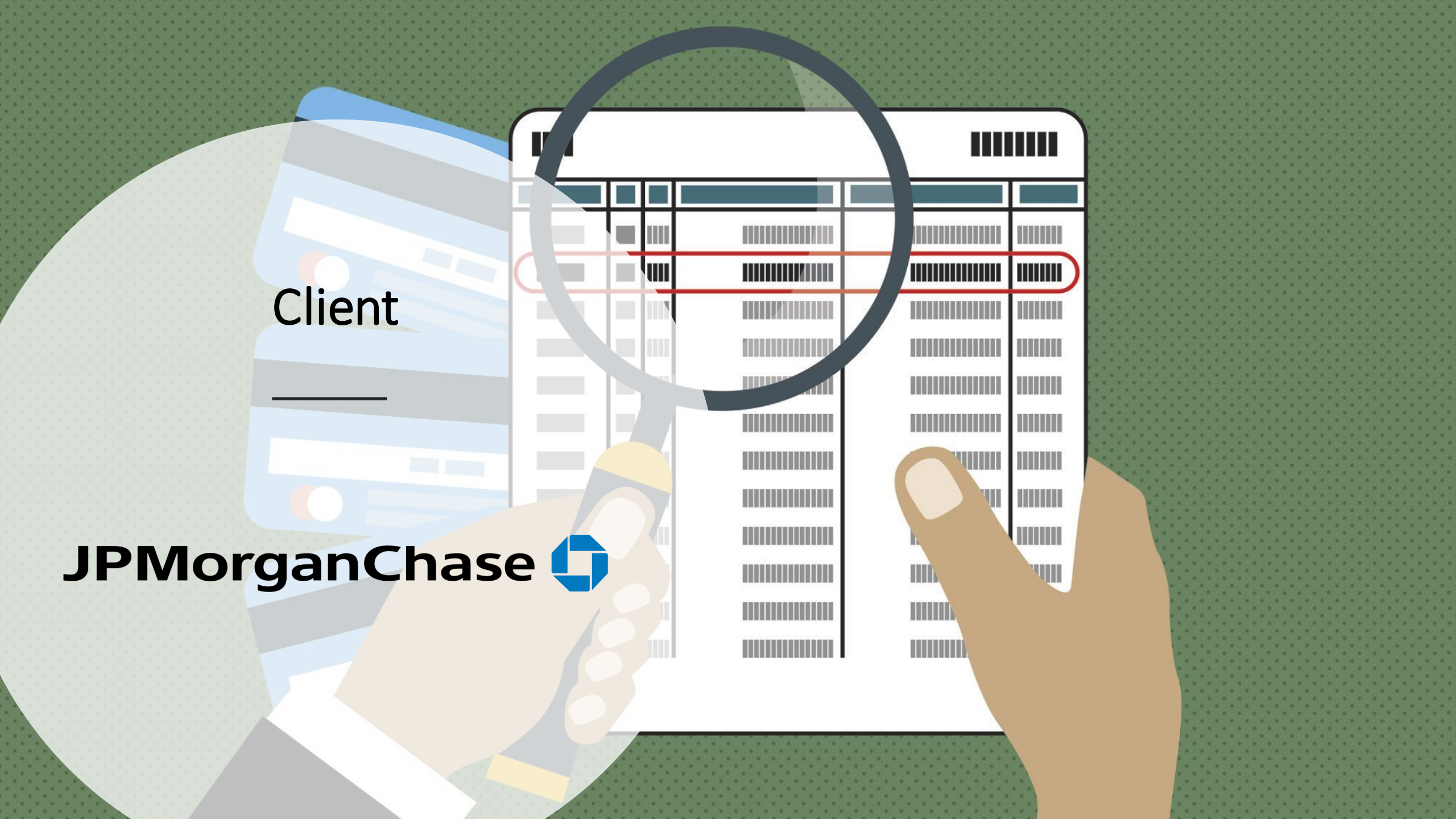
3/23/2022



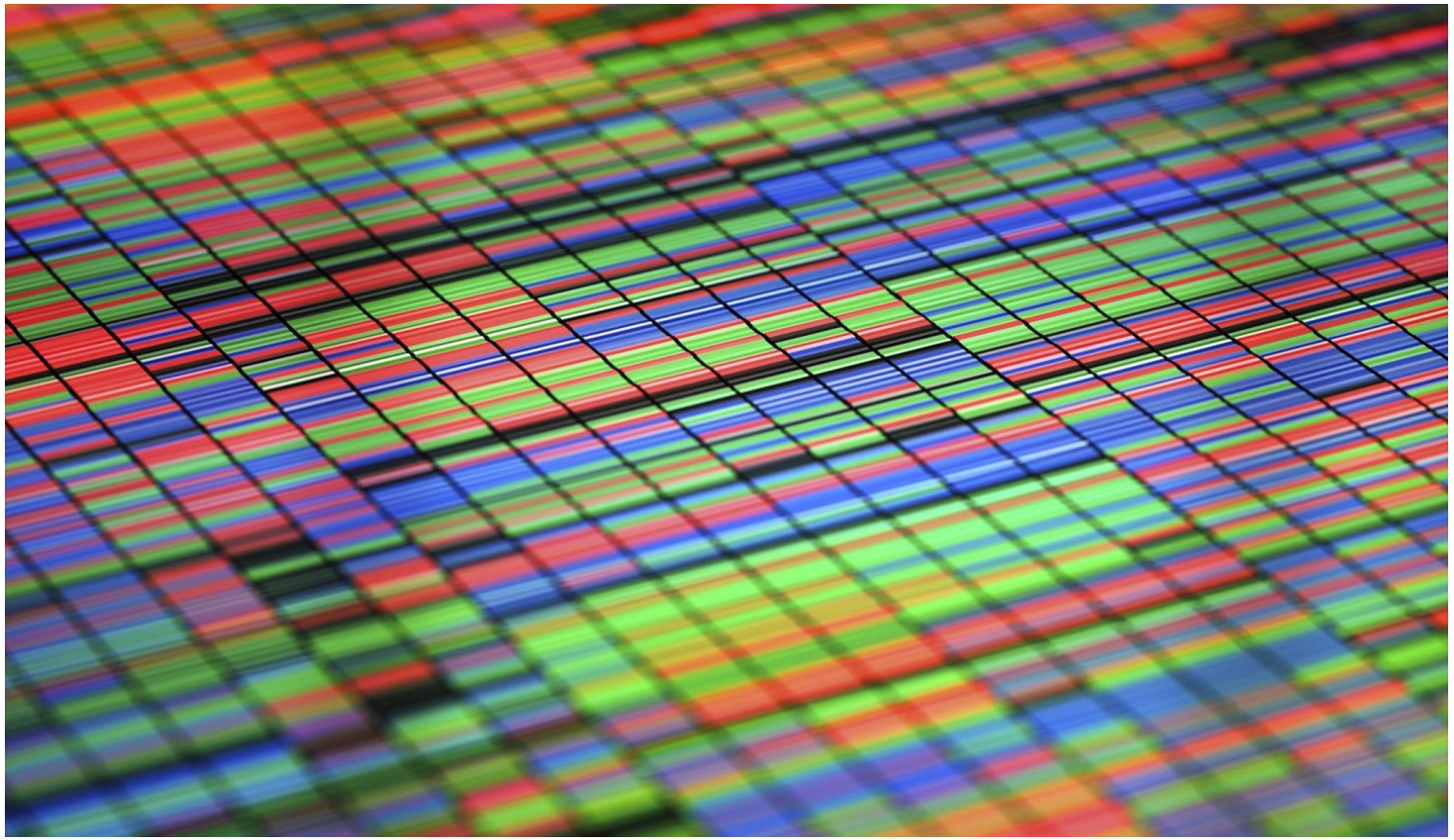
161%

of credit card fraud occurrences have increased between 2015 – 2020, according to Atlas VPN.




Client
JPMorganChase

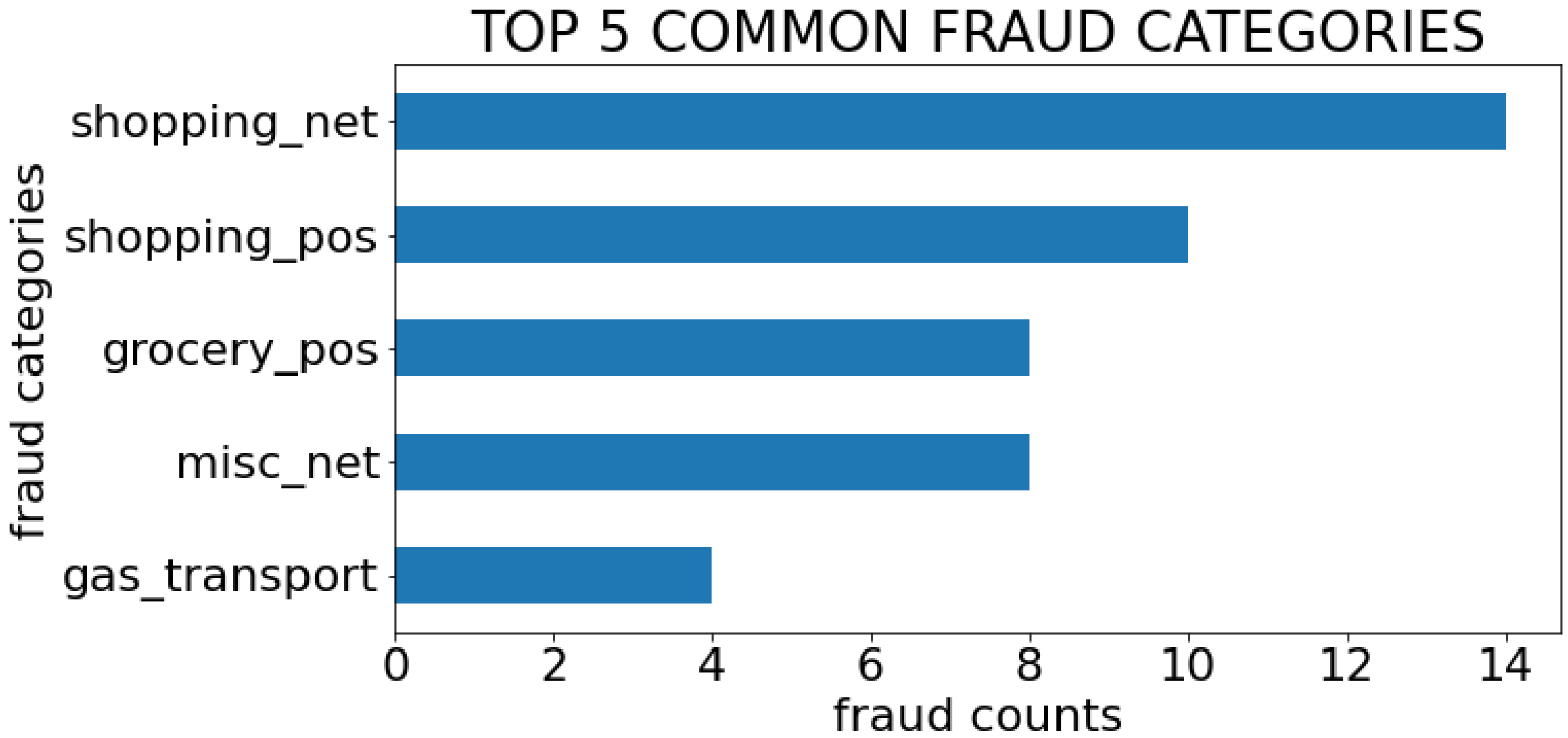
Data



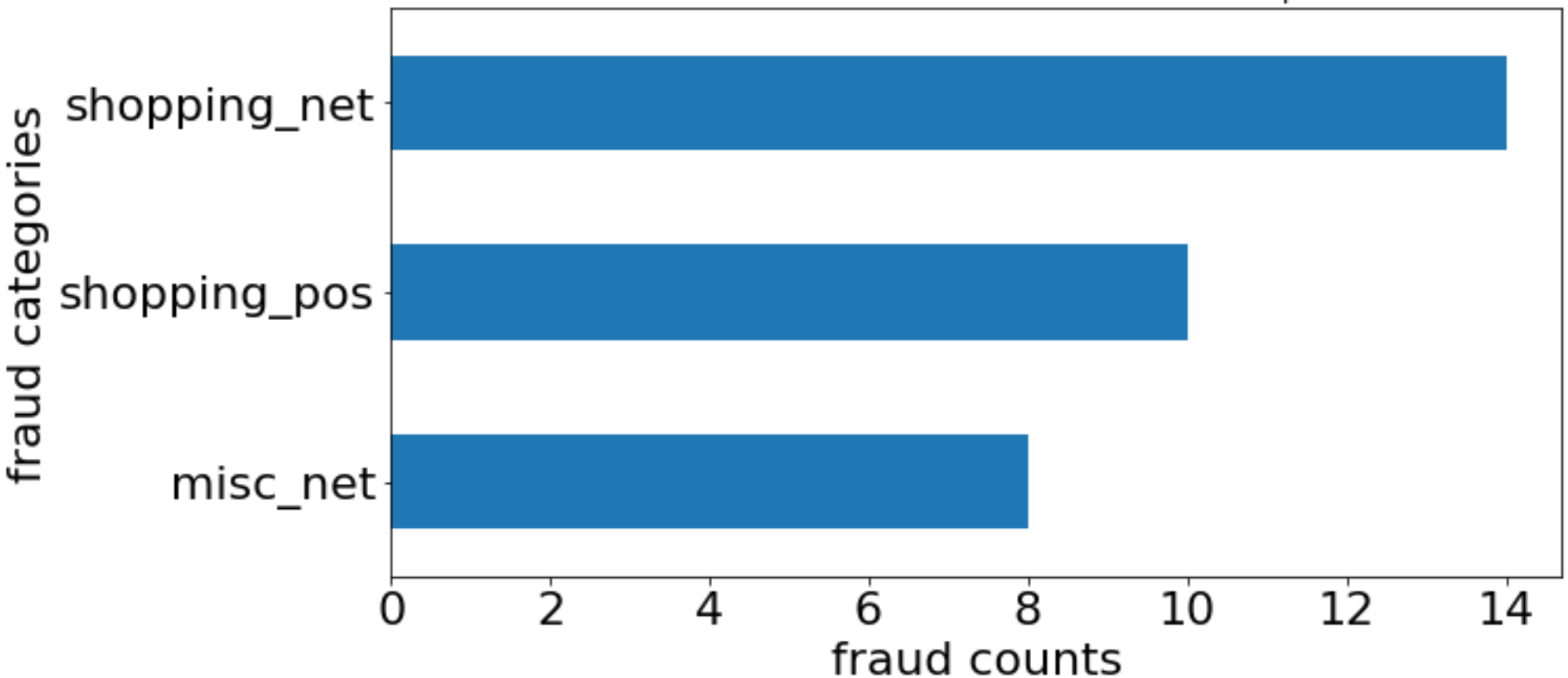
- Simulated data from Kaggle
- Highly imbalanced with 0.6% fraud rate
- 10,000 rows with 21 features

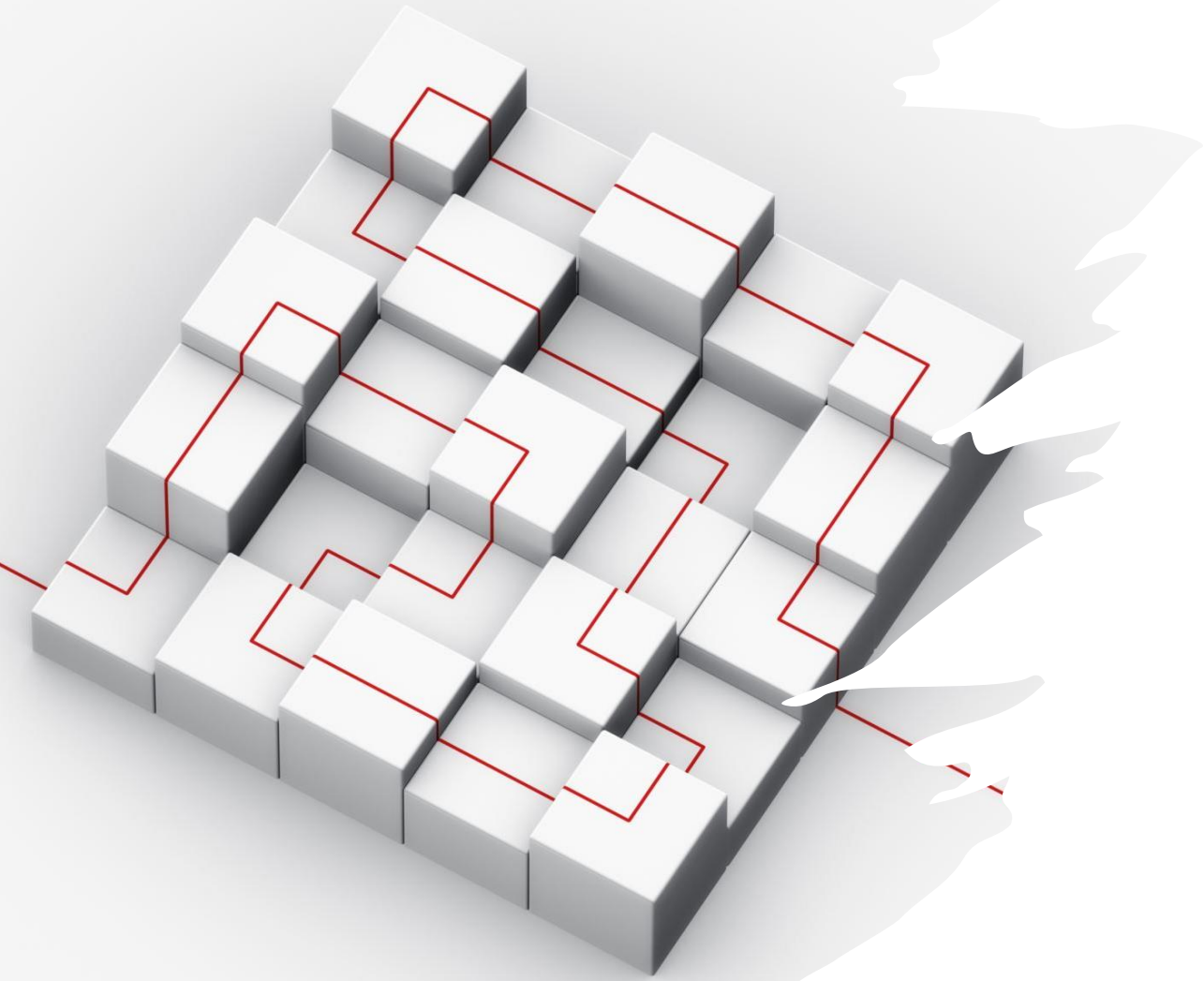
Fraudulent Transaction Statistics

Average	Median	Minimum	Maximum	Mode	Percent Fraud
$558.71	$719.50	$6.90	$1376.04	$333.28	0.6 %


TOP 5 COMMON FRAUD CATEGORIES
shopping_net
shopping_pos
grocery_pos
misc_net
gas_transport
fraud categories
0
2
4
6
8
10
12
14
fraud counts

COMMON FRAUD CATEGORIES HIGHER THAN $550 PER TRANSACTION





Project Design

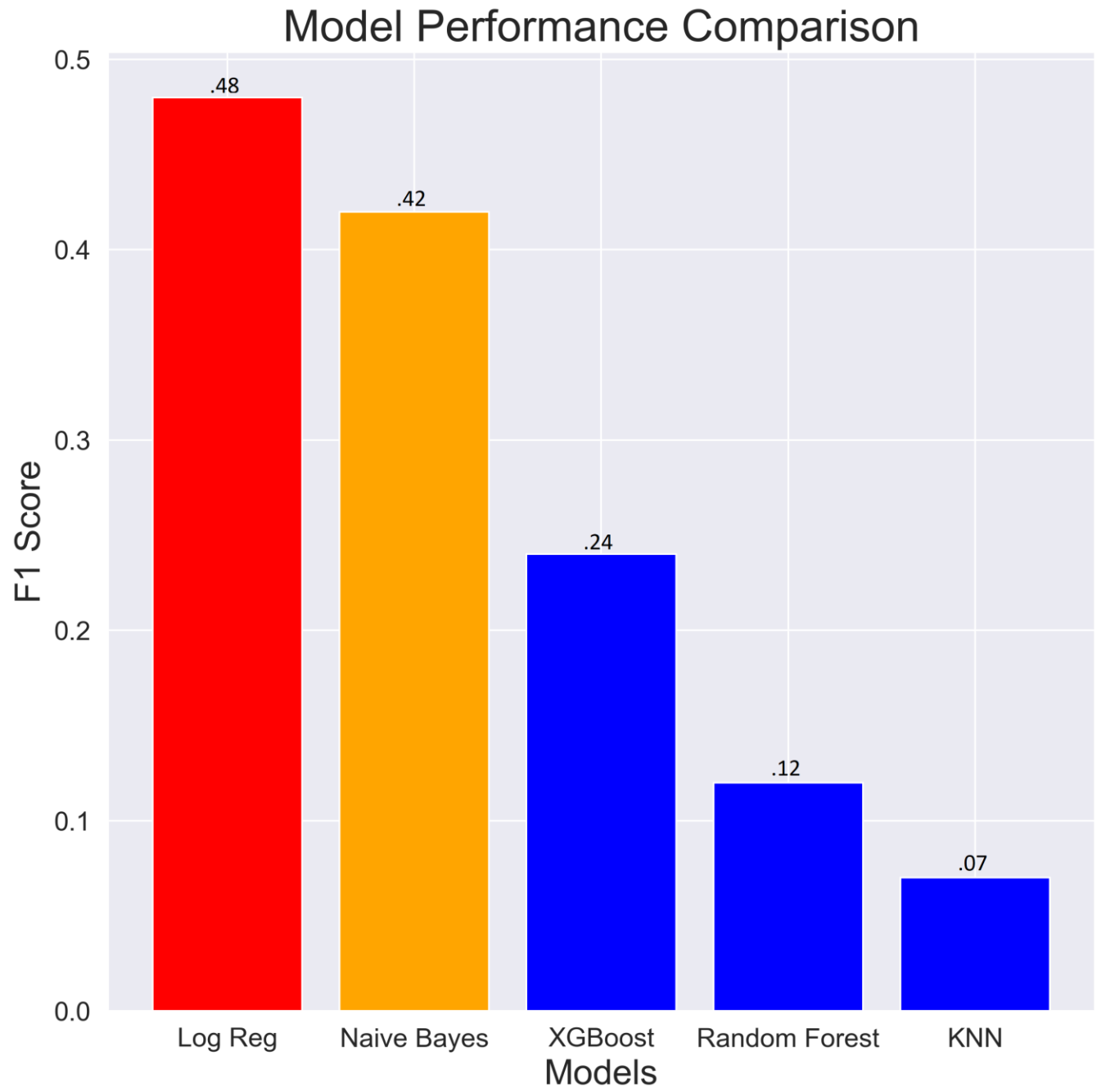
1. Metric selection
2. Baseline Model – Logistic Regression
    1. Stratified sampling
    2. Class weight balancing
    3. Hyperparameter tuning
3. Compare baseline with 4 other models

Metric

- F1 score was selected
- Vital financial assets – no compromise between recall and precision!
- High recall + high precision = high F1 score
- (fewer false negatives – those real malicious fraud charges that were identified as valid charges) + (fewer false positives – those falsely flagged fraud charges that were actually legitimate) = no one losing money + no one being annoyed and upset

Logistic Regression Tuning Results

Logistic Regression	F1 score
1) LR with class imbalance	0.0
2) LR with class_weight='balanced'	0.06
3) LR with best hyperparameters	0.18
4) LR with both best hyperparameters and class weights	0.47
5) LR with scaled features	0.48


Model Performance Comparison
0.5
0.4
0.3
0.2
0.1
0.0
F1 Score
.48
.42
.24
.12
.07
Log Reg
Naive Bayes
XGBoost
Random Forest
KNN
Models

Future Work

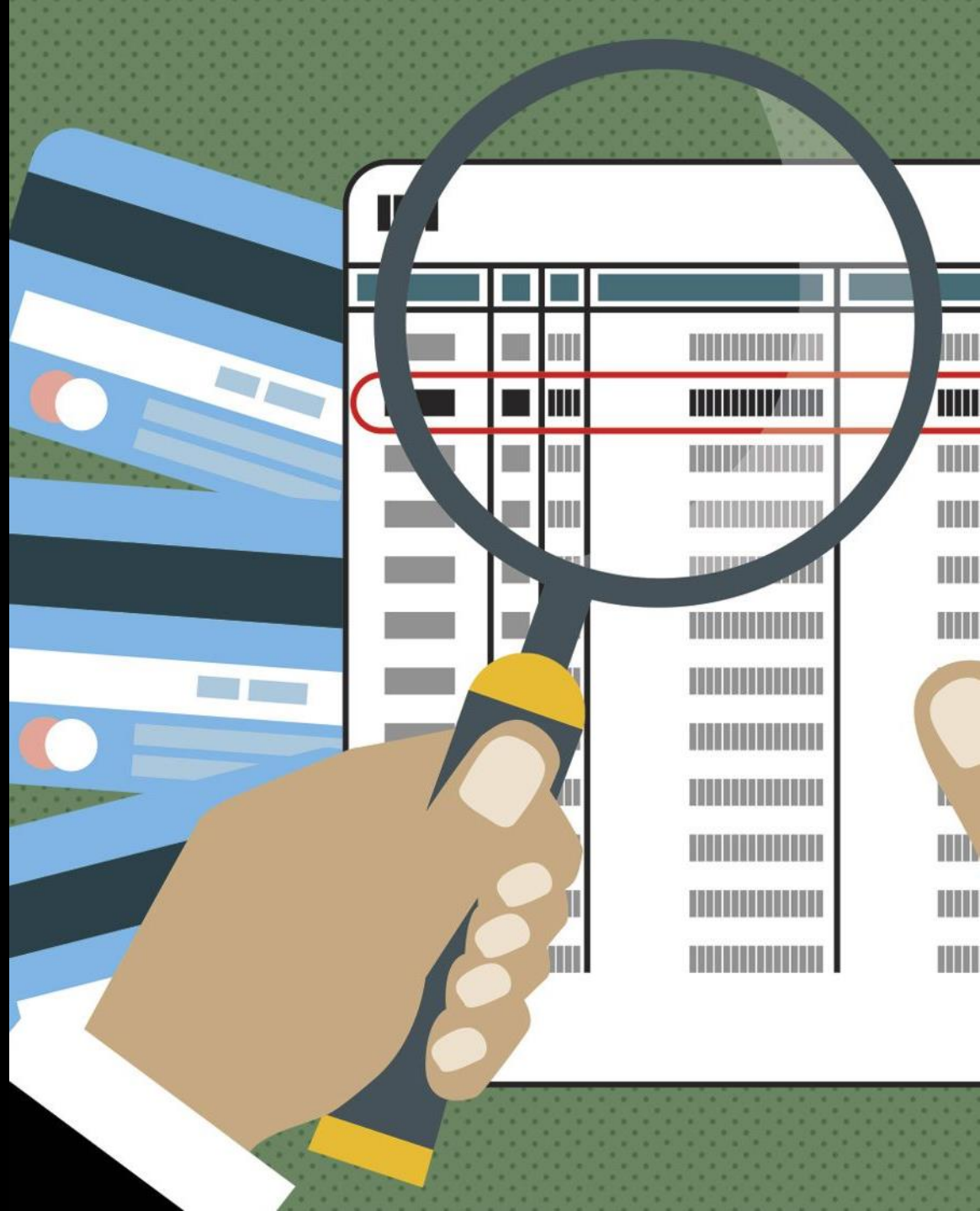
- Apply resampling techniques to reduce class imbalance: under-sampling, over-sampling, and SMOTE (synthetic minority oversampling technique)
- Integrate all 5 models into an ensemble classifier such as a voting ensemble or a stacked ensemble
- Time series analysis

Thank you!

Appendices

Biggest Incident in History

In mid-2000s, more than 32,000 credit card information was stolen resulting in total financial losses of 17 million Great British Pound (GBP) or 22 million US Dollars (USD).

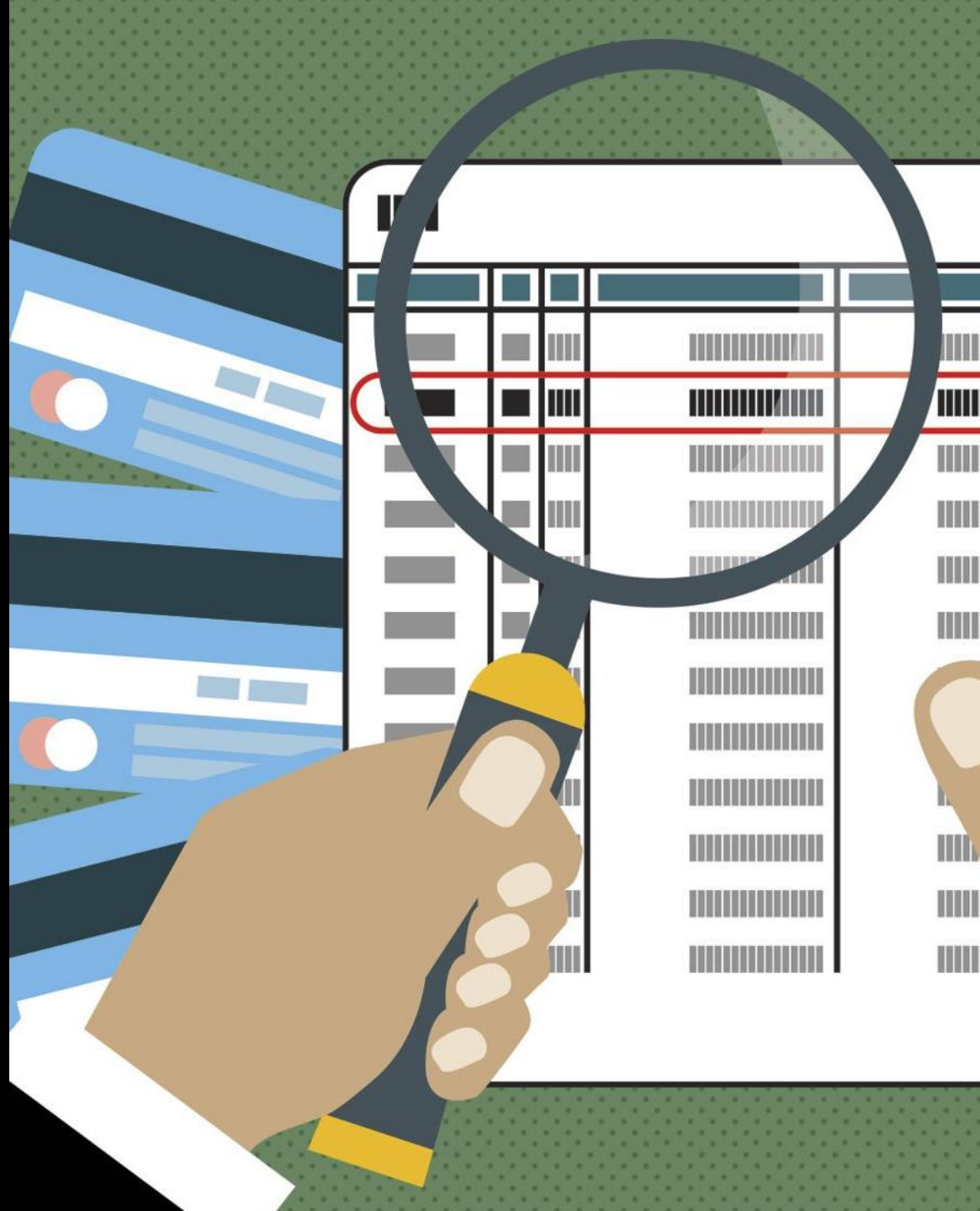


Problems

1) how to catch more seemingly legitimate transactions that are actually legitimate?

2) how to catch more seemingly fraud transactions that are actually fraud?

3) Imbalanced data -- 0.6% fraud



Cost Benefit Analysis

- Cost – Positive Number
- Benefit – Negative Number

TN (Normal Transaction):	C_{TP} = -1 x Transaction Amount x Merchant Fee
FP (Falsely Flagged Fraud):	C_{FP} = Intervention Cost – Transaction Amount x Merchant Fee+ Customer Frustration
FN (Undetected Fraud):	C_{FN} = Transaction Amount
TP (Detected Fraud):	C_{TP} = Intervention Cost

Source: RCG Global Services