Introduction to fraud detection

FRAUD DETECTION IN PYTHON

Charlotte Werger
Data Scientist

datacamp

Meet your instructor

Hi my name is Charlotte and I am a Data Scientist

What is fraud?

Examples of fraud: insurance fraud, credit card fraud, identify theft, money laundering, tax evasion, product warranty, healthcare fraud

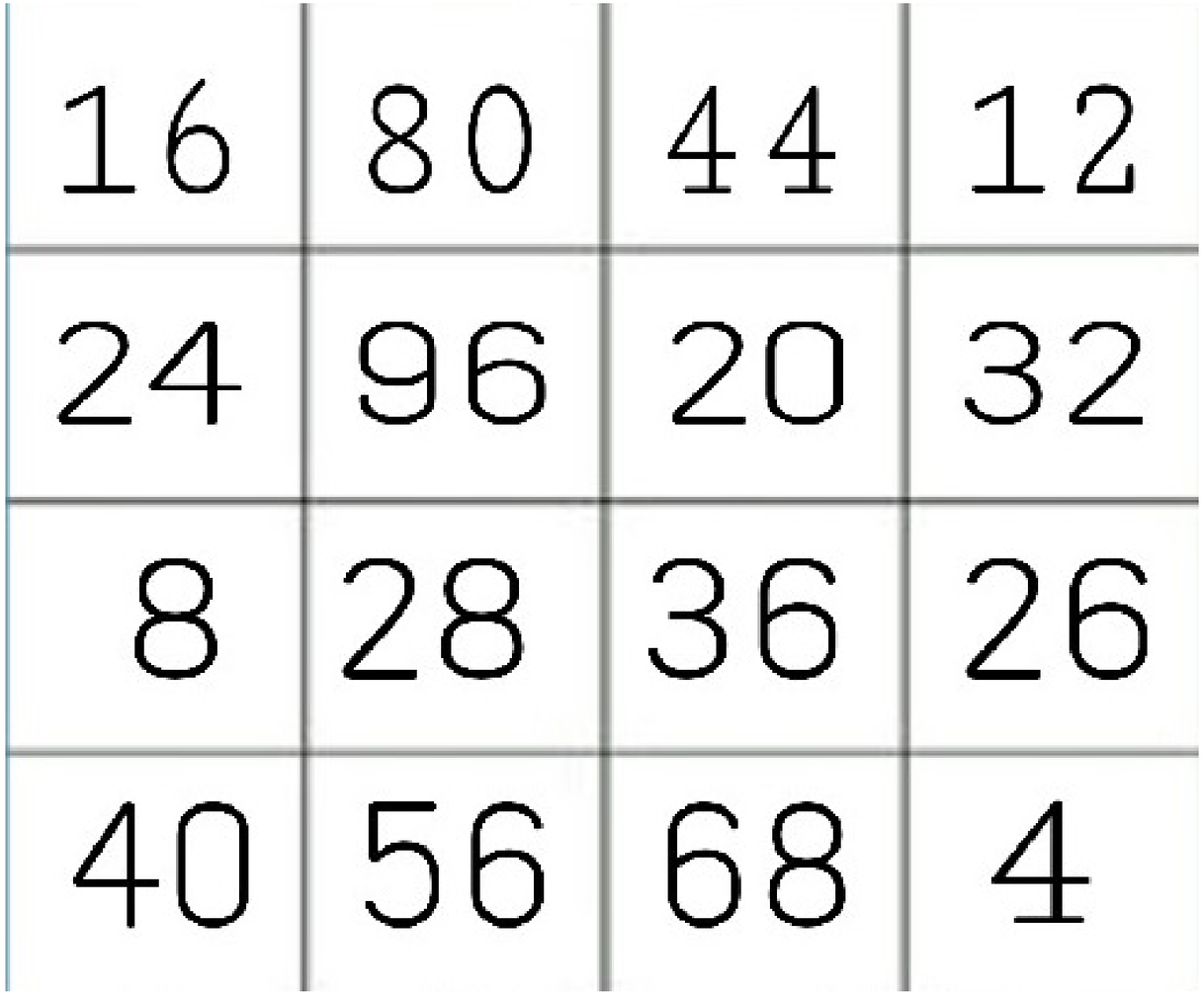
Fraud is

- uncommon
- concealed
- changing over time
- organized

Fraud detection is challenging

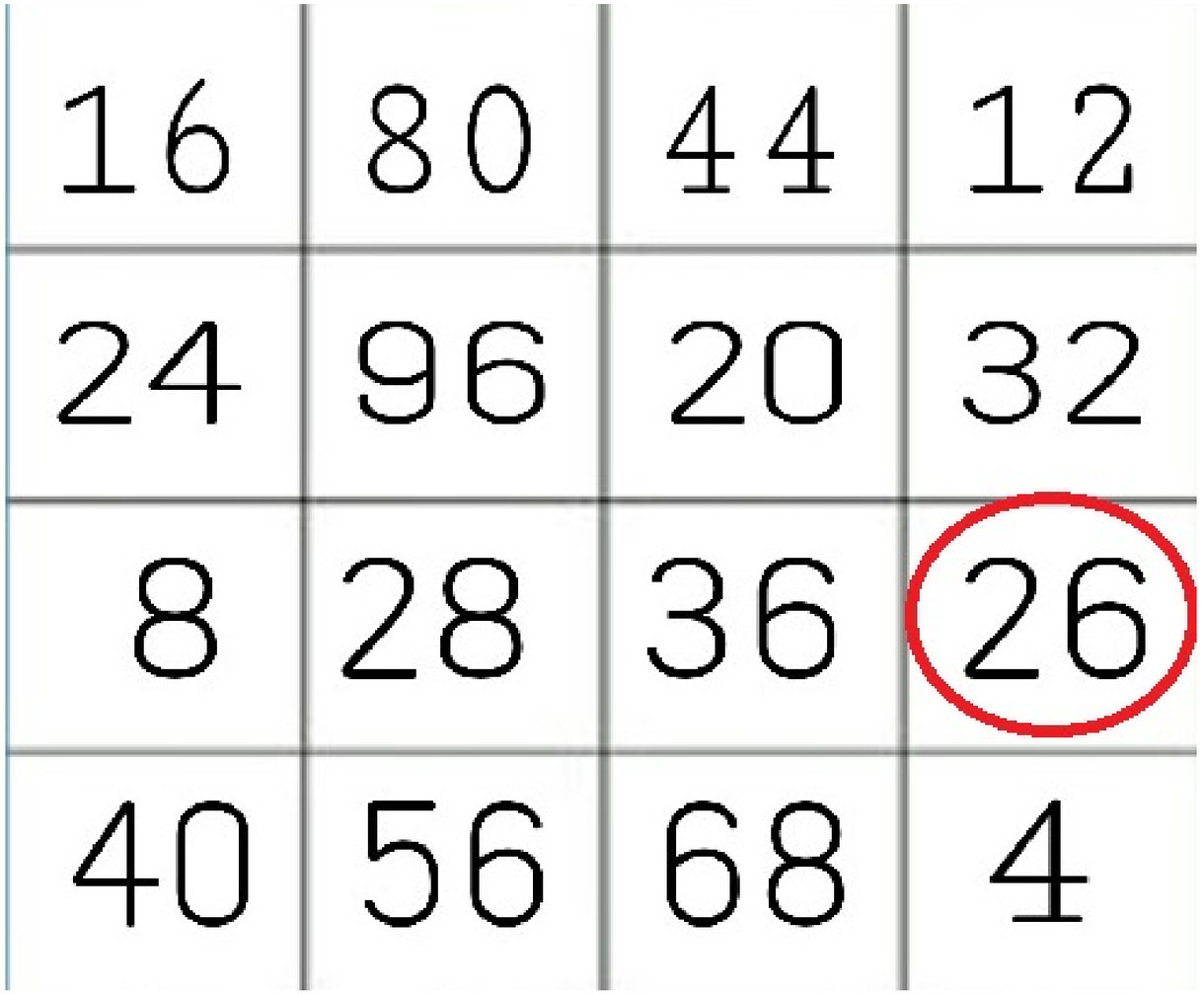
Fraud detection is challenging

Fraud detection is challenging



16	80	44	12
24	96	20	32
8	28	36	26
40	56	68	4

Fraud detection is challenging



How companies deal with fraud

Fraud analytics teams:

1. Often use rules based systems, based on manually set thresholds and experience
2. Check the news
3. Receive external lists of fraudulent accounts and names
4. Sometimes use machine learning algorithms to detect fraud or suspicious behavior

Let's have a look at some data

```
df=pd.read_csv('creditcard_data.csv')


df.head()
```

```
    V1          V2       ...    Amount    Class
0  -0.078306    0.025427 ...     1.77      0
1   0.000531    0.019911 ...     30.90     0
2   0.015375   -0.038491 ...     23.57     0
3   0.137096   -0.249694 ...     13.99     0
4  -0.014937    0.005771 ...     1.29      0
```

```
df.shape
```

```
(5050, 30)
```

Let's practice!

FRAUD DETECTION IN PYTHON

Increasing successful detections using data resampling

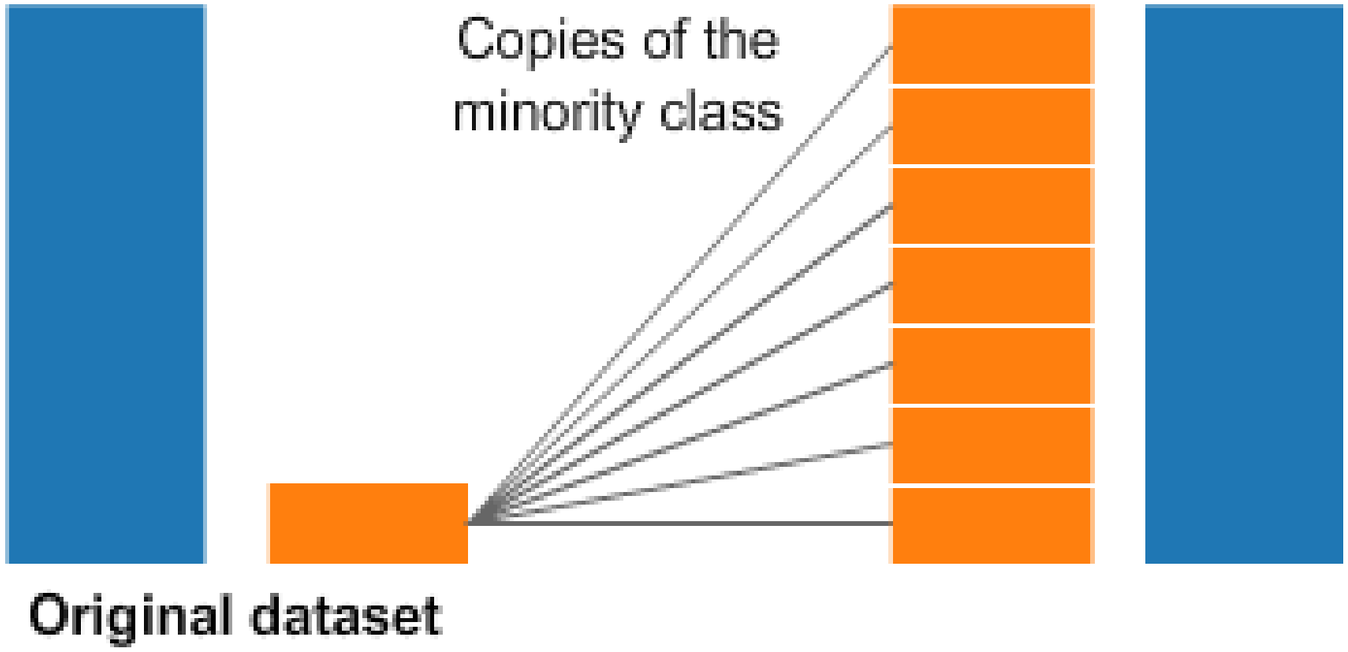
FRAUD DETECTION IN PYTHON

Charlotte Werger
Data Scientist

datacamp

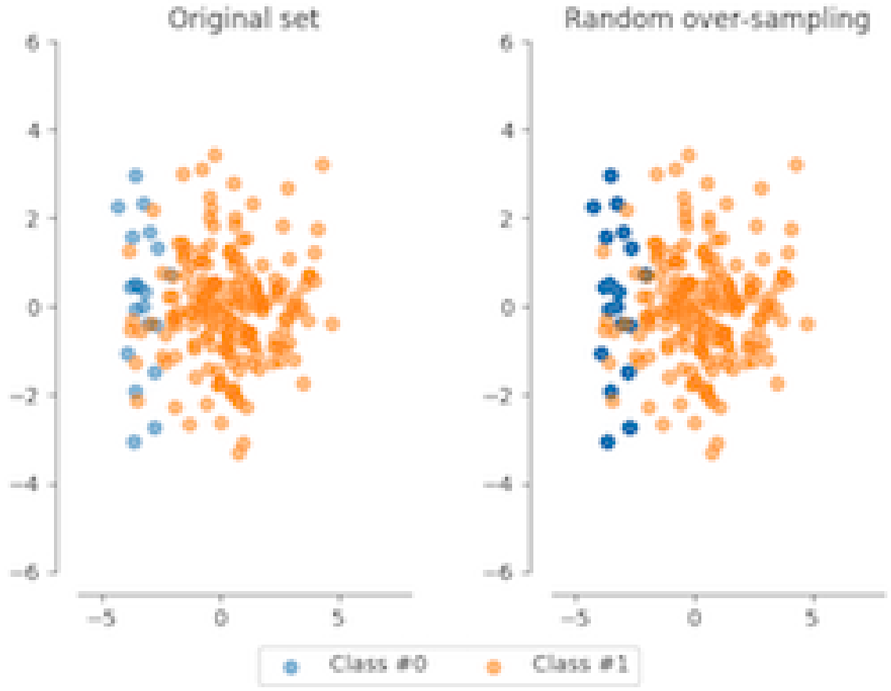
Undersampling

Oversampling

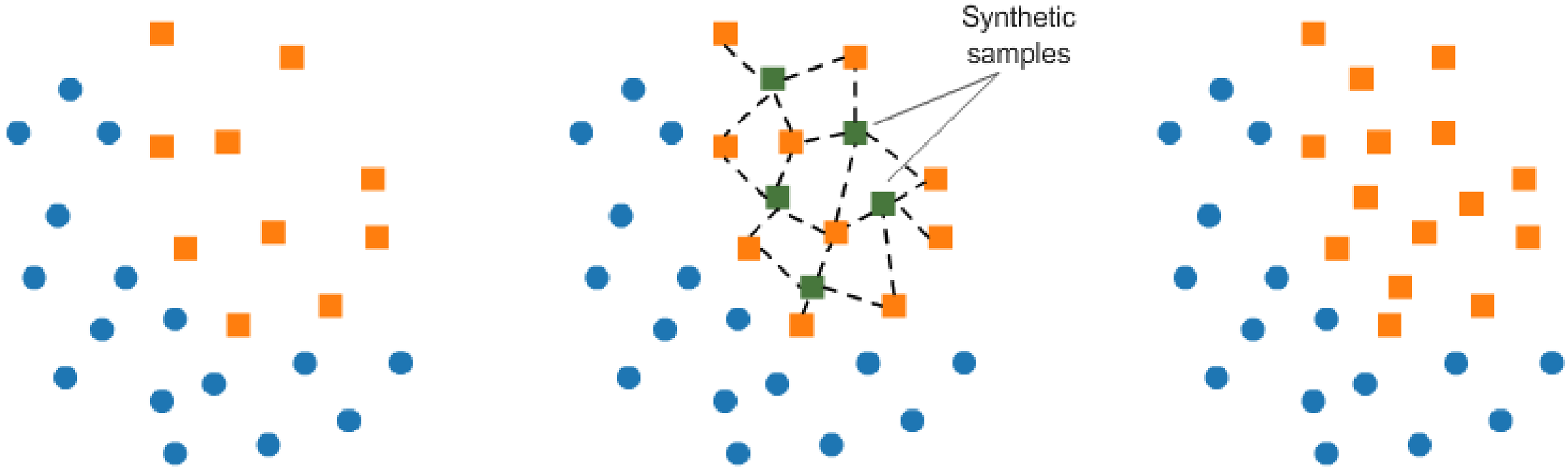


Oversampling in Python

```
from imblearn.over_sampling import RandomOverSampler
method = RandomOverSampler()
X_resampled, y_resampled = method.fit_resample(X, y)
compare_plots(X_resampled, y_resampled, X, y)
```



Synthetic Minority Oversampling Technique (SMOTE)



[1] https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Which resampling method to use?

- Random Under Sampling (RUS): throw away data, computationally efficient
- Random Over Sampling (ROS): straightforward and simple, but training your model on many duplicates
- Synthetic Minority Oversampling Technique (SMOTE): more sophisticated and realistic dataset, but you are training on "fake" data

When to use resampling methods

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
# Define resampling method and split into train and test
method = SMOTE()
X_train, X_test, y_train, y_test = train_test_split(X, y,
 train_size=0.8, random_state=0)
# Apply resampling to the training data only
X_resampled, y_resampled = method.fit_resample(X_train, y_train)
# Continue fitting the model and obtain predictions
model = LogisticRegression()
model.fit(X_resampled, y_resampled)
# Get your performance metrics
predicted = model.predict(X_test)
print (classification_report(y_test, predicted))
```

Let's practice!

FRAUD DETECTION IN PYTHON

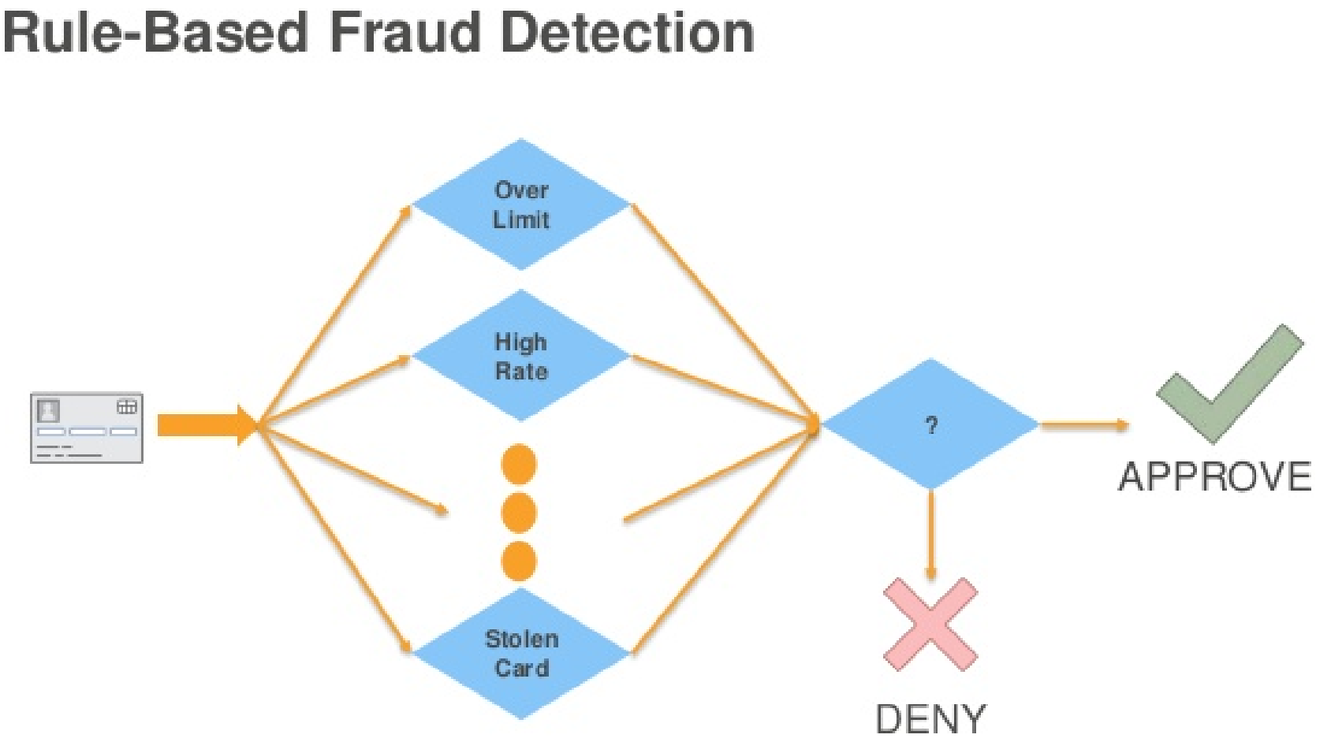
Fraud detection algorithms in action

FRAUD DETECTION IN PYTHON

Charlotte Werger
Data Scientist

datacamp

Traditional fraud detection with rules based systems



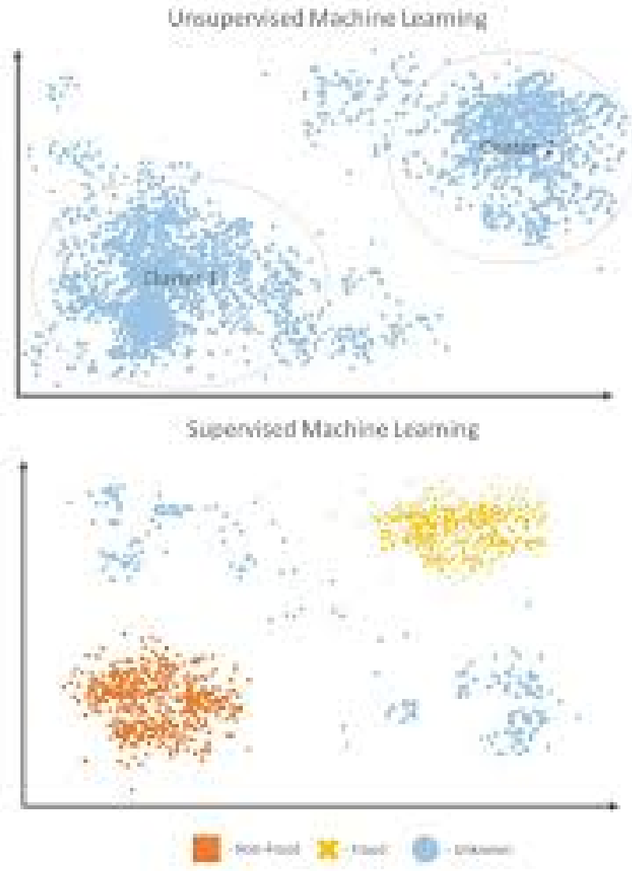
Drawbacks of using rules based systems

Rules based systems have their limitations:

1. Fixed thresholds per rule to determine fraud
2. Limited to yes/no outcomes
3. Fail to capture interaction between features

Why use machine learning for fraud detection?

1. Machine learning models adapt to the data, and thus can change over time
2. Uses all the data combined rather than a threshold per feature
3. Can give a score, rather than a yes/no
4. Will typically have a better performance and can be combined with rules



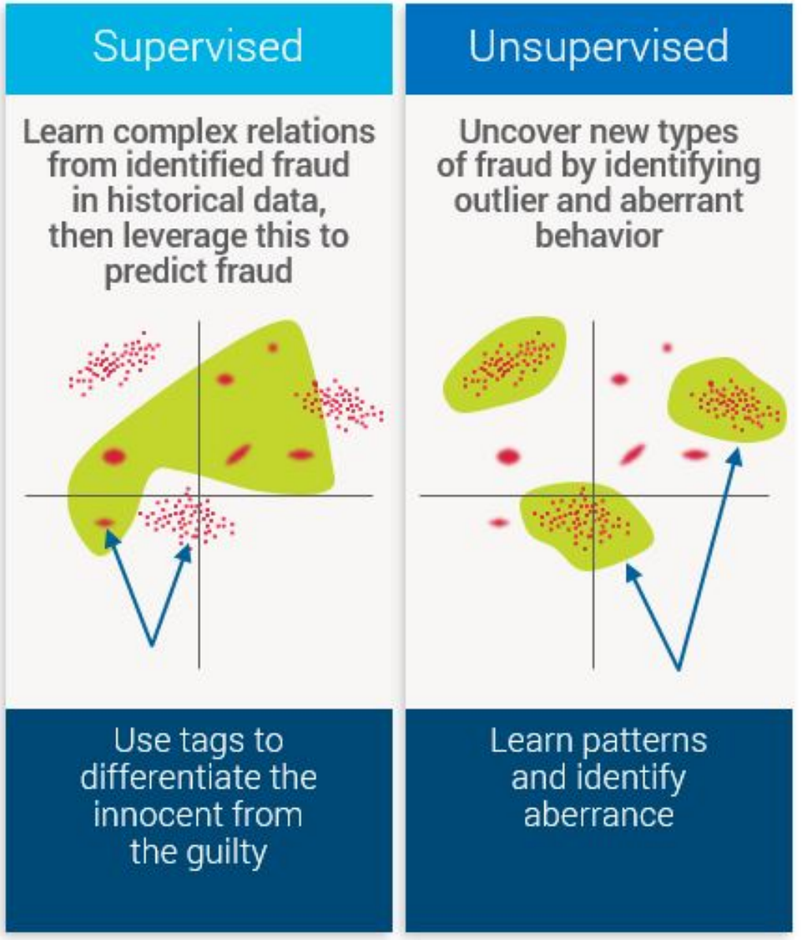
Refresher on machine learning models

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn import metrics
# Step 1: split your features and labels into train and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
# Step 2: Define which model you want to use
model = LinearRegression()
# Step 3: Fit the model to your training data
model.fit(X_train, y_train)
# Step 4: Obtain model predictions from your test data
y_predicted = model.predict(X_test)
# Step 5: Compare y_test to predictions and obtain performance metrics
print (metrics.r2_score(y_test, y_predicted))
```

```
0.821206237313
```

What you'll be doing in the upcoming chapters

- Chapter 2. Supervised learning: train a model using existing fraud labels
- Chapter 3. Unsupervised learning: use your data to determine what is 'suspicious' behavior without labels
- Chapter 4. Fraud detection using text data: Learn how to augment your fraud detection models with text mining and topic modeling



Source: FICO Blog

Let's practice!

FRAUD DETECTION IN PYTHON