

**Analyzing Unstructured Data – Final Report**



GROUP NAME BCG

TEAM 4

Yuvraj Gupta, Shubham Gupta, Vijaya Rani, Juilee Bhosale, Deepika Jindal

# **Introduction**

With the penetration of the internet and increasing popularity of online platforms, online advertising has grown to become one of the most important forms of advertising. The global online advertising market grew 21% to $87 billion in 2017 and is expected to grow to $117 billion by 2021. Online classified ad websites such as Craigslist ([www.craigslist.org](http://www.craigslist.org)), Backpage ([www.backpage.com](http://www.backpage.com)), Oodle ([www.oodle.com](http://www.oodle.com)), and eBay Classifieds ([www.ebayclassifieds.com](http://www.ebayclassifieds.com)) provide a convenient & popular way to sell goods or services and their popularity is continuing to increase. The World Wide Web provides a convenient and easily accessible medium for users to list and browse advertisements when compared to more traditional media such as newspapers and printed booklets. The widespread accessibility of the web has an unwanted effect of attracting online scammers who pose as genuine sellers by posting fake advertisements in an effort to defraud would be buyers. Scammers have the ability to steal millions of dollars from unsuspecting users and threaten the reputation and utility of online ad services.

## **What is Craigslist?**

Craigslist is a massive international classifieds website. It’s very much like the classifieds section in a newspaper except it’s free, you can use photos and it reaches an audience of millions of people in minutes. The site has a range of broad categories, and each category has several subcategories, making it easy to navigate to the appropriate section quickly. The website currently provides service across 450 cities. It also includes discussion forums and a local events calendar.

One can use Craigslist to buy and sell merchandise, look for a job, look for things happening in your area, find an apartment, open a discussion, etc. Craigslist was started by a software engineer named Craig Newmark in 1993, as a way to let people know about events happening in and around the San Francisco Bay Area. To his surprise people quickly began using the service for things other than events, like posting jobs and selling things. Through word of mouth the number of subscribers to the site exploded and in 1999 it expanded to Boston. Later that same year Craigslist incorporated, simultaneously expanding into nine major U.S cities. The service continued its amazing growth and in 2003 began to go global. Today Craigslist receives over nine billion page-views per month and it is the ninth most visited website in the United States. Newmark remains active as an employee and major shareholder; his primary focus today is keeping Craigslist free of spammers and scammers.

## **Why Housing Rental?**

Whether you’re a college student or a retiree, there are many advantages to renting a home rather than purchasing one, and in recent years the United States rental market has seen an increase in demand for apartment and house rentals. In 2016, about 27 percent of Americans [rented their home](https://www.statista.com/statistics/561961/homeownership-usa/). Although renting is seen as an affordable alternative to purchasing a home, finding a suitable rental has become a major challenge for many people living and working in major cities.

Housing Rental is an in-demand category in Craigslist where one can find various classified advertisements basis your needs & preferences . Craigslist has become the single largest information exchange about rental housing market in the USA. With an average American moving 11.4 times in their lives, millions of new listings are added monthly and this sector is growing year on year

## 

## **Problem Statement**

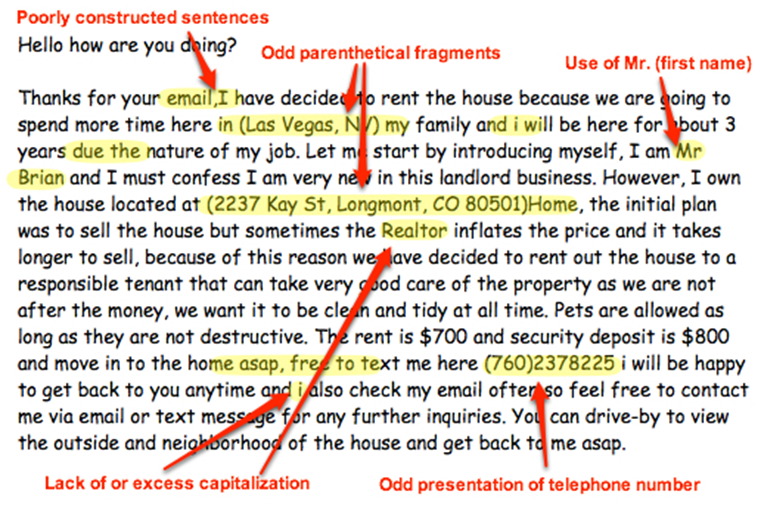
Millions of ads are posted every single day on Craigslist worldwide to a large extent anonymously. Millions of housing listings are posted in a month. It is difficult to check the listings for the people looking for a new house. As per trends 6% of housing ads are spams. However, they can’t run all around the world policing and prosecuting people. This project intends to solve the house hunt problem by sending the updates of new listings as per the selection criteria of the user by filtering spam in housing listings.

Classified ad sites routinely process hundreds of thousands to millions of posted ads, and only a small percentage of those may be fraudulent. Online scammers often go through a great amount of effort to make their listings look legitimate. Examples include copying existing advertisements from other services, tunneling through local proxies, and even paying for extra services using stolen account information. Given below is on such listing that appears more like a work of art.

This project would try to provide value to both its client(Craigslist) & its users by solving some of the key issues. High volumes of rental scams damages the reputation of Craigslist & increases its user drop-off rates. Users have to spend hours finding legitimate ads and it takes a lot of time & resources to select a genuine listing from thousands of existing listings.

This project consists of applying data analysis and text analysis concepts & techniques towards the detection of online, classified fraud for housing ad listings and building an automated notification system to send new listings as per user’s search keywords. Traditional data mining is used to extract relevant attributes from an online classified advertisements database and machine learning algorithms are applied to discover patterns and relationships of fraudulent activity. With our proposed approach, we will demonstrate the effectiveness of applying data mining techniques towards the detection of fraud in online classified advertisements for housing ads in major cities.

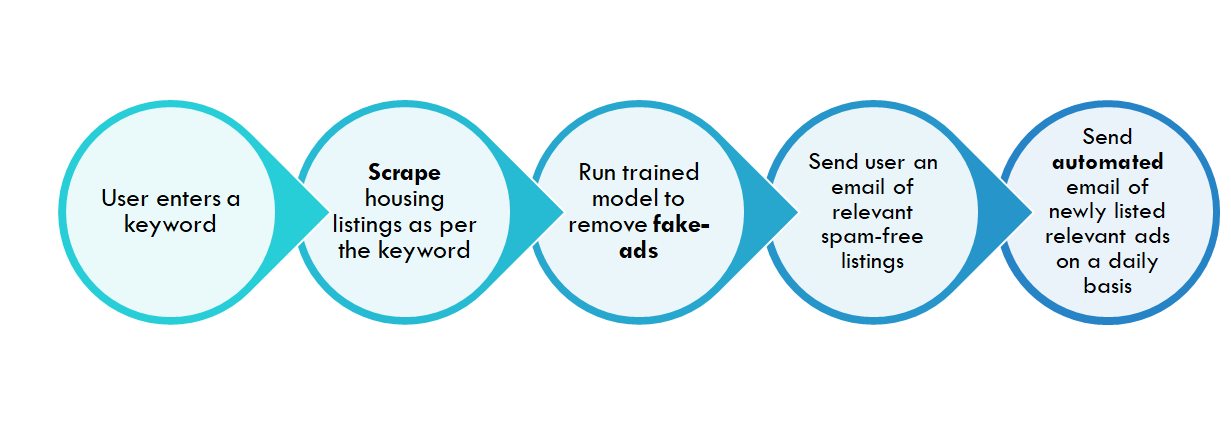
**Ways to identify a fraud ad**



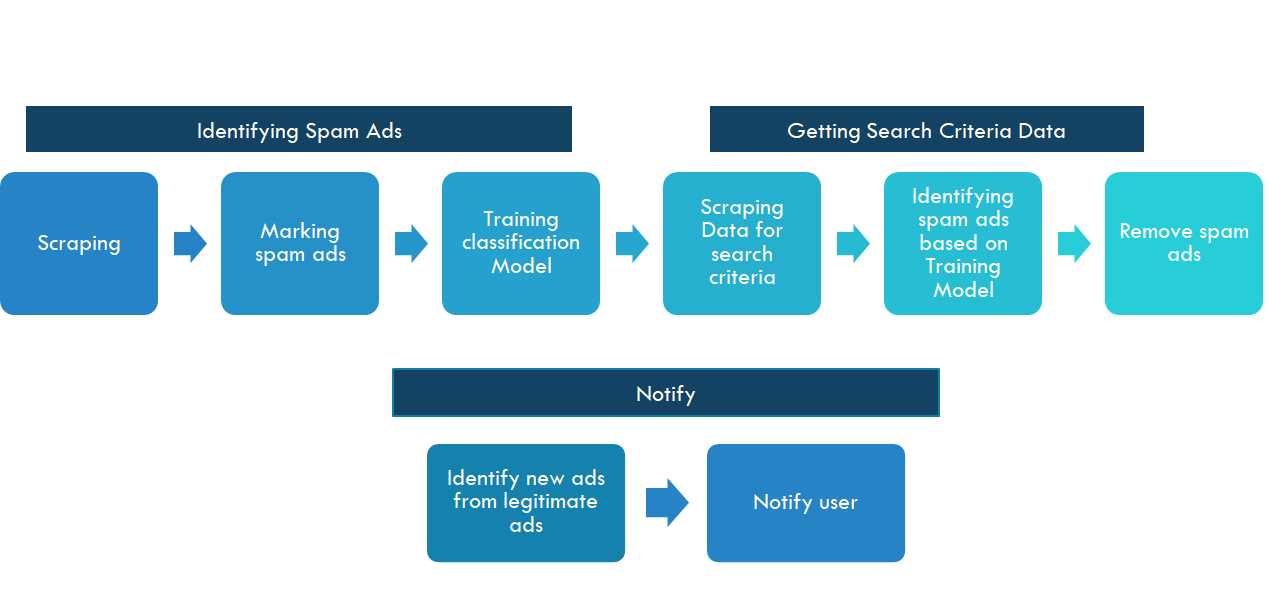
**Project Objective**

* Build a notification system to show spam free ads for housing rentals
  + Robust model for fraud advertisement detection to ensure safety of users
  + Automatically notify users about new listings as per the keywords specified

**Project Methodology**

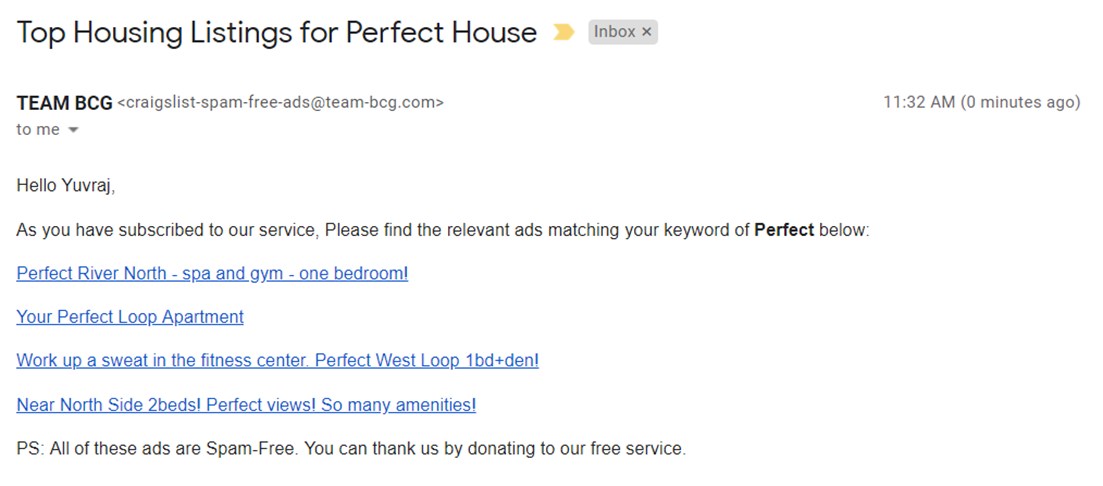
Simplistic Approach for the solution

In this project, the user enters a keyword for which he wants to view the ads. Then the project scrapes the housing rental posts from Craigslist based on the user-provided keyword, fetch features from the post's components (i.e. amenities, price, location), from the text in the title and body of the post, and several time-based features. Then after scraping the data we run our trained machine learning model which removes the fake-ads for the user. Post that, we send the user an email of the relevant spam-free ads. Also, we automate the process for the user so that everyday the user receives the email of relevant spam-free new listings as per the keyword mentioned.

Step by Step Process of Modelling the Solution

In our modelling solution, we have 3 steps:

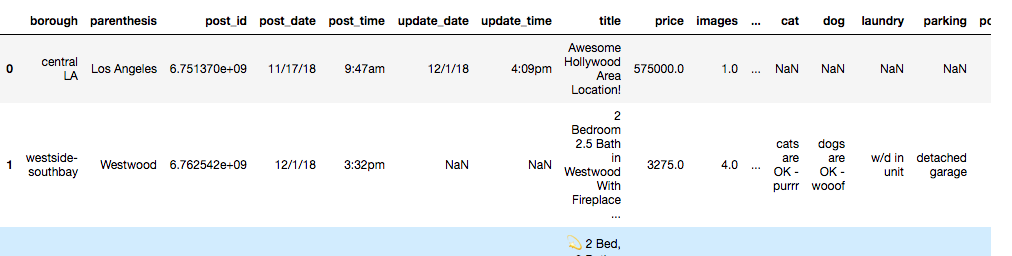
1. Identification of Spam Ads
   1. In this first we scrape the data from craigslist. We create 2 files
      1. urls\_city.csv - This file contain the url of the listings captured from the craigslist
      2. city\_data\_complete.csv - This file contains the features extracted from all the listings as captured in the urls file.
   2. Then, we needed to create a training data-set containing the labels of whether the ad was SPAM or not. For that, we re-checked the captured urls (after 7 days, 14 & 30 days) to validate whether the listing has been marked as spam or not. Based on that we created a final file which contained the features of the listing and added a label of SPAM with values as True or False.
   3. This data was used as a training dataset for training the various machine learning models.
2. Getting Search Criteria Data
   1. In this we scraped the data from craigslist based on the keyword specified by the user.
   2. Then we passed the listings to our trained model which detects the spam-ads.
   3. Based on detection, it removes all the spam-ads and creates a list of spam-free ads.
3. Notification System
   1. After identification of spam-free ads, we identify the most relevant ads from the legitimate listings data which has been curated.
   2. We send a notification email to the user which contains the relevant spam-free ads.
   3. This process is automated and the user receives an e-mail on daily basis containing the new listings of relevant spam-free ads.

The sample e-mail generated by the system is:

**Data Analysis**

The problem statement involved building a solution to **automate the process of sending notifications of genuine housing listings to a user based on his/her search keywords on Craigslist**. The analysis consisted of the following key steps:

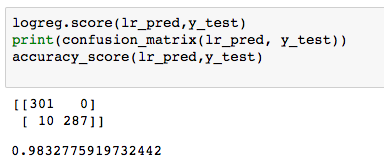
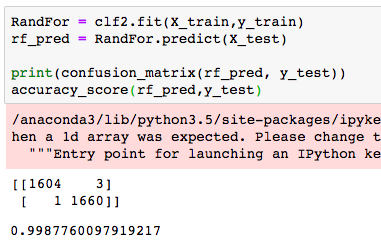
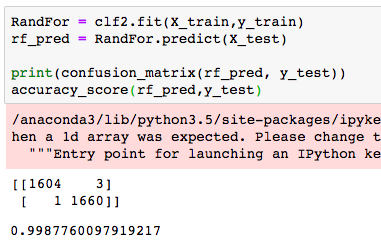
1. **Data Scraping**: We scraped close to 2732 housing rental postings for Los Angeles city with features of an advertisement like borough, post\_date, post\_time ,price, number of images, square\_footage, housing\_type, description , bedroom, bathroom, cat allowed, dogs allowed, laundry, parking, smoking,furnished etc. which will taken as inputs for building fake advertisement classification model.



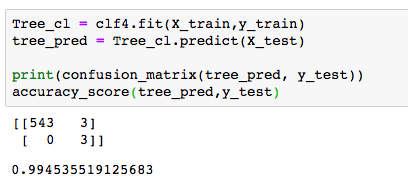
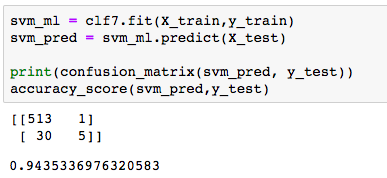
2. **Spam Tagging:** The ads were identified as spam by rechecking the same postings a week later to identify which ads were removed. Ads which had been ‘flagged for removal’ were tagged as Spam ads. The training dataset was fed into a predictive model to identify fake advertisement listings using few features/predictors of the given ad.

3. **Statistical Modelling:** The training dataset was used to build a classification model to classify ads into spam or not. The model uses input parameters like features of advertisement (price, bedrooms, number of images, text description) to identify whether an ad is spam or not. The following were the steps followed for model building:

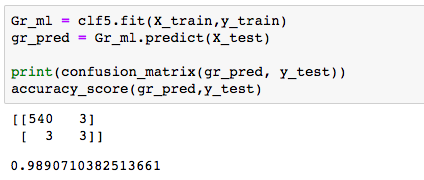
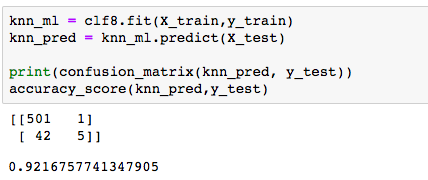
* **Data Pre-processing**:
  + Removed special characters, stopwords, punctuations
  + Replaced missing value treatment with appropriate values or eliminated them
  + Additional columns creation: For eg created a column to understand if a ad has been reposted or not
  + TF-IDF vectorisation of description of ad
  + Creation of dummy variables for categorical variables
* **Feature Selection**: We used Random Forest classifier with 100 estimators to select relevant features out of all variables to build the statistical model
* **SMOTE OverSampling**: Since only very less percentage (~2%) of ads are spam, this is an unbalanced classification problem. Hence we have performed synthetic minority oversampling technique (SMOTE) to over sample and balance the data. Using SMOTE improved our classification accuracy by a significant amount
* **Train - Test Split:** To test the classification model we have split the data into training and test sets in 70:30 proportion
* **Classification Model:** We experimented with any classification models like Logistic Regression, Random Forest, SVM, Decision Tree etc and used accuracy, precision and recall metrics to evaluate these models. Ultimately we used a combination of these models and built an ensemble model by assigning equal weightage to all classification models and using logistic regression as meta classifier.
* **Validation:** The statistical fake advertisement detection model developed needs validation to estimate the accuracy of the model. A **cross-validation** method was used to divide the data into two portions, one for model development(training set) and the other for model validation(test set). It helps in estimating how well the predicted classification agree with the responses within the independent sample. We evaluated various models on the basis of accuracy and **confusion matrix**, which shows the correct and incorrect predictions for each class.

 Logistic Regression Random Forest Classifier

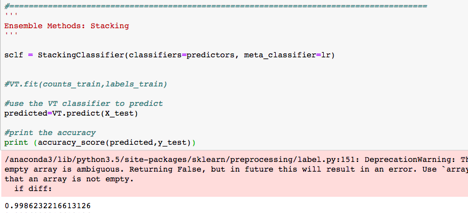
Tree Classifier SVM



Gradient Boosting KNN



Ultimately we decided to use a combination of all models and built an ensemble model using logistic regression as a meta classifier. The ensemble model uses a series of classifiers to consecutively build classification results and predicts based on results made by previous model. The ensemble gives the best results



**Summary of results**:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 98.3% |
| Random Forest | 99.6% |
| Tree Classifier | 99.4% |
| Gradient Boosting | 98.9% |
| KNN | 92.1% |
| SVM | 94.3% |
| Ensemble Modelling | 99.9% |

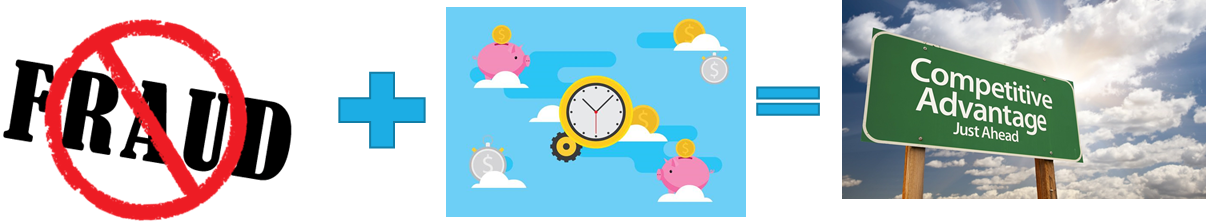
4. **Search Results Scraping**: Whenever an user searches for a housing related keyword, data for that search criteria is scraped and saved. The output of that search keyword is saved and fed to the trained model to classify ads as fake/non-fake.

5. **Spam Ads filtering**: All spam ads identified are removed. Also all the new listings which have been previously shown to the user are removed. New search result queries are sent to the user via email. The process is detailed as following. All the new genuine listings which didn’t appear in user’s last search results would then be used in the automated email communication sent to the user.

6. **Notify user** - An automatic email is sent to the user with a list of all urls identified in the system & detail of the properties/listings. SMTP(Simple mail transfer protocol) library was used to add subject, mail id. See

**Conclusion**

Our project provides multifold benefits not only to the craigslist but to the users as well which provides a competitive advantage to Craigslist for creating a spam-free ad portal.



**Craigslist(Fraud prevention) User safety Competitive advantage**

Our project provides added value to the Craigslist by:

* Say No to Frauds by providing a Spam-free Ad listing experience
* Building a loyal customer base for Craigslist
* Enhances the customer lifetime value
* Enhances the reliability of Craigslist
* Increases the reputation of Craigslist

Our project provides value to Craigslist users by:

* Ensuring user safety and mitigating the risk of fraudulent advertisements
* Narrowing down the selection for users
* Reduce time of users to find legitimate ads from hours to mins
* Providing convenience by sending them automated Notification of relevant set of spam-free ads
* Improved customer satisfaction

As a result our project provides a one-stop solution to find a spam-free ad and notify the users of relevant ads. It provides a competitive advantage to Craigslist by providing a platform which has zero tolerance to Fraud-Ads and providing a spam-free ad platform.

**Future scope**

The project has a few additional steps that have yet to be implemented. These ideas include

* Extraction of phone numbers from the craigslist ad listing
* Breaking location down by neighborhood
* Identifying price outliers and calculating whether a post falls within a normal range based on its location (since prices vary widely)
* Text based classifications such as number of (or percentage of) upper case letters, special characters, and length of text.
* Expanding the scope of project to other countries of Craigslist which have different languages.

**Appendix**

**Why ad spam may be a problem for house seeker?**

According to the popular stats blog, FiveThirtyEight, the average American will move 11.4 times in their lives. This means we can assume 11 homes searches are done in over the course of an American's lives. That's a lot of homes—especially considering most realtors suggest only living in three homes throughout your life. Some 37 million Americans move every year and many of them make mistakes in the process. The Pew Research Center’s analysis of U.S. Census Bureau data shows that more U.S. households are renting than at any point in 50 years, with 36.6 percent of households renting their home. The average rent for a two bedroom apartment in Manhattan is $3,895, according to the January 2015 Citi habitat market report. Craigslist’s New York apartment classifieds are a con artist favorite: Fraud in New York City’s Craigslist classifieds has become so pervasive that Craigslist has considered charging a fee for its ads. Their hope is that by putting a charge in place, they will discourage phony listings. Most of these fraudulent postings are common bait and switch schemes. However, some of the cases reported involved more elaborate schemes run by professional criminals. These scam artists have managed to bilk apartment seekers for thousands of dollars. Everyone knows how competitive the New York apartment market is, with too many people looking for far too few apartments. Some bold con artists have capitalized on this situation and used it to their advantage.

Of 37 million people one third check listings on craigslist if we charge them $10 for sending genuine listings and one third people buy this service then we can earn $100 million for just housing adds. This will increase the people’s trust in our website and reduce the risk of losing customers.

We intend to extend this project for all kinds of listing on Craigslist.

**Various types of Fraud Ads**

Phony Internet ads selling big-ticket items, from cars to recreational vehicles to boats, led to nearly 7,000 complaints to the FBI's Internet Crime Complaint Center (IC3) and accounted for more than $20 million in losses, From June 2009 to June 2014. Victims also made payments for what they thought were real lawn mowers, tractors, heavy equipment and more after being lured in by fake ads. Here's how the scam works: After seeing a fake ad, which typically has a photo and a deeply discounted selling price, the victim responds to a phone number and has to leave a message. The message is replied to with a text in which an email address is sent for further contact.

Some US$7.2 billion in ad spending was wasted on fraudulent robot-viewing last year, said White Ops and the United States' Association of National Advertisers (ANA). This is based on a study of 49 ANA members' digital advertising activity between October last year and January. Cybercrime is global, and digital advertising is global. No market, is immune to ad fraud. Here are some statistics from ad frauds:

* Marketers lost $7.2 billion to digital ad fraud in 2016 (WhiteOps, 2016)
* 1 in 5 ad-serving websites are visited exclusively by fraud bots (The Verge, 2017)
* For every $3 spent on digital ads, fraud takes $1 (Adage.com, 2015)