**CHAPTER 6**

**GATHERING DATA**

Data collection is the process of gathering and measuring information from countless different sources. In order to use the data we collect to develop practical [artificial intelligence (AI)](https://www.datarobot.com/wiki/artificial-intelligence/) and [machine learning](https://www.datarobot.com/wiki/machine-learning/) solutions, it must be collected and stored in a way that makes sense for the business problem at hand.

**6.1 Data Collection**

The set of phishing URLs are collected from opensource service called PhishTank. This service provide a set of phishing URLs in multiple formats like csv, json etc. that gets updated hourly. To download the data: https://www.phishtank.com/developer\_info.php. From this dataset, 5000 random phishing URLs are collected to train the ML models.

The legitimate URLs are obtained from the open datasets of the University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html. This dataset has a collection of benign, spam, phishing, malware & defacement URLs. Out of all these types, the benign url dataset is considered for this project. From this dataset, 5000 random legitimate URLs are collected to train the ML models.

**6.1.1 Column Definitions**

1. **phish\_id:** The ID number by which Phishtank refers to a phish submission. All data in PhishTank is tied to this ID. This will always be a positive integer.
2. **phish\_detail\_url:** PhishTank detail url for the phish, where you can view data about the phish, including a screenshot and the community votes.
3. **url:** The phish URL. This is always a string, and in the XML feeds may be a CDATA block.
4. **submission\_time:** The date and time at which this phish was reported to Phishtank. This is an ISO 8601 formatted date.
5. **verified:** Whether or not this phish has been verified by our community. In these data files, this will always be the string 'yes' since we only supply verified phishes in these files.
6. **verification\_time:**The date and time at which the phish was verified as valid by our community. This is an ISO 8601 formatted date.
7. **online:** Whether or not the phish is online and operational. In these data files, this will always be the string 'yes' since we only supply online phishes in these files.
8. **target:** The name of the company or brand the phish is impersonating, if it's known.

**6.2 Full Dataset Content**

There are 5 zip files in total and range from ~2 gb to 3 gb in size. Additionally, we randomly sampled 5% of these datasets and created a smaller dataset for use in Kernels. The random sample contains 5000 sites.