



1

# Leveraging Machine Learning in the Age of Information



– Kevin Jacob

Business/ Mathematics Major 1<sup>st</sup> Year, University of Texas, Austin, United States

 <https://orcid.org/0000-0003-2429-9775>  [kevin.jacob@utexas.edu](mailto:kevin.jacob@utexas.edu)

– Kevin Li

Computer Science Major 1<sup>st</sup> Year, University of Texas, Austin, United States

 <https://orcid.org/0000-0002-7041-1070>  [kxl4126@utexas.edu](mailto:kxl4126@utexas.edu)

In the Information Age, user data has become a valuable asset to any business. Data science is such a growing field due to the importance of data, as machine learning models can be generated to increase productivity and performance of service, improve user experience, and increase revenue streams.

Large companies have invested billions into this field already, giving them an edge in their respective industries, but how smaller companies join this AI revolution and gather data without investing so much money?

## Keywords

- Machine Learning
- HTTP
- AI
- Web Scraping
- Data

## ARTICLE HISTORY

### Paper Nomenclature:

Experiential Research Papers (ERP)

Paper Code: CYBNMV2N1JAN2020ERP1

Submission Online: 04-Jan-2020

Manuscript Acknowledged: 05-Jan-2020

Originality Check: 07-Jan-2020

Originality Test Ratio: 0% (Urkund)

Peer Reviewers Comment: 12-Jan-2020

Blind Reviewers Remarks: 14-Jan-2020

Author Revert: 18-Jan-2020

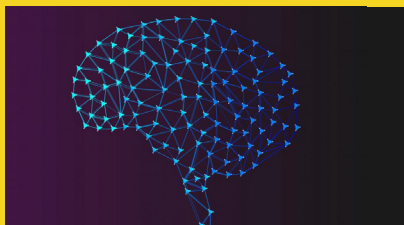
Camera-Ready-Copy: 20-Jan-2020

Editorial Board Citation: 31-Jan-2020

Published Online First: 31-Mar-2020

## Introduction

The solution is automated data collection via processes like collecting user data and web scraping. Web scraping is able to transcribe specified information using the Hypertext Transfer Protocol (HTTP) to a central database. Tools such as web and image crawlers are freely available to the public, and allows the developer to scrape and collect large amounts of data from the web. The developer can then comb through the information and use it to optimize there required task.



## What is Good Data?

Before we discuss the process of automated data collection, we must first determine what data we should be targeting. Manually analyzing large data sets can be inefficient, leading to some people neglecting to filter out “bad data”. However, without this processing, there could be a multitude of negative ramifications. When collecting data, a variety of characteristics must be met. For example, it must be representative and non-biased. If it doesn’t meet those criteria, it could lead to inaccuracies that interfere with your task, or particular groups could be unfairly targeted. The resulting complications could lead to a significant loss of revenue. For example, in California, USA, mathematical algorithms were used to determine a “risk assessment” for whether an inmate should be released from prison. The data sets that were scraped to train this model

were “highly correlated with race and class” (Alexander, *The Newest Jim Crow* 1). This led to the biased treatment of certain minorities in California, USA, when it came to their release. To avoid outcomes similar to these events, we must ensure that we have “good data”.

## Web Scraping

These days, a large majority of available public information is stored on websites. Technically, this data is accessible by everyone and can be exploited by anyone. However, actually sifting through the abundance of data on the internet can pose a significant challenge. The most common and effective way to do targeted data collection over the web is through a technique called web scraping, or web crawling. The basic process of scraping a webpage for data using automation is detailed below:

- A desired site is picked to scrape. The site should first be perused by a developer to get an idea for the general pattern in the layout of the desired webpages. The developer should identify which selectors (a code snippet which specifies the location of an element on a webpage) are necessary for selecting the proper data elements in the html code. This will allow the process of web scraping to be automated.
- The desired site is picked to scrape and load up the URL with a get request. Then, any forms required for the scraping bot to access the data are filled out. These forms could include search fields or login pages. Both of these tasks can be completed with something like the Python library, Selenium.
- The process of navigating forms and fields is continued until reaching the webpage which contains data that needs to be scraped. This page is now downloaded with an html parser, such as the one offered by the Python library, BeautifulSoup.
- The proper selectors are used to target the desired elements of the web page and extract the data. Data can be in the form of text or image links. The data is then stored in some sort of database for further use.
- This process is repeated across multiple webpages and even websites. Some more complex APIs, like Python's Scrapy, are able to dynamically go through links on webpages to find information on sites not initially specified by the developer.

```
from selenium import webdriver
from selenium.webdriver.support.ui import Select
import time
from bs4 import BeautifulSoup
import math

import HTMLParser
outfile = open("data.csv", "a") #opens a data file to get ready to write data to it
url = "https://www.amity.edu/"
browser = webdriver.Chrome() #opens a chrome browser
browser.get(url) #gets the website and loads it up

selector = Select(browser.find_element_by_name("title")) #selects the title of the website
selector.select_by_visible_text("Careers") #selects the option titled Careers
button = browser.find_element_by_xpath("/html/body/div/form/input").click() #clicks the button with this xpath

soup = BeautifulSoup(browser.page_source, 'html.parser') #downloads html file of the page
options = soup.findAll("table") #finds all elements that tables
```

Web scraping, when applied properly, allows developers to fully leverage the power of a computer in automated data collection. The most efficient and well structured web scraping architectures can parse millions of websites in a small time frame. Through this, any developer can build a large database for a variety of uses.

### Collecting User Data

Another viable method for automating data collection is through tracking user data.

The implementation of some underlying software which records how a certain service or application is used can be immensely beneficial. Such software can allow organizations to gain insight into who is using their product and how they are doing so. One particular benefit of this method of data collection is that the data is effectively proprietary. This means that no other party should have free access to it and the data collector has full rights over the data. Having exclusive access to datasets can give organizations an edge over the competition and is an important addition to the publicly available data.

### Making the Data Worthwhile

After a sufficient amount of data is collected, developers need to make sure the data is used properly and efficiently. Although machine learning algorithms will usually perform better with more available data to train on, they experience diminishing returns

until this may no longer be the case. Eventually, the algorithm may reach a threshold where the negligible benefit from just having more raw data renders focusing on mass data collection impracticable. This is where focusing on the actual quality of data that is being collected becomes much more valuable. Some useful techniques to improve the efficacy of collected data are detailed below:

- Data aggregation can be used to lessen the complexity of the data, making machine learning models more generalizable. If all the data initially collected is thrown immediately into a machine learning model, then an issue called overfitting could arise. This occurs when the model is very accurate in the isolated training environment, but loses much of its accuracy in the real world. Some ways to ensure this doesn't happen include:
  - PCA - reduces dimensionality of the data so there are less variables for the machine learning model to train on
  - Significance testing - used to remove irrelevant variables that do not correlate sufficiently enough with the target that is being predicted
- Dropping outliers and/or leads to a more stable machine learning model as extrema may skew the model in a disproportionate way.

- Normalizing the columns so certain variables do not have a disproportionately large impact on the prediction. For example, if a dataset had one column in the order of thousands and another column was in the order of millions, one column may have an excessive effect on the prediction model. Basing the columns on its own normal distribution can fix this issue.
- Making sure the data is representative is important in a statistical and social setting. In terms of statistics, a dataset must be representative of the target population in order for the resulting predictive model to be accurate. In the social sense, ensuring representation decreases the likelihood of any undesired discrimination.
- Establish checks for inconsistent or unexpected data for manual investigation in order to ensure that the data collection process is running smoothly. This is commonly accomplished through running frequent reports on the data to detect any anomalies.

- Certain pieces of data may contain missing data in some columns. This can be addressed by filling out the missing column with the mean of that column.

By following these techniques, it can be ensured that the data gathered is being put to actually improve machine learning models.

### Privacy Issues

Data collection is a concern for many organizations and corporations looking to maximize their efficacy in the market, and as a result, benefits the general population in many ways. Artificial intelligence can make services more accurate, personalized, and efficient, but it often comes at a cost on behalf of the common consumer. Nothing comes for free, and some privacy may be the cost of a better user experience.

Another issue is the accessibility of this information. Because so much data is collected and stored in a central database, this increases the severity of any data breaches. If a hacker gains access to this database, they would have important information such as passwords, medical records, credit-card information, and more for a plethora of customers.

Whether we like it or not, the age of information is upon us, and will not go away soon. In order to keep up with the data revolution, developers must fully embrace the importance of data collection and take advantage of the automated techniques discussed in this paper.

### Sources:

- 9 Tips to Improve Data Quality. *Riskconnect* Available at: <http://www.riskconnectclearsight.com/9-tips-to-improve-data-quality/>. (Accessed: 27th January 2020)
- Alexander, M. The Newest Jim Crow. *The New York Times* (2018). Available at: <https://www.nytimes.com/2018/11/08/opinion/sunday/criminal-justice-reforms-race-technology.html>. (Accessed: 27th January 2020)
- Deagon, B. Big Data Companies And Data Analytics: The Stocks To Buy And Watch. *Investor's Business Daily* (2020). Available at: <https://www.investors.com/news/technology/big-data-companies-data-analytics-stocks-to-watch/>. (Accessed: 27th January 2020)
- Deepak & Deepak. Data Preprocessing in Data Mining. *GeeksforGeeks* (2019). Available at: <https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/>. (Accessed: 27th January 2020)



**Kevin Jacob** is a student of the University of Texas, Austin in United States pursuing his Bachelors of Science in Business and Mathematics. He has designed a web app for the company Texas Board X, along with reinforcement learning models. Kevin was a finalist in the International Career DECA Conference for creating his own business. He wants to use his skills in both math and computer science to solve problems in his community.

[kevin.jacob@utexas.edu](mailto:kevin.jacob@utexas.edu)



**Kevin Li** is a student of the University of Texas, Austin in United States pursuing his bachelors of science in computer science. He's interested in artificial intelligence and application based mathematics. He has developed an app to help people with their speech impediment and has used machine learning to solve real world environments. He also works to use his software development skills to improve the efficacy of the education system. Kevin looks forward to studying and applying machine learning to create innovative solutions to pressing world issues.

[kxl4126@utexas.edu](mailto:kxl4126@utexas.edu)

### Annexure I

Submission Date	Submission Id	Word Count	Character Count
07-Jan-2020	1247060068 (Turnitin)	2158	11804

ORIGINALITY REPORT			
0%	0%	0%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
PRIMARY SOURCES			

Note: The Cybernomics had used the turnitin plagiarism [http://www.turnitin.com] tool to check the originality



### Reviewers Comment

**Review 1:** The improvements in hardware and software in the superseding years have led to far better-quality presentation for that traditional style of data detection.

**Review 2:** The major extension was supported by the understanding of the importance of metadata. By manually defined relationships among fields in multiple sources.

**Review 3:** The article has covered various aspects of automatic data collection such as web scraping or crawling as an efficient way to gather users data without investing so much money.



### Editorial Excerpt

This article has 0% plagiarism which is accepted as per the standards of publication for the magazine. The authors (Kevin Jacob, Kevin Li,) have covered all the facts in current scenario. As Artificial intelligence (AI) has converted the basis of everyday technologies including phones, cars, banking apps, home devices and more. The development and success of companies including Amazon, Facebook, Google and Uber which operate on digital platforms raise fundamental challenges for administrators of established companies. Hence after review and comment it is decided to marked under “Experiential Research Papers (ERP)” category.

### Acknowledgement

**Kevin Jacob:** I want to use my skills in both math and computer science to solve problems in the community. I want to give a special thanks to my parents (Mrs. Nirmala Jacob & Mr. Winston Jacob) and all his faculty mentors who have always supported me, especially Rajbala ma’am for providing the opportunity to write the article “Leveraging Machine Learning in the Age of Information” for Cybernomics.

**Kevin Li:** I want to give special thanks to my parents (Mr. Gary Li and Mrs. Holly Cao) who have always supported me, in my interests. Special thanks to Rajbala ma’am for giving me an ability to write the article “Leveraging Machine Learning in the Age of Information” for Cybernomics.

### Disclaimer

All views expressed in this paper are my/our own. Some of the content is taken from open source websites & some are copyright free for the purpose of disseminating knowledge. Those some We/I had mentioned above in the references section and acknowledged/cited as when and where required. The author/s has cited their joint own work mostly, Tables/Data from other referenced sources in this particular paper with the narrative & endorsement has been presented within quotes and reference at the bottom of the article accordingly & appropriately. Finally some of the contents which are taken or overlapped from open source websites for the knowledge purpose. Those some of i/we had mentioned above in the references section.



Kevin Jacob & Kevin Li  
“Leveraging Machine Learning  
in the Age of Information”  
Volume-2, Issue-1, Jan 2020.  
(www.cybernomics.in)

Frequency: Monthly, Published: 2020  
Conflict of Interest: Author of a Paper  
had no conflict neither financially  
nor academically.

