Name:	Bikis Muhammed
Email:	bmpb8@mst.edu
Course:	CS 5402
Assignment:	What is data mining/What is artificial intelligence
Course:	CS 5402

In [67]:
```
# Imported for searching google with python
from googlesearch import search
# Imported for http request
import requests
# For parsing html and xml data/files
from bs4 import BeautifulSoup
# For date and time
from datetime import datetime
# For using csv file
import pandas as pd
# For generating word clouds Image
from wordcloud import WordCloud, STOPWORDS
# For extracting a domain name of a website
import tldextract
# For removing punctuation
import string
# For counting the frequncy of words in a string
from collections import Counter
# For ploting the frequency of common words
import matplotlib.pylab as plt
# For minimal natural language processing
import nltk
# To get extended list of words from wordnet and words
from nltk.corpus import wordnet, words
```

Concept/Objctive Description:

Analyze the frequency/infrequency usage of words by different web resources to explain data mining and artificial intelligence.

Data Collection:

Data collection is done in an automated fashion with the help of different python packages. Google python package is used to search google and get the first fifteen URLs for both "What is data mining" and "What is artificial intelligence." Each URL is then placed on two lists, and this list of URLs is searched for the text that contains using the technique of web scraping or crawling. All the extracted text is then placed into two separate string variables for both "What is data mining" and "What is artificial intelligence. All data found is placed in a CSV file for further processing.

Example Description:

Category
This is just the query value. It will have either "What is data mining" or "What is artificial intelligence."

PullDate
This the date when the data is web-scrapped.

Source
This is the general name instance of each website. It is not the title of an article on a website. For instance, if https://www.wikipedia.org is visited, Wikipedia would be extracted as the name of the site.

Link
This example contains all URLs that have been used for web scraping.

Text
This contains all the generated and compiled and appended text data.

Web Searching and Scraping:

Web Scraping

In [2]:
```
# URL get request
# beautifulsoup html parsing
# finding all paragraph tags
# get the text undder that tag
def webScrap (URL):
    text = ''
    result = requests.get(URL)
```

```
    soup = BeautifulSoup(result.content, features="html.parser")
    paragraphs = soup.findAll('p')
    for par in paragraphs:
        text += par.getText()
    return text
```

Google Searching and Minimal Language processing

Validating includes elemntaing websites from being using more than one time in the search results.

In [3]:

```
# Google search
    # Quiry:  the query text which is being searched
    # tld:  the domain .com, .org
    # lang: languge
    # num:  number of requests per page
    # stop determine the number of urls that will be generated
# Append the result to a list

# Google will provide 45 search result and 15 will valid websites will be selected for scrapping
# webscarapped text will be cleaned
def getWebData ( query):
    URLs = []
    HOSTNAME = []
    TEXT = []
    limit = 15
    counter = 0
    for i in search (query,  tld='co.in', lang='en', tbs='0', safe='off', num=3, start=0, stop= 50, pause=2.0, country='', ex
        subdomain, domain, suffix = tldextract.extract(i)
        text = str(webScrap (i))

        # Removing newline char
        text =  text.replace('\n', '')

        # Remove unnecessary holder
        text =  text.replace('\\x', '')

        # Removing punctuation from strings
        for punk in string.punctuation:
            text = text.replace(punk,'')

        # Removing number from strings
        number = '0123456789'
        for num in number:
            text =  text.replace(num, '')
        #-------------------------------------------------------------------------
        #THis taks for ever
        # Remove any number left between strings
#         text = ''.join ([i for i in text if not i.isdigit()])

        # Removing non English words from text (simple natural language processing)
        # It uses both words and wordnet from nltk library, and the custom string contains 384042 words
#         customWord = (' '.join(word for word in wordnet.words()).split()) + words.words()
#         text = ' '.join([word for word in nltk.wordpunct_tokenize(text) if word.lower() in customWord or not word.isalpha()

        #-------------------------------------------------------------------------


        # Check if the domain is not empty and the hostname is already selected (which is a part of normalization process)
        if domain != '' and domain not in HOSTNAME and text and ("Data mining is" in text or "Artificial intelligence is" in
            if counter < limit:
                HOSTNAME.append(domain)
                URLs.append(i)
                TEXT.append(text)
                counter += 1
            else:
                return  HOSTNAME, URLs, TEXT

    # return Hostname, URLs
```

Word Cloud

Wordcloud and collection python libraries will be used for text abalytics perposes.

In [4]:

```
# makeWordCloud method has text and image name parameters.
# STOPWORD library is also used to remove unnecessary words
def makeWordCloud (text, imagename, sw):
    # Generating the word cloud file
    mywordcloud=  WordCloud (
    background_color= 'white',
    stopwords= STOPWORDS.update(sw),
    height= 800,
    width=600,
    collocations= False,
    )
    mywordcloud.generate(text)
    mywordcloud.to_file(imagename)
```

Exploratory Data Analysis:

Not applicable.

Mining or Analytics:

In [5]:
```
# Adding a question mark at the end of a search may help searching
query = ['What is data mining', 'What is artificial intelligence']
# Ask google to find websites that has this exact match
query1 = ['"Data mining is"', '"Artificial intelligence is"']
```

Table Data

What is data mining

In [6]:
```
category = [query[0]]*15
pullDate = pulldate = [datetime.today().strftime('%m/%d/%Y')]*15
source, link, text1 = getWebData(query1[0])
```

In [7]:
```
pd.DataFrame({'Category':category, 'PullDate': pulldate, 'Source':source, 'Link': link, 'Text':text1}).to_csv('result1.csv',
df = pd.read_csv("result1.csv")
df
```

Out[7]:

	Category	PullDate	Source	Link	Text
0	What is data mining	06/16/2021	wikipedia	https://en.wikipedia.org/wiki/Data_mining	Data mining is a process of extracting and dis...
1	What is data mining	06/16/2021	sas	https://www.sas.com/en_us/insights/analytics/d...	Skip to main contentAmericasEuropeMiddle East ...
2	What is data mining	06/16/2021	talend	https://www.talend.com/resources/what-is-data-...	Data mining isn't a new invention that came wi...
3	What is data mining	06/16/2021	investopedia	https://www.investopedia.com/terms/d/dataminin...	Data mining is a process used by companies to ...
4	What is data mining	06/16/2021	indiatimes	https://economictimes.indiatimes.com/definitio...	How they can help in wealth creationTomorrow i...
5	What is data mining	06/16/2021	britannica	https://www.britannica.com/technology/data-mining	Our editors will review what you've submitted ...
6	What is data mining	06/16/2021	tableau	https://www.tableau.com/learn/articles/what-is...	Data mining is the process of understanding da...
7	What is data mining	06/16/2021	techtarget	https://searchsqlserver.techtarget.com/definit...	Data mining is the process of sorting through ...
8	What is data mining	06/16/2021	dbta	https://www.dbta.com/Editorial/Trends-and-Appl...	The exponentially increasing amounts of data b...
9	What is data mining	06/16/2021	sisense	https://www.sisense.com/glossary/data-mining-b...	Explore DashboardBy submitting this form I agr...
10	What is data mining	06/16/2021	utexas	https://www.laits.utexas.edu/~norman/BUS.FOR/c...	by Doug Alexanderdeatracorcom Data mining is a...
11	What is data mining	06/16/2021	xplenty	https://www.xplenty.com/blog/what-is-data-mini...	SolutionsIndustriesChoose the solution that's ...
12	What is data mining	06/16/2021	iberdrola	https://www.iberdrola.com/innovation/data-mini...	Search suggestionsSearch suggestions Leaders ...
13	What is data mining	06/16/2021	logianalytics	https://www.logianalytics.com/resources/bi-enc...	See how you can create deploy and maintain a...
14	What is data mining	06/16/2021	guru99	https://www.guru99.com/data-mining-tutorial.html	Data Mining is a process of finding potentiall...

In [8]:
```
# Wordcloud
# Manual cleaning of the data
stopwords = {'need', 'part', 'allows', 'create', 'wheather', 'benefits', 'determine', 'using', 'past', 'want', 'understand',
clean_list =[]
STOPWORDS.update(stopwords)
bigword = ''.join(text1)

# # Removing newline char
# bigword =  bigword.replace('\n', '')

# # Remove unnecessary holder
# bigword =  bigword.replace('\\x', '')

# # Removing punctuation from strings
# for punk in string.punctuation:
#     bigword = bigword.replace(punk,'')

# # Removing number from strings
# number = '0123456789'
# for num in number:
#     bigword =  bigword.replace(num, '')


clean_text1 =''
for word in bigword.split():
    if str(word).lower() not in list(STOPWORDS):
        clean_text1 += ' '+  word

makeWordCloud(str(clean_text1), 'result1.png', STOPWORDS)
```

What is artificial intelligence

In [9]:
```
category = [query[1]]*15
pullDate = pulldate = [datetime.today().strftime('%m/%d/%Y')]*15
source, link, text2 = getWebData(query1[1])
```

In [10]:
```
pd.DataFrame({'Category':category, 'PullDate': pulldate, 'Source':source, 'Link': link, 'Text':text2}).to_csv('result2.csv',
df = pd.read_csv("result2.csv")
df
```

Out[10]:

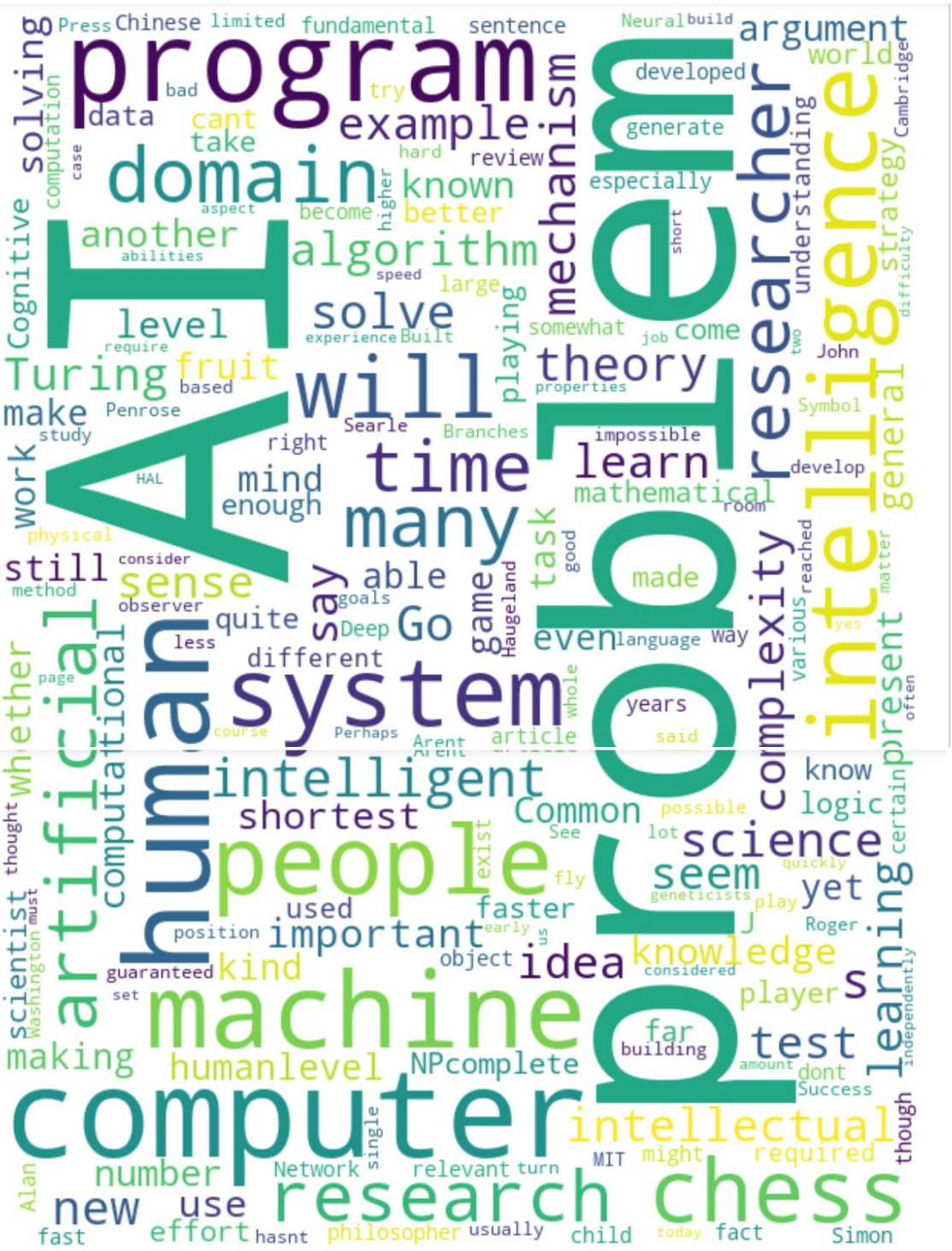
	Category	PullDate	Source	Link	Text
0	What is artificial intelligence	06/16/2021	investopedia	https://www.investopedia.com/terms/a/artificia...	Gordon Scott has been an active investor and t...
1	What is artificial intelligence	06/16/2021	wikipedia	https://en.wikipedia.org/wiki/Artificial_intel...	Artificial intelligence AI is intelligence dem...
2	What is artificial intelligence	06/16/2021	builtin	https://builtin.com/artificial-intelligence	Subscribe to Built In to get tech articles jo...
3	What is artificial intelligence	06/16/2021	brookings	https://www.brookings.edu/research/what-is-art...	Guidance for the Brookings community and the p...
4	What is artificial intelligence	06/16/2021	ibm	https://www.ibm.com/cloud/learn/what-is-artifi...	While a number of definitions of artificial in...
5	What is artificial intelligence	06/16/2021	sas	https://www.sas.com/en_us/insights/analytics/w...	Skip to main contentAmericasEuropeMiddle East ...
6	What is artificial intelligence	06/16/2021	techtarget	https://searchenterpriseai.techtarget.com/defi...	Artificial intelligence AI is the simulation o...
7	What is artificial intelligence	06/16/2021	futureoflife	https://futureoflife.org/background/benefits-r...	"Everything we love about civilization is a pr...
8	What is artificial intelligence	06/16/2021	zdnet	https://www.zdnet.com/article/what-is-ai-every...	An executive guide to artificial intelligence ...
9	What is artificial intelligence	06/16/2021	stanford	http://jmc.stanford.edu/artificial-intelligenc...	Q What is artificial intelligenceA It is the s...
10	What is artificial intelligence	06/16/2021	accenture	https://www.accenture.com/us-en/insights/artif...	return to previous buttonThis will navigate yo...
11	What is artificial intelligence	06/16/2021	forbes	https://www.forbes.com/sites/forbesbusinesscou...	Jody Glidden is the CoFounder and CEO of Intro...
12	What is artificial intelligence	06/16/2021	europa	https://www.europarl.europa.eu/news/en/headlin...	Artificial intelligence AI is set to be a defi...
13	What is artificial intelligence	06/16/2021	smithsonianmag	https://www.smithsonianmag.com/innovation/arti...	Save off the newsstand priceIn June of A few...
14	What is artificial intelligence	06/16/2021	louisiana	https://userweb.ucs.louisiana.edu/~isb9112/dep...	What is Artificial IntelligenceByIstván S N Be...

In [11]:

```
# Wordcloud
# Manual cleaning of the data
stopwords = {'may', 'issue', 'called', 'may', 'best', 'now', 'well', 'every', 'major', 'high', 'complex', 'agent', 'form', 'r
clean_list =[]
STOPWORDS.update(stopwords)
bigword = ''.join(text2)

clean_text2 =''
for word in bigword.split():
    if str(word).lower() not in list(STOPWORDS):
        clean_text2 += ' '+  word

makeWordCloud(str(clean_text2), 'result2.png', STOPWORDS)
```



Evaluation:

Frequency Analysis of words

What is data mining and artificail intelligence(The most frequent 100 words)

In [64]:

```
DM = {}  AI = {}

# Frequency of the most common 100 words
for word, count in Counter(clean_text1.lower().split()).most_common(100):
    DM[word] =  count

for word, count in Counter(clean_text2.lower().split()).most_common(100):   #give a number value for most commpn argument
    AI[word] =  count
# for letter, count in countedWords.most_common(30):
#     wordCount[letter.lower()] = count

# for i,j in DM.items():
#         print('{0}: {1}'.format(i.upper(),j))

print("-----------------------------------------------------------")
print('| {:<14} | {: <4} | {: <13} | {: <5} |'.format("Nominal Data", "FreqInDM", "Nominal Data", "FreqInAI" ))
print("-----------------------------------------------------------")
for x, y in zip(DM.keys(), AI.keys()):
    print('| {: <14} | {: <8} | {: <13} | {: <8} |'.format(str(x), DM[str(x)],str(y), AI[str(y)]))
print("-----------------------------------------------------------")
```

Nominal Data	FreqInDM	Nominal Data	FreqInAI
data	3889	ai	2387
mining	2613	human	699
helps	707	people	689
information	450	intelligence	679
customer	412	problem	649
techniques	407	problems	639
used	352	computer	633
business	350	programs	613
process	329	chess	574
customers	328	computers	561
companies	290	will	553
predict	285	machine	544
example	274	many	436
analysis	266	program	391
new	255	research	389
credit	248	humans	386
patterns	246	artificial	383
technique	232	intelligent	376
may	230	system	359
use	213	domains	343
service	206	systems	342
different	200	go	320
tools	198	solve	319
statistical	181	machines	311
etc	181	science	306
identify	180	theory	303
model	175	new	297
ecommerce	173	researchers	297
oracle	173	complexity	282
analytics	165	intellectual	273
difficult	165	mechanisms	272
learn	163	algorithms	271
prediction	149	turing	259
understanding	146	example	253
offer	145	s	252
likely	145	sense	245
work	139	learning	243
insurance	137	important	243
detection	135	time	236
retail	135	knowledge	236
make	134	another	234
providers	131	level	229
students	131	shortest	229
tool	130	known	222
crime	129	use	221
classification	128	solving	218
set	127	computational	209
offers	127	learn	205
marketing	125	test	205
card	125	idea	199
rules	124	making	198
help	123	common	198
training	122	argument	196
future	121	able	194
results	120	still	191
many	120	work	177
behavior	120	yet	175
usage	118	general	174
trends	117	says	174
database	114	mind	173
phase	114	whether	173
pattern	111	number	173
software	110	even	170
users	109	humanlevel	168
revenues	106	mathematical	167
products	106	cant	164
items	106	kind	163
knowledge	105	quite	163
sequential	104	npcomplete	162

operations	103	far	161
check	103	data	160
analyze	101	logic	160
production	101	enough	158
well	100	take	157
company	100	game	155
regression	100	strategy	153
association	99	cognitive	152
distance	98	j	151
outer	96	made	148
requires	94	faster	148
organizations	94	playing	145
store	94	times	145
databases	93	better	144
automated	93	present	136
makes	93	fruit	135
age	92	used	134
existing	91	come	134
industries	91	required	133
search	90	different	132
deployment	89	understanding	132
education	88	world	131
websites	88	developed	130
manufacturing	88	chinese	128
systems	87	might	127
clustering	87	certain	126
profitable	87	relevant	126
supermarkets	87	try	125
banking	87	seems	124
investigation	87	deep	122
advance	87	dont	122

Common words

In [65]:
```
AIC = {}
MDC = {}
for commonkey in set(DM).intersection(set(AI)):
    AIC[commonkey] = AI [commonkey]
    MDC[commonkey] = DM[commonkey]
    print (commonkey)
```

```
learn
work
use
used
new
systems
example
many
knowledge
different
data
understanding
```

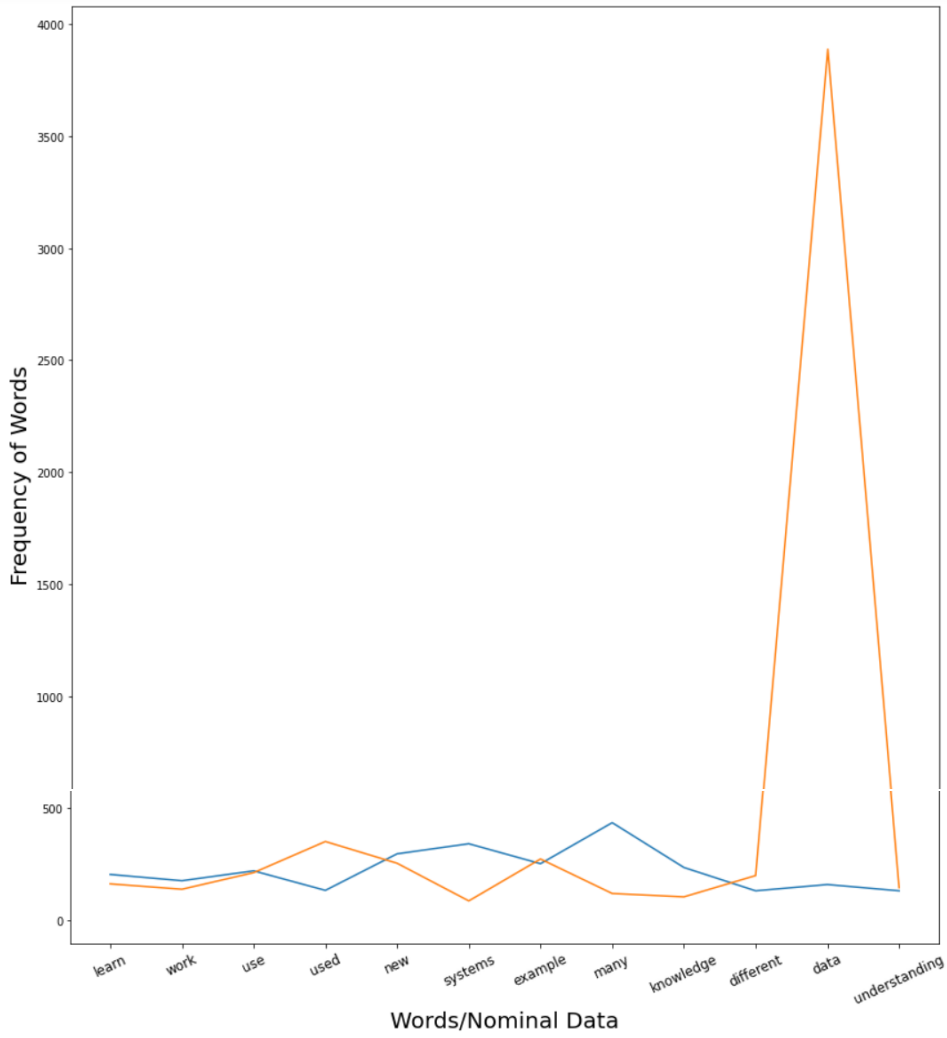
Frequency Graph

In [77]:
```
x = list(AIC.keys())
y = list(AIC.values())

k = list(MDC.keys())
l = list(MDC.values())

plt.plot(x, y)
plt.plot(k,l)

# plt.bar(range(len(DM)), DM.values(), align='center', width=0.5)
plt.xlabel('Words/Nominal Data', fontsize=20)
plt.ylabel('Frequency of Words', fontsize=20)
plt.xticks(ticks = x, rotation=25, labels = x, fontsize =12)
plt.rcParams["figure.figsize"] = (14,15.5)
plt.show()
# fig = plt.figure()
# fig.savefig('plot.png')
```



Results:

After selecting the most frequented 100 words from both data mining and artificial intelligence, the only common words found were 12 in count. As we can see from the graph, the word data is used more frequently than other words, especially in Data mining.

References

GeeksforGeeks (2021). Python Libraries. Retrived (2021, Jun 14) from https://www.geeksforgeeks.org/

JUpyter (2021). The Jupyter Notebook. Retrived (2021, June 10) from https://jupyter-notebook.readthedocs.io/en/stable/notebook.html

Schafer, Corey (2016). Jupyter Notebook Tutorial: Introduction, Setup, and Walkthrough. Retrived (2021, June 10) from https://www.youtube.com/watch?v=HW29067qVWk.

Stackoverflow (2021). Python Libraries. Retrived (2021, Jun 14) from https://stackoverflow.com/.