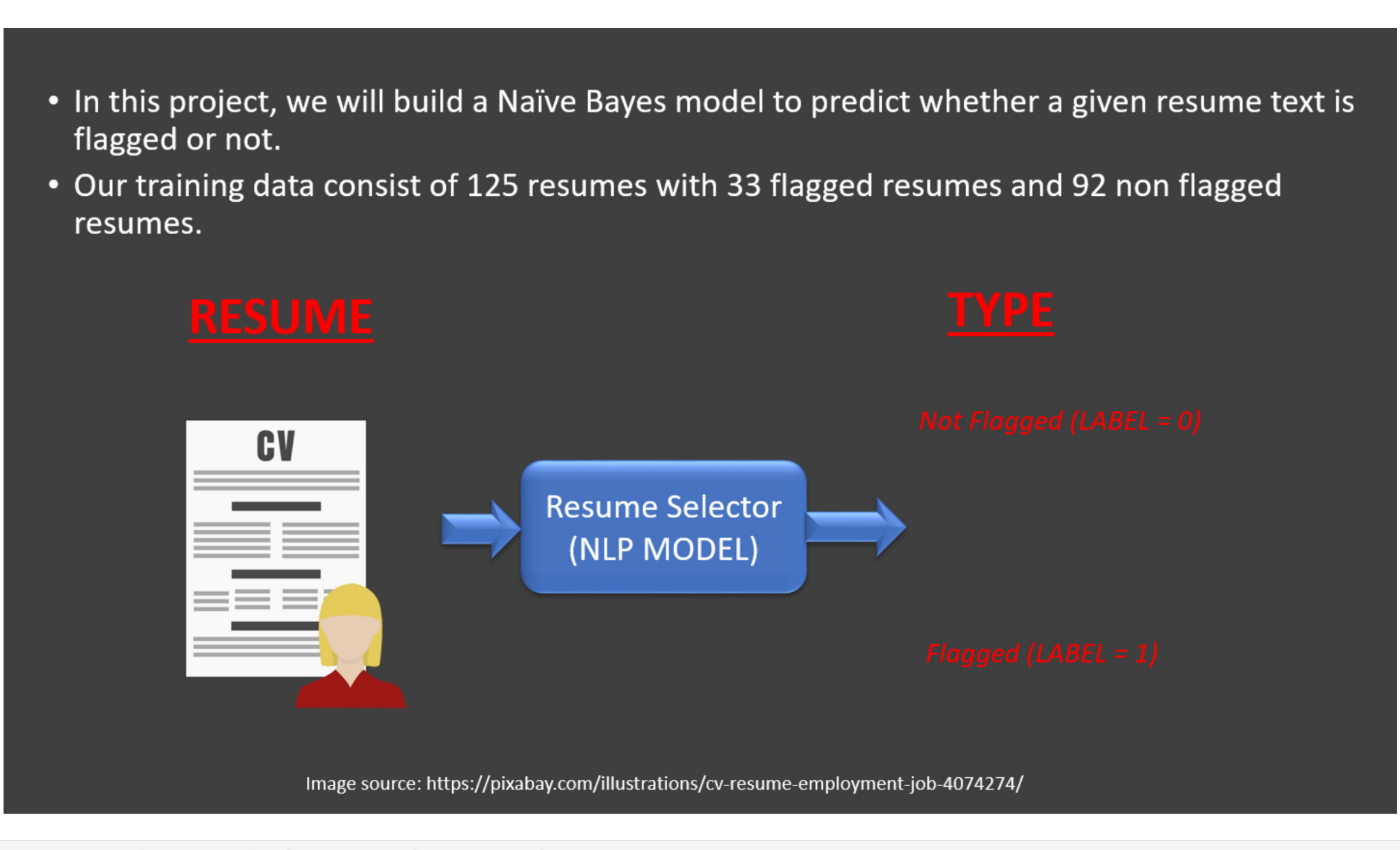
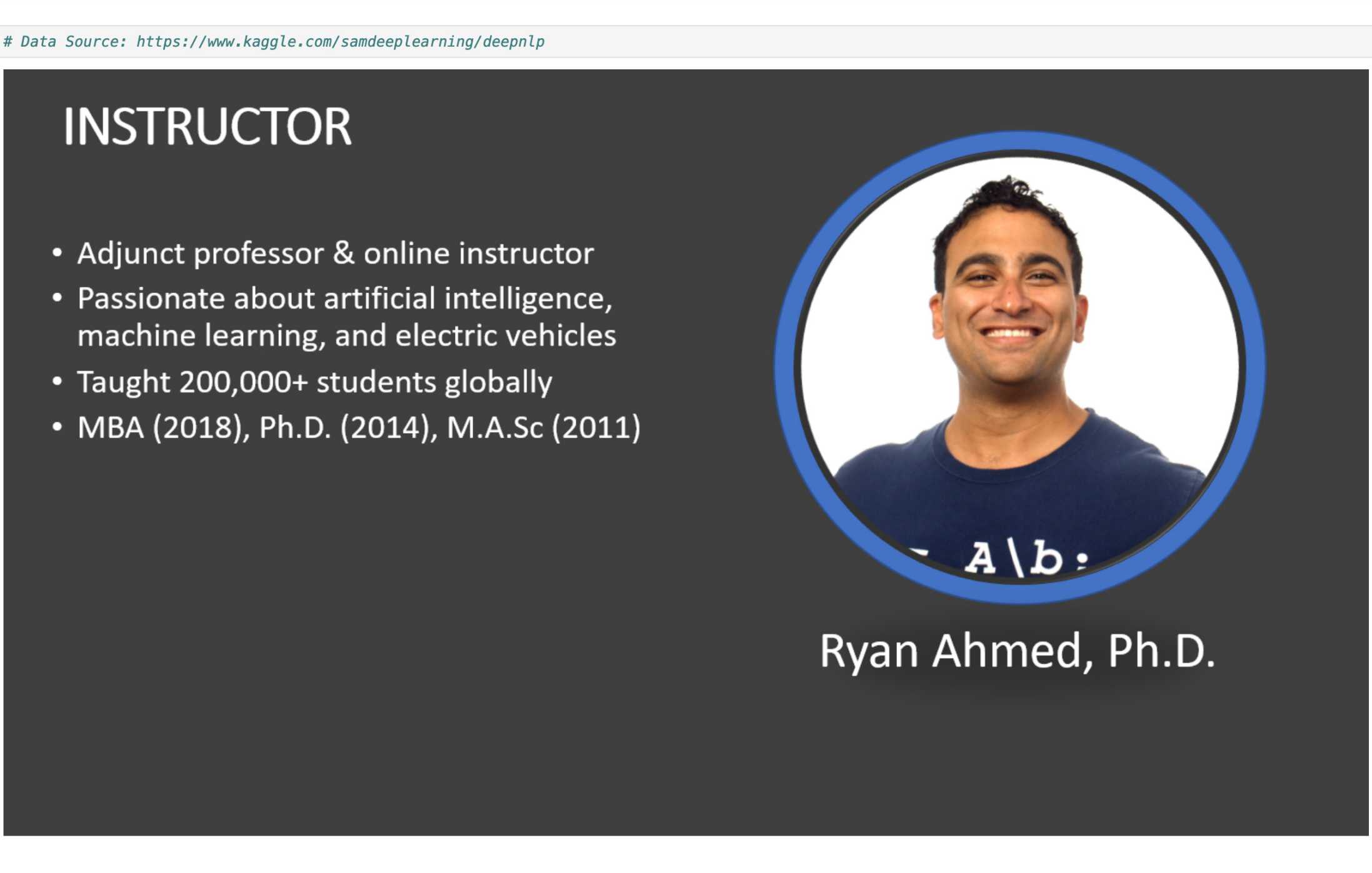
TASK #1: UNDERSTAND THE PROBLEM AND BUSINESS CASE



- In this project, we will build a Naïve Bayes model to predict whether a given resume text is flagged or not.
- Our training data consist of 125 resumes with 33 flagged resumes and 92 non flagged resumes.

Image source: https://pixabay.com/illustrations/cv-resume-employment-job-4074274/

```
In [1]: # Data Source: https://www.kaggle.com/samdeeplearning/deepnlp
```



INSTRUCTOR

- Adjunct professor & online instructor
- Passionate about artificial intelligence, machine learning, and electric vehicles
- Taught 200,000+ students globally
- MBA (2018), Ph.D. (2014), M.A.Sc (2011)

Ryan Ahmed, Ph.D.

TASK #2: IMPORT LIBRARIES AND DATASETS

```
In [2]: # install nltk
        !pip install nltk
```

```
Requirement already satisfied: nltk in /opt/conda/lib/python3.10/site-packages (3.8.1)
Requirement already satisfied: click in /opt/conda/lib/python3.10/site-packages (from nltk) (8.1.3)
Requirement already satisfied: joblib in /opt/conda/lib/python3.10/site-packages (from nltk) (1.2.0)
Requirement already satisfied: regex>=2021.8.3 in /opt/conda/lib/python3.10/site-packages (from nltk) (2024.5.15)
Requirement already satisfied: tqdm in /opt/conda/lib/python3.10/site-packages (from nltk) (4.64.1)

[notice] A new release of pip is available: 24.0 -> 25.2
[notice] To update, run: pip install --upgrade pip
```

```
In [3]: # install gensim
        !pip install gensim
```

```
Requirement already satisfied: gensim in /opt/conda/lib/python3.10/site-packages (4.3.2)
Requirement already satisfied: numpy>=1.18.5 in /opt/conda/lib/python3.10/site-packages (from gensim) (1.23.5)
Requirement already satisfied: scipy>=1.7.0 in /opt/conda/lib/python3.10/site-packages (from gensim) (1.9.3)
Requirement already satisfied: smart-open>=1.8.1 in /opt/conda/lib/python3.10/site-packages (from gensim) (7.0.4)
Requirement already satisfied: wrapt in /opt/conda/lib/python3.10/site-packages (from smart-open>=1.8.1->gensim) (1.14.1)

[notice] A new release of pip is available: 24.0 -> 25.2
[notice] To update, run: pip install --upgrade pip
```

```
In [4]: !pip install wordcloud
```

```
Requirement already satisfied: wordcloud in /opt/conda/lib/python3.10/site-packages (1.9.3)
Requirement already satisfied: numpy>=1.6.1 in /opt/conda/lib/python3.10/site-packages (from wordcloud) (1.23.5)
Requirement already satisfied: pillow in /opt/conda/lib/python3.10/site-packages (from wordcloud) (9.2.0)
Requirement already satisfied: matplotlib in /opt/conda/lib/python3.10/site-packages (from wordcloud) (3.6.2)
Requirement already satisfied: contourpy>=1.0.1 in /opt/conda/lib/python3.10/site-packages (from matplotlib->wordcloud) (1.0.6)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.10/site-packages (from matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in /opt/conda/lib/python3.10/site-packages (from matplotlib->wordcloud) (4.38.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/conda/lib/python3.10/site-packages (from matplotlib->wordcloud) (1.4.4)
Requirement already satisfied: packaging>=20.0 in /opt/conda/lib/python3.10/site-packages (from matplotlib->wordcloud) (21.3)
Requirement already satisfied: pyparsing>=2.2.1 in /opt/conda/lib/python3.10/site-packages (from matplotlib->wordcloud) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in /opt/conda/lib/python3.10/site-packages (from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.10/site-packages (from python-dateutil>=2.7->matplotlib->wordcloud) (1.16.0)

[notice] A new release of pip is available: 24.0 -> 25.2
[notice] To update, run: pip install --upgrade pip
```

```
In [5]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        from wordcloud import WordCloud, STOPWORDS
        import nltk
        from nltk.stem import PorterStemmer, WordNetLemmatizer
        from nltk.corpus import stopwords
        from nltk.tokenize import word_tokenize, sent_tokenize
        import gensim
        from gensim.utils import simple_preprocess
        from gensim.parsing.preprocessing import STOPWORDS
        from sklearn.metrics import classification_report, confusion_matrix
```

```
In [6]: from jupyterthemes import jtplot
        jtplot.style(theme='monokai', context='notebook', ticks=True, grid=False)
        # setting the style of the notebook to be monokai theme
        # this line of code is important to ensure that we are able to see the x and y axes clearly
        # If you don't run this code line, you will notice that the xlabel and ylabel on any plot is black on black and it will be hard to see them.
```

```
In [7]: # load the data
        resume_df = pd.read_csv("resume.csv", encoding = 'latin-1')
        resume_df
```

Out[7]:

	resume_id	class	resume_text
0	resume_1	not_flagged	\rCustomer Service Supervisor/Tier - Isabella ...
1	resume_2	not_flagged	\rEngineer / Scientist - IBM Microelectronics ...
2	resume_3	not_flagged	\rLTS Software Engineer Computational Lithogra...
3	resume_4	not_flagged	TUTOR\rWilliston VT - Email me on Indeed: ind...
4	resume_5	flagged	\rIndependent Consultant - Self-employed\rBurl...
...	...	...	...
120	resume_121	not_flagged	\rBrattleboro VT - Email me on Indeed: indeed....
121	resume_122	not_flagged	\rResearch and Teaching Assistant - University...
122	resume_123	not_flagged	\rMedical Coder - Highly Skilled - Entry Level...
123	resume_124	flagged	\rWaterbury VT - Email me on Indeed: indeed.co...
124	resume_125	not_flagged	\rResearch and Development Scientist - Burling...

125 rows × 3 columns

```
In [8]: # data containing resume
        resume_df = resume_df[['resume_text', 'class']]
        resume_df
```

Out[8]:

	resume_text	class
0	\rCustomer Service Supervisor/Tier - Isabella ...	not_flagged
1	\rEngineer / Scientist - IBM Microelectronics ...	not_flagged
2	\rLTS Software Engineer Computational Lithogra...	not_flagged
3	TUTOR\rWilliston VT - Email me on Indeed: ind...	not_flagged
4	\rIndependent Consultant - Self-employed\rBurl...	flagged
...	...	...
120	\rBrattleboro VT - Email me on Indeed: indeed....	not_flagged
121	\rResearch and Teaching Assistant - University...	not_flagged
122	\rMedical Coder - Highly Skilled - Entry Level...	not_flagged
123	\rWaterbury VT - Email me on Indeed: indeed.co...	flagged
124	\rResearch and Development Scientist - Burling...	not_flagged

125 rows × 2 columns

MINI CHALLENGE #1:

- Print the first and last elements in the dataframe.

```
In [9]: resume_df.head(10)
```

Out[9]:

	resume_text	class
0	\rCustomer Service Supervisor/Tier - Isabella ...	not_flagged
1	\rEngineer / Scientist - IBM Microelectronics ...	not_flagged
2	\rLTS Software Engineer Computational Lithogra...	not_flagged
3	TUTOR\rWilliston VT - Email me on Indeed: ind...	not_flagged
4	\rIndependent Consultant - Self-employed\rBurl...	flagged
5	\rPoultney VT - Email me on Indeed: indeed.com...	not_flagged
6	\rMedical Laboratory Scientist (Special Chemis...	not_flagged
7	Statistician\rBurlington VT - Email me on Inde...	flagged
8	Research technician\rBurlington VT - Email me ...	not_flagged
9	Barbara Hennessey-Elzohairy\rNewport VT - Emai...	not_flagged

```
In [10]: resume_df.tail(10)
```

Out[10]:

	resume_text	class
115	\rEssex Jct VT - Email me on Indeed: indeed.co...	flagged
116	Postdoctoral Associate - University of Vermon...	not_flagged
117	\rSpatial Analyst and UAV Flight Operator - Sp...	flagged
118	| Bioinformatics Analyst\rBurlington VT - Ema...	flagged
119	Professional\rNewbury VT - Email me on Indeed:...	not_flagged
120	\rBrattleboro VT - Email me on Indeed: indeed....	not_flagged
121	\rResearch and Teaching Assistant - University...	not_flagged
122	\rMedical Coder - Highly Skilled - Entry Level...	not_flagged
123	\rWaterbury VT - Email me on Indeed: indeed.co...	flagged
124	\rResearch and Development Scientist - Burling...	not_flagged

TASK #3: PERFORM EXPLORATORY DATA ANALYSIS

```
In [11]: # obtain dataframe information
         resume_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 125 entries, 0 to 124
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   resume_text  125 non-null    object
 1   class        125 non-null    object
dtypes: object(2)
memory usage: 2.1+ KB
```

```
In [12]: # check for null values
         resume_df.isnull().sum()
```

```
Out[12]: resume_text    0
         class          0
         dtype: int64
```

```
In [13]: resume_df['class'].value_counts()
```

```
Out[13]: not_flagged    92
         flagged        33
         Name: class, dtype: int64
```

```
In [14]: resume_df['class'] = resume_df['class'].apply(lambda x:1 if x == 'flagged' else 0)
         resume_df
```

Out[14]:

	resume_text	class
0	\rCustomer Service Supervisor/Tier - Isabella ...	0
1	\rEngineer / Scientist - IBM Microelectronics ...	0
2	\rLTS Software Engineer Computational Lithogra...	0
3	TUTOR\rWilliston VT - Email me on Indeed: ind...	0
4	\rIndependent Consultant - Self-employed\rBurl...	1
...	...	...
120	\rBrattleboro VT - Email me on Indeed: indeed....	0
121	\rResearch and Teaching Assistant - University...	0
122	\rMedical Coder - Highly Skilled - Entry Level...	0
123	\rWaterbury VT - Email me on Indeed: indeed.co...	1
124	\rResearch and Development Scientist - Burling...	0

125 rows × 2 columns

MINI CHALLENGE #2:

- Divide the DataFrame into two, one that belongs to class 0 and 1. Do we have a balanced dataset?

```
In [15]: class_0_df = resume_df[resume_df['class'] == 0]
         class_0_df
```

Out[15]:

	resume_text	class
0	\rCustomer Service Supervisor/Tier - Isabella ...	0
1	\rEngineer / Scientist - IBM Microelectronics ...	0
2	\rLTS Software Engineer Computational Lithogra...	0
3	TUTOR\rWilliston VT - Email me on Indeed: ind...	0
5	\rPoultney VT - Email me on Indeed: indeed.com...	0
...	...	...
119	Professional\rNewbury VT - Email me on Indeed:...	0
120	\rBrattleboro VT - Email me on Indeed: indeed....	0
121	\rResearch and Teaching Assistant - University...	0
122	\rMedical Coder - Highly Skilled - Entry Level...	0
124	\rResearch and Development Scientist - Burling...	0

92 rows × 2 columns

```
In [16]: class_1_df = resume_df[ resume_df['class'] == 1]
         class_1_df
```

Out[16]:

	resume_text	class
4	\rIndependent Consultant - Self-employed\rBurl...	1
7	Statistician\rBurlington VT - Email me on Inde...	1
10	Barbara Kurth\rResearch Assistant Professor Cl...	1
15	Caroline Clauson Work Study Student\rBrattlebo...	1
17	Chelsea Martin\rEnvironmental Scientist - Vana...	1
18	Christopher Fusting Data Science Consultant\rB...	1
23	David Grass\rCity Research Scientist - Environ...	1
24	David Lucero Public Health Analyst\rBurlington...	1
27	Drew Burkhard\rEnvironmental Scientist Meets M...	1
31	Eric Anderson Data Scientist\rBrattleboro VT -...	1
32	Eric Hedeman\rBurlington VT - Email me on Inde...	1
33	Erik McCullen Senior Scientist/Engineer\rFairf...	1
38	Geoffrey Abbott Software / Research Engineer\r...	1
39	George Gallant Staff Engineer / Scientist\rUnd...	1
42	Guang Zeng\rResearch Scientist/Imaging Special...	1
49	Jake Kittell\rPrincipal - Happiness Tool Co. I...	1
52	John Klein\rCastleton VT - Email me on Indeed:...	1
56	Joseph Lea\rUndergraduate Research Assistant -...	1
60	Kenneth Sikora Entrepreneur research scientist...	1
70	Mark Mckenna\rChief Scientist Unmanned Systems...	1
75	Melissa Burbank\rBrattleboro VT - Email me on ...	1
79	Michael Fink\rData Scientist/TPM - Smart Resou...	1
87	Nisha Chaube Graduate Teaching Assistant\rColc...	1
101	Shannon Warburton Senior Scientist - MERCK AND...	1
104	Steven Brady\rPostdoctoral Fellow - VT Coopera...	1
106	Swaminathan Prasanna Senior UI Developer - ATS...	1
107	Tam Tran\rShelburne VT - Email me on Indeed: i...	1
112	William Sribney\rStatistical Software Develope...	1
114	Ph.D.\rSouth Royalton VT - Email me on Indeed...	1
115	\rEssex Jct VT - Email me on Indeed: indeed.co...	1
117	\rSpatial Analyst and UAV Flight Operator - Sp...	1
118	| Bioinformatics Analyst\rBurlington VT - Ema...	1
123	\rWaterbury VT - Email me on Indeed: indeed.co...	1

TASK #4: PERFORM DATA CLEANING

```
In [17]: resume_df['resume_text'] = resume_df['resume_text'].apply(lambda x: x.replace('\r', ''))
         resume_df
```

Out[17]:

	resume_text	class
0	Customer Service Supervisor/Tier - Isabella Ca...	0
1	Engineer / Scientist - IBM Microelectronics Di...	0
2	LTS Software Engineer Computational Lithograph...	0
3	TUTORWilliston VT - Email me on Indeed: indee...	0
4	Independent Consultant - Self-employedBurlingt...	1
...	...	...
120	Brattleboro VT - Email me on Indeed: indeed.co...	0
121	Research and Teaching Assistant - University o...	0
122	Medical Coder - Highly Skilled - Entry LevelSu...	0
123	Waterbury VT - Email me on Indeed: indeed.com/...	1
124	Research and Development Scientist - Burlingto...	0

125 rows × 2 columns

```
In [18]: # download nltk packages
         nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /home/jovyan/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
Out[18]: True
```

```
In [19]: # download nltk packages
         nltk.download("stopwords")
```

```
[nltk_data] Downloading package stopwords to /home/jovyan/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
Out[19]: True
```

```
In [20]: # Get additional stopwords from nltk
         from nltk.corpus import stopwords
         stop_words = stopwords.words('english')
         stop_words.extend(['from', 'subject', 're', 'edu', 'use', 'email', 'com'])
```

```
In [21]: # Remove stop words and remove words with 2 or less characters
         def preprocess(text):
             result = []
             for token in gensim.utils.simple_preprocess(text) :
                 if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) > 2 and token not in stop_words:
                     result.append(token)

             return ' '.join(result)
```

```
In [22]: # Cleaned text
         resume_df['cleaned'] = resume_df['resume_text'].apply(preprocess)
```

```
In [23]: resume_df
```

Out[23]:

	resume_text	class	cleaned
0	Customer Service Supervisor/Tier - Isabella Ca...	0	customer service supervisor tier isabella cata...
1	Engineer / Scientist - IBM Microelectronics Di...	0	engineer scientist ibm albert gregoritsch ecaw...
2	LTS Software Engineer Computational Lithograph...	0	lts software engineer computational lithograph...
3	TUTORWilliston VT - Email me on Indeed: indee...	0	tutorwilliston alec schwartz awork college bio...
4	Independent Consultant - Self-employedBurlingt...	1	independent consultant self alex reutter fefwo...
...	...	...	...
120	Brattleboro VT - Email me on Indeed: indeed.co...	0	brattleboro bcc skilled presenter trainer micr...
121	Research and Teaching Assistant - University o...	0	research teaching assistant university cdd gra...
122	Medical Coder - Highly Skilled - Entry LevelSu...	0	medical coder highly skilled entry levelsudbur...
123	Waterbury VT - Email me on Indeed: indeed.com/...	1	waterbury bec fcwilling relocate work employer...
124	Research and Development Scientist - Burlingto...	0	research development scientist burlington cda ...

125 rows × 3 columns

```
In [24]: print(resume_df['cleaned'][0])
```

```
customer service supervisor tier isabella catalog companysouth burlington aecf work service supervisor tierisabella catalog company shelburne august present
customer service visual set display website maintenance supervise customer service team popular catalog company manage day day issues resolution customer ups
et ensure customer satisfaction troubleshoot order shipping issues lost transit order errors damages manage resolve escalated customer calls ensure customer
satisfaction assist customers order placing cross selling upselling catalog merchandise set display sample merchandise catalog library customer pick area fac
ility website clean adding images type product information assistant events coordinator office services assistanteileen fisher irvington february july suppor
t director architecture architecture coordinator daily activities including preparing monthly expense reports scheduling calendar maintenance arranging aspec
ts travel logistics catering interior design research projects manage event set ups entire process eileen fisher corporate locations catering overseeing set
walk space client review event forms facilities team daily management professional calendars require heavy scheduling office services include companywide roo
m reservations office supply orders filtered calls chief creative officer owner companytemp white plains december february office services assistant receptio
nist managed heavy volume orthopedic specialty benefit management company directed heavy daily incoming mail flow processed daily checks entered data excel g
enerate totals accounting personal january december home office assistant personal assistant provided professional office support established psychologists n
ew york area carefully handled personal confidential patient information organized uncluttered simplified office space create user friendly atmosphere coordi
nated researched travel related details flights hotels visas cars managed personal errands phone calls emails responsible mail processing bank deposits psych
ologists service representative account managercm almy sons greenwich january january greenwich january january customer service representative account manag
er provided high level customer service clergy church members denominations answered heavy volume assisted customers highly efficient manor assisted customer
s overall design garments final decision making church item purchases managed maintained large account database daily phone calls customer accounts responsib
le tracking large shipments replacement lost damaged items dorset street south burlington aimeerblair gmail assistant chief financial salute americas heroes
ossining january january ossining january january administrative assistant chief financial officer interviewed military veterans families considered financia
l aid reviewed highly confidential database candidate mediated discussions military veterans collectors arrange final payouts debt incurred time injury final
ized paperwork award payouts coordinated travel logistics large sponsored events assisted disabled veterans events provided basic administrative assistant sa
les team trade coordinatorleo electron microscopy thornwood august thornwood august administrative assistant sales team trade coordinator communicated genera
l information provided quotes high end buyers worked closely team sales associates arranging meetings potential buyers prepared final proposals closing sale
information purchased electron microscopes arranged aspects travel logistics trade shows united states canada attended trade shows sales associates scientist
s insure electron microscopes arrived safely set assisted demonstrations close sales trade floorartist charles fazzino pop artistcharles fazzino new rochelle
freelance artist assembled dimensional piece art weekly basis home office responsible detailed finishing work making pieces presentable purchase galleries wo
rld visual artswestchester community college new york school years combined office services focus administrative assistance customer service event coordinati
on trade coordination facilitating
```

```
In [25]: print(resume_df['resume_text'][0])
```

```
Customer Service Supervisor/Tier - Isabella Catalog CompanySouth Burlington VT - Email me on Indeed: indeed.com/r//49f8c9aecf490d26WORK EXPERIENCECustomer Se
rvice Supervisor/TierIsabella Catalog Company - Shelburne VT - August 2015 to Present2 Customer Service/Visual Set Up & Display/Website Maintenance¢ Supervis
e customer service team of a popular catalog company¢ Manage day to day issues and resolution of customer upset to ensure customer satisfaction¢ Troubleshoot
order and shipping issues: lost in transit order errors damages¢ Manage and resolve escalated customer calls to ensure customer satisfaction¢ Assist customer
s with order placing cross-selling/upselling of catalog merchandise¢ Set up and display of sample merchandise in catalog library as well as customer pick-up
area of the facility ¢ Website clean-up: adding images type up product information proofreadingAdministrative Assistant /Events Coordinator/Office Services A
ssistantEileen Fisher Inc - Irvington NY - February 2014 to July 2015Support to Director of Architecture and Architecture Coordinator in all daily activities
including: preparing monthly expense reports scheduling calendar maintenance arranging all aspects of travel/logistics catering interior design research proj
ects¢ Manage event set ups through entire process for two Eileen Fisher corporate locations¢ Catering overseeing set up walk-thru of space with client review
event forms with facilities team ¢ Daily management of two professional calendars that require heavy scheduling¢ Office services that include: companywide ro
om reservations office supply orders¢ Filtered calls to the Chief Creative Officer/Owner of the companyTemp AssignmentOrthoNet - White Plains NY - December 2
013 to February 2014Office Services Assistant/Receptionist¢ Managed heavy call volume for orthopedic specialty benefit management company¢ Directed heavy dai
ly incoming mail flow¢ Processed daily checks and entered data into Excel to generate totals for accounting reportsExecutive Personal AssistantWestchester NY
- January 2012 to December 2013Home Office Assistant/ Personal Assistant¢ Provided professional office support to three established Psychologists in the New
York area ¢ Carefully handled personal and confidential patient information¢ Organized uncluttered and simplified office space to create a more user-friendly
atmosphere ¢ Coordinated and researched all travel related details (flights hotels visas cars etc.)¢ Managed personal errands phone calls and emails.¢ Respon
sible for mail processing and bank deposits while Psychologists were travelingCustomer Service Representative/ Account ManagerCM Almy & Sons Inc - Greenwich
CT - January 2007 to January 2012 Greenwich CT January 2007 - January 2012Customer Service Representative/ Account Manager¢ Provided a high level of customer
service to clergy and church members of all denominations¢ Answered heavy call volume and assisted customers in a highly efficient manor¢ Assisted customers
with overall design of garments final decision making of church item purchases ¢ Managed and maintained a large account database with daily phone calls to cu
stomer accounts¢ Responsible for tracking large shipments and also replacement of lost or damaged items435 Dorset Street * South Burlington VT 05403
_ 914.564.4381 _ Aimeerblair319@gmail.comAdministrative Assistant to Chief Financial OfficerCoalition to Salute Americas Heroes - Ossining NY - January 2005
to January 2007Ossining NY January 2005 - January 2007Administrative Assistant to Chief Financial Officer¢ Interviewed military veterans and their families t
o be considered for financial aid ¢ Reviewed a highly confidential database for candidate¢ Mediated discussions between military veterans and collectors¢ Arr
ange final payouts for debt incurred during time of injury¢ Finalized paperwork for award payouts¢ Coordinated travel and logistics for large sponsored event
s¢ Assisted disabled veterans during events¢ Provided basic administrative supportAdministrative Assistant to Sales Team/ Trade Show CoordinatorLeo Electron
Microscopy - Thornwood NY - May 2000 to August 2003Thornwood NY May 2000 - August 2003Administrative Assistant to Sales Team/ Trade Show Coordinator¢ Communi
cated general information and provided quotes to high end buyers¢ Worked closely with a team of sales associates arranging meetings with potential buyers¢ Pr
epared final proposals and closing sale information on purchased electron microscopes¢ Arranged all aspects of travel and logistics for trade shows within th
e United States and Canada.¢ Attended trade shows with sales associates and scientists to insure all electron microscopes arrived safely for set up¢ Assisted
with demonstrations and close of sales on trade show floorArtist Charles Fazzino 3D Pop ArtistCharles Fazzino - New Rochelle NY - 1993 to 1996and 2003-2005Fr
eelance Artist¢ Assembled 3 dimensional piece-art on a weekly basis from home office¢ Responsible for detailed finishing work and making pieces presentable f
or purchase in galleries world-wideEDUCATIONAAS in Visual ArtsWestchester Community College - New York NY School knowledgeADDITIONAL INFORMATIONProviding mor
e than 15 years of combined office services with a focus on Administrative Assistance Customer Service Event Coordination Trade Show Coordination and Facilit
ating
```

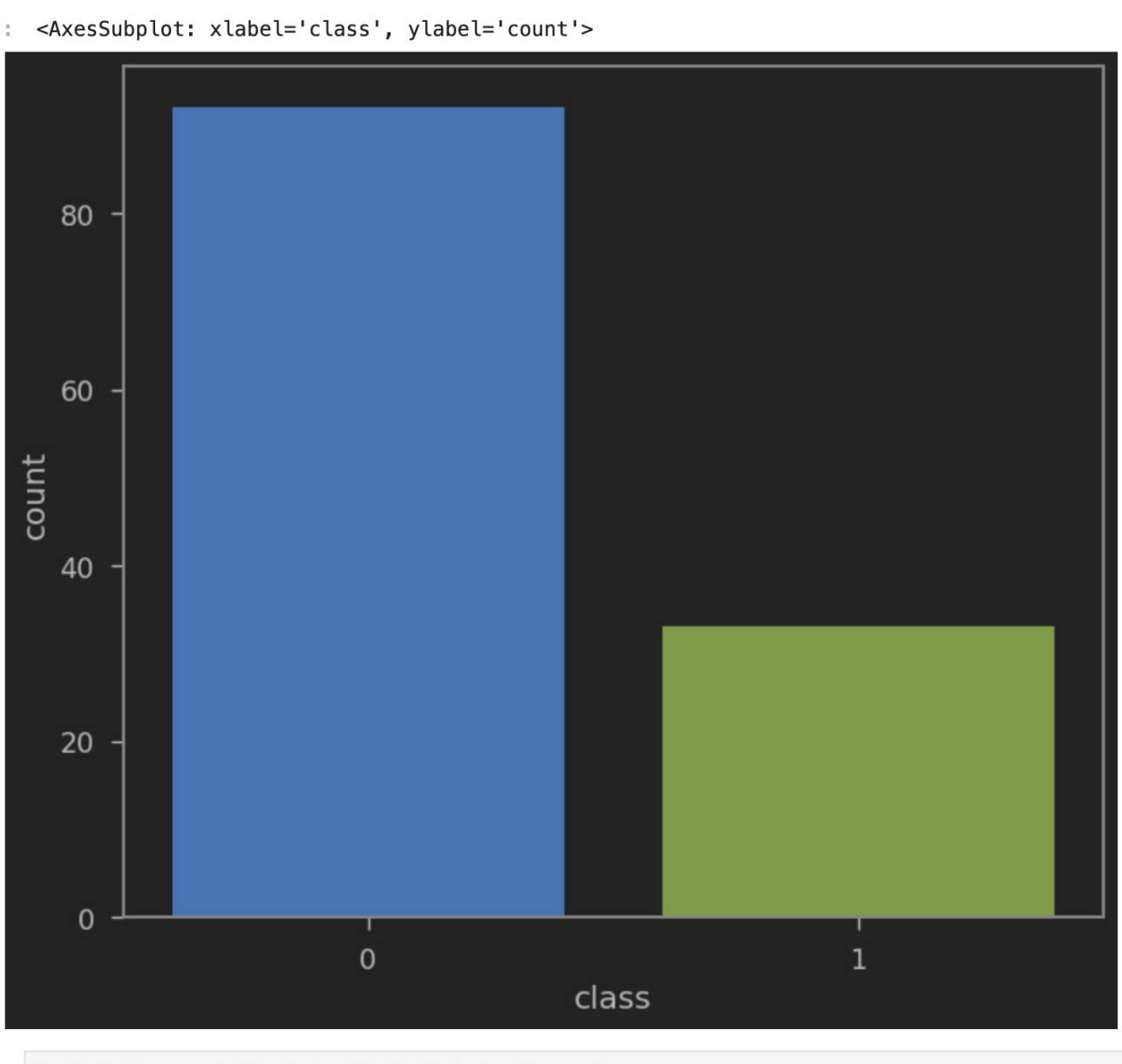
TASK #5: VISUALIZE CLEANED DATASET

```
In [26]: resume_df['class'].value_counts()
```

```
Out[26]: 0    92
         1    33
         Name: class, dtype: int64
```

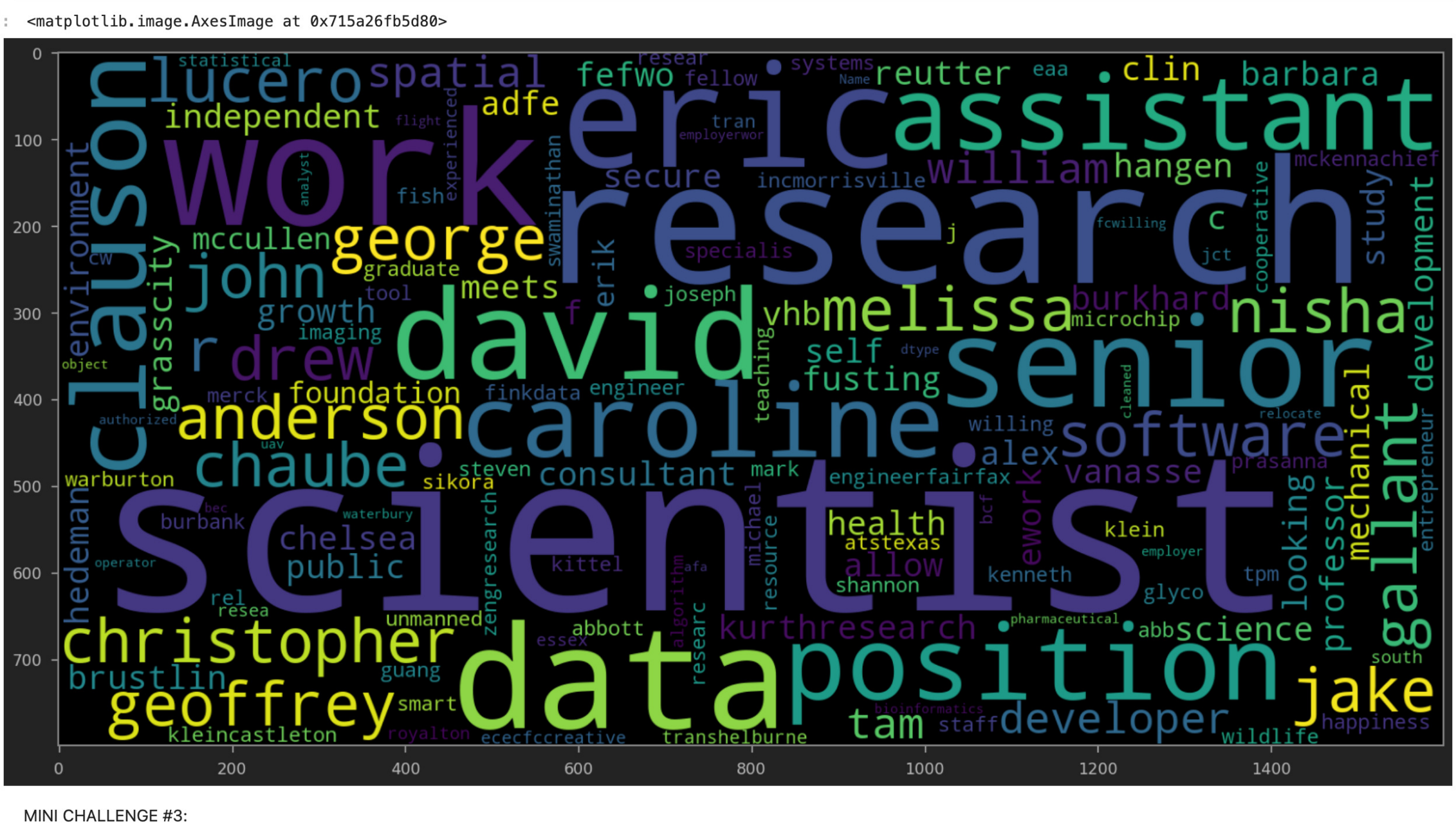
```
In [27]: # Plot the counts of flagged vs not flagged
         sns.countplot(x='class', data=resume_df, label = 'Count Plot')
```

```
Out[27]: <AxesSubplot: xlabel='class', ylabel='count'>
```



```
In [28]: # plot the word cloud for text that is flagged
         plt.figure(figsize = (20,20))
         wc = WordCloud(max_words = 2000, width = 1600, height = 800, stopwords = stop_words).generate(str(resume_df[resume_df['class']==1].cleaned))
         plt.imshow(wc)
```

```
Out[28]: <matplotlib.image.AxesImage at 0x715a26fb5d80>
```

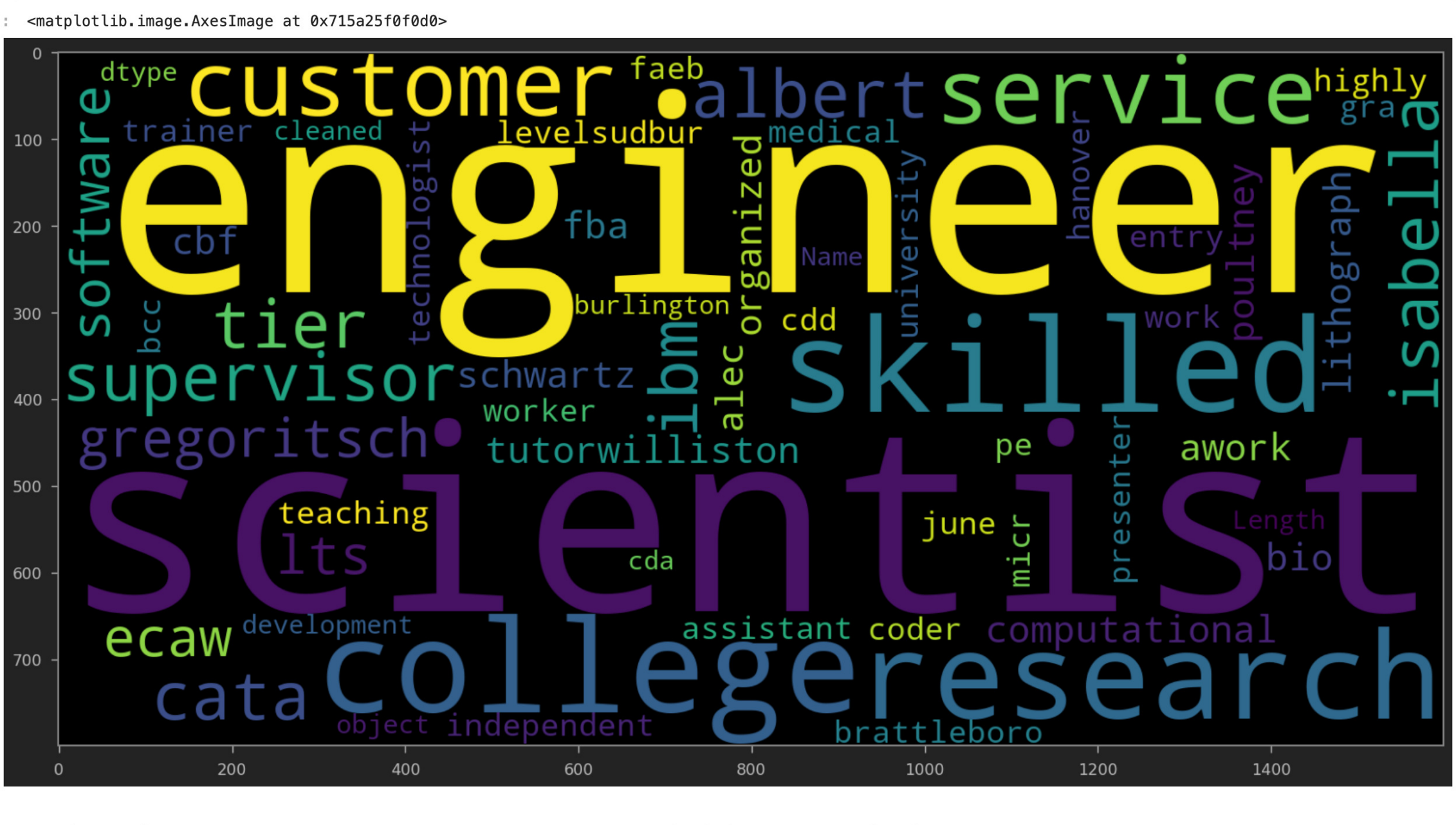


MINI CHALLENGE #3:

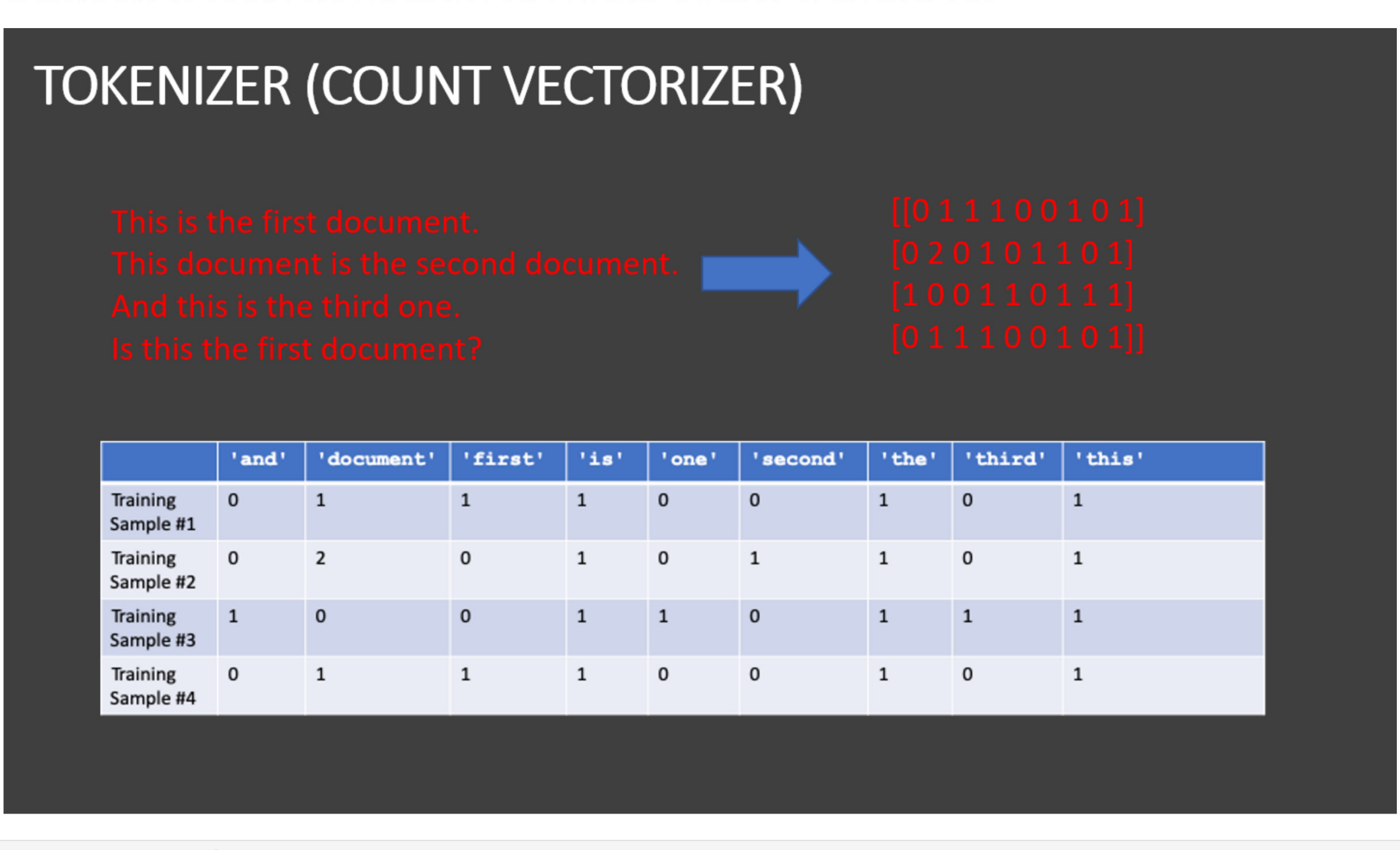
- Plot the wordcloud for class #0

```
In [29]: plt.figure(figsize = (20,20))
         wc = WordCloud(max_words = 2000, width = 1600, height = 800, stopwords = stop_words).generate(str(resume_df[resume_df['class']==0].cleaned))
         plt.imshow(wc)
```

```
Out[29]: <matplotlib.image.AxesImage at 0x715a25f0f0d0>
```



TASK #6: PREPARE THE DATA BY APPLYING COUNT VECTORIZER



TOKENIZER (COUNT VECTORIZER)

This is the first document.
This document is the second document.
And this is the third one.
Is this the first document?

[[0 1 1 1 0 0 1 0 1]
[0 2 0 1 0 1 1 0 1]
[1 0 0 1 1 0 1 1 1]
[0 1 1 1 0 0 1 0 1]]

	'and'	'document'	'first'	'is'	'one'	'second'	'the'	'third'	'this'
Training Sample #1	0	1	1	1	0	0	1	0	1
Training Sample #2	0	2	0	1	0	1	1	0	1
Training Sample #3	1	0	0	1	1	0	1	1	1
Training Sample #4	0	1	1	1	0	0	1	0	1

```
In [30]: # CountVectorizer example
         from sklearn.feature_extraction.text import CountVectorizer
         sample_data = ['This is the first document.','This document is the second document.','And this is the third one.','Is this the first document?']

         vectorizer = CountVectorizer()
         X = vectorizer.fit_transform(sample_data)
```

```
In [31]: print(vectorizer.get_feature_names_out())
```

```
['and' 'document' 'first' 'is' 'one' 'second' 'the' 'third' 'this']
```

```
In [32]: print(X.toarray())
```

```
[[0 1 1 1 0 0 1 0 1]
 [0 2 0 1 0 1 1 0 1]
 [1 0 0 1 1 0 1 1 1]
 [0 1 1 1 0 0 1 0 1]]
```

```
In [33]: # Applying CountVectorizer to the cleaned text
         vectorizer = CountVectorizer()
         countvectorizer = vectorizer.fit_transform(resume_df['cleaned'])
```

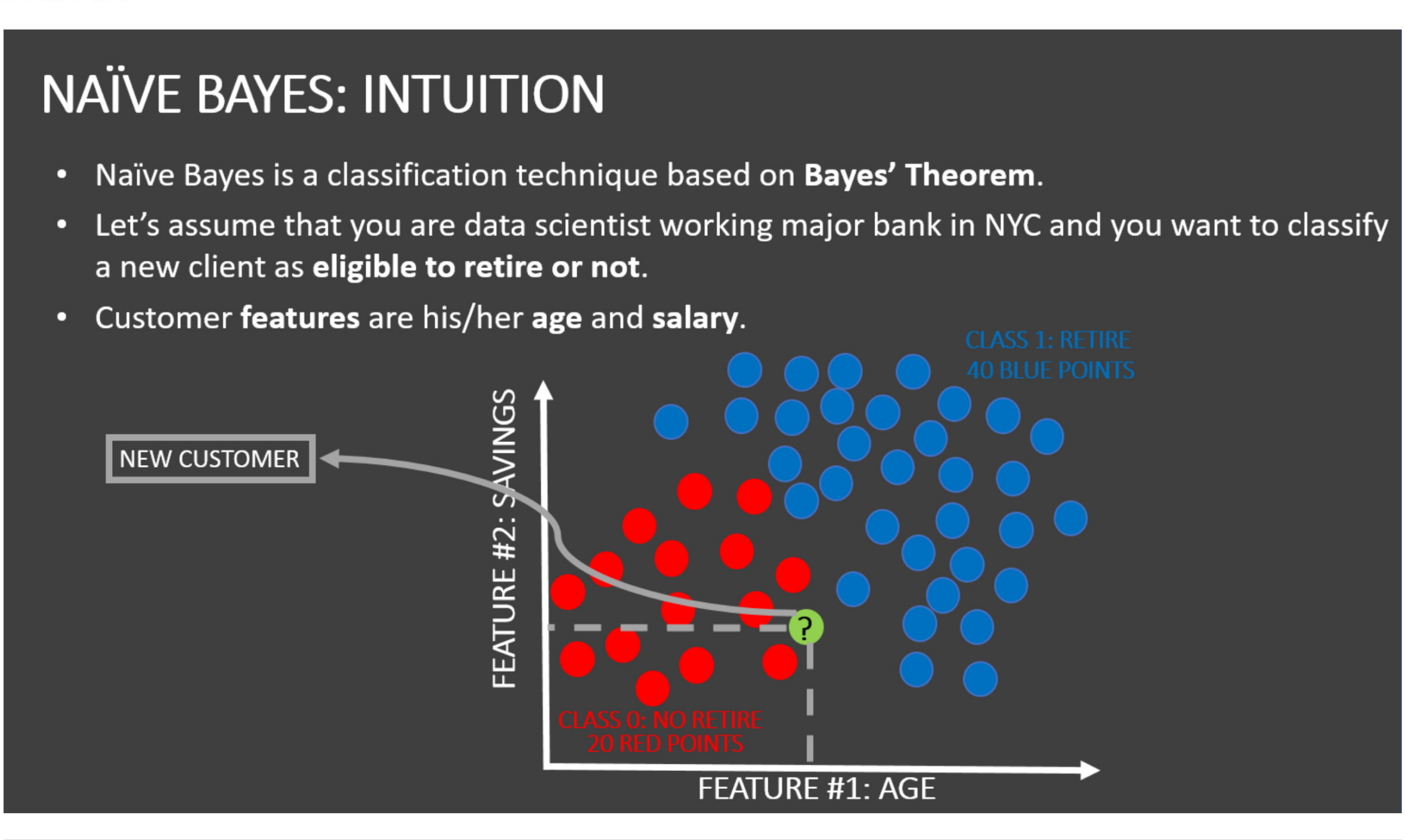
```
In [34]: print(vectorizer.get_feature_names_out())
```

```
['aaalac' 'aabb' 'aac' ... 'ãæcomputer' 'ètravel' 'ô_torrent']
```

```
In [35]: print(countvectorizer.toarray())
```

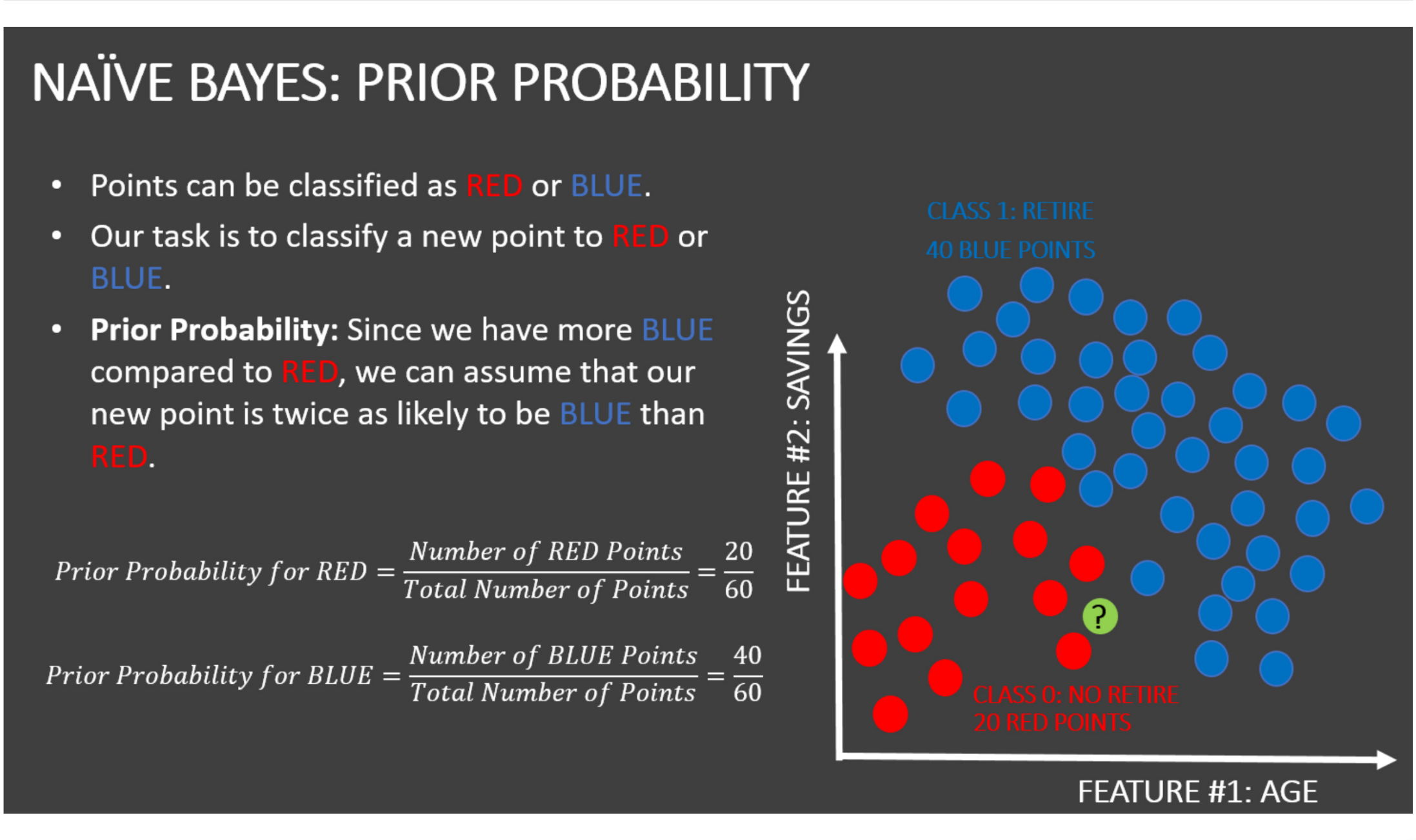
```
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

TASK #7: UNDERSTAND THE THEORY AND INTUITION BEHIND NAIVE BAYES CLASSIFIERS - PART #1



NAÏVE BAYES: INTUITION

- Naïve Bayes is a classification technique based on **Bayes' Theorem**.
- Let's assume that you are data scientist working major bank in NYC and you want to classify a new client as **eligible to retire or not**.
- Customer **features** are his/her **age** and **salary**.

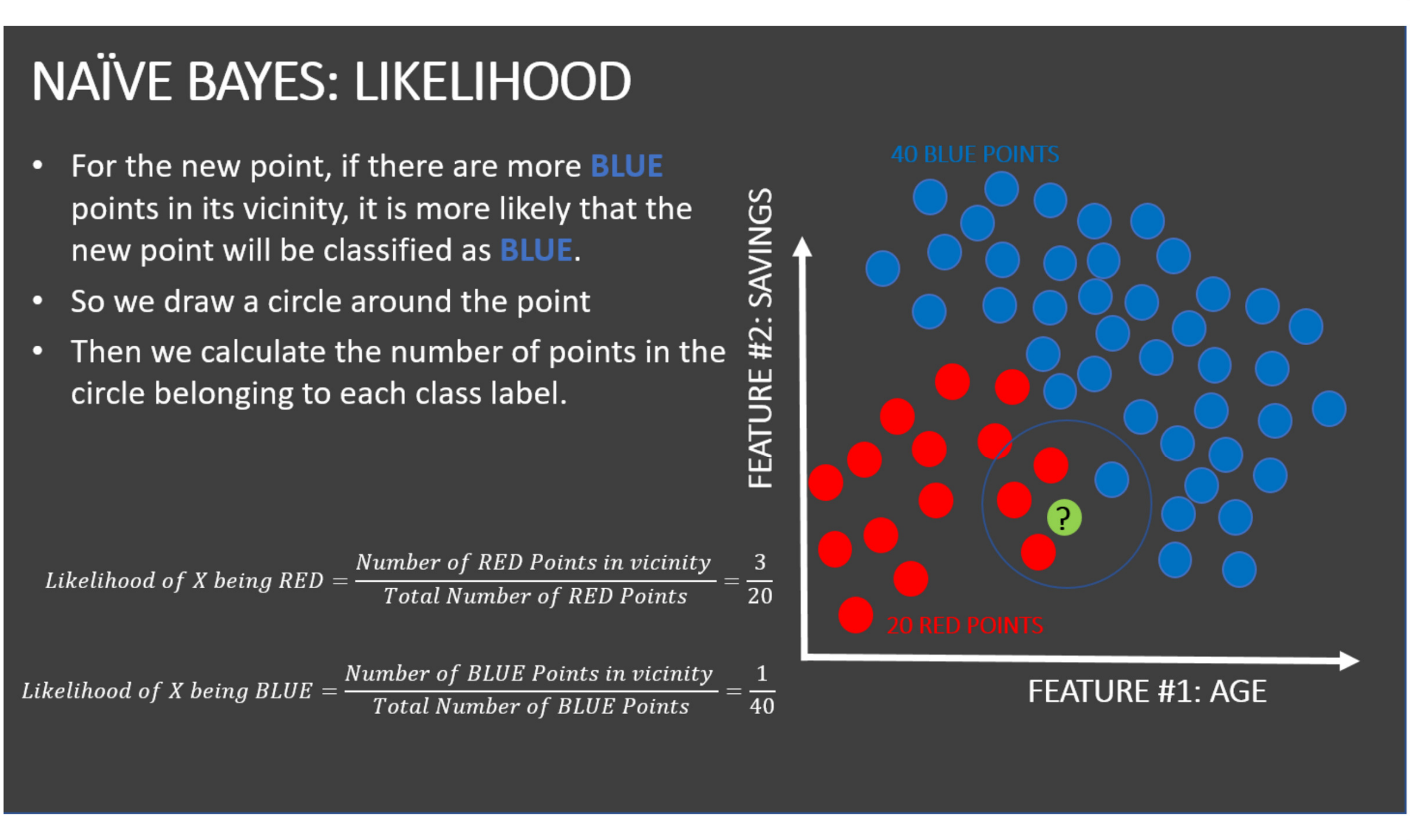


NAÏVE BAYES: PRIOR PROBABILITY

- Points can be classified as RED or BLUE.
- Our task is to classify a new point to RED or BLUE.
- **Prior Probability:** Since we have more BLUE compared to RED, we can assume that our new point is twice as likely to be BLUE than RED.

$$Prior\ Probability\ for\ RED = \frac{Number\ of\ RED\ Points}{Total\ Number\ of\ Points} = \frac{20}{60}$$

$$Prior\ Probability\ for\ BLUE = \frac{Number\ of\ BLUE\ Points}{Total\ Number\ of\ Points} = \frac{40}{60}$$

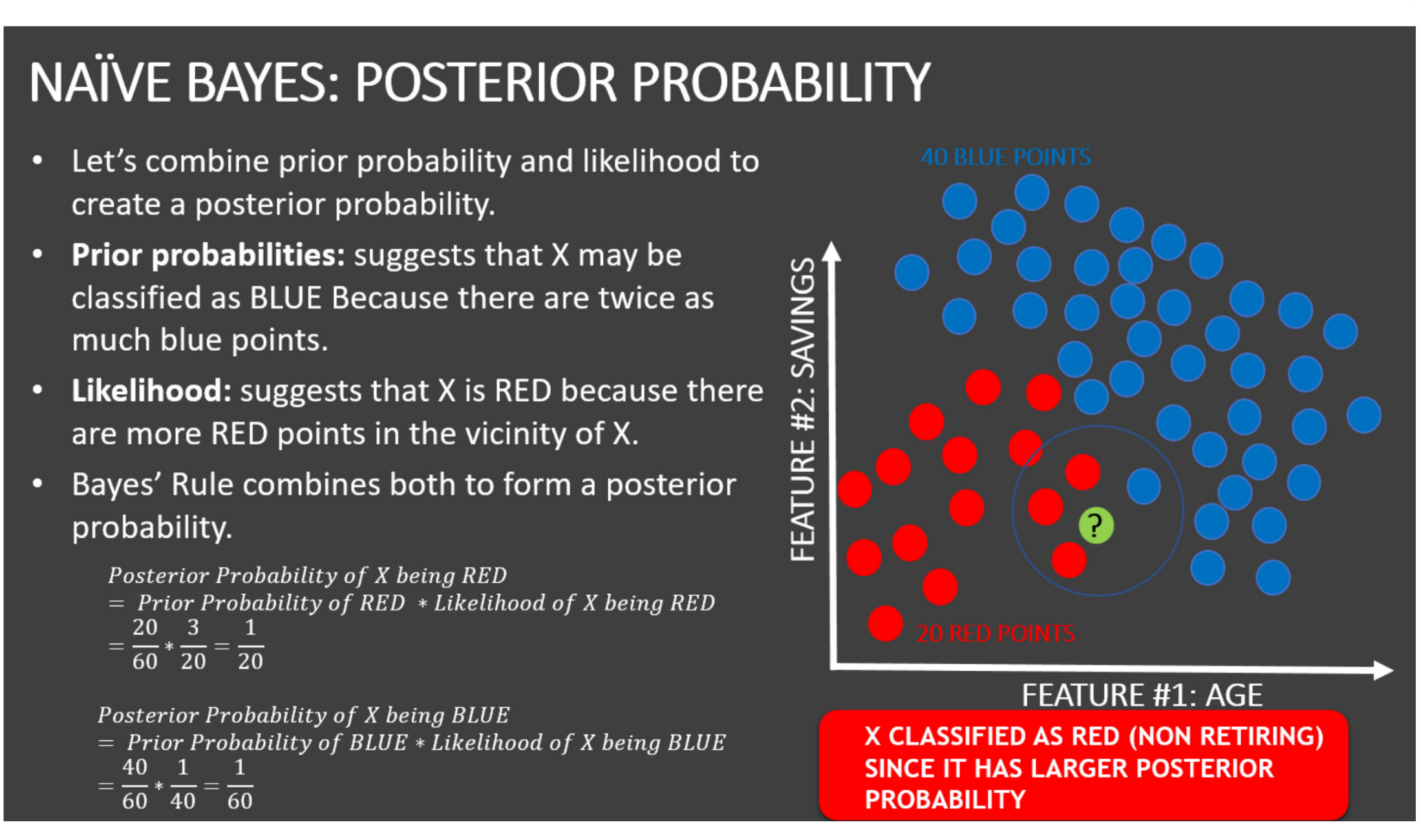


NAÏVE BAYES: LIKELIHOOD

- For the new point, if there are more **BLUE** points in its vicinity, it is more likely that the new point will be classified as **BLUE**.
- So we draw a circle around the point
- Then we calculate the number of points in the circle belonging to each class label.

$$Likelihood\ of\ X\ being\ RED = \frac{Number\ of\ RED\ Points\ in\ vicinity}{Total\ Number\ of\ RED\ Points} = \frac{3}{20}$$

$$Likelihood\ of\ X\ being\ BLUE = \frac{Number\ of\ BLUE\ Points\ in\ vicinity}{Total\ Number\ of\ BLUE\ Points} = \frac{1}{40}$$



NAÏVE BAYES: POSTERIOR PROBABILITY

- Let's combine prior probability and likelihood to create a posterior probability.
- **Prior probabilities:** suggests that X may be classified as BLUE Because there are twice as much blue points.
- **Likelihood:** suggests that X is RED because there are more RED points in the vicinity of X.
- Bayes' Rule combines both to form a posterior probability.

$$Posterior\ Probability\ of\ X\ being\ RED$$
$$= Prior\ Probability\ of\ RED\ * Likelihood\ of\ X\ being\ RED$$
$$= \frac{20}{60} * \frac{3}{20} = \frac{1}{20}$$

$$Posterior\ Probability\ of\ X\ being\ BLUE$$
$$= Prior\ Probability\ of\ BLUE * Likelihood\ of\ X\ being\ BLUE$$
$$= \frac{40}{60} * \frac{1}{40} = \frac{1}{60}$$

X CLASSIFIED AS RED (NON RETIRING) SINCE IT HAS LARGER POSTERIOR PROBABILITY