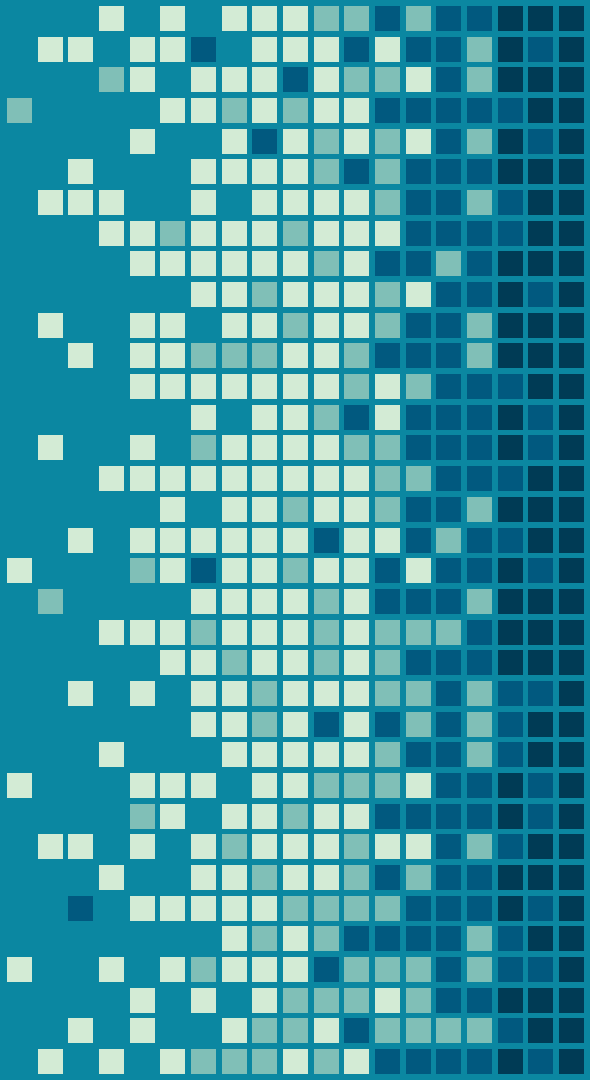
User Identification Through Machine Learning

Lydia Chen, Alexis Fryc, Thomas Hontz, Christie Hung, Mofeyi Oluwalana, Anisha Shin, Shreyas Srinivasan, Armaan Tobaccowalla, Timothy Topolski, James Tsatsaros



What is Machine Learning?

- Predict or Classify future data
- Process large datasets quickly
- Applied to various fields:
  - Gender Classification
  - Speaker Recognition

Gender Classification Through Social Media

Overview

- Classifying a user's gender through predictors in their tweets
- Probability classification algorithm
- Comparing the performance of the machine to human performance

Data Collection

- Obtained data from *CrowdFlower*
- Included gender of user, random tweet, location, link color, etc.
- Cleaning Data
  - Resulted in 9,991 tweets

Preprocessing

- Randomized order of tweets
- Removed non-letter characters

“#One_Direction” → “OneDirection”

“Bag of Words” Model

- Problem: how can a list of phrases be represented?
- Example:
  - 1: “I like cats”
  - 2: “I hate cats”

“Bag of Words” Model

	I	like	hate	cats
1:	1	1	0	1
2:	1	0	1	1

Weighting Words: tf-idf

- How can the computer determine which words are more important?

Term Frequency - Inverse Document Frequency:

$$\frac{\text{Frequency in each tweet}}{\text{Frequency in the data set}}$$

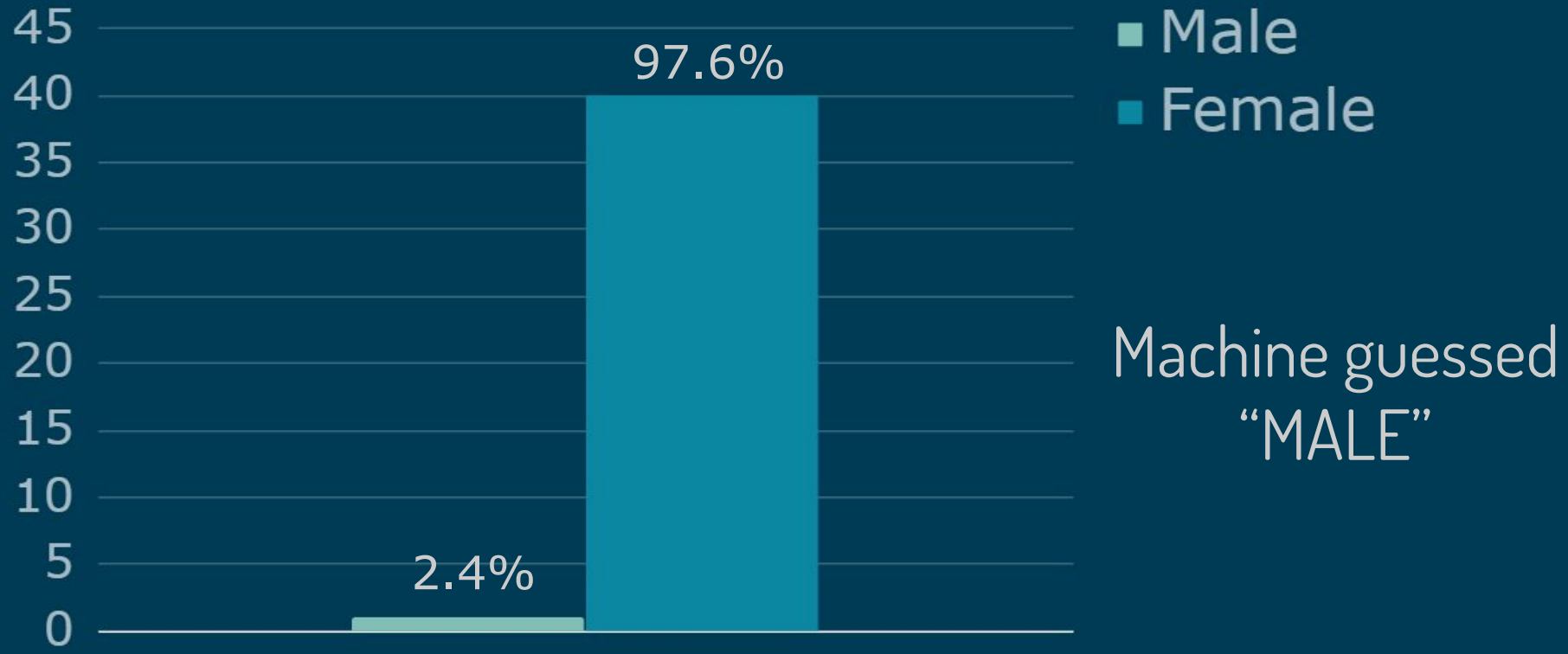
Machine Learning Algorithm

- Multinomial Naive Bayes Classifier
  - Uses conditional probabilities

P(M | “Wrestling”) > P(F | “Wrestling”)

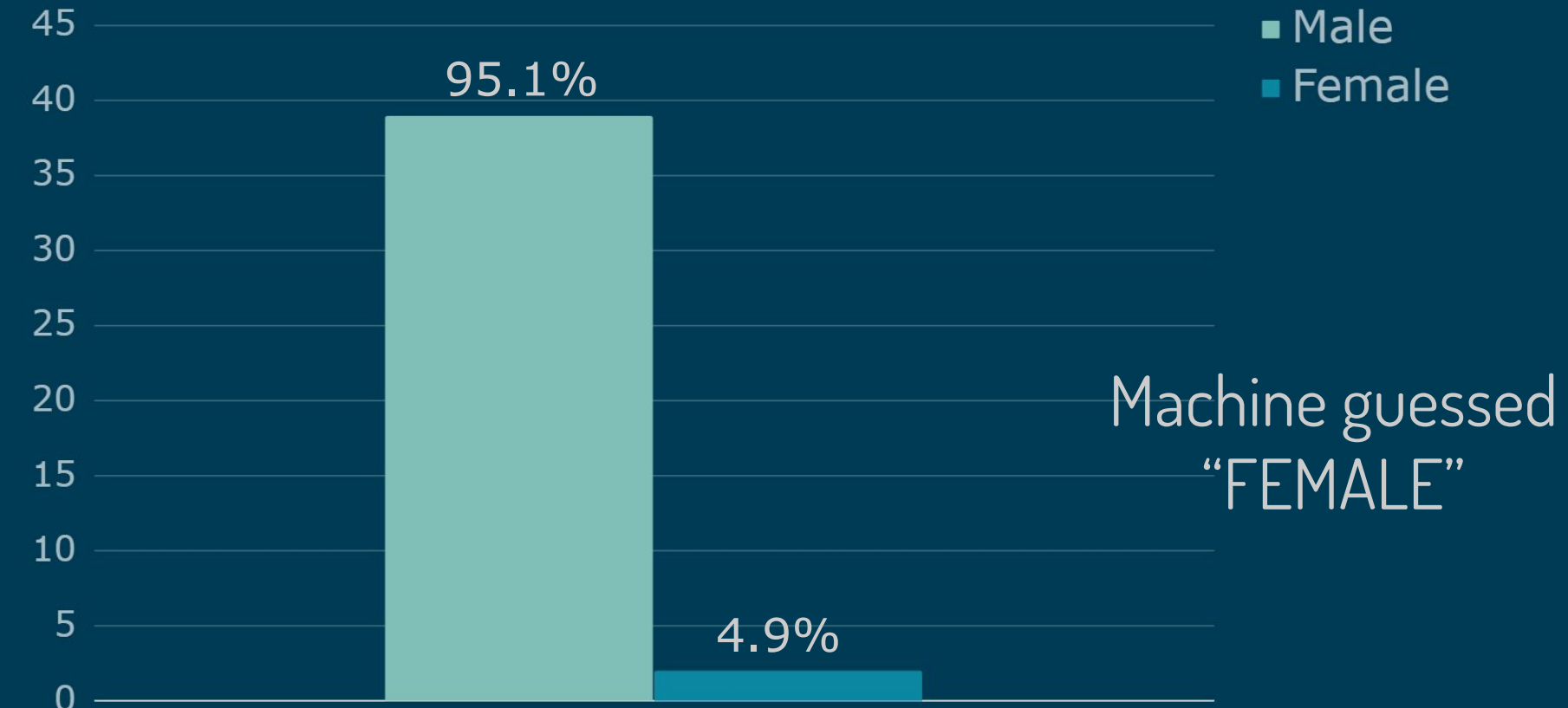
	Predicted: Male	Predicted: Female
Actual: Male	542	1254
Actual: Female	333	1862

"Got some new quail today - cute chirpy little birds, and loads of tiny free eggs :-)"



Machine guessed "MALE"

"If you scored a touchdown on sunday and didnt dab, hit them folks, or do that hotline bling dance, it shouldn't have counted."



Machine guessed "FEMALE"

Discussion

- Machine was right about 60% of the time
- Survey results:
  - Humans were 57-65% accurate
  - Overall, no significant difference between the machine and humans for this project

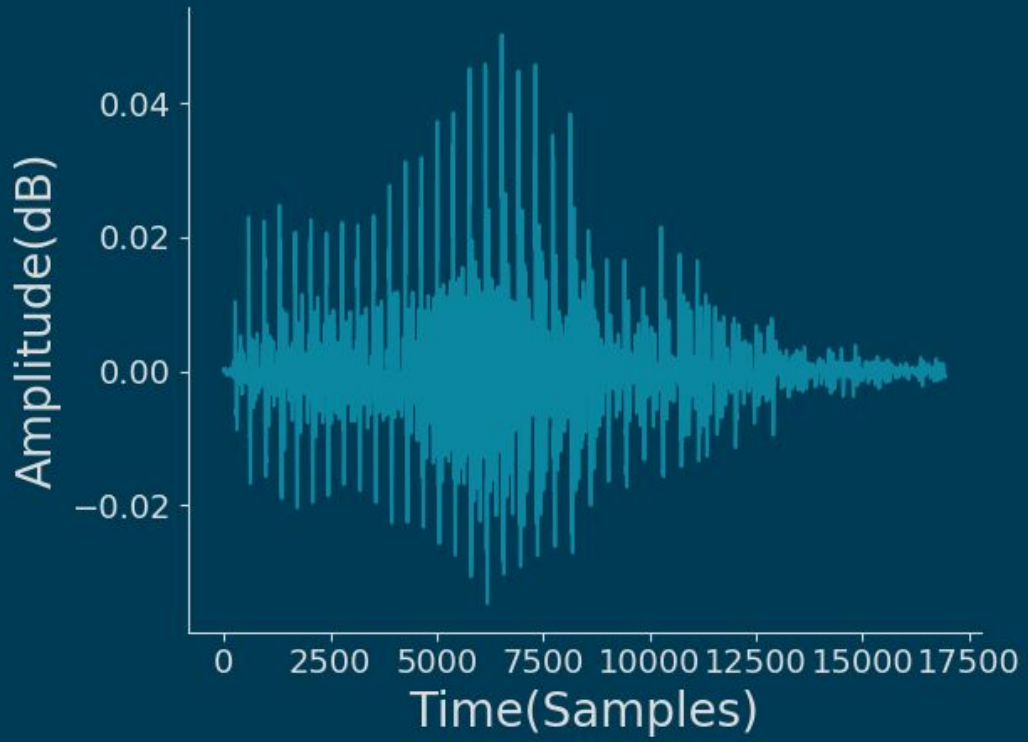
Speaker Recognition

Overview

- Identifying speaker based on individual information
  - Power and Frequency
- Dataset of known speakers
- Compare identity with machine learning algorithm

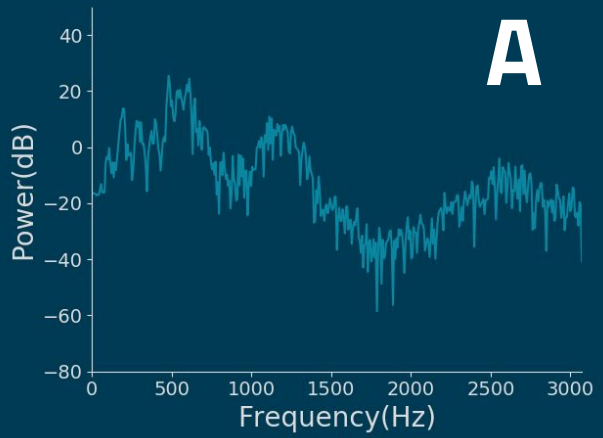
Data Collection

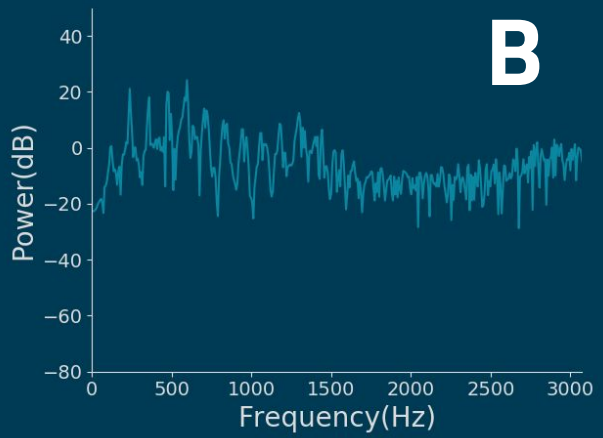
- 6 Audio Recordings each from 11 participants
  - Exported as .wav files
- Recording conditions kept constant
- 2,261,690 total data points

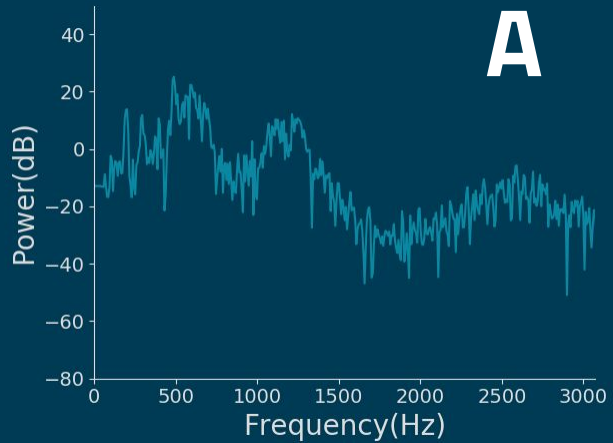

Amplitude(dB)
0.04
0.02
0.00
−0.02
0
2500
5000
7500
10000
12500
15000
17500
Time(Samples)

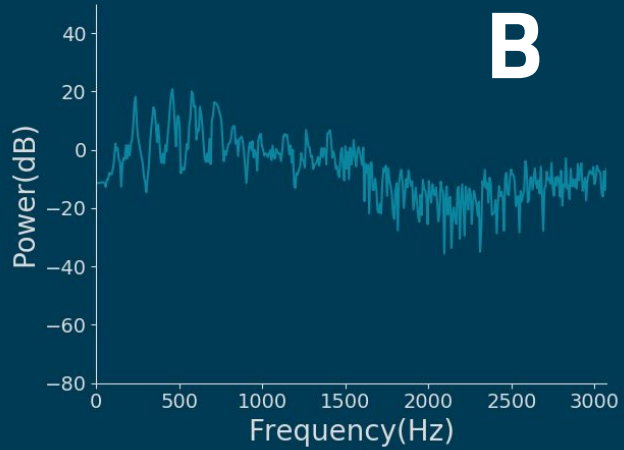
Preprocessing

- Clip trimming
- Normalize volume
  - Divide by magnitude to create unit vector
- Fast Fourier Transform
  - Creates a unique voice profile for each user
- High frequency noise removal


A
Power(dB)
40
20
0
−20
−40
−60
−80
0
500
1000
1500
2000
2500
3000
Frequency(Hz)

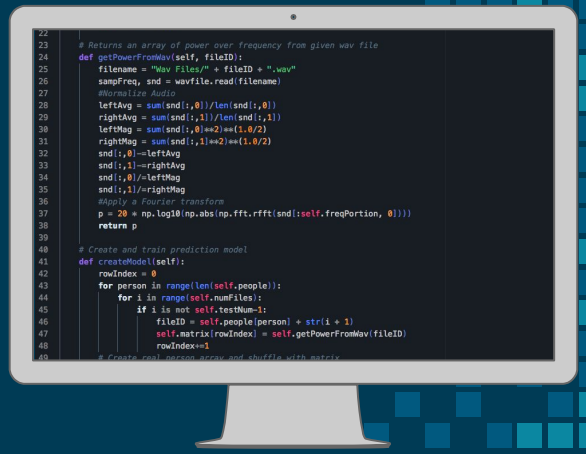

B
Power(dB)
40
20
0
−20
−40
−60
−80
0
500
1000
1500
2000
2500
3000
Frequency(Hz)


A
Power(dB)
40
20
0
−20
−40
−60
−80
0
500
1000
1500
2000
2500
3000
Frequency(Hz)


B
Power(dB)
40
20
0
−20
−40
−60
−80
0
500
1000
1500
2000
2500
3000
Frequency(Hz)

Machine Learning Algorithm

- Multi Class Classifier
  - OneVsRestClassifier
- Analysis of Amplitude
  - 18% Accuracy
- Analysis of Frequency
  - “Frequency vs Power”



```
# Returns an array of power over frequency from given wav file
def getPowerFromWav(self, fileID):
    filename = "Wav Files/" + fileID + ".wav"
    sampFreq, snd = wavfile.read(filename)
    #Normalize Audio
    leftAvg = sum(snd[:,0])/len(snd[:,0])
    rightAvg = sum(snd[:,1])/len(snd[:,1])
    leftMag = sum(snd[:,0]**2)**(1.0/2)
    rightMag = sum(snd[:,1]**2)**(1.0/2)
    snd[:,0]-=leftAvg
    snd[:,1]-=rightAvg
    snd[:,0]/=leftMag
    snd[:,1]/=rightMag
    #Apply a Fourier transform
    p = 20 * np.log10(np.abs(np.fft.rfft(snd[:self.freqPortion, 0])))
    return p

# Create and train prediction model
def createModel(self):
    rowIndex = 0
    for person in range(len(self.people)):
        for i in range(self.numFiles):
            if i is not self.testNum-1:
                fileID = self.people[person] + str(i + 1)
                self.matrix[rowIndex] = self.getPowerFromWav(fileID)
                rowIndex+=1
    # Create real person array and shuffle with matrix
```

Results

- 100% accuracy with predictions
- Confusion matrix
  - Detailed description of computer's accuracy
  - Values on diagonal: correct results

Conclusion

- Identified users based on their unique characteristics
- At least as accurate as humans
- Countless applications

Live Demo

- Presentation template by SlidesCarnival
- Photographs by Unsplash
- Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal. 1959 [accessed 2017 Jul 26];3(3):535–554. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.368.2254&rep=rep1&type=pdf
- Domingo P. A Few Things to Know about Machine Learning. University of Washington. 2012 [accessed 2017 Jul 26]. https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
- Sigmund M. Speaker Recognition [thesis]. Brno University of Technology; 2000. http://www.imm.dtu.dk/~lfen/Speaker%20Recognition.pdf
- Press WH. Numerical Recipes in C. 2nd ed. New York, NY: Cambridge University Press; 1992.
- Pedregosa F. Sci-Kit Learning: Machine Learning in Python. Journal of Machine Learning Research. 2011 Oct 12 [accessed 2017 Jul 26]. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C. 2nd ed. New York, NY: Cambridge University Press; 1992.
- Grus J. Data Science from Scratch. 1st ed. Beaugureau M, editor. Cambridge: O'Reilly Media, Inc.; 2015.
- Drygajlo A. Automatic Speaker Recognition for Forensic Case Assessment and Interpretation. Springer Link. 2017 [accessed 2017 Jul 26]. https://link.springer.com/chapter/10.1007%2F978-1-4614-0263-3_2
- Social Media Fact Sheet. Pew Research Center. 2017 Jan 12 [accessed 2017 Jul 26]. http://www.pewinternet.org/fact-sheet/social-media/
- Tenuto J. Using machine learning to predict gender. CrowdFlower. 2015 Nov 6 [accessed 2017 Jul 16]. https://www.crowdflower.com/using-machine-learning-to-predict-gender/
- Raschka S. 2016. Applying machine learning to sentiment analysis. Python machine learning. Birmingham (UK): Packt Publishing. p. 233-250.
- Ray S, Rizvi MSZ, Shaikh F, N, Jain S. 6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python). Analytics Vidhya. 2017 May 2 [accessed 2017 Jul 24]. https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/`

Acknowledgements

- Overdeck Foundation
- State of New Jersey
- Drew University
- Johnson & Johnson
- Novartis
- Bayer Healthcare
- Independent College Fund of New Jersey
- NJGSS Alumnae and Parents of Alumnae
- Allergan
- Celgene

PRESENTATION DESIGN

This presentation uses the following typographies and colors:

- Titles: **Dosis**
- Body copy: **Titillium Web**

You can download the fonts on these pages:

http://www.impallari.com/dosis

http://www.campivisivi.net/titillium/

#d3ebd5 **#80bfb7** **#0b87a1** **#01597f**

You don't need to keep this slide in your presentation. It's only here to serve you as a design guide if you need to create new slides or download the fonts to edit the presentation in PowerPoint®