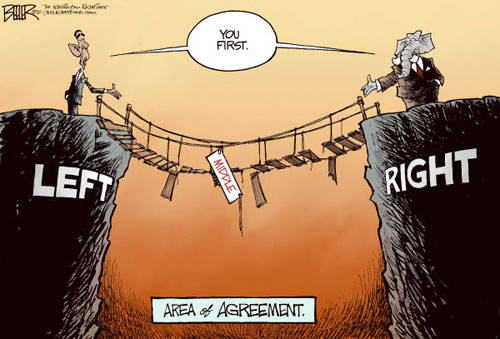
## horizontal line



PuddleNet

09.04.20XX

**─**

Captain Bouncy and the Rodneys

Goonies, Inc.

368 38th Street,

Astoria, OR. 97103

# 

# Intro Text

Have you ever wondered why Democrats and Republicans have such polarized views on every aspect of social life? Do they even talk to each other? That's what we explore...  
 <br>Created by Captain Bouncy, the Rodneys, Boss & Co.

# Overview

“PuddleNet.” It was to be our answer to a question we had not yet found. This project started with a desire to understand the political climate and the language being consumed by the masses. We wanted to be explorers of a polarized lexicon that seemed to be dividing this country.

As our project unfolded, we identified 20 sources from which to gather language. We chose an equal number of popular left- and right-leaning news media sites to scrape, and began our process of gathering human thought that has been recorded in sentences.

As the data began rolling in, millions of words per site, we needed to find a way to start making sense of it all. We chose a two-pronged approach. First, we would begin grouping and inspecting to look for patterns, find relationships, further direct our exploration. Second, we we wanted a machine learning model to learn what left- or right-leaning text would look like so that we could use it to identify political leanings of a text, or even build text blocks with perspective.

And so we got to work...

# Lexicography

Word clouds are very subjective and require data cleaning in order to extract insights from them. It relies heavily on the human brain to identify patterns and/or relationships. In its simplest form, the more frequent words are illustrated with bigger size in the cloud. Common grammar words are typically removed so “THE” does not steal the show.

In looking at the individual word clouds (select from drop downs), we began inspecting the language. We began to have an inkling that there was definitely something different between the lexicons used by each side. However, we were unable to put a finger on it. We then rolled the word clouds up into groups of left and right, and the pattern began to take shape.

The Conservative Lexicon seemed to be a little more focused on political terms, whereas the Liberal Lexicon almost seemed a little erratic.

This was surprising. So, we needed to explore the anatomy of the articles…

These interactive word clouds allow you to select and compare the 10 different sites from each side.

We concluded there was a disparity between the left and right centered around topic focus.

# Article Anatomy

This simplest possible box plot displays the full range of variation (from min to max), the likely range of variation (first and third quartile), and a typical value (the median). Upper and Lower Fences help identify outliers in the data set. In short, the box plot provides a simple visual of the dimensions of the data.

Our box plots were very insightful. We noticed almost immediately that the article anatomies were very different. Liberal media tended to have longer articles by nearly twice as many words, on average, than the Conservative Media. This began to explain some of the disparity between the word clouds, but we simply needed more information.

We wanted to explore why longer articles that clearly had more information, seemed to not be heard as loudly in the political arena.

And so we looked for a way to visualize the interconnectedness of the sites, to see if there was a proportionate flow between the sides...

This graph shows the liberal articles were, on average, twice as long as the Conservative articles. This lead us to explore the interconnectedness of the sites to see if there was a proportionate flow of data between the sides.

# Crossing the Divide

A Chord plot is used to display the inter-relationships between data. It is arranged in a circle and arcs will show links to other sites, and the thickness of the link represents volume of interconnectedness.

We went back to our raw data and began scouring the pages for links to external sites. This would show definitively how frequently the sites referred to each other, and if the sites crossed the political chasm seemingly between the sides.

We found there is a chasm. The left and right do not tend to share. This was not too surprising, though still disappointing.

However, we did find what we call “hub” sites. Only three seemed to cross the divide heavily, and in both directions.

So, there we were, looking at all but insulated sides, and we noticed there simply was more volume on the liberal side, they wrote longer articles, and they covered more topics of interest…

At this point, our assumption would be the liberal media should be dominating the political ecosphere.

This was discordant with what we see everyday, so we went back to the data…

This Chord Diagram shows there is clearly a divide between the media outlets. However, we noticed there were several “hub” sites connecting to all other sites. This also showed the volume of Liberal media outpaced conservative media significantly. So, we went back to the data.

# Word Cloud (US All)

This cloud spanned both sides of the stage. It clearly shows it is more similar to the Conservative Word Cloud than it is to the Liberal Word Cloud. Despite what we discovered, the volume and quality of the Left, the Right-Leaning media was dominating the space.

# Conclusion

Our exploration of the media lexicon has taught us 4 very important lessons:

1. Starting the adventure without knowing the destination leads you to better discoveries than knowing where you want to be, and hacking your way to the end.
2. When taken one at a time, the data is not that informative, but following the threads and stitching a story together is better way to find truth, and not just a visualization
3. The truth that we found is: Short, hyper-focused, political, hot-button titles will dominate the media over structured, logical, and factual argument and variety.
4. Machine Learning is utterly awesome, yet exceedingly difficult to master.

# Goals

1. Explore and determine commonalities (middle ground) for the right and left wing
2. Explore the interconnectedness of liberal and conservative news sites to determine whether they share the opposite views through linking or not and keep the reader firmly in their sphere
3. Explore and determine the commonalities of speech patterns independently and between the liberal and conservative sphere
4. Create a machine learning model that can:
   1. Take text and determine if it is leaning right or left
   2. Enter text and let the machine construct the paragraph based on what it has learned from either right or left leaning sites
5. Explore data dimensions to determine if events or topics sway the liberals/conservative one way or another.

# Milestones

## Identify Liberal and Conservative sites

We intend to find the top 10 sites for each side. We will use a reputable ranking site to help guide our decisions on which sites to include. The goal is to keep the left and right sites as similar as possible in terms of intensity and extremism.

## Scrape Sites

We intend to scrape the sites for their html in order to gather the raw data. We will use beautiful soup and significant error checking. Our goal is to scrape a minimum of 10 million words from these sites, with a preference for more.

## Machine Learning Model

We intend to use deep learning, unsupervised machine learning for our natural language processing.

## Analysis - Data Cleanup

We intend to inspect the data and begin analysis. This will be an iterative process where we experiment, learn, visualize, and experiment again.

## Web Site

We intend to build an interactive website that will showcase our analysis, tell our story, and still provide enough interactivity for the user to create their own story

## Presentation

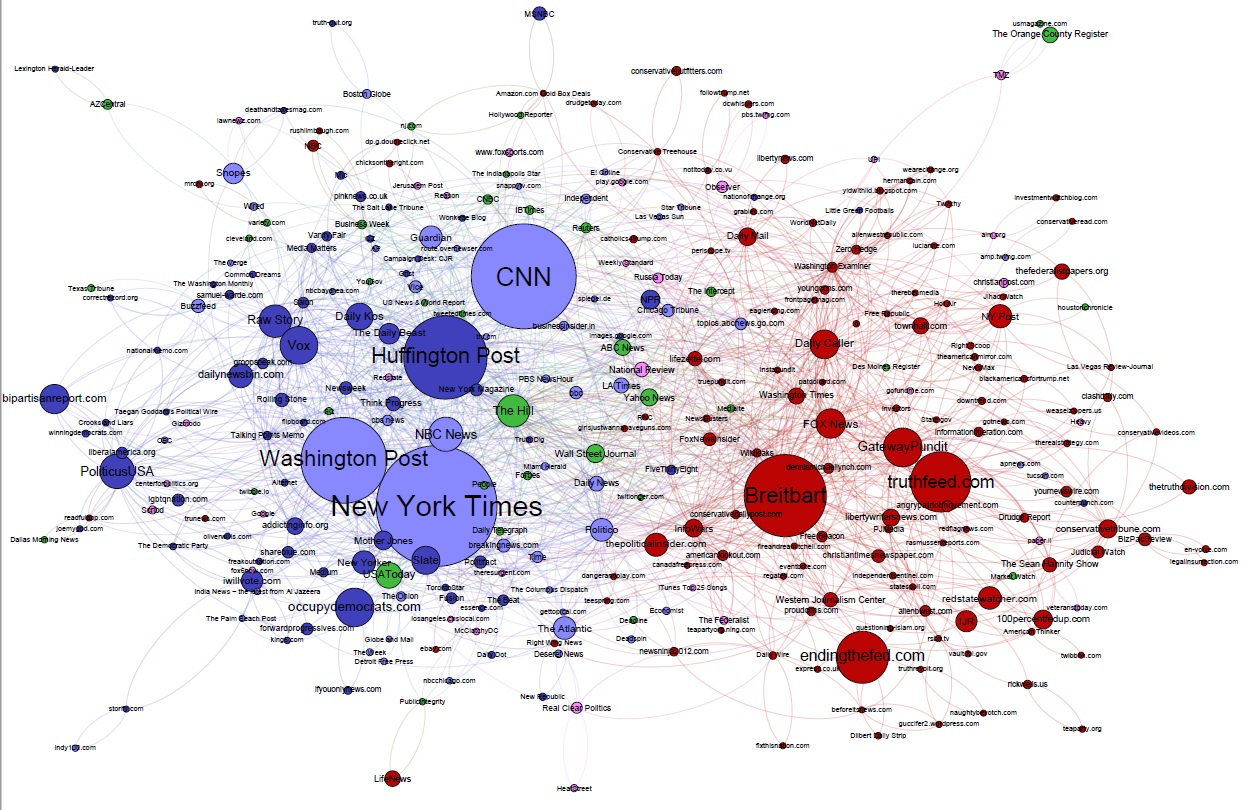
We intend to build a prezi presentation to help tell our story.

# Inspirations

We were inspired by several topics and graphics. First, we spent an afternoon discussion the possibilities of what Machine Learning could do if it could categorize left-leaning versus right-leaning text. Second, we gave ourselves a proof-of-concept deadline and set ourselves to prove the feasibility of our ideas. The below are some of the things we came across. It should also be noted that this project evolves as we explore the data.

## Liberal/Conservative Interconnectedness graph

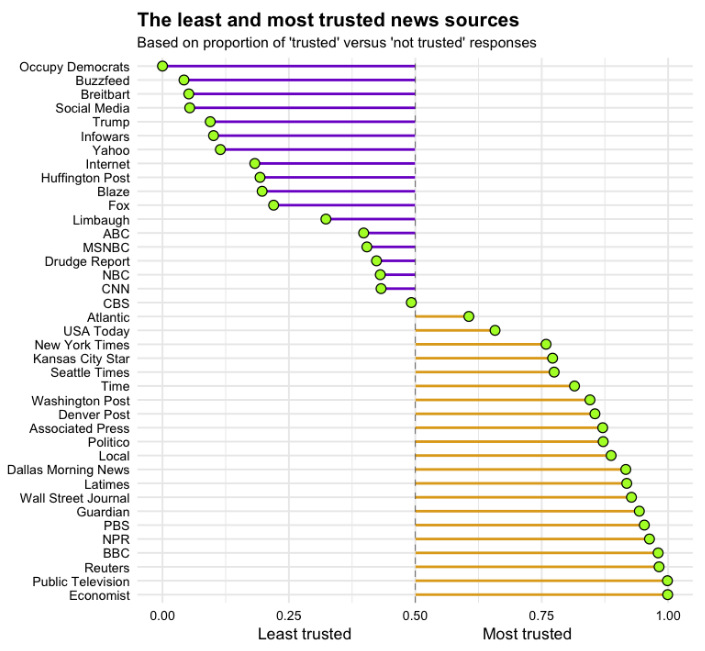
The below graph inspired to see if there was a “great divide” between the new sites, creating reinforcing spheres of influence, or if they would provide the other views on the story by linking to a site that leaned the opposite direction.

[](https://cdn.cjr.org/wp-content/uploads/2017/03/Election-Facebook-1.jpg)

## Intensity or Extremity graph

The below graph inspired us in multiple ways.

1. We could measure the extremity of the sites through scores produced by the ML model
2. If we extracted dates from articles, we could determine extremity based on topic
3. Furthermore, with dates, we could potentially find events that brought the liberals and conservatives together, or push them apart



.

## Word or Phrase Analysis graph

The below graph inspired us in multiple ways.

1. We could see what buzzwords are being used most by either side
2. Creating a word cloud would give us insights on topic focus

