The Truth in the Age of Misinformation

Can we detect fake news?

Ryan Cornelius
Ayan Khalif
Aaron Otto
Kirsten Rain

Table of contents

Tools and Libraries Used

- **# Libraries for data cleaning**
- import itertools
- from sklearn.feature_extraction.text import TfidfVectorizer
- from sklearn.linear_model import PassiveAggressiveClassifier
- import re
- from textblob import TextBlob
- from sklearn.feature_extraction.text import ENGLISH_STOP_WORDS
- from sklearn.model_selection import train_test_split
- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- import os
- from wordcloud import WordCloud
- **#Libraries for NN W2V model**
- import tensorflow as tf
- from sklearn.metrics import accuracy_score,confusion_matrix
- from gensim.models import Word2Vec
- from nltk.tokenize import word_tokenize
- import multiprocessing
- import nltk

- Jupyter Notebooks
- Tableau
- yData-Profiling tool

Data Set and Sources

We got the idea from “12 Data Science Projects for Beginners and Experts”

https://builtin.com/data-science/data-science-projects

This is where we got the data set. It is from 2019.

This site also included starter code, but no one was willing to use it without understanding it so they started from scratch!

Index	key	title	text	label
0	8476	You Can Smell Hillary’s Fear	And it’s an interesting question. Hillary’s old strategy was to lie and deny that the FBI…	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Political Suicide At A Trump Rally (VIDEO)	— MSC News Politics (@MSCPolitics) November 9, 2016 The Democratic Party couldn’t have asked for a better m… The ringing endorsement of the man he clearly hates on	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	Among roughly 40 leaders who did attend was Israeli Pri…	REAL
3	10142	Bernie supporters on Twitter erupt in anger against the DNC: 'We tried to warn you!'	— Ana Navarro (@ananavarro) November 9, 2016 Popular left-wing Facebook page The Other 98%, which wa…	FAKE
4	875	The Battle of New York: Why This Primary Matters	Trump needs to capture more than 50 percent of the vote…	REAL
5	6903	Tehran, USA	being tormented because of my ethnicity and religion in… Through patience, humor, and understanding, I was able …	FAKE
6	7341	Girl Horrified At What She Watches Boyfriend Do After He Left FaceTime On	Baylee had been discussing things with her boyf… According to KRON , Baylee was mid-conversation with Ya… Admitting that she first thought it was a joke, seconds	FAKE
7	95	‘Britain’s Schindler’ Dies at 106	A Czech stockbroker who saved more than 650 Jewish chil…	REAL
8	4869	Fact check: Trump and Clinton at the 'commander-in-chief' forum	In 1984, for example, Ronald Reagan landed in Beijing a…	REAL

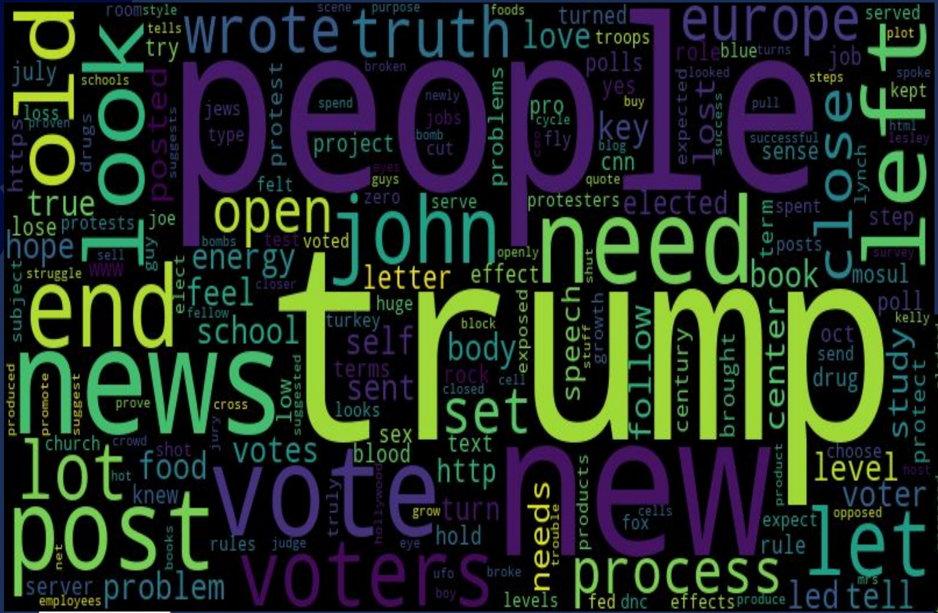
01 Cleaning the Data

It took us several iterations to clean the data:

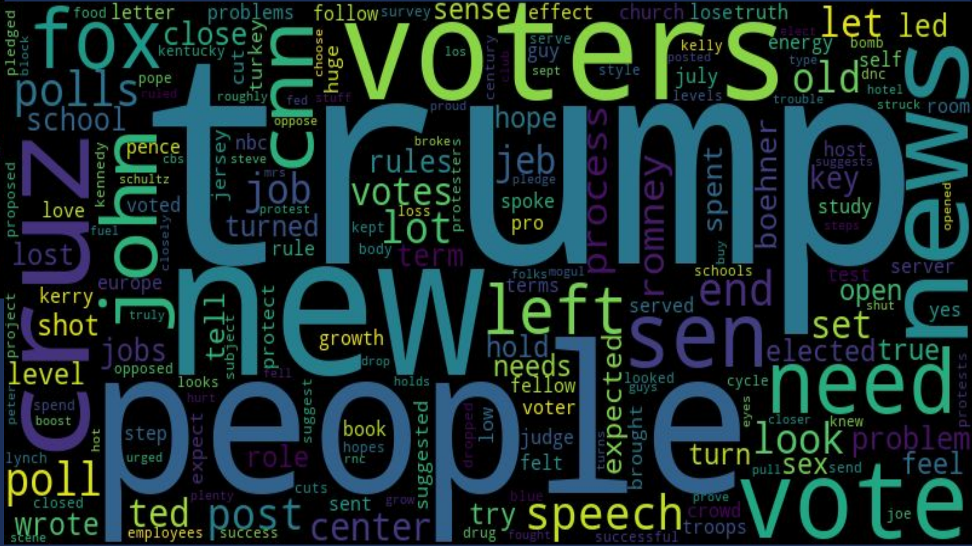
- Figure out the best format to ingest (to prevent weird characters)
- Tokenize the words (split on space)
- Use sklearn's ENGLISH_STOP_WORDS to exclude words like the, of, are, by...
- Use str.len()>2 to get rid of little fragments left over
- Remove numbers

Then we started to get excited about having good text to work from!

Real vs Fake Word Counts



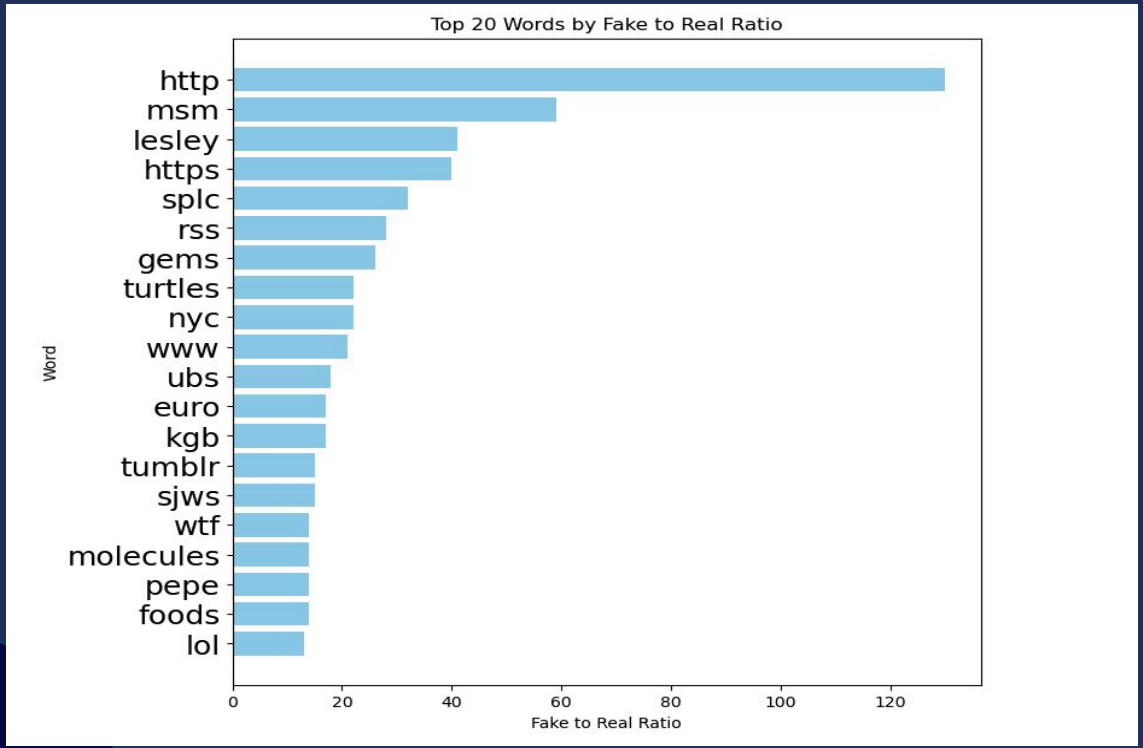
We had a panic moment as we were cleaning the data when we started to realize that the words with a high frequency on the REAL list were nearly identical to the words on the FAKE list.
Uh Oh! Will our model be able to pick up other differences????



Pretty similar at first glance!

Differences in the Fake to Real Ratio

I started to get excited when I calculated the ratios between the FAKE and REAL word counts. Sorting by those revealed some big differences. Is this what the model would pick up on?



Notice:

- http and https
- msn
- www

Is fake news more likely to site websites?

Also:

- wtf
- lol

Those should be giveaways!

Sentiment Analysis

Still curious about how this was going to work,

I decided to do sentiment analysis on the Fake and Real News.

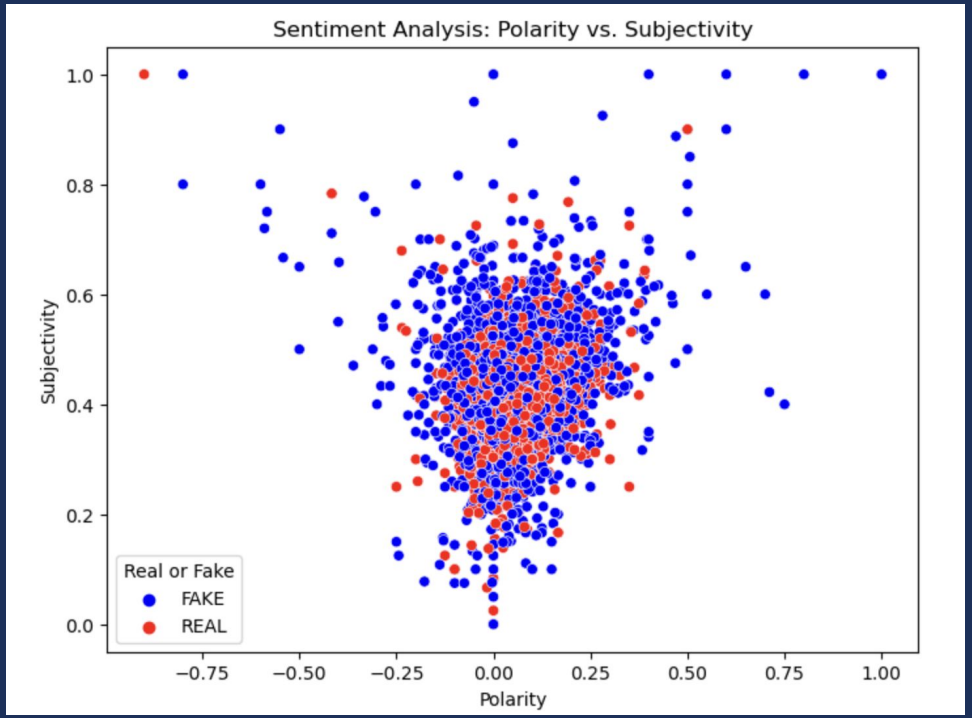
It is daunting to see that even most of the fake news falls within a “neutral zone” in the center of the graph. This implies that it would be extremely difficult for a casual reader to detect fake news on their own.

Higher on the y-axis implies more opinionated phrasing and less factual.

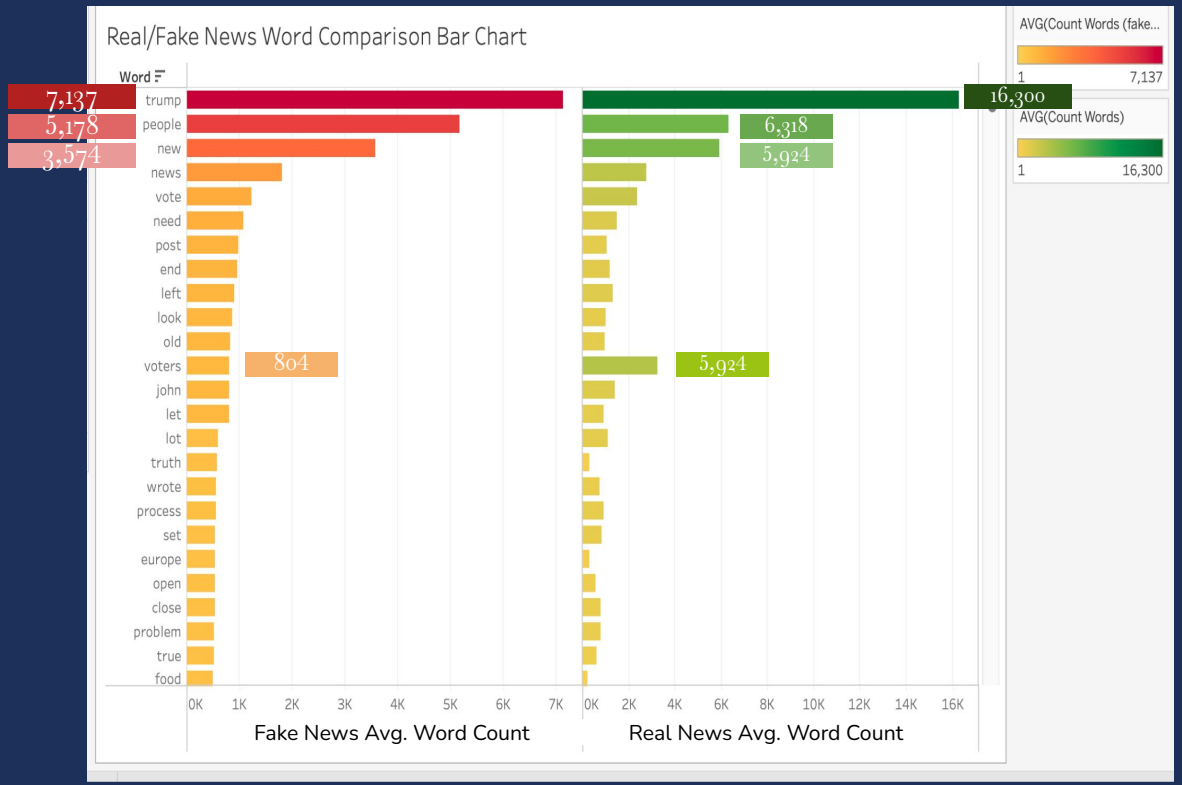
Further from the center on the x-axis are more positive or negative in their tone.

Articles in these two areas are more likely to be Fake (blue dots).

Is this what our model would pick up on????

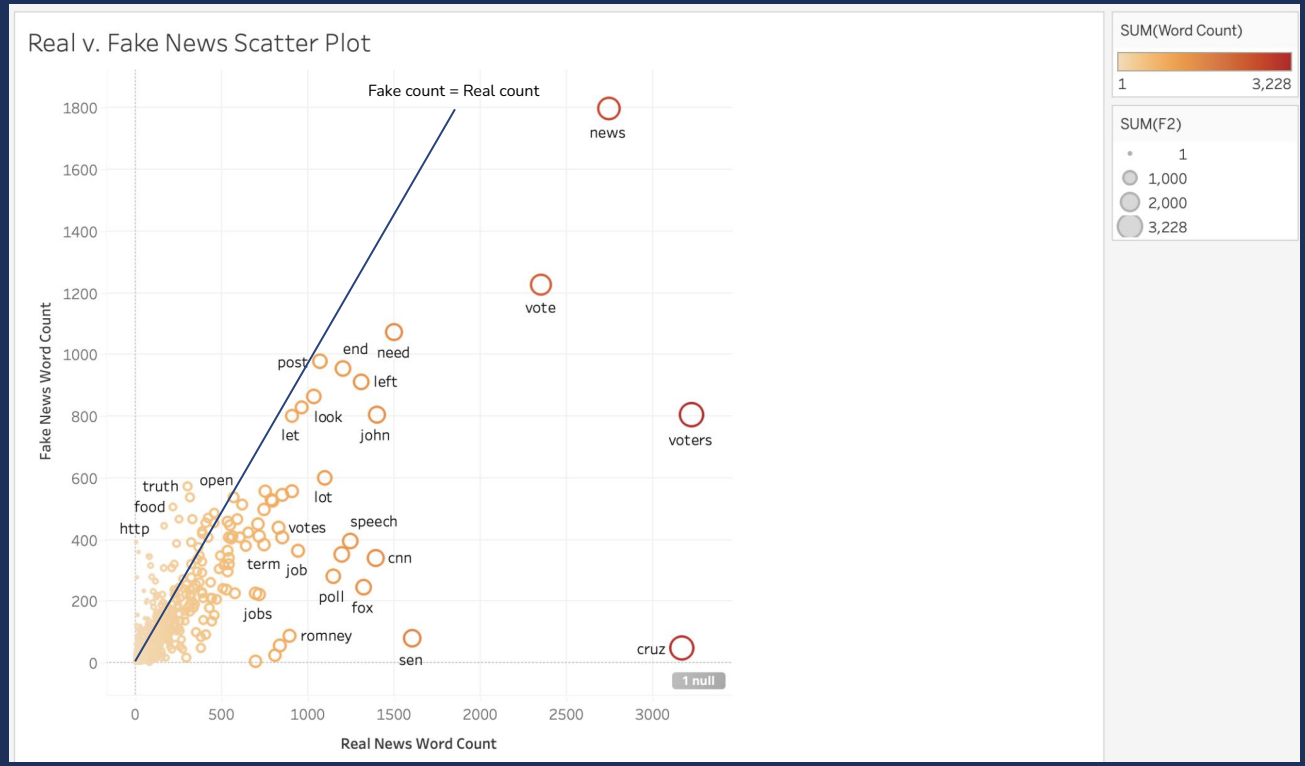


Real vs Fake Word Counts



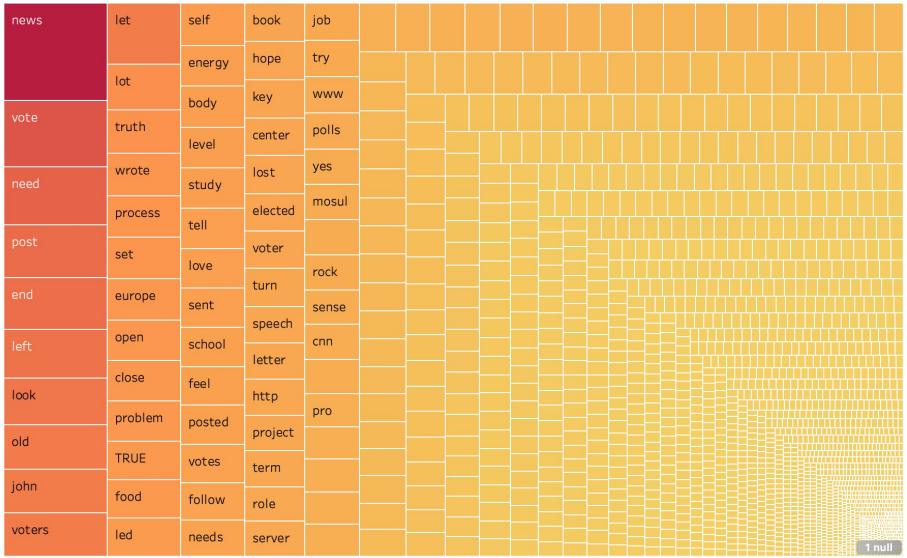
Scatter Plot

Scatter Plot with 3 Outliers Removed

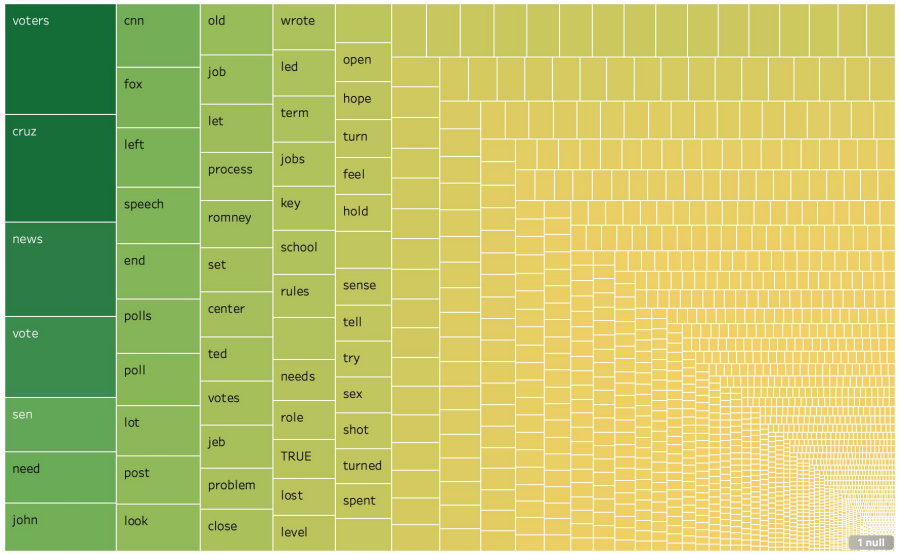


Word Freq. Treemap without Outliers

Fake News Word Freq. Treemap



Real News Word Freq. Treemap



03 The Model

Primarily used Word2vec, one of the simpler natural language processing neural net model designs.

- Tokenizes words and forms vectors to describe sentences
- Alternatives
  - Bag of Words
  - Doc2vec
  - Syntax/Semantic analysis (probability context free grammars)
  - Discourse analysis (rhetorical structure theory)
- Achieved 87% accuracy on test data split, 99.6% on train split


```
[52]: # Initialize and train the Word2Vec model
      w2v_model = Word2Vec(sentences = X,
          vector_size=init_vectorsize,          # Size of word vectors
          window=window,                  # Context window size
          min_count=min_count,              # Minimum word frequency
          workers=cores,            # Number of CPU cores to use
          epochs=w2v_epochs,                  # Number of training epochs
          max_vocab_size=max_vocab_size
      )
```

03 The Model

Attempted multiple different hidden layer orientations

- Standard input layer
  - Non-trained layer intaking a weight matrix
- Long short-term memory (LSTM) layer
  - Is useful for slowly improving models with backpropagation
  - Showed the best results, capable of **87% accuracy alone**
- Standard dense layers, varying activation
- Standard sigmoidal output layer, loss function, and optimizer

```
Layer2 = True
LSTM_units = 512 #LSTM layer units

Layer3 = True
layer3Act = "relu" #Hidden layer 3 activation
hidden_nodes_layer3 = 512 #Hidden layer 3 node count

Layer4 = True
layer4Act = "relu" #Hidden layer 4 activation
hidden_nodes_layer4 = 128 #Hidden layer 4 node count

outputAct = "sigmoid" #Output layer activation
```

```
[59]: # initializing the W2V model weight matrix to input into the NN initialization layer.
      weight_matrix = w2v_model.wv.vectors
```

```
# First input hidden layer using model weight matrix, set to untrainable.
nn.add(tf.keras.layers.Embedding(num_keys, output_dim=init_vectorsize, weights=[weight_matrix], input_length=longest_vector, trainable=False))
```

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 1000, 150)	6415950
lstm_1 (LSTM)	(None, 128)	142848
dense_1 (Dense)	(None, 2048)	264192
dense_2 (Dense)	(None, 512)	1049088
dense_3 (Dense)	(None, 1)	513

```
Epoch 23/25
158/158 [==============================] - 91s 578ms/step - loss: 0.0159 - accuracy: 0.9952
Epoch 24/25
158/158 [==============================] - 91s 577ms/step - loss: 0.0133 - accuracy: 0.9958
Epoch 25/25
158/158 [==============================] - 91s 578ms/step - loss: 0.0189 - accuracy: 0.9962
```

4 Conclusion

It was cool to see the machine learning model give such successful results: **87% accuracy.** Post clean and analysis of labeled data set the model produced very interesting results. **High frequency words were nearly identical in both classifications**: ‘fake’ & ‘true’.

Additionally, the clustering of fake articles shows centralized nesting inside the ‘true’ data set.

Given the context of the prior two observations, a model producing 87% accuracy is impressive. Our team feels there are disturbing similarities to in word associations found between ‘true’ and ‘false’ news articles.

4 Conclusion

Given our current understanding of the data we would encourage consumers of news to read with a guarded point of view.

Chances of publicly available info being misleading are equally as high as not of being fraudulent. Unless it contains the word truth; then it is FAKE ;)

Finally, we acknowledge that it would be an improvement to have more than a single data source.

Thanks!

Any questions?
Comments?
Concerns about prevalence of fake/real news?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Information accessibility

Technology

Mercury is the closest planet to the Sun and the smallest of them all

Telephony

Venus has a beautiful name and is the second planet from the Sun

Digitization

Despite being red, Mars is actually a cold place. It’s full of iron oxide dust

Intentional manipulation

Beliefs

Mars is actually a very cold place

Emotions

Venus has extremely high temperatures

Opinions

Jupiter is the biggest planet of them all

Actions

Saturn is a gas giant and has several rings

Rhetorical procedures

Reassumption

Mars is actually a very cold place

Lies

Venus has extremely high temperatures

Hoaxes

Neptune is the farthest planet from the Sun

Fallacies

Jupiter is the biggest planet of them all

Generalization

Saturn is a gas giant with several rings

Obscurantism

Mercury is the closest planet to the Sun

Awesome words

“This is a quote, words full of wisdom that someone important said and can make the reader get inspired”

—Someone Famous

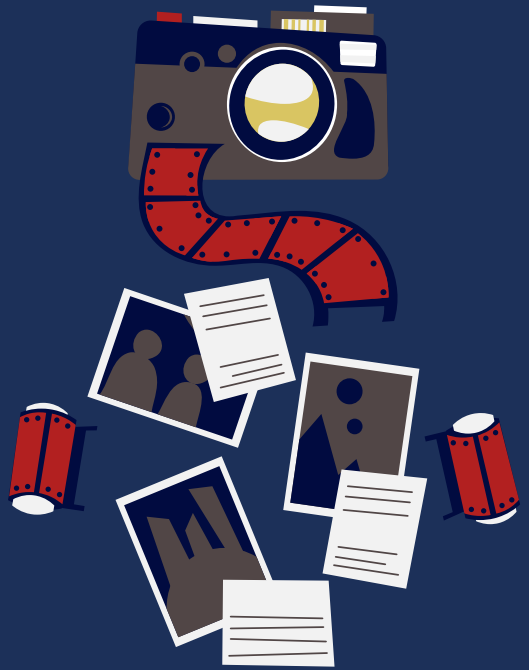
A picture is worth a thousand words

A picture always reinforces the concept

Images reveal large amounts of data, so remember: use an image instead of a long text. Your audience will appreciate it

300,000

Big numbers catch your audience’s attention



9h 55m 23s

Jupiter’s rotation period

333,000

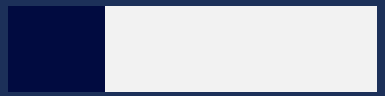
The Sun’s mass compared to Earth’s

386,000 km

Distance between Earth and the Moon

Let’s use some percentages

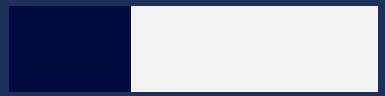
20%



Mercury

Mercury is the closest planet to the Sun and the smallest of them all

30%



Venus

Venus has a beautiful name and is the second planet from the Sun

50%

Mars

Despite being red, Mars is actually a cold place. It’s full of iron oxide dust

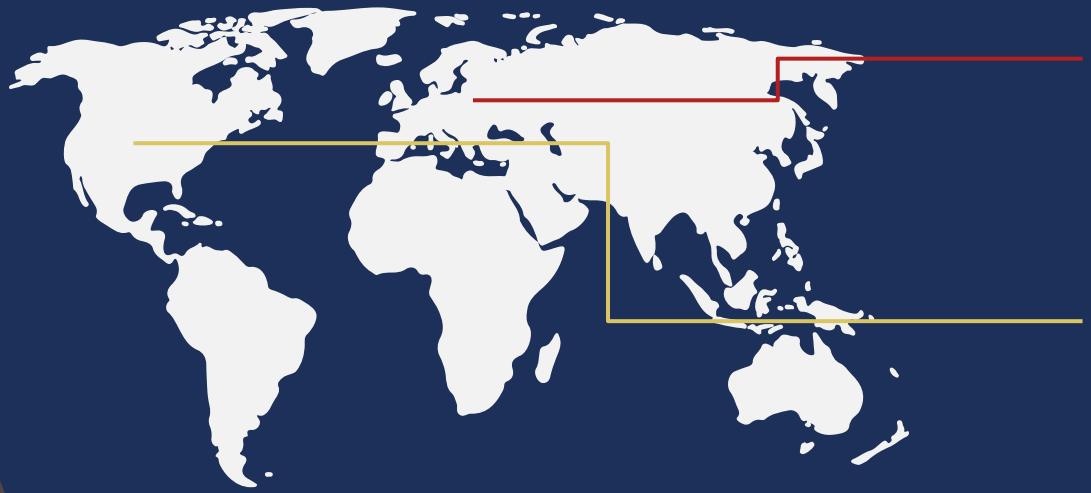
Tablet mockup

You can replace the image on the screen with your own work. Just right-click on it and select “Replace image”

Phone mockup

You can replace the image on the screen with your own work. Just right-click on it and select “Replace image”

This is a map



Venus

Venus is the second planet from the Sun

Mercury

Mercury is the closest planet to the Sun

The most widespread hoaxes

Source	News
Source 01	Mars is full of iron oxide dust
Source 02	Jupiter doesn’t have a solid surface
Source 03	Saturn was named after a Roman god
Source 04	Neptune is far away from us

Extra information

Mars

Mars is actually a very cold place

Venus

Venus has extremely high temperatures

Neptune

Neptune is the farthest planet from the Sun

Mercury

Mercury is the closest planet to the Sun

Saturn

Saturn is a gas giant with several rings

Most used media

01

Venus

Venus has a toxic atmosphere

02

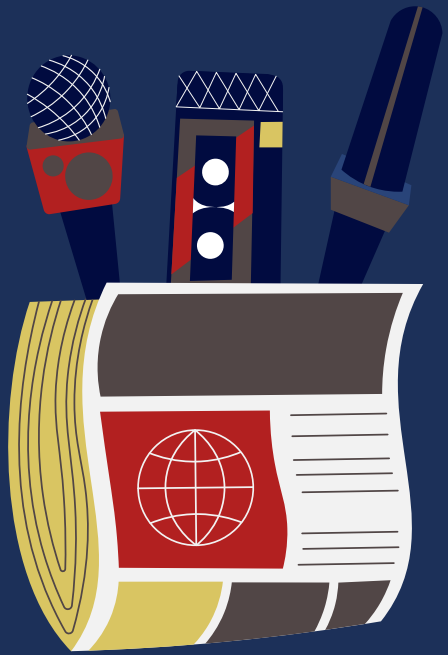
Mars

Despite being red, Mars is very cold

03

Jupiter

Jupiter doesn’t have a solid surface



History of misinformation

Venus is the second planet from the Sun

XXXX

Mars is full of iron oxide dust

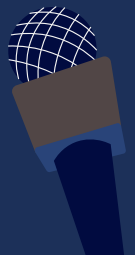
XXXX

XXXX

Earth is the third planet from the Sun

XXXX

Jupiter doesn't have a solid surface



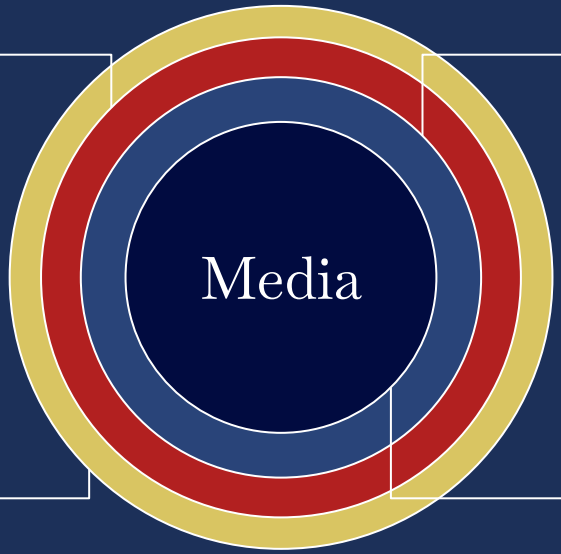
This is an infographic

Mars

Mars is actually a very cold place

Mercury

Mercury is the closest planet to the Sun



Venus

Venus has extremely high temperatures

Saturn

Saturn is a gas giant and has several rings

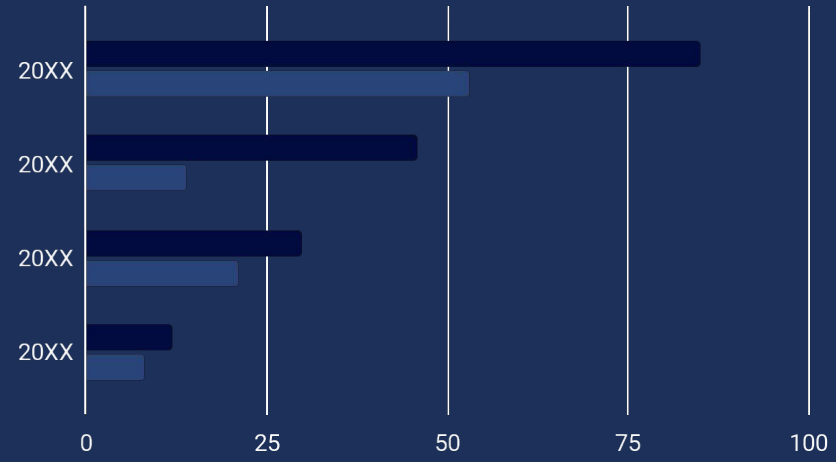
You can use this graph

Venus

Venus is the second planet from the Sun

Mercury

Mercury is the closest planet to the Sun

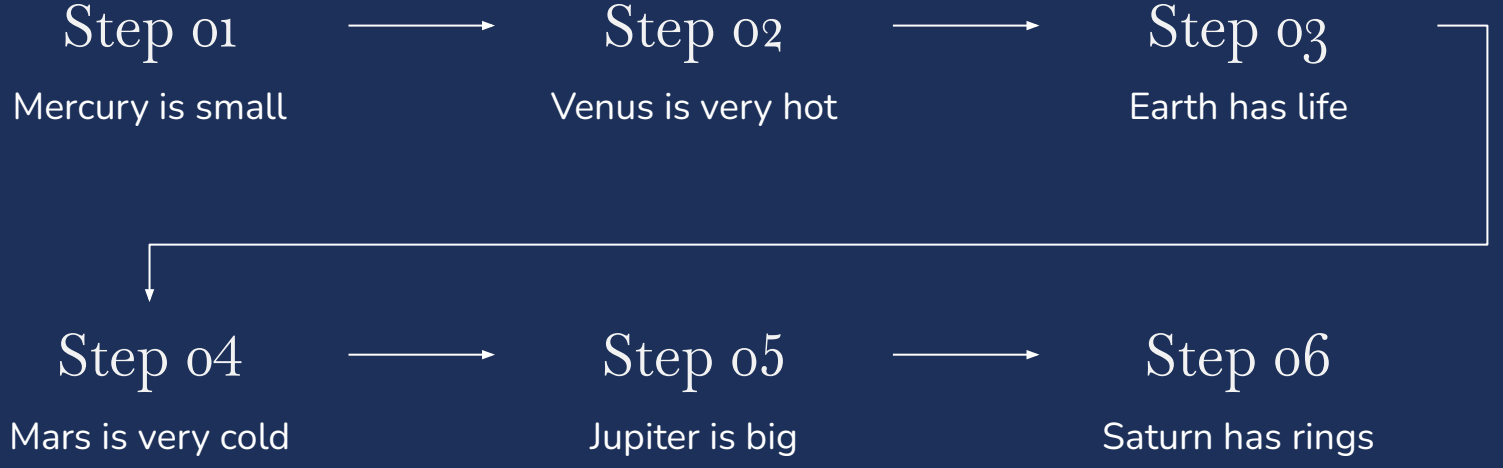


Follow the link in the graph to modify its data and then paste the new one here. **For more info, click here**

Tables represent your data

	Mass	Diameter	Gravity
Mercury	0.06	0.38	0.38
Mars	0.11	0.53	0.38
Saturn	95.2	9.4	1.16

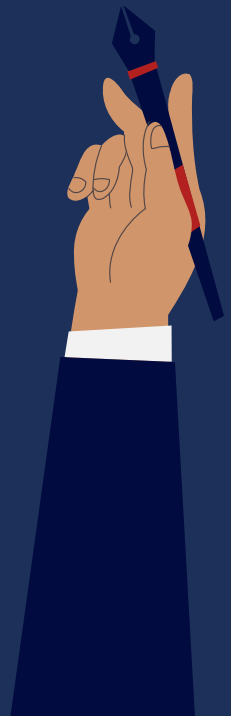
Fabrication of a lie



Information vs misinformation

Information

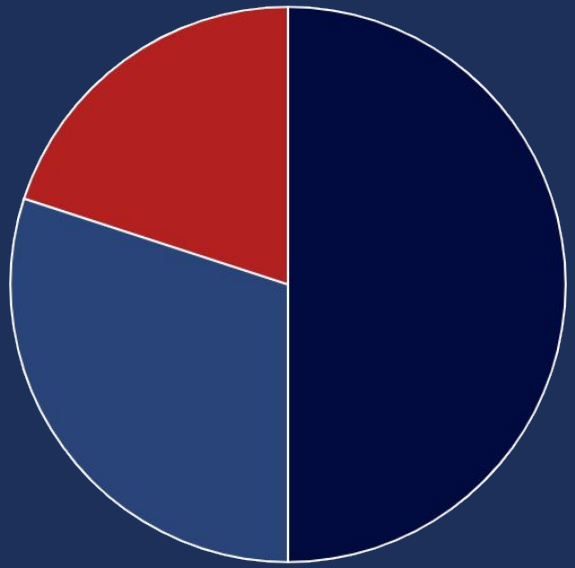
- Jupiter is a huge gas giant
- Mars is made of basalt
- Saturn is the ringed planet
- Mercury is a small planet



Misinformation

- Venus has high temperatures
- Earth is a planet that has life
- Pluto is now a dwarf planet
- Neptune is far away from us

You can use this graph



50%

Mercury

Mercury is very small

30%

Venus

Venus is a hot planet

20%

Saturn

Saturn is a gas giant

Follow the link in the graph to modify its data and then paste the new one here. **For more info, click here**

Conclusions

Mercury is the closest planet to the Sun and the smallest one in the entire Solar System. This planet's name has nothing to do with the liquid metal, since Mercury was named after the Roman messenger god. Despite being closer to the Sun than Venus, its temperatures aren't as terribly hot as that planet's

Alternative resources

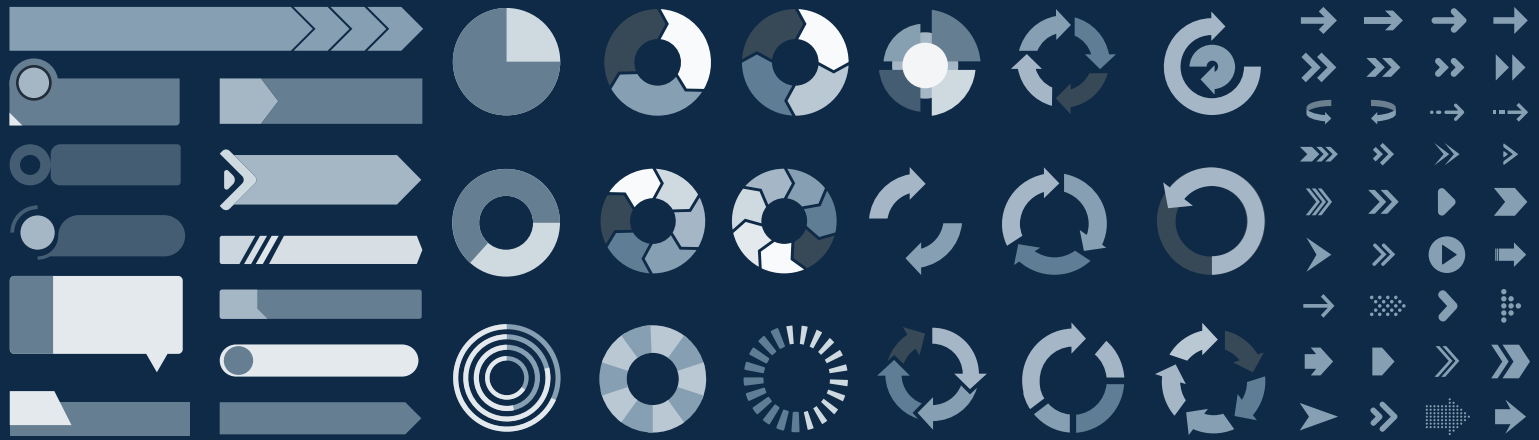
Here’s an assortment of alternative resources whose style fits that of this template:

Vectors

- Flat dia del periodista instagram stories collection
- Dia del periodista hand drawn flat illustration

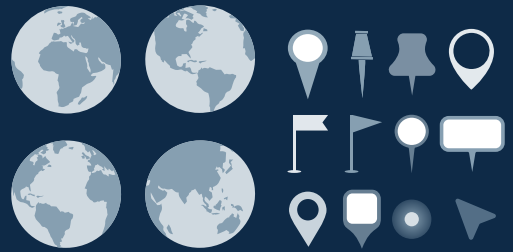
Use our editable graphic resources...

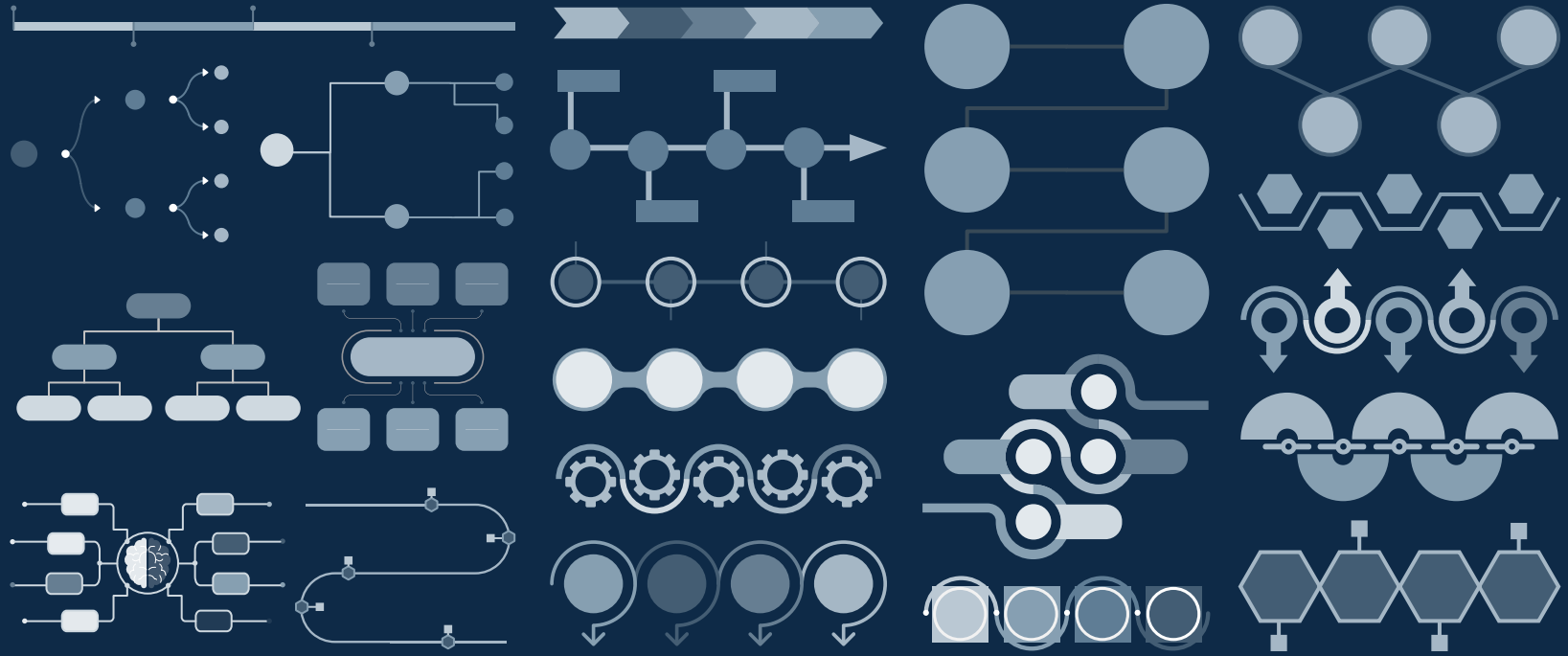
You can easily **resize** these resources without losing quality. To **change the color**, just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want. Group the resource again when you're done. You can also look for more **infographics** on Slidesgo.

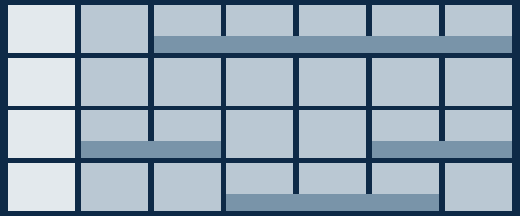


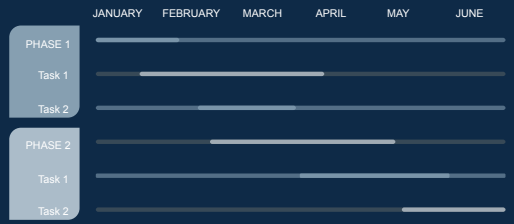


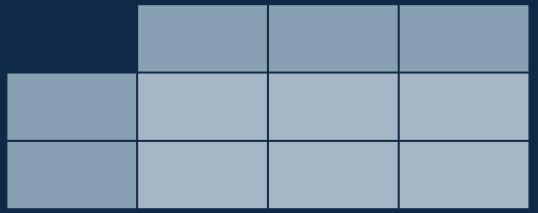


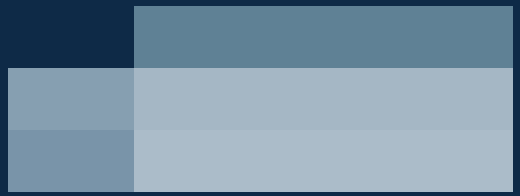



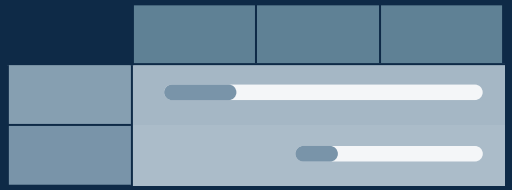


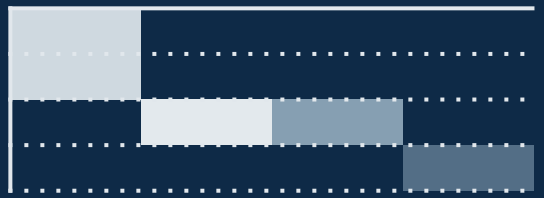


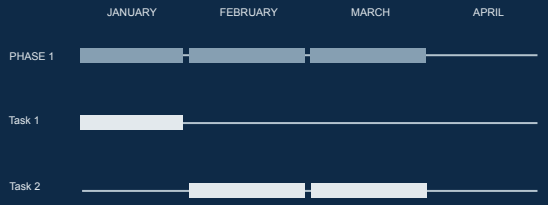

JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
PHASE 1
Task 1
Task 2
PHASE 2
Task 1
Task 2

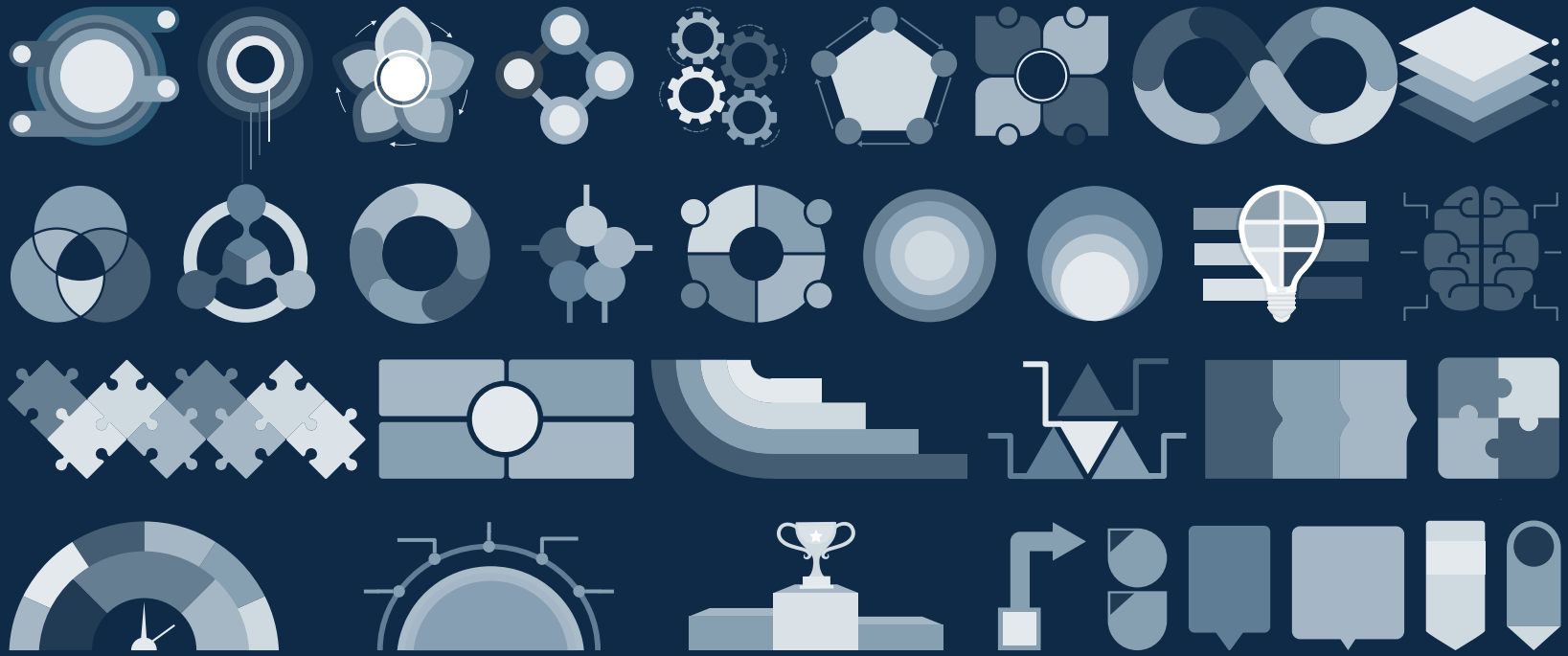









JANUARY
FEBRUARY
MARCH
APRIL
PHASE 1
Task 1
Task 2



...and our sets of editable icons

You can **resize** these icons without losing quality.
You can **change the stroke and fill color**; just select the icon and click on the **paint bucket/pen**.
In Google Slides, you can also use **Flaticon's extension**, allowing you to customize and add even more icons.

Educational Icons

Medical Icons

Business Icons

Teamwork Icons

Help & Support Icons

Avatar Icons

Creative Process Icons

Performing Arts Icons

Nature Icons

SEO & Marketing Icons