Tentative Project Title: **Fantastic Facts and How to Fake Them**

Abstract:

Misinformation and disinformation are powerfully destructive forces in this age of global communication, when a false idea can instantly spread to many vulnerable ears. While they are both forms of false information, the difference between the two boils down to intent: misinformation is false information created and spread regardless of an intent to harm or deceive, whereas disinformation is a type of misinformation created to deliberately deceive the reader and shape a false narrative. Both forms are often shared widely, regardless of whether or not the sharer knows the information is wrong.

We are looking to analyze misinformation, how it spreads, how it is created, who creates it, who are the main consumers, and who benefits. We’ll investigate fact checking platforms, the speed at which lies spread vs truth, the effort required to disprove a false claim compared to the effort required to create one, the various social media sites and their role in disseminating these lies, and the thin line between satire and fake news.

There is a huge amount of data out there, especially regarding the spread of misinformation with regards to COVID-19, U.S. politics, climate change, and current events like the war in Ukraine.

A few datasets we're already looking at:

Ukraine:
https://ieee-dataport.org/documents/propaganda-and-fake-news-war-ukraine
Politics:
https://data.world/d1gi/ten-gop-twitter-top-social-interactions
https://data.world/d1gi/ten-gop-assorted-impact-metrics-and-analytics-data
General signs of credibility:
https://data.world/credibilitycoalition/2019-study
https://data.world/credibilitycoalition/credibility-factors2020
COVID-19
https://esoc.princeton.edu/publications/esoc-covid-19-misinformation-dataset
COVID-19 Vaccine Misinformation
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8648668/

Some Articles:

Articles about combating misinformation:

- Shifting attention to accuracy can reduce misinformation online
- Scaling up fact-checking using the wisdom of crowds
- The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings

NY times article

- How Disinformation Splintered and Became More Intractable

Team Agreement:

Misinformation C.R.A.P. Project

Team Agreement

- We will communicate project development through the Slack channel we have set up, and regular day-to-day communications through our WhatsApp group.
- We will strive for an equal distribution of technical work among all members of the group. All code should be documented well.
- Final design decisions will be discussed among all members; fair compromises should be made when necessary.

- Work hours should be split as evenly as possible (actual task output may differ based on an individual's ability / previous experience). This ensures not only fairness but also learning opportunities for everyone. Our regular communications will be used to ensure everybody is doing their part.
- Work will not necessarily be done together through virtual meetings, but good communication via WhatsApp is expected in a timely manner. Given the nature of distance learning and Extension School study, it is expected that everyone will work within their time constraints to meet previously agreed deadlines.

Signatures:
Ronan Fonseca
Ian Kelk
Pablo Moreno Montes

Date: October 23, 2022

Detailed Project Plan:

Basic Info:

- **Project title**: “Fantastic Facts and How to Fake Them”
- **Team members:** Ian Kelk, Ronan Fonseca, Pablo Moreno Montes
- **Emails:** iak415@g.harvard.edu, rod257@g.harvard.edu, jom1271@g.harvard.edu
- **Team Name:** Misinformation C.R.A.P.

Background and Motivation:

Misinformation has affected all of us. From elections to health issues with the spread of COVID, it has done a huge amount of damage. Never before has it travelled so fast, and so easily as it can with the incredible reach of social media.

As a specific bit of inspiration, Ian Kelk published a paper on automatic fake news detection and presented it at ACL 2022 (The Association for Computational Linguistics) this past spring, and as a result has done a lot of investigating the various problems of recognizing fake news and disinformation - some of it even philosophical in nature. This is briefly covered in the first 3 minutes of the conference video, and more information in the paper itself.

Related Work:

“Fantastic Facts and How to Fake them” is a work title partly inspired by the guides available online on how to spot fake news, some of which include visualizations, such as this example

from Visual Capitalist. One interesting point covered by this infographic is that not all fake news are the same, and there is subtlety and nuance inherent in discourse that may make the decision of labeling something as fake news or not difficult.

This decision has been made by all participants of this group when participating in sections for this course, and particularly when answering questions about integrity in data visualization. Often, the examples shown were published in mainstream media outlets (one notable offender being Fox News), and this also served as a motivation for us to add to Ian's past work by creating visualizations related to the theme.

Naturally, the datasets that we will build our visualizations upon are also a source of inspiration for this work, as the questions we are able to answer will be directly related to what sort of data is available to substantiate our visualizations.

Data:

We have already begun the process of finding a number of datasets to build the visualizations with, and it appears that we won't have to do much in the way of data collection. Misinformation is a hot topic and there is a lot of ongoing research, which makes finding datasets less challenging. Here is a list we have so far:

Ukraine War:
https://ieee-dataport.org/documents/propaganda-and-fake-news-war-ukraine
Politics:
https://data.world/d1gi/ten-gop-twitter-top-social-interactions
https://data.world/d1gi/ten-gop-assorted-impact-metrics-and-analytics-data
General signs of credibility:
https://data.world/credibilitycoalition/2019-study
https://data.world/credibilitycoalition/credibility-factors2020
COVID-19
https://esoc.princeton.edu/publications/esoc-covid-19-misinformation-dataset
COVID-19 Vaccine Misinformation
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8648668/

As well, we are collecting news articles for inspiration on visualizations and what sorts of materials we can cover:

Articles about combating misinformation:

- Shifting attention to accuracy can reduce misinformation online
- Scaling up fact-checking using the wisdom of crowds

- The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings

NY times article that was published Oct 20.

- How Disinformation Splintered and Became More Intractable

Data Cleanup:

We are still in the process of data collection, and have found several readily available datasets relative to the theme, as indicated above. The level of data cleaning required may vary significantly based on the datasets we choose to adopt for this project, and as a result, one must take into account the goals of answering the questions we want to answer and the time required to prepare the data to perform the visualizations, given the time constraints of the project. For that reason, the choice of dataset will be a joint decision between the group members.

On the plus side, both Ian and Ronan are in the data science program and have extensive practice in data cleaning using both python and R, and will be able to perform and document any data cleaning in a reproducible way with either Jupyter notebooks or R markdown.

Project Map

Audience:

Our project deals with the creation and spread of fake news, which is a topic that affects all of us. It is, however, a topic that affects certain groups of people in different ways, and as a result, the delivery and tone of our project can be different if we choose to focus on any particular group. For instance, if we were to target children and teenagers, we could approach this topic based on how it is present and affects them on platforms that they use, such as TikTok, and how certain YouTube channels specializing in viral videos make fake do-it-yourself videos. For the elderly, we could make visualizations that explain in further detail how new technologies such as deep fakes are affecting what we can consider as true or false.

So, if we were to choose between teens, the elderly, and everyone, we would not necessarily have fewer things to say by picking the first two, but the emphasis and delivery would likely be different. However, limiting the scope of such an important yet poorly understood topic feels like missing out on the opportunity to reach out to the largest number of people possible, and it is for that reason that we will not choose any subset of population. We can assume an average familiarity of visualizations, but we can also assume previous knowledge of fake news and propaganda. Everyone believes it exists, even those that are ironically getting this same information from fake news itself.

Once we have established that we will choose “everyone” as a target audience, it is important to define how we can reach out in a way that as many people as possible find easy to understand, but also in ways that they can relate to the visualizations exposed. For instance, the examples of potential topics that were previously discussed could be gathered in a sort of visualization that indicates how fake news affects people at various ages, and how it differs as a result. We can expect a reasonable level of data literacy, in the sense that the classic data visualization graphs should be easily understood, but also original visualizations, provided that they are delivered in a simple, non-technical way that is also engaging and fun. As a result, the data visualizations will not necessarily be extremely detailed, focusing more on explaining the relevance of the problem, its different manifestations, and how it affects our lives.

We’ve also found a detailed timeline of fake news pieces throughout history that are discussed only in text form. We could create a map and timeline to show their occurrences in history and geography.

Questions Related to the Data:

Making visualizations about fake news in an objective way is a particularly tough challenge because one has to try to define what constitutes fake news and what doesn’t. At the same time, because of this nature of ambiguity in what can be classified as fake news, we also have to respect the autonomy of our target audience and make them feel better prepared to, given their own understanding and world view, recognize and avoid fake news. With this in mind, some of the questions that would help in that sense would be:

1. In what ways can fake news be presented?
2. Who are the creators and disseminators of fake news?
3. What is the role mainstream media outlets play in creating spreading or containing the spread of fake news?
4. How to draw the line between opinions and factually inaccurate statements? (As a fictional example of one such ambiguous statement: “The government’s actions are preventing GDP growth”).
5. What *is* fake news? There is really a spectrum of truth to news, because rarely are headlines purely objective. For example, right now the headline on Fox News is **“Abrams' ties to defund police movement surface during debate during attack on sheriff's department”**. Is the emotionally loaded word “attack” really needed there? But does it render a headline real or fake? Is the phrase “**TWISTED TIES:New information revealed about Paul Pelosi's attacker, his ex”** real or fake news based on the use of “twisted ties” which is clearly designed to paint Paul Pelosi in a bad light?
6. How do we spot fake news?
7. Is fake news always presented in an emotionally charged fashion?
8. What is the history of fake news, and how did it spread historically compared to today?
9. What is the motive of fake news?

10. What channels distribute fake news the most?
11. What types of fake news are there?
12. Which type of article has the highest percentage of articles considered as “clickbaity”?
13. Which type of article has the highest level of credibility?
14. Which type of article shows the highest percentage of subheadlines with negative language?
15. Do the narratives of fake news change over time for a given event?
16. What are the real-world effects of fake news?
17. Do spreaders of fake news spread many types of fake news?
18. What countries generate the most fake news?

Describing the Data:

There are several data sources that we listed in the “Data” section of this Process Book, out of which only one or a few will be selected for building visualizations.

Credibility Study 2019, by the Credibility Coalition group. It contains the results of over 2000 articles annotated for credibility, each being rated by at least 3 annotators who committed to an Annotator Code of Conduct. There are several columns in the dataset, some of which are not relevant for this project (such as when each rating was made), which do not need to be detailed, but the main structural columns can be described as:

- Annotator: the code of the specific annotator for that new article
- Media_content: a small snippet of the news article
- Media_url: the url for the news article
- Report_title: the title of the news article
- Task_n_question: several questions about the article, that are answered in a form. The ratings how credible the article is perceived to be, if it seems “clickbaity”, if the headlines match the content, and if the language is extremely positive or negative. The “n” in the colum goes from 1 to 9.
- Task_n_question: The **categorical** answer to the question, what was marked in the form. So there are 9 categorical variables in this dataset that are relevant.
- Task_n_comments: Free comments made by the narrator in text form.

ESOC COVID-19 Misinformation Dataset Fake news and false information on COVID-19 can spread just as quickly as the virus itself. On March 16, the Empirical Studies of Conflict Project in collaboration with Microsoft Research began cataloging misinformation efforts around the pandemic. The dataset has 32 columns, and I will list the ones I think are relevant to our project.

- Reported_On: URL of where the misinformation was reported (in a news article, blog, report). **Categorical**.
- Publication_Date: Date that the news article, blog, report was published. **Ordinal.**
- Twitter_Reference: Captures whether the story mentions Tweets or Twitter. Equal to 1 if yes, 0 if no. **Categorical**.

- Title: Main title of the news article, blog, report from the previous field. **Categorical.**
- Primary_Country: Where the disinformation is spread according to the article. For Tweets or specific social media posts, it's the location the article references as being the targeted location or the coder's best guess as to what that is if none is referenced. If it is ambiguous, place "Ambiguous" as an option. **Categorical**.
- Secondary_Country: Include all other countries or regions mentioned in the article (if any), separated by commas. E.g. Canada, United States. **Categorical.**
- Primary_Language: What language is the piece of disinformation being spread? Limit one language for this field. If a specific screenshot or social media post is available, use the language that is provided. If the article mentions the disinformation spread in a different language (but with no example, like in the previous case), use this language. **Categorical**.
- Secondary_Languages: Include all other languages mentioned in the article (if any), separated by commas. E.g. Spanish, Italian. **Categorical**.
- Main_Narrative: The type of story the piece of misinformation is pushing. What's the content of the story? Note: limit to one narrative per entry. This part is lengthy but interesting, and I include it because I think an interactive visualization about narrative could be fascinating. **Categorical**.
    - **False cures and preventative measures** – the story reports unproven cures or procedures to take to fight COVID-19, including misinformation pertaining to vaccine development and trials.
    - **Nature of the virus** – the story refers to specific characteristics about the virus, who it affects (age or ethnic groups) or how it can be spread (e.g., elderly people are more hit, young people are less susceptible, it cannot be spread on surfaces, etc.). Includes stories which:
        - claim the virus has a supernatural origin, e.g. "COVID is a punishment by God for our sins."
        - predict the virus (e.g. claims the coronavirus occurs every 100 years or that a book predicted the pandemic);
        - claims that the virus is not real;
        - attempt to rename the virus (e.g. referring to the virus as the "Wuhan Pneumonia");
        - spread misinformation on the mortality rate of the virus; or
        - claim that the virus is not a threat or not as serious as credible public health sources suggest it is (e.g. stories discounting WHO or CDC reporting on fatality rates).
    - **Origin of the virus** – the story reports unverified origins of the virus (e.g., it came from consuming bat soup, or it came from eating meat) or false information about sources of infection. Excludes stories that suggest COVID-19 is a designed organism (e.g. bio-weapon, originated from a U.S. lab, was brought over to China by the U.S. Army).
    - **COVID-19 status of individuals and groups** – the story speculates on patient zero identities, individual or group status to include infection or

fatality rates within specific populations, and celebrity test as well as recovery results (e.g. unconfirmed reports regarding Patient Zero in Kenya, avoid certain areas because there's a large Chinese population that have the virus, or Tom Hanks using COVID-19 as an excuse to stay in Australia because he's part of a child sex ring associated with Weinstein). Includes misinformation on the outcome of individuals diagnosed with the virus (e.g. burial areas or funeral homes overwhelmed by the number of deaths, overfilled hospitals, misreported mortality rates in specific countries, etc.), as well as misinformation on the recovery status of individuals (e.g. large numbers of recovered patients are donating their plasma).

- **Government responses** – the story reports false information regarding government or political responses to fight the outbreak (e.g. US national guard locking down the US, or Italian biocontainment protocols). The responses can be emergency, short-term responses or anything related to reopening, stimulus, new rounds of lockdowns, a second wave of the virus, etc. This category should also include any praises or criticisms regarding government responses from the government side. Stories about a return to normal fall under this category if the government is a clear actor (e.g. China announces a return to normal, Taiwan reopens closed stores). A general "return to normal" with no clear order from the government falls under the "other" category.
- **Non-Government responses** – the story reports false information regarding individual, nonprofit, social, and business responses pertaining to COVID-19. Includes misinformation on positive or negative responses taken by non-government actors. For example, any misinformation about reopening by such entities, the breaking of lockdown compliance by non-government actors (e.g. breaking quarantine or social distancing rules, refusing to wear masks, or holding religious gatherings during lockdown), or misinformation about individual protests and backlashes, as well as fake praise for fabricated about non-governmental responses to the pandemic, or claims of the community helping out individuals (e.g. supermarkets providing assistance or restaurants offering free deliveries), and so on etc. fall into this category.
- **False diagnostic procedures** – the story reports unproven diagnostic procedures for the virus (e.g. holding your breath is a test for carrying COVID-19).
- **Weaponization or design** – the story describes COVID-19 as a designed organism or an intentional part of a country's bio-weapons plot (e.g. Chinese, American, Russian plots), or an unintentional consequence of genetically-engineered organisms. Also, includes claims that the virus is being used as a weapon, even if the narrative admits that the virus is naturally occurring.

        - **Other** – for stories which may not fall in any of the previous narrative definitions.
    - Narrative_Description: Adds another level of detail to describe the entry in the Main_Narrative field, according to the proposed guidelines above. Provide specifics on what the content is about with details from the story. For example, if the main narrative is “Weaponization or design”, this could be “US registered a patent in 2003 and used the virus as a bioweapon against China”. Mention if the narrative frames the virus in religious terms and how. **Freeform text field**, not really any normal type of data but useful for exposition..
    - Motive: What is the disinformation trying to accomplish? Given the content and context, record subjective judgement about what the producer of the information is trying to achieve with promoting the disinformation. **Categorical**.
        - **Fear** – disinformation with no discernible purpose beyond stoking fear among the public.
        - **Profit** – stories which are linked to a financial motive, e.g. selling a cure or preventative measure for COVID-19.
        - **Politics** – stories that target specific political actors or groups (e.g., President Trump, Italians in the US, Asian-Americans) to either:
            - Discredit and weaken them (e.g. Xenophobic remarks, smears); or
            - Help their political standing (e.g. COVID-19 is a deep-state plot).
        - **Undermine target country institutions** – efforts to attack governments, international organizations, or specific ministries/bureaucracies (e.g, UN, EU, NATO, Italian healthcare system, etc.).
        - **Downplay Severity** – stories which try to paint COVID-19 and the pandemic as normal or low-consequence (e.g. it’s just like the flu, hospitals are totally fine, it’s no more infectious than other diseases).
        - **False Hope** – stories trying to instill a sense of hope (e.g. X celebrity is turning their house into a hospital to take in patients, there’s already a cure, a vaccine has already existed).
        - **Help** – stories spread by the source with the goal of being helpful (e.g. false cures and preventative measures spread with good intent). Use this category only when the reporting and/or context clearly indicate the misinformation narrative was spread with good intentions.
        - **Other** – content which does not fit into the above, including sarcastic commentary and individuals trying to sell fake stories for the sake of it (e.g. the Tom Hanks story from above, Nostradamus predicted this years ago, Author X predicted this in their book).
    - Motive_Description: Provides greater detail to justify the entry from the Motive field. Write down specifically who the actors are and what they are doing given the coding and the context (one example of “Undermine target country institutions” is “Russian efforts to weaken the US administration and create chaos in the US by passing false information regarding the US response”). **Freeform text field**, not really any normal type of data but useful for exposition.

- Source: How do we characterize the group that is pushing this misinformation? **Categorical.**
  - **Individual actor** – individuals or groups (“common people”) who push misinformation through phishing scams, social media.
  - **State sponsors** – states who are pushing misinformation efforts. Companies – known companies who exacerbate or house misinformation.
  - **Media** – known news websites, TV networks that push misinformation.
  - **Political actor** – known individuals or groups who support a certain politician or party pushing the misinformation.
- Source_Description: Who specifically is pushing this misinformation? **Free form, possibly usable as categorical.**
  - If key field Source contains: **Individual actor**
    - If we know the specific actor or group, name them.
    - If we do not know the specific actor or group, place “General Public.” Typically, we use this for reports that contain chain messages or phishing emails because we are mostly unsure who started the chain message.
  - **State sponsors** - If we know the specific country or state, name them.
  - **Companies** - If we know the specific company, name them.
  - **Media** - If we know the specific news or TV network, name them.
  - **Political actor**
    - If we know the specific government representative (House representative, Senator, Member of Parliament, Minister), name them.
    - If we know only of groups or supporters for a particular government representative, name them as such “<Government representative name> supporter.”
- Distrib_Channel: On which platforms (social media or not) is the misinformation being spread? Record the platform(s) that were mentioned in the articles. Facebook, Twitter, WhatsApp, Kakao Talk, TikTok, Youtube, Podcasts, 4Chan, Website (Personal, Blog) etc. **Categorical**.

Disinformation “half-dozen” dataset. https://github.com/gi-ux/Disinformation-Dozen-TweetIDs

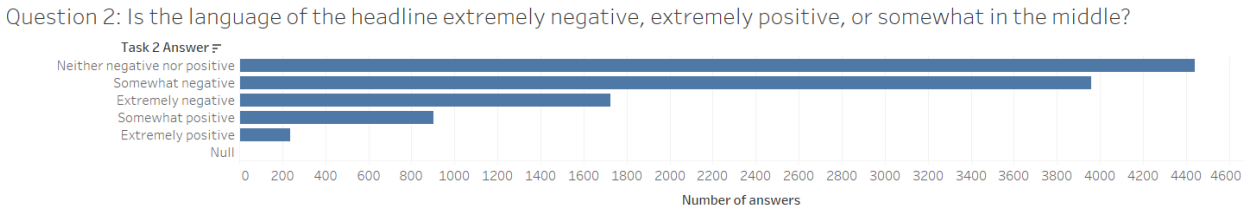
In May of 2021 the Center for Countering Digital Hate produced a report condemning 12 individuals for producing 65% of misinformation spread about COVID-19.

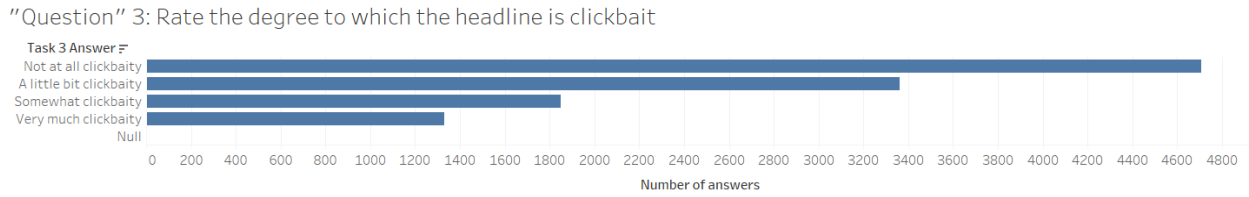
We were able to find a recreation of this data, and most interestingly, of the original 12, only 6 remain with unsuspended Twitter accounts. We downloaded the tweets using the tweet IDs and an app called Twitter Hydrator, and have dubbed the remaining 66 the “disinformation half dozen”. It comprises 15,648 tweets from the 6 individuals, with all trackable information covered for them.

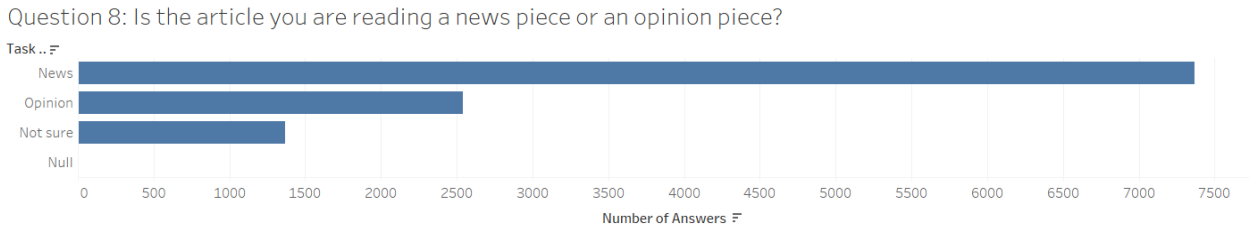
Data visualization with Tableau:

Ronan Fonseca

These are the visualizations made for the **Credibility Study 2019** dataset.
Note: The visualizations are all bar charts because they are (in my opinion) the most suitable form of visualization for the categorical answers that were provided in the chart.



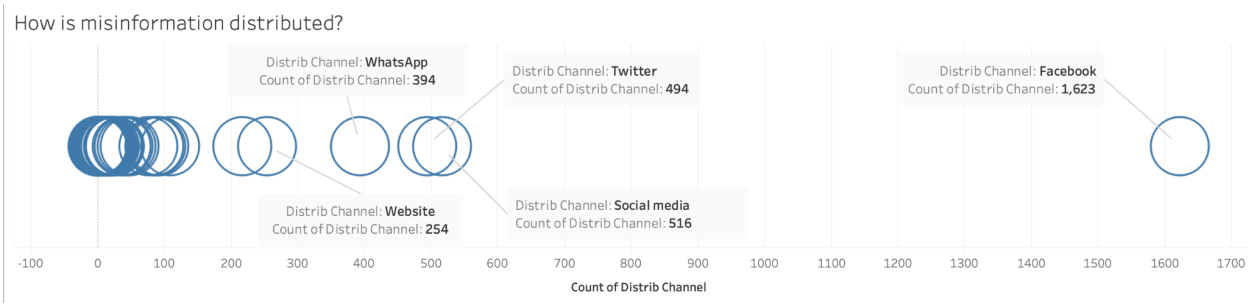




I found that the bar chart for Question 2 differs from the questions proposed because it brought up an interesting point that we had not originally thought of, which is: **How can language be used to change the user's perception of an event, even if it is described with no falsehoods in it?** The bar chart for "Question" 3 is closest to the original question that was made, relative to the ways in which fake news can be presented. Clickbait, at least in my opinion, can be considered a form of fake news, and as someone who has received several links of news that had alarmist headlines of articles with misleading content, I believe that it may be worth it to mention clickbait in this project. Question 8 exposes of this form exposes the importance of understanding if a certain statement is one of opinion or a news piece. Several of the respondents couldn't tell if the article that they were reading were news or opinion, and this may be done intentionally to confuse readers. So, when it comes to the question that was originally made of "how to draw the line", it may be ideal to have media outlets directly label opinion pieces as such.

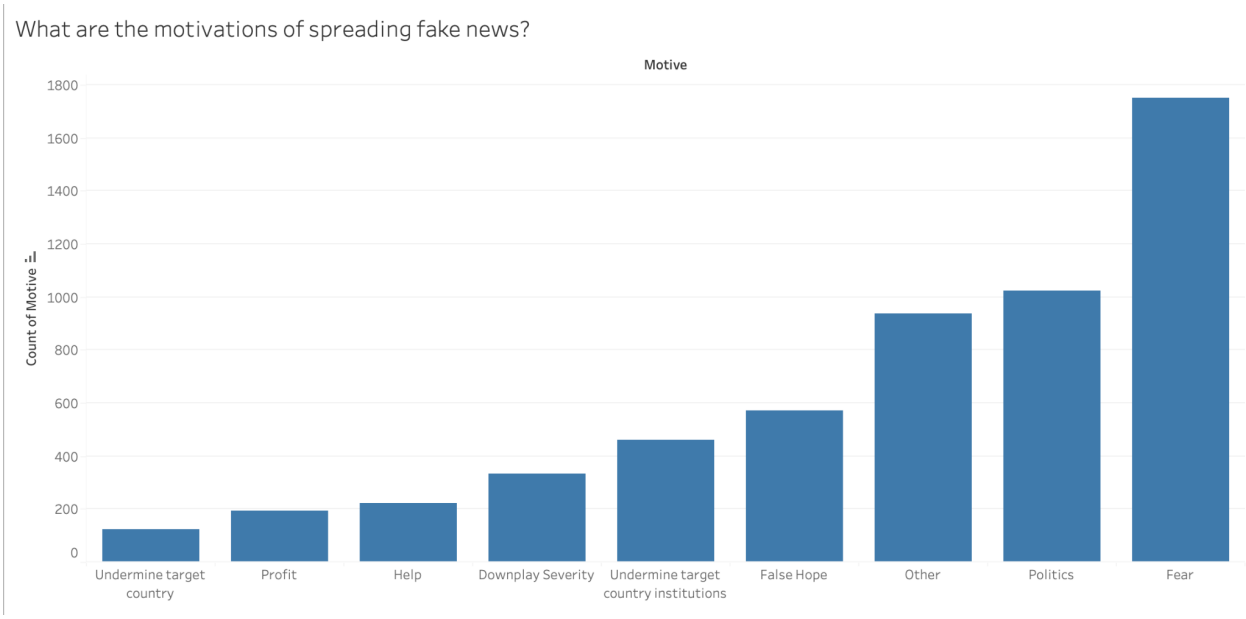
Ian Kelk

These are the visualizations made for the **Covid-19 Misinformation** dataset.
Question 10 - What channels distribute fake news the most?

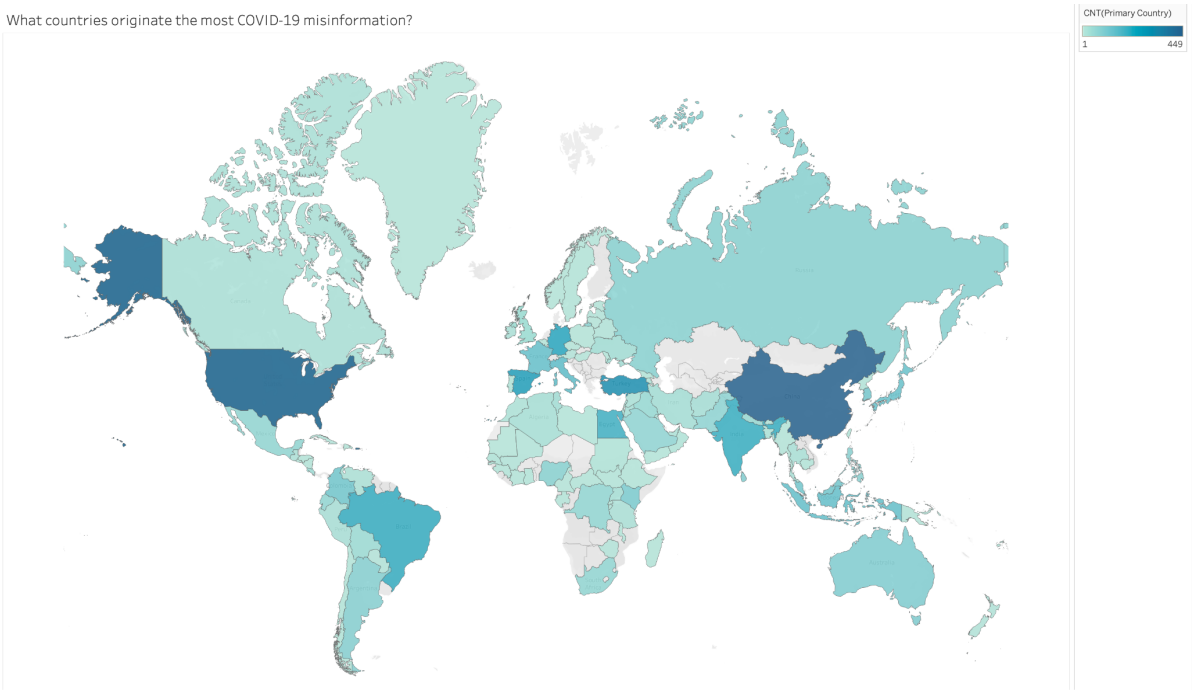


This one was especially interesting to me. I suspected that social media was the biggest culprit on spreading covid misinformation, but I thought Twitter would have just as large an involvement, when in fact Facebook blows them all out of the water. This data does need to be refined and cleaned a bit though, since it does have sub-categories which include Facebook and Twitter combined. As well, WhatsApp was a major player, and also owned by Meta, the parent company of Facebook. It appears that a single company has a huge responsibility on spreading disinformation. I should note that the dataset has 725 categories, so the prominence of Facebook is notable.

Question 9 - What is the motive of fake news?

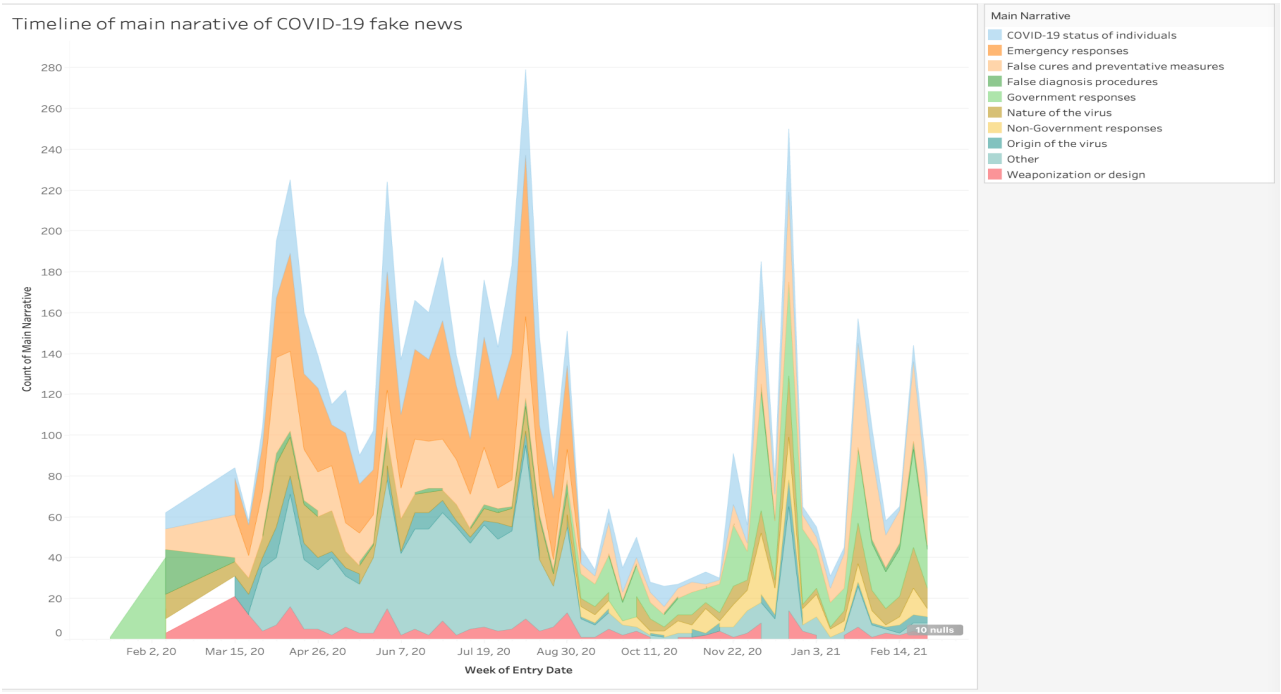


This was a surprise to me as well, as I assumed that politics would be the largest driving hope behind fake news, even COVID-19 misinformation.



The largest creators of COVID-19 misinformation in the dataset are the USA and China, with Turkey, Spain, Germany also significant. In South America, Brazil played the biggest role, and in Africa it was by far Egypt. I thought it interesting that Russia, which has played major roles in political misinformation, was under-represented in this data.

One more for fun - the evolution of the narrative of fake news over time:
15- Do the narratives of fake news change over time for a given event?



It’s interesting to see that the narratives did change over time, with the mysterious “other” taking up much of the narrative from March to September. Otherwise it seems somewhat consistent.

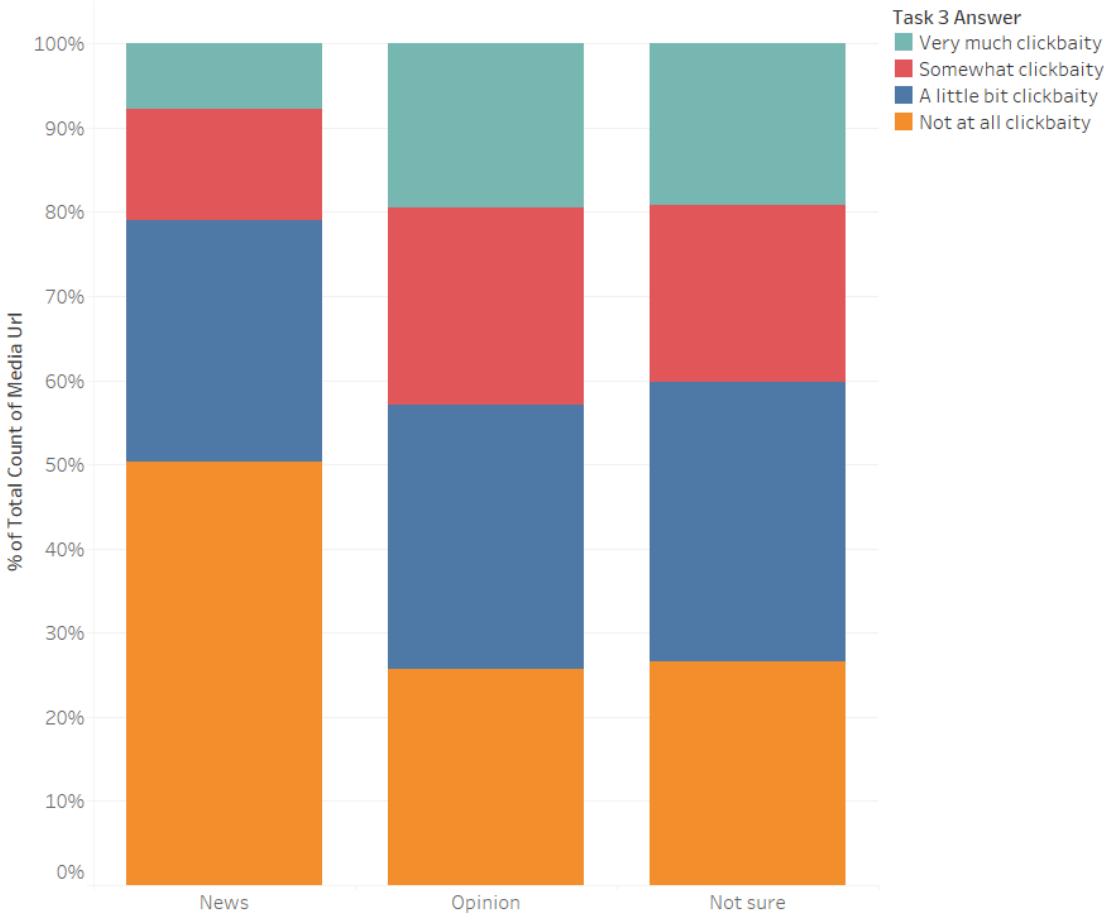
The questions I answered in Tableau were directly related to the questions we came up with as a team, and in fact the visualizations revealed details I wouldn’t have expected - ie that Meta as a company is incredibly responsible for the dissemination of misinformation (much more than any other) and that Russia had a smaller role in COVID-19 misinformation than I thought..

Pablo Moreno Montes

These are the visualizations made for the **Credibility Study 2019** dataset.
I mainly focused on how different types of articles compare to each other on measurements related to credibility. I used stacked charts that add up to 100% because I was interested in seeing each category's share in the different types of articles, regardless of the total volume of articles.

Question 12 - Which type of article has the highest percentage of articles considered as “clickbaity”?

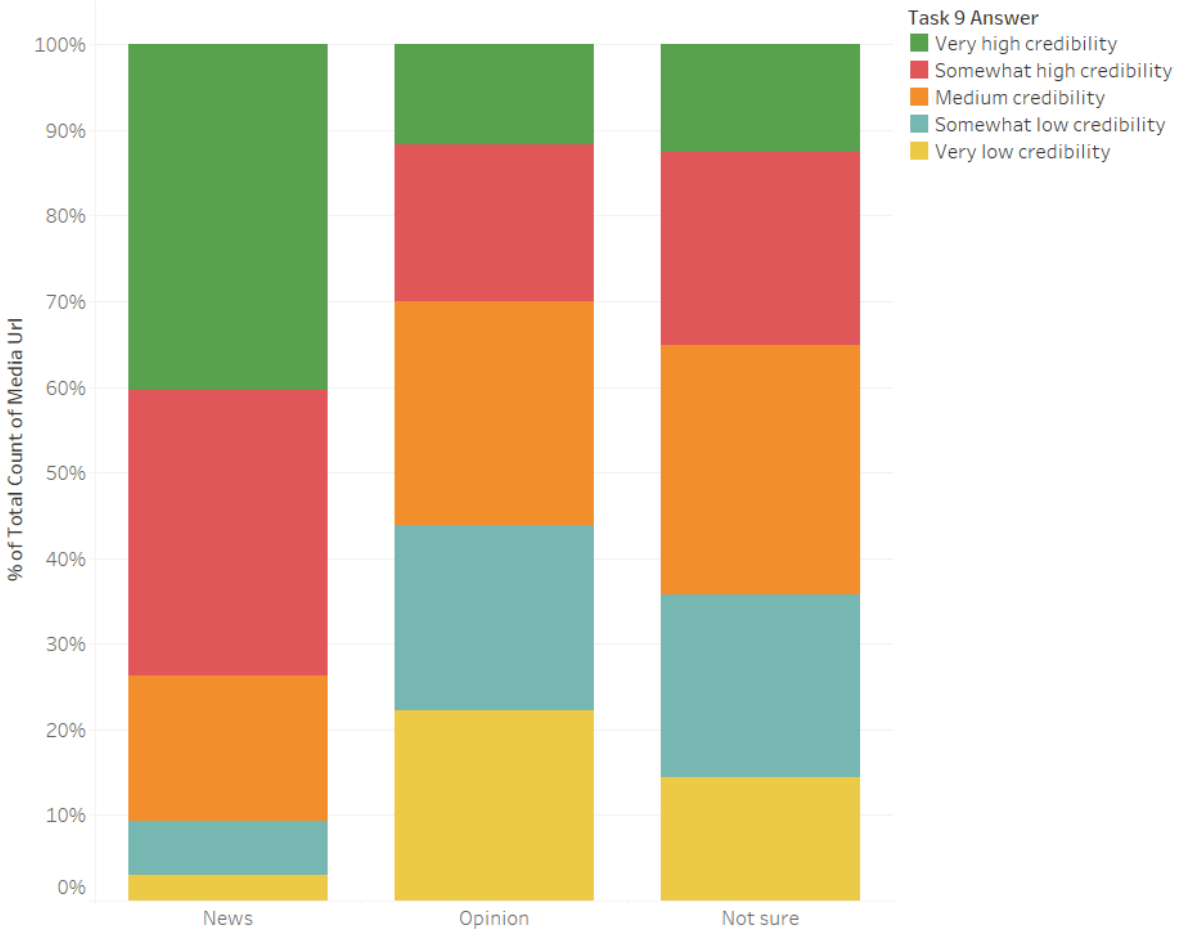
Which type of article has the highest percentage of articles considered as “clickbaity”?



% of Total Count of Media Url for each Task 8 Answer. Color shows details about Task 3 Answer. The view is filtered on Task 8 Answer and Task 3 Answer. The Task 8 Answer filter keeps News, Not sure and Opinion. The Task 3 Answer filter keeps A little bit clickbaity, Not at all clickbaity, Somewhat clickbaity and Very much clickbaity.

Question 13 - Which type of article has the highest level of credibility?

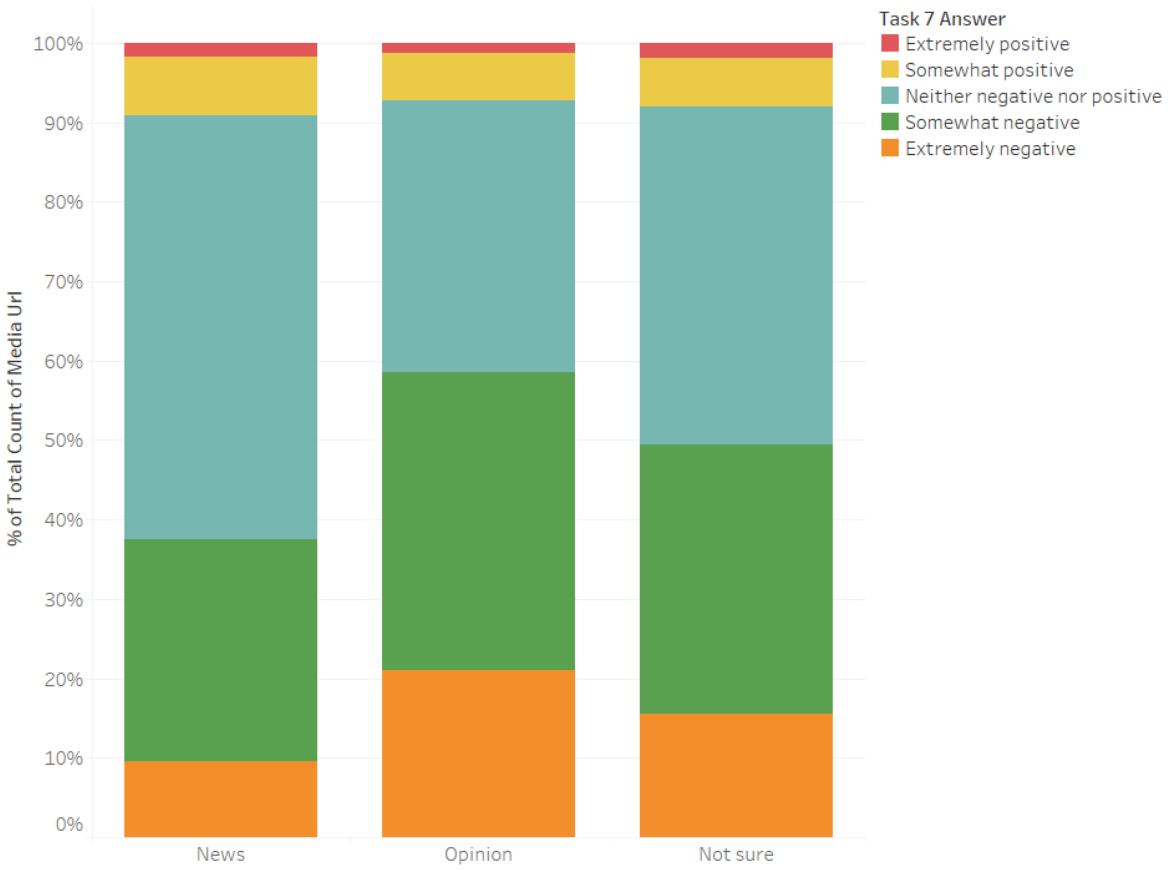
Which type of article has the highest level of credibility?



% of Total Count of Media Url for each Task 8 Answer. Color shows details about Task 9 Answer. The view is filtered on Task 8 Answer and Task 9 Answer. The Task 8 Answer filter keeps News, Not sure and Opinion. The Task 9 Answer filter keeps Medium credibility, Somewhat high credibility, Somewhat low credibility, Very high credibility and Very low credibility.

Question 14 - Which type of article shows the highest percentage of subheadlines with negative language?

Which type of article shows the highest percentage of subheadlines with negative language?



% of Total Count of Media Url for each Task 8 Answer. Color shows details about Task 7 Answer. The view is filtered on Task 8 Answer and Task 7 Answer. The Task 8 Answer filter keeps News, Not sure and Opinion. The Task 7 Answer filter keeps Extremely negative, Extremely positive, Neither negative nor positive, Somewhat negative and Somewhat positive.

These are some of the findings we obtained from the three charts above:

- Opinion articles are considered twice as much "clickbaity" as news articles (43% vs. 21%).
- News articles are twice more credible as opinion articles (73% vs. 31%).
- While news articles and opinion articles have a similar share of headlines with positive language, opinion articles have 20% more negative language headlines.

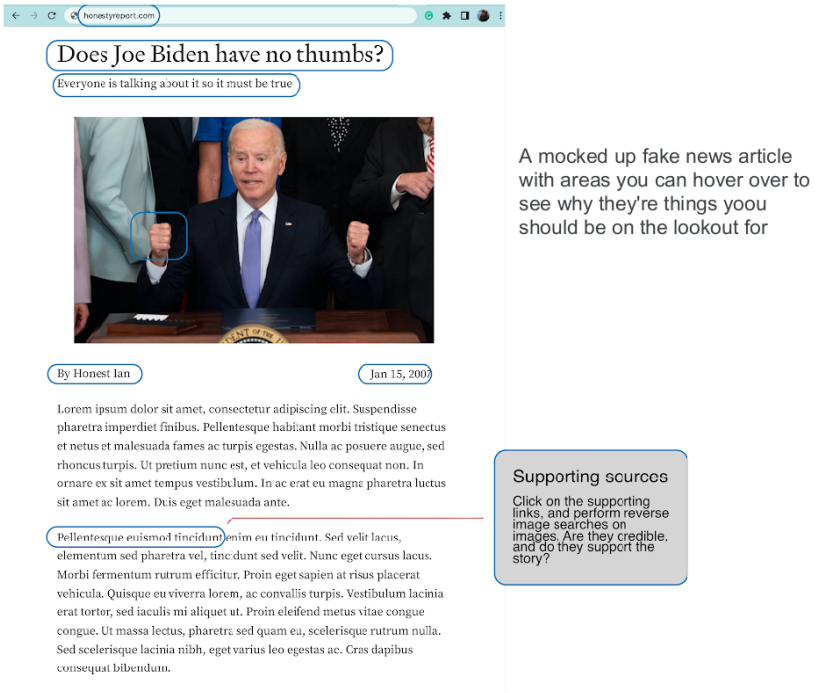
Overall, these charts confirm what the team discussed: there's a clear tendency for opinion articles to be more prone to deliver fake news.

Data, Sketches, Design, Storyboard

Important note: A new dataset has been added, the “disinformation half-dozen” dataset, which is described on the Data section.

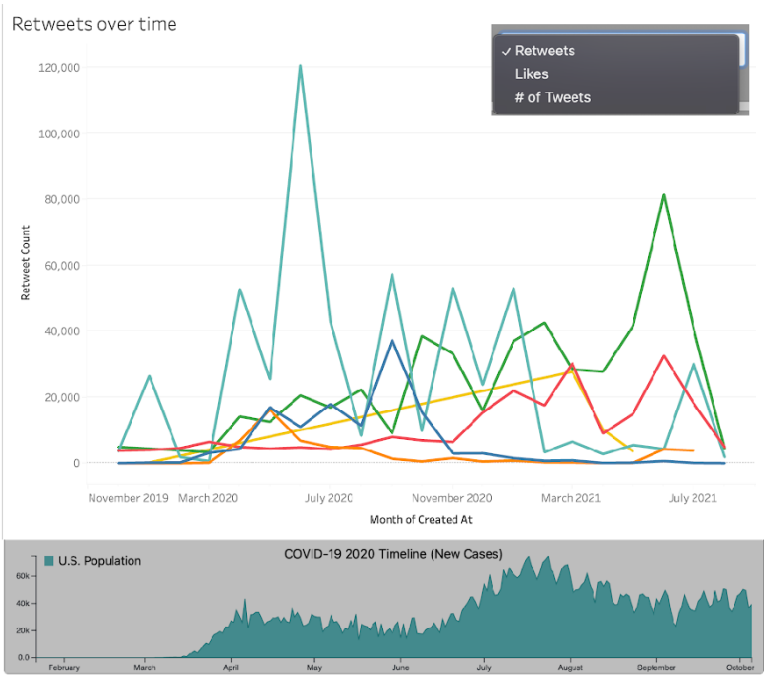
Ian sketches:

1. Q6 How do we spot fake news?



2. Q16 What are the real-world effects of fake news?

   Idea: Grey out other colors while focusing on one color



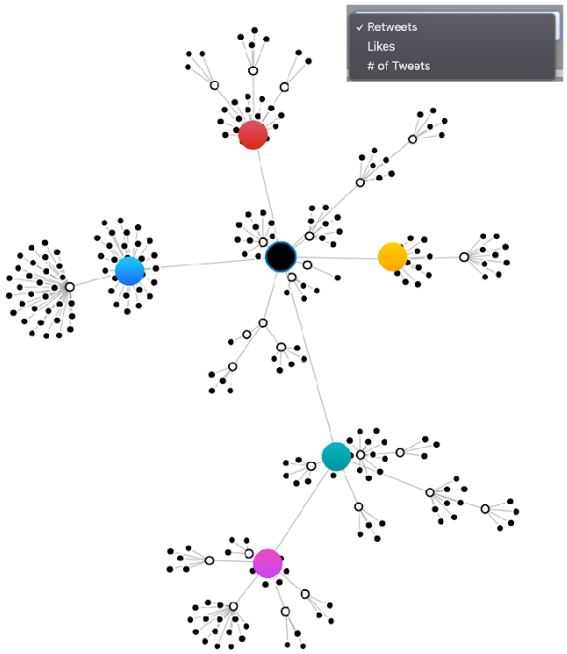
Brushable area chart of covid cases and deaths to compare to retweets, likes, and # of tweets of disinformation half-dozen.

Each color is one of the half-dozen spreaders

3. Q17 Do spreaders of fake news spread many types of fake news?

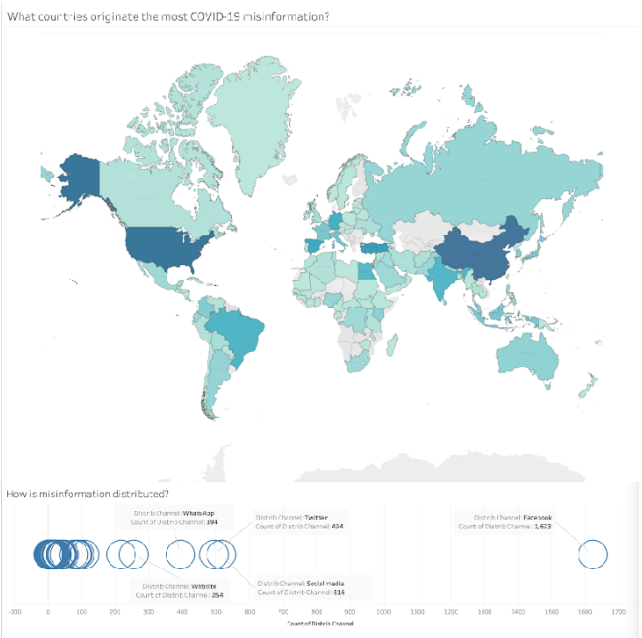
Word cloud as a static visualization of how the spreaders of misinformation are not just limited to one topic. They spread misinformation about MANY conspiracy theories.

4. Q2 Who are the creators and disseminators of fake news?



Rough draft of using a force directed tree to illustrate number of retweets, likes, and tweet quantities to number of recipients

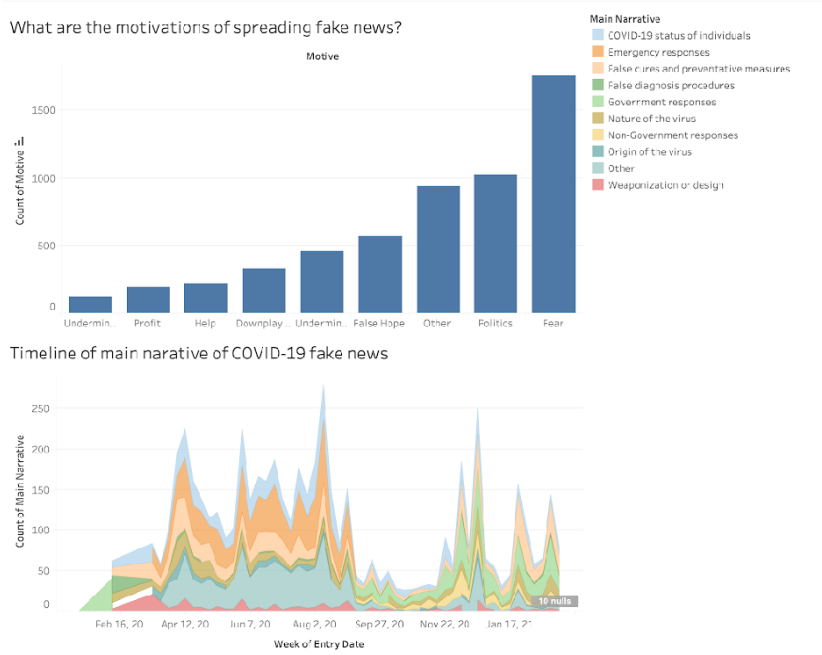
5. Q2 Who are the creators and disseminators of fake news?
   Q18 What countries generate the most fake news?
   Q10 What channels distribute fake news the most?



A choloropleth map and circle view showing originations and distribution channels of fake news

6. Q9 What is the motive of fake news?

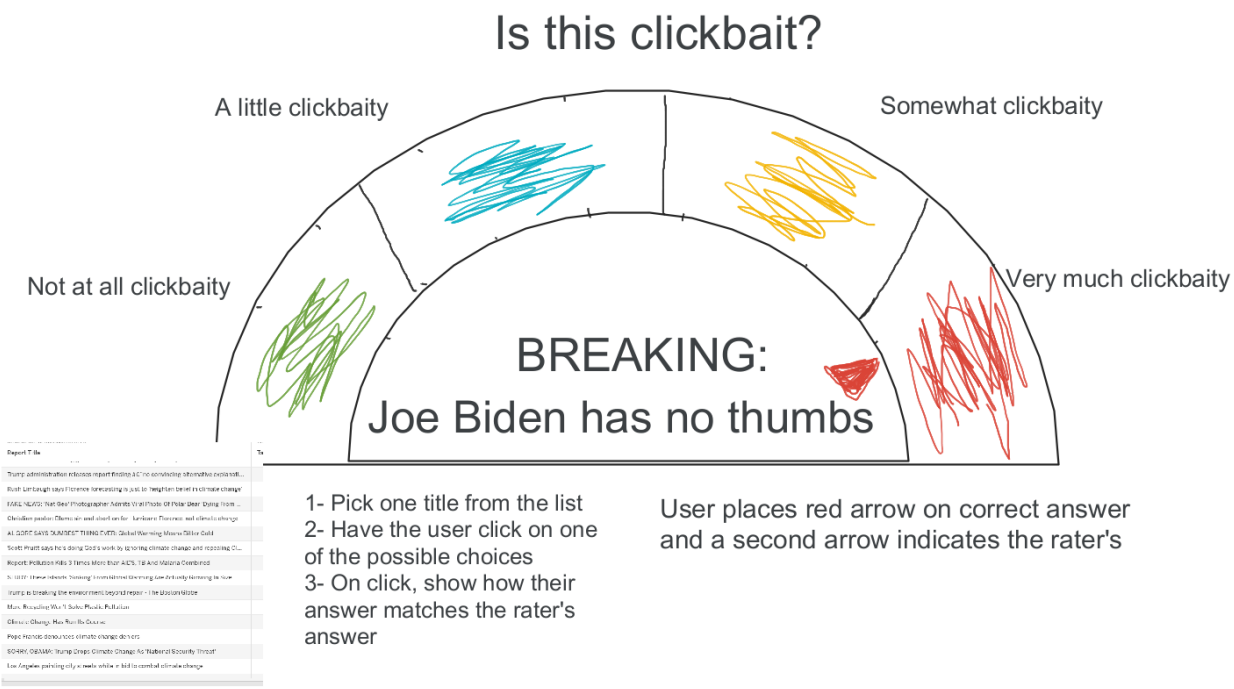
Q10 What channels distribute fake news the most?



Narrative over time of covid misinformation

Ronan Sketches:

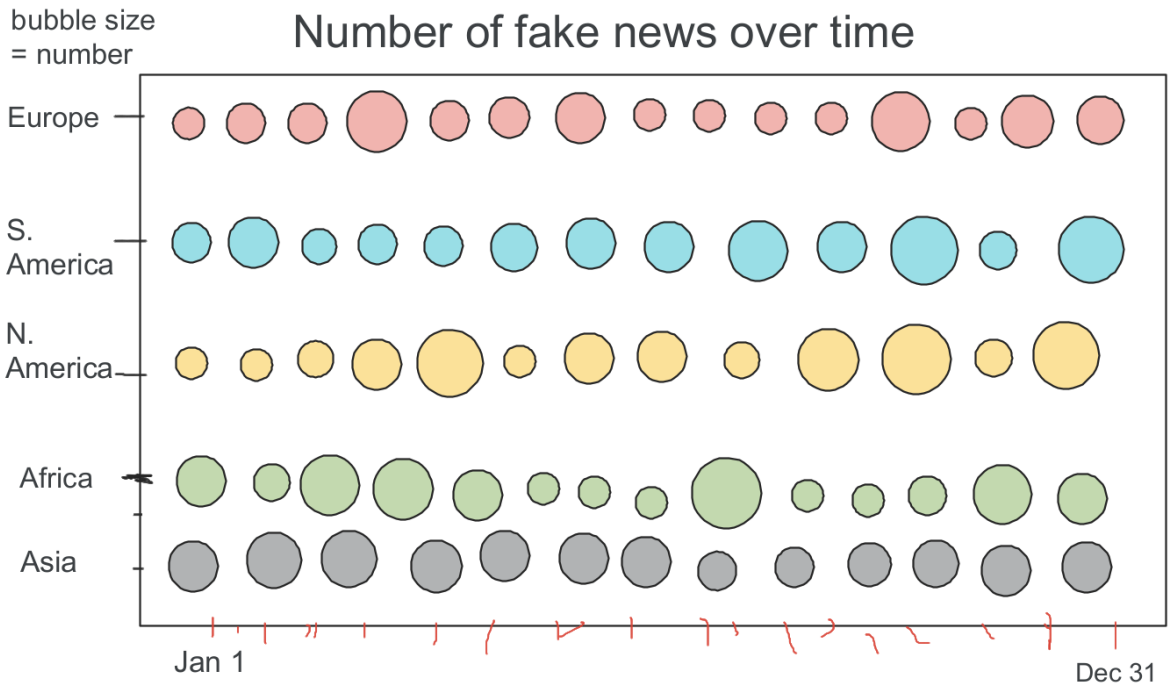
7. Q1 In what ways can fake news be presented?
   Q3 What is the role mainstream media outlets play in creating spreading or containing the spread of fake news?



1- Pick one title from the list
2- Have the user click on one of the possible choices
3- On click, show how their answer matches the rater's answer

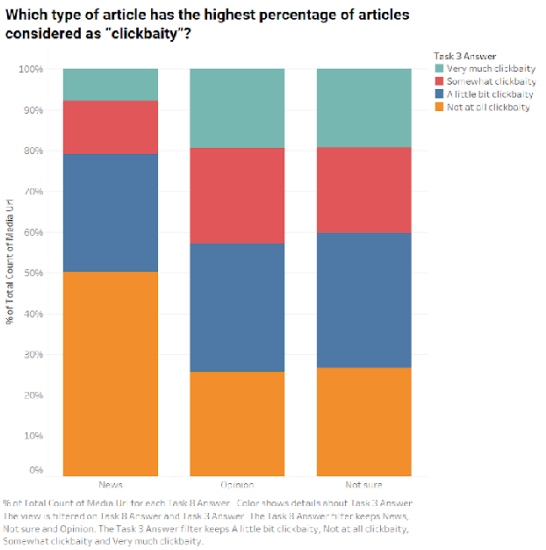
User places red arrow on correct answer and a second arrow indicates the rater's

8. Q8 What is the history of fake news, and how did it spread historically compared to today?



9. Q1 In what ways can fake news be presented?
   Q2 Who are the creators and disseminators of fake news?
   Q3 What is the role mainstream media outlets play in creating spreading or containing the spread of fake news?
   Q4 How to draw the line between opinions and factually inaccurate statements? (As a fictional example of one such ambiguous statement: “The government’s actions are preventing GDP growth”).
   Q6 How do we spot fake news?
   Q7 Is fake news always presented in an emotionally charged fashion?

Opinions, news, or neither?

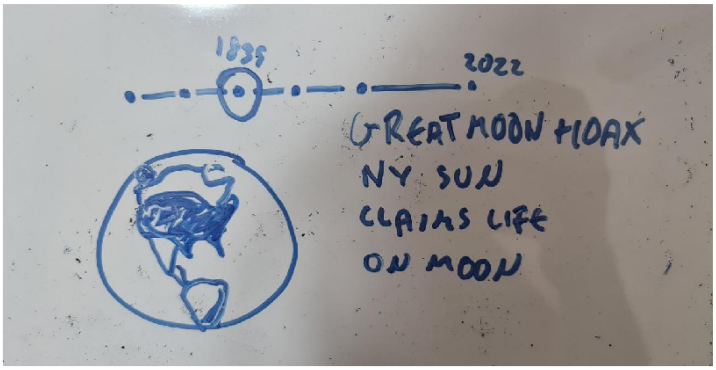


Hover over each subsection and grab a random title from that subsection

In bullet points:
Title of paper
Link (maybe broken?)
Was it considered overly negative/positive?
Other form answers

10. Q8 What is the history of fake news, and how did it spread historically compared to today?
    Q7 Is fake news always presented in an emotionally charged fashion?

Tale as old as time... fake as it can be



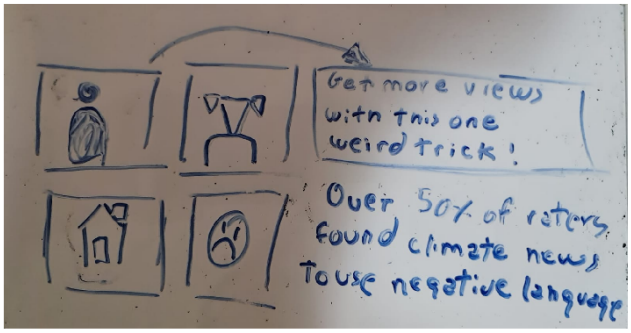
What this is: a timeline of historical fake news events. The user clicks on events on the timeline, this shows the event on the right, and the globe spins to show where the fake news event happened.

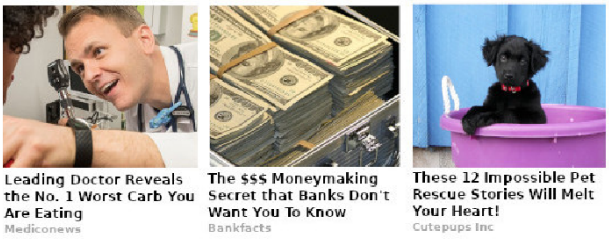
Possible sources:

https://www.commonsense.org/sites/default/files/pdf/2017-08/newsmedialit-fakenewstimeline-85x11.pdf
https://european-seed.com/2022/09/fake-news-throughout-history/

11. Q7 Is fake news always presented in an emotionally charged fashion?
    Q10 What channels distribute fake news the most?

But first, a word from our shady sponsors!





Based on the "chum box" clickbait ads, this would be in the middle of our presentation, each image would show one data pointon fake news related to the credibility dataset.

Example: Get more views with this one weird trick!

The trick: negative language
Over 50% of raters found the news to use negative language.

Voting

For some context, we lost our group member Pablo just before needing to complete this milestone, and as a result the voting has been compromised. Ronan keeps declaring himself the winner, but this is fake news because clearly Ian is the winner since mail-in votes don't count. Jokes aside, we realized that most of our visualizations can be included and some can be combined into small dashboards; this is made more clear with the whiteboard storyboard.

Sketch ID	Question ID	Author
1	6	Ian X X
2	16	Ian X
3	17	Ian X
4	2	Ian X
5	2, 10, 18	Ian X
6	9, 10	Ian X
7	1,3	Ronan X
8	8	Ronan
9	1,2,3,4,6,7	Ronan

10	7,8	Ronan X
11	7,10	Ronan X

We believe that the chosen sketches are the best ones possible to draw the narrative storyline that we are aiming for, being informative while also playful, with a degree of interactivity, and also encompassing the different datasets we covered. The reason we chose many of the visualizations was twofold. One is that we are down to two people, so we only had to choose from ten instead of the expected 15, and the other is that a number of these visualizations can be combined into mini-dashboards since they use the same source data.

Sketch data storyboard:

Identify Insights:

Ronan Fonseca insights:

1- Fake news can come from anywhere and can take several forms.
This comes from observing the credibility data and seeing news outlets such as the BBC publishing articles that were considered of “medium credibility” according to the rater. It can also be spread by individuals on social media, also with significant impact.
2- There is a certain amount of nuance to considering something fake news.
What I consider to be a clickbait title for a news piece with extreme language may not be what another person considers it to be. This is evident from the credibility dataset.
3- Misinformation kills.
We were observing the covid retweets dataset and the covid deaths dataset from the homework and couldn’t help but draw this obvious, self-evident insight.

Ian Kelk insights:

1-The tops spreaders of fake news also spread pretty much all the other conspiracy theories. This was observed from looking at the hashtags most used by the “disinformation half-dozen” that we analized and created a word cloud from.
2- Fear is the top motive of fake news. This was quite disturbing. The creators of fake news are malevolent bad actors who seek to harm people, and this fact really drives home the point that we need to address this problem.
3- Twitter is far less of a problem than Meta platforms. This may change in the future, as Twitter is becoming an unknown quantity as the ownership has changed and seems to be far less motivated to moderate fake news.

4- Fake news has real consequences. The rise of covid misinformation is directly responsible for increased deaths. While this has happened in previous pandemics, never before with the speed that it does via social media.

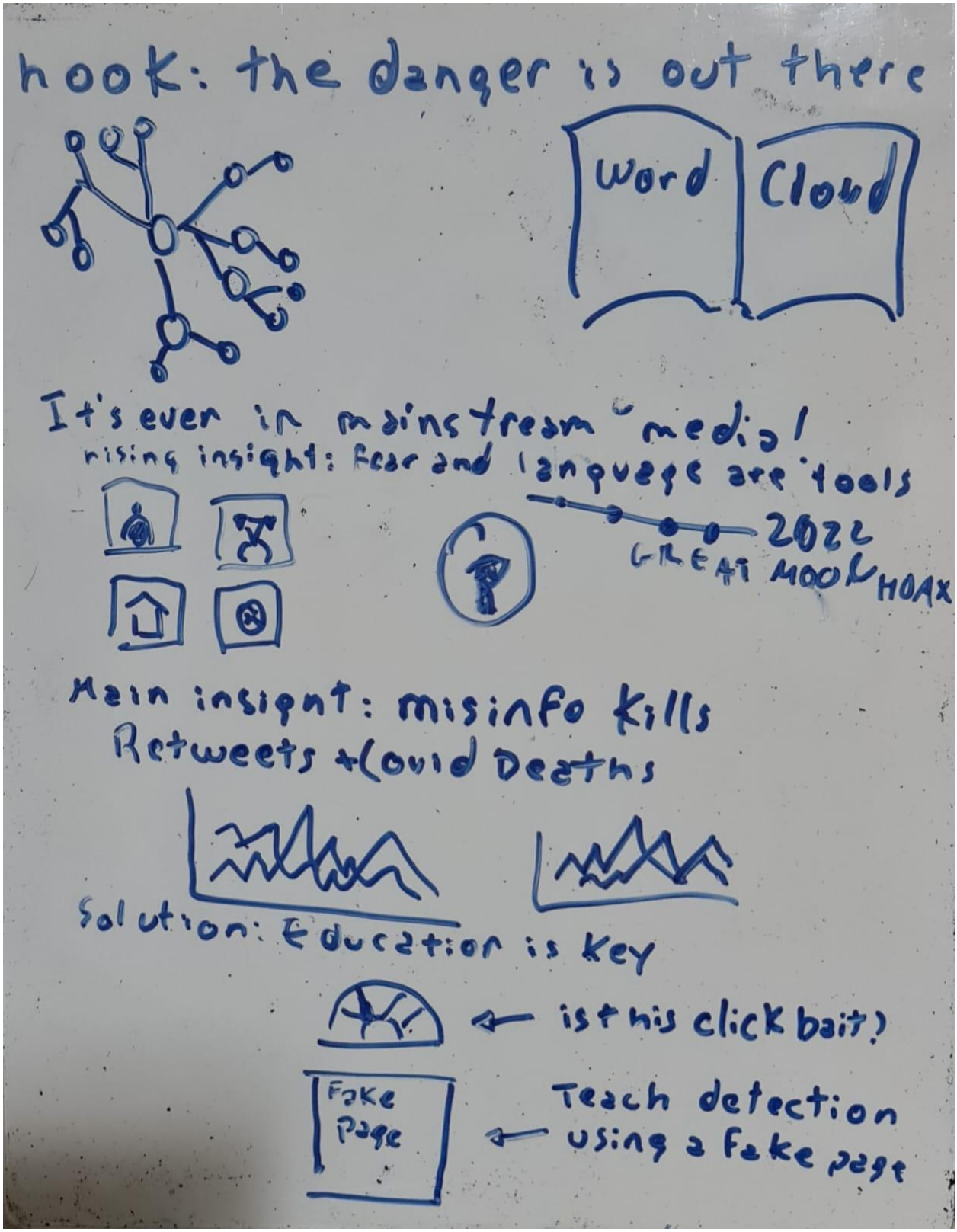
Pick a Main Insight:

The one main insight we picked is simple. **Misinformation kills**. It is concise, informative, impactful. The group members felt repulsed as we were exploring the data and deeper into the rabbit hole of the fake news network, particularly the “Misinformation Dozen”, and particularly “The Defender”, a “news outlet” specializing in misinformation run by one of the people in charge of mass spreading misinformation on Twitter. Explaining the importance of detecting and combatting fake news is as simple as pointing out: it kills.

Each visualization has been resized and reorganized to fit the whiteboard, and the actual visualizations will take more space on the window and likely not be placed side by side. For instance, the fake page visualization is much larger than all others but it only takes a very small space on the whiteboard, and the Covid Deaths + Retweets timelines will likely be stacked as shown on the visualization sketch. We likely will have more static pages in between visualizations with more exposition, but this short storyboard outlines what we see at a minimum.

An important note, chum boxes are well known but nobody seems to know what they are called: https://en.wikipedia.org/wiki/Chumbox

Storyboard


hook: the danger is out there
Word
Cloud
It's even in mainstream media!
rising insight: fear and language are tools
2022
GREAT MOON HOAX
Main insight: misinfo kills
Retweets + Covid Deaths
Solution: Education is key
← is this click bait?
Fake Page
← Teach detection using a fake page

Prototype Stage

Outline of Steps Taken:

During this stage, the remaining team members (Ian and Ronan) have done the following:

1) **Created an initial design** for the project with the final storyboard, in the form of a Webstorm project called projectScaffold. **Interactions are clearly described** (mouseovers, clicks) in an area called “Description” below each visualization placeholder. Titles and subtitles are further used to indicate the **Storytelling** aspect of the project, in a clear “So what” manner.
2) **Created each one visualization**, Ian’s being related to Spread of Covid News over Time and Ronan’s being related to the timeline of fake news. Ronan’s visualization in particular has the i**nnovative** aspect of tying a draggable globe and a timeline in order to display the history of fake news in a way that could not be (as far as the extent of my knowledge) done in Tableau, while Ian’s visualization has animation elements and selectable patterns that are very interesting visually and cannot be done without the help of the D3 framework or (presumably) some other JavaScript visualization framework.
3) Each has **edited/cleaned the datasets** that are/will be used in the project, using Python and Jupyter notebooks, in a reproducible form. As is common in our lab and homework assignments, further adjustments not related to cleaning may be necessary to make the data more convenient to displaying them, like making objects with properties for each observation, but this will be done during the creation of the visualizations as part of the natural process of building them.

To make things easy to view, we’ve uploaded our project scaffold and first 2 visualization prototypes to Github:

Project Scaffold

- Shows proposed layout. These images are just placeholders; the final visualizations will share a common color theme and be a cohesive whole. We’re considering making it resemble a newspaper.

  https://iankelk.github.io/prototype_v1/project_scaffold/

Visualization Prototype 1

- Fake news over time using force directed visualization. You can change the type from “Global” to “By Category” and the animations are rather nice.

https://iankelk.github.io/prototype_v1/vis1_covid/

Visualization Prototype 2

- Draggable globe and a timeline in order to display the history of fake news. Currently it displays random colorings of the map and imagery and text from what we've collected online. We plan to make the globe link to the timeline and display and rotate as they're displayed.

  https://iankelk.github.io/prototype_v1/vis2_credibility/

Cleaning

The folders **vis1_covid_cleaning** and **vis2_credibility_cleaning** contain Jupyter Notebooks with how we cleaned and prepared the data. There was quite a bit of aggregating for the covid data that was needed to take a 5.5 mb Excel file and extract aggregated JSON files containing weekly data for the various categories.

Prototype Stage 2

Progress Report:

Our visualization project is ambitious, and we have not yet completed absolutely everything we consider to satisfy our vision for the project. However, this was not due to a lack of work. Quite on the contrary. This week, we have done:

- Completed the timeline visualization, with a spinning globe, animated timeline and transitions.
- Completed the bubble chart visualization.
- **Created an intermediate "chum box" visualization with interesting facts on fake news using event listeners to create html elements.**
- **Created a completely new visualization with a force-directed graph**
- **Updated our scaffold to use a "slide-show" format that allows the user to scroll down to see each visualization.**
- Started work to place all newly created visualizations there.

Some of the work that is on the works is adding supporting text between visualizations, which will serve to further drive our points and connect one visualization to another. Some of the visualizations we proposed may not be completed (such as the "gauge" visualization), but we

are confident that we will be able to create an entire storyline with proper supporting text using the (several) visualizations we already have.

We have continued work on Github:

Updated Project Scaffold:

https://iankelk.github.io/project/

This is the updated version of our project scaffold, including the structure with scrolling down that we are aiming for. The implemented visualizations that are interactive there are the “globe and timeline (visualization 10 in the process book)”, “chumbox” (visualization 11), and “bubbles” (visualization 8). We encountered an issue with variables sharing the same names and HTML elements sharing IDs that will require further refactoring work, which is why this part is not fully complete, and the Force-Directed graph is not available there. There are also placeholders, and we may cut some visualizations out if required by time constraints, without hurting the overall narrative that we want to make.

For convenience, we also have the visualizations separately accessible:

Updated Visualization Prototype 2:

https://iankelk.github.io/vis1/

New Force-Directed Twitter Interaction Graph:

https://iankelk.github.io/vis2/

Updated Visualization Prototype 1:

https://iankelk.github.io/globe/

New “Chum box” fun visualization:

https://iankelk.github.io/chumBox/

Data Cleaning Files

Our submission also includes additional data cleaning files, which were included in covid_cleaning_python_notebooks.

Think-Aloud Study

Form

Tester Name: Jerry Asuncion
Tester Email: jea374@g.harvard.edu

General Observations from the think-aloud study:

Our very first visualization, the force-directed graph regarding the Misinformation 6, is not very clear in terms of content. Some explanation needs to be made regarding Twitter Interactions, and why we chose to display these particular people.

What does the tester like about your data story?
The tester liked the development of our data story in the scaffold and particularly mentioned the quality and content of our visualizations. He liked how the final page ends with a positive/empowering message ("This is it! Fake news is out there, but you CAN detect and avoid them now!).

What improvements does the tester point out?
Make the Twitter-related part of our visualization more clear by providing context (we have not described the report from the Center for Countering Digital Hate in our visualization, for instance). He naturally suggested we complete our visualizations, and suggested we work on improving aesthetics to make it more pleasing.

Discussion

Discuss the results of the think-aloud study in your team. In your process book, answer the following questions:

- Based on the results of your 'think aloud' study, what would you improve in your data story?

We should make it more clear why we are displaying information regarding the misinformation 6 and what role they play in the spread of misinformation on social media.

One thing that the tester did was scroll past the “Chum box”, which meant that it was not clear enough that it was interactive, despite the text warning.

- Are there any additional insights and visualizations you would use? Would you amplify or change your message? Did your narrative work? Did the tester get your takeaways?

The tester got our takeaways and narrative, but the message can be clarified with further clarifications and supporting text.

- Decide as a team which of these improvements you will implement and write down your decisions and why you made them in your process book as a numbered list.

1. Add clarifying supporting text (especially for the Twitter interactions graph) to improve the narrative.
2. Improve the “chum box” to make it draw more attention.
3. Complete scaffolding
4. Make refinements to visualizations in terms of text and styling

- Implement the intended changes and check them off your list (e.g., adding “done”). You can distribute the tasks among your team members. If you are unable to implement specific changes, please explain why and describe the expected results in your process book.

☑ ~~Add clarifying supporting text (especially for the Twitter interactions graph) to improve the narrative.~~

Done. Actions taken:
Introductory slides were added explaining the context of fake news and how they tie into the whole context of fake news and the rest of our storytelling. One slide was specifically devoted to explaining the importance of the Misinformation 6 in spreading misinformation.

☑ ~~Improve the “chum box” to make it draw more attention.~~

Done. Actions taken:
Added introductory slide with “clickbait style” context just before the “call to action” part of our visualization, to act as a sort of comic relief with additional educational content for fake news. Revised the explanatory content with additional explanations in a playful style, and red blinking text that will definitely catch the attention of the viewer.

☑ ~~Complete scaffolding.~~

Done. Actions taken:

All visualizations have been incorporated into our single-page index.html and placeholders have been removed. The storyline is enhanced and supported with additional slides as needed.

☑ ~~Make refinements to visualizations in terms of text and styling.~~

Done. Actions taken:
Solved issues with conflicting HTML IDs and reduced the size of some visualizations to make them better fit smaller laptop screens. Removed CSS styling relative to containers in each visualization which significantly reduced the digital ink used in our visualizations. This greatly improved visualization quality, both in terms of Tufte's guidelines and C.R.A.P principles, by relying more on alignment and similarity and less on lines separating rows and containers with borders.

Final Submission

Final changes made to the project:

A lot has changed from the initial scaffold we submitted in the Prototype Stage. Some notable changes were:

- The inclusion of a "Story" visualization based on the code for the visualization "Philippines - Plastic Waste and its Sachet Economy" by Ryan Joshua Liwag. It explains the distribution of fake news from the Covid Misinformation Dataset internationally and the impact of the participation of individuals in spreading fake news.
- Turning the static word cloud (formerly a static image) into a full visualization implemented using D3.js
- Implementing the fullpage.js and aos.js libraries. These libraries allow all of the visualizations to be gathered as a single page, with texts fading in at specified intervals.
- Unified the styling for all visualizations, being mindful of data-ink ratios and C.R.A.P. principles.
- Added background and supporting images to slides to make them more informative and visually appealing.
- Added help tooltips or comments to all visualizations as needed.

The entire project underwent several refinements to the page structure, code, visualizations, storytelling, and aesthetics. It has been hosted on GitHub, and its use has been essential in coordinating the work between the 2 project members. The links for the project video and its website are available below.

Data and Descriptions

Our project, “Fantastic News and How to Fake Them”, has used multiple datasets, and within those datasets, we have manipulated them in many ways to make them usable for each visualization. This document describes the datasets used and details the columns for each of the files that were used in the visualizations.

Disinformation 6 Dataset

The “Disinformation 6” Dataset was extracted from Twitter using the list of IDs of tweets made by the Disinformation Dozen, available on GitHub. Since Twitter’s Terms of Service do not allow sharing of complete JSON datasets of tweets, only the Tweet IDs were available on the Github repository. Then, we needed to use a tool called Twitter Hydrator to obtain the data currently available on Twitter for our project. Half of the members were suspended from the platform, which is why we could only obtain the interactions of the remaining ones. This information was then used to create three files, each being used in a different visualization.

misinfo6.json

After cleaning, this dataset is used as a JSON file in the force-directed graph titled “Who are the Disinformation 6 Contacting?”, and its columns are:

"username": the user’s username
"name": the user’s real name
"verified": boolean variable if the user is verified on Twitter
"description": the user’s self-description in their “Bio” section
"protected": false,
"location": the user’s self-described location.
"created_at": date/time of the creation of account.,
"followers_count": how many users follow this account.,
"following_count": number of users this account follows,
"tweet_count": number of tweets,
"listed_count": number of Twitter lists,
"group": number used to group the user according to the interaction of the Disinformation 6 members.

covid_vs_tweets.json

The timeline of COVID-19 deaths vs Disinformation 6 tweets was used by combining the Covid-19 deaths dataset provided to us for Homework 10 and the tweets by the Disinformation 6. The file is organized as follows:

“covid”: contains a list with an index, a date, the number of cases, and the number of deaths for that day.
“tweets”: contains lists for the number of favorited tweets, retweets, and tweets for each of the Disinformation 6 members for each day of the pandemic in the dataset.

Screenshots of the nested data structure of this JSON file are added for further clarification:

wordcloud.csv

The top 150 words on the tweets made by the Disinformation 6 were made into a word cloud, in which the bigger words were the ones used most frequently. This file is used for that visualization, and only has two columns:

“key”: the word tweeted.
“value”: the number of times it was tweeted.

ESOC COVID-19 Misinformation Dataset

This dataset, available here, contains data collected by the researchers with misinformation efforts on social media relative to COVID-19. It has categorical data relative to the motives for this misinformation, the narrative being pushed, and the region that misinformation is related to. This data was then cleaned and separated into 3 separate .json files for the “Misinformation Over Time” bubble visualization. Information from the ESOC dataset was also used in the “Disinformation: A Story Across the Globe” visualization, in a CSV file.

motive.json, narrative.json, region.json

For the “bubbles” visualization, the ESOC dataset was broken down into three JSON files, named motive.json, narrative.json, region.json, each of which shares both a date and a count of tweets for that date, as follows:

"date": "2020-01-27T00:00:00.000Z",
"count": 1

Depending on the file chosen, the JSON object will contain one of the columns below, which are used in the category selector of the visualization and serve to aggregate the data:

"motive": "Downplay Severity"
"narrative": "COVID-19 status of individuals"
"region": "Africa"

fake_news_story.csv

This file is used in the visualization to count the number of documented fake news cases across the world. It has the following columns:

“country”: the country from which the fake news originated
“country_code”: a three-letter abbreviation of the country name. Example: Brazil = BRA.

“lat”: the country’s latitude
“long”: the country’s longitude
“num_fake_news”: the number of fake news events recorded for that country

Historical Fake News Dataset

This dataset was manually created by us by collecting images and information from several online sources. It contains a list of historical fake news ordered by year, with images, a describing text, and sources for the images and text. It is available as a tab-separated document.

historyfake.tsv

This dataset, in tab-separated format, contains the following columns:
“id”: used for uniquely identifying each fake news event
“year”: the year in which the event took place
“name”: the name of the country in which the event occurred. Called “name” to match how the country names are referenced in the GeoJSON data used in the globe for this visualization.
“image_code”: the name of the .jpg file to be loaded in the visualization.
“Img_source”: the source of the image.
“source”: the source of the news content.
“alt”: the description of the image used in the “alt” property of the HTML element.
“event”: the name of the event.
“news”: a text describing the event.

Project Video

- https://www.youtube.com/watch?v=V8gTSvInKDA

Project Website

- iankelk.github.io/project/