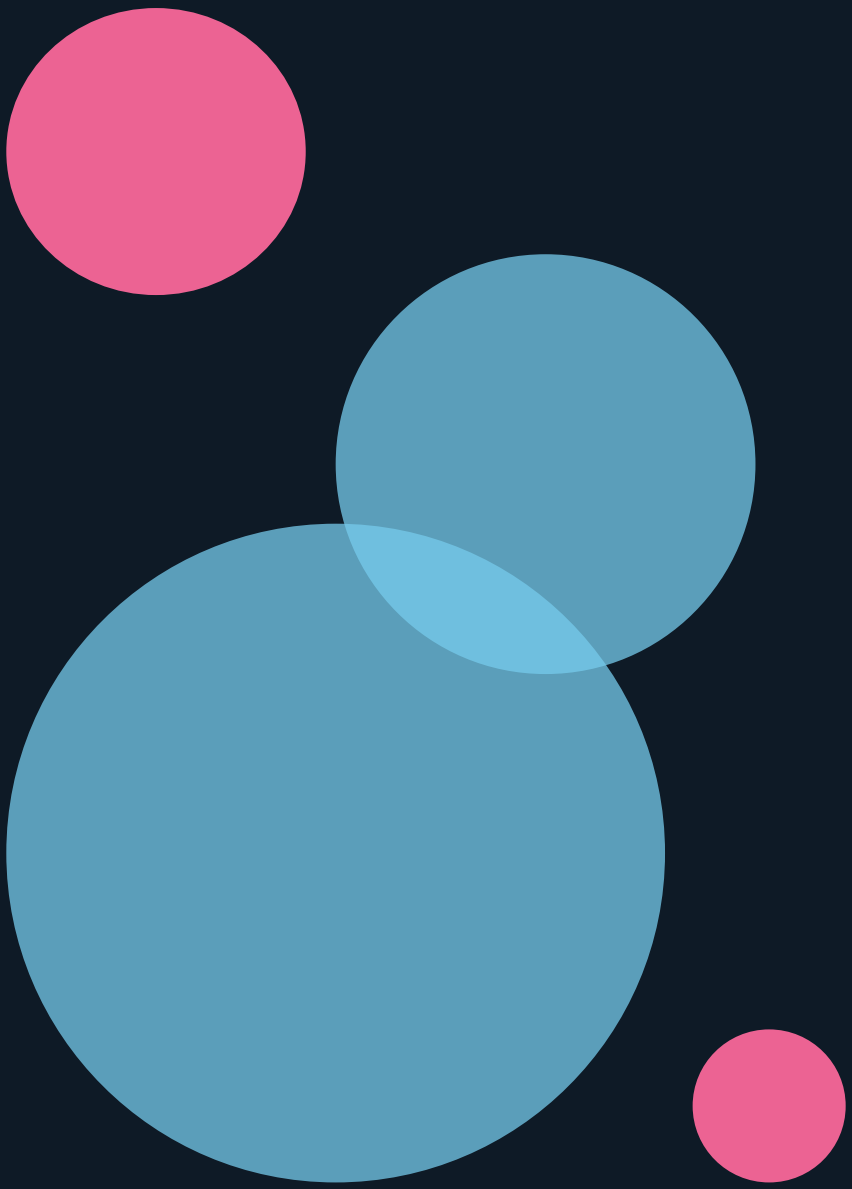


COVID-19- TWEETS VIZ

THE INFLUENCE OF FAKE NEWS IN 2020

2020 | IMAC ENGINEERING PROGRAM

LANDRODIE Nils
LEFEVRE Erwann

QUIRIN Cloé
SCAVINNER Vincent

Université Gustave Eiffel

Chapter

TABLE OF CONTENTS

Chapter

SUMMARY

MediaFutures/

Applying for the track

Artist for media

Covid-19 TWEETS VIZ is a data visualisation project that answers MediaFutures Open Call, following the Artist for media track. The various projects answering the call are based on the use of data on climate, health, ecology and science to offer experiences that encourage citizens to become more involved in these causes.

As the COVID-19 virus quickly spreads around the world, unfortunately, misinformation related to COVID-19 also gets created and spreads like wildfire. Such misinformation has caused confusion among people, disruptions in society, and even deadly consequences in health problems.

To understand and analyse the diffusion and the impact of such information, we have decided to link the information, true or false, relayed by Twitter users to their popularity and the impact they could have. This impact is mainly linked to current data linked to an account on a social network: number of subscribers, average number of retweets, etc.

Chapter

INNOVATION

1. Overall project idea and objectives/

COVID-19 and misinformation

Data set **here**

For the past year and a half, Covid-19 has been one of the most talked about topics, especially on social media where information as well as misinformation can be shared easily with millions of people. Such misinformation has caused confusion among people, disruptions in society, and even deadly consequences in health problems. To be able to understand, detect, and mitigate such COVID-19 misinformation, therefore, has not only deep intellectual values but also huge societal impacts.

Our idea was to see and to analyse what was the actual impact of spreading misinformation.

For that we used a Data set, *COVID-19 Fake News Dataset from Kaggle*, composed of millions of tweets, divided into 4 different categories : Fake News and Real News on the one hand (this mainly concerns news channel or newspaper) and Fake Claims and Real Claim on the other hand (this concerns average people who share their opinions). In addition, *Hydrator*, allows us to retrieve information related to the tweets contained in the data set.

1. Overall project idea and objectives/

COVID-19 and misinformation

Data set **here**

In order to measure the impact of this misinformation on Twitter we decided to ask ourselves how to grade and represent the influence of this information on the users of the social network. Based on our dataset presented previously, we found several data interesting to exploit : the number of followers a user has, the number of accounts they follow, the number of likes on the post as well as the number of times the post was retweeted. We also wanted to play on the fact that an account can be verified or not and thus we could conclude that if the account was verified, it should be more influential than a non-verified account.

Our first idea was to process these data independently in order to have the most information we could present to the users of our innovation. However, we quickly realised that this might not be interesting enough and that we might need to process the dataset first to bring out the most relevant **News** and **Claims**.

Therefore we had the idea to create an influence factor, a mathematical formula which allows us to sort our data but also to visually determine, afterwards, the impact that these tweets could have had on Twitter. This mathematical formula represents the essence of our project because it is thanks to this factor that we were able to create a design to visualize our data. In order to elaborate this formula, we based ourselves on the data that could represent the influence of those tweets as we saw previously (i.e. the number of likes, retweets, followers,...)

1. Overall project idea and objectives/ COVID-19 and misinformation

The number of retweets and likes are multiplied by a factor that represents how influential each of those factors are. A retweet has much more reach than a like as a retweeted post is shown on each followers timeline, which is not the case for a liked post.

As for the amount of followers, if an account has a large amount of followers but is also following a large amount of people, this account might not have that much of an influence. That is why we decided to group those two elements together.

Finally we decided to use a logarithmic function, as when we use it the coefficient of determination is very high (which is a good proof that our formula is quite accurate). In the end, the data is processed according to their influence which is previously determined by the mathematical formula we created.

$$\log_{10}(1+ (\text{retweet nb} * 4.2)) * \log_{10}(1+(\text{favorites nb} * 1.2)) * \log_{10}(1+(\text{follower nb}/(1+\text{following nb})))$$

retweet nb : number of retweets
favorite nb : number of likes
follower nb : number of account's followers
following nb : number of following of the account

2. Relevance to the call/

Behavior on socials networks

COVID-19, which is constantly evolving, raises questions, concerns and misunderstandings. The media allows us to share news, articles and our opinions on the subject, but do we really measure their impact?

The impact of these tweets on public opinion directly echoes MediaFutures' track. Indeed, we are presenting an innovation, a representation of data in a digital space showcasing the influence of information shared on one of the most used social networks, Twitter.

In addition, we use science to question individuals and society on the importance of the veracity of the information we disclose. Our project, therefore, allows the user to question his behavior on the networks but also to question the content they consume. Finally, we could ask ourselves with the technological evolution and the rise of social networks, how could we treat the news to avoid misinformation?

3. Outcomes/

Prototypes and deliverables

In order to be able to grasp the project easily and to ensure a correct rendering, we made several prototypes, both for the design and for the technologies. Our final project is therefore a data visualization project representing the influence of COVID-19 related information during the year 2020 on Twitter.

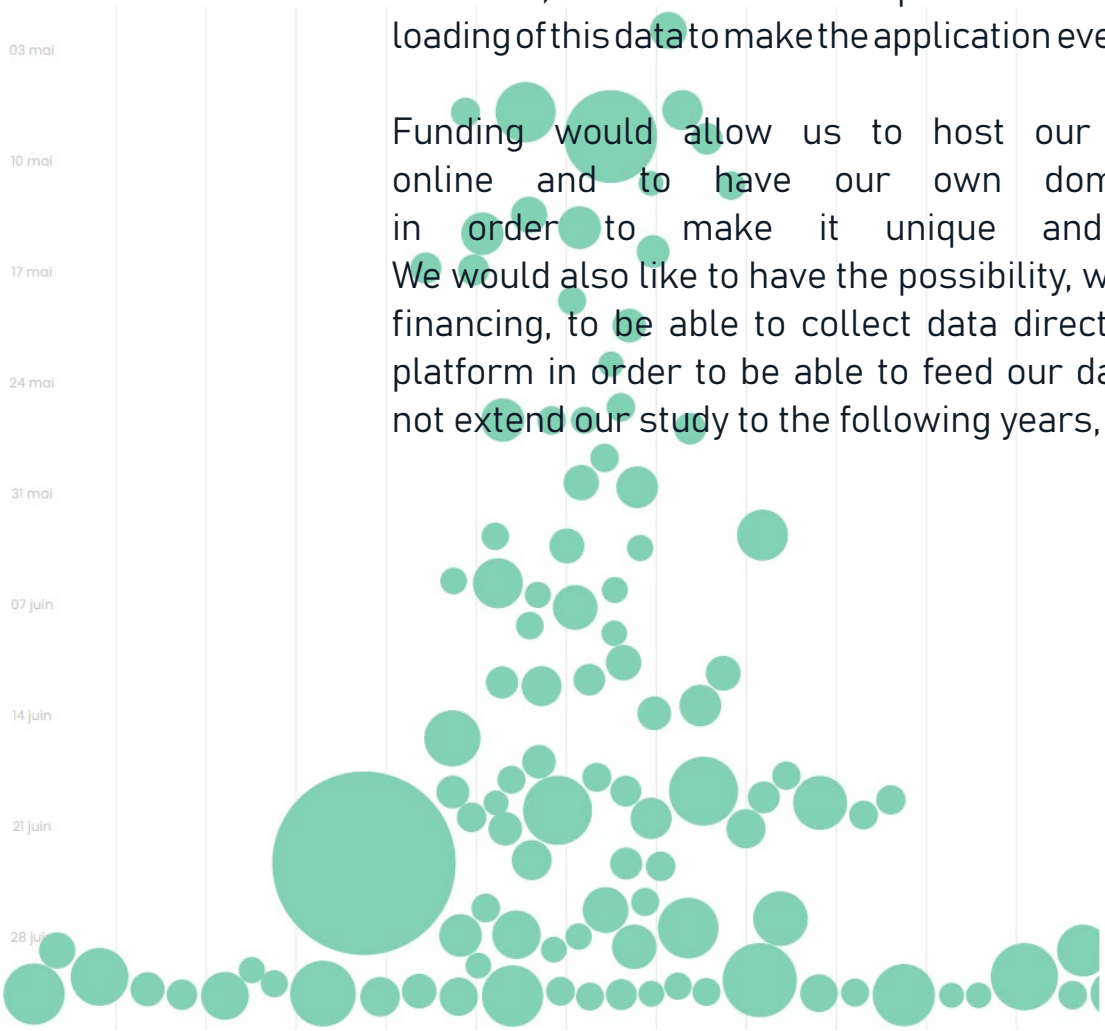
As we have seen, the COVID-19 Fake News Dataset from Kaggle refocuses misinformation related to COVID-19. This misinformation has confused people, disrupted society and even had deadly consequences. That is why this dataset helps researchers detect and minimize this misinformation by refocusing it in one place.

3. Outcomes/
Prototypes and deliverables

If our project were to be funded, we hope to be able to implement a deeper interactivity, on several levels, in our application. Indeed, due to a lack of time we have been mainly concerned with the issue of data representation.

Moreover, we realized during the implementation of our application that the loading of data was not optimized. Indeed, as our data set provided us with millions of tweets in several CSVs, it was complicated to load everything quickly. Therefore, using a Python script and the Pandas library (sorting CSVs), we were able to merge all the tweets filtered by our formula into a single CSV. However, with more time we hope to further optimize the loading of this data to make the application even smoother.

Funding would allow us to host our application online and to have our own domain name in order to make it unique and concrete. We would also like to have the possibility, with possible financing, to be able to collect data directly from our platform in order to be able to feed our data and why not extend our study to the following years, 2021, 2022...



Chapter

DESIGN

1. Use of data/

Mapping (annex 1)

The design represents a big part of our work. Indeed, in order to represent at best the influence of the veracity of the information around COVID-19 in 2020, we had to think about how to depict this influence.

Firstly, we wanted to play with colors as we are here talking about true or false facts. However, we did not want to use the classic red and green combination. We also wanted to represent a network of information, a sort of constellation echoing the social networks and the internet. That is why we chose to use a light blue and light pink combination to recreate this feeling of true/false while remaining in this idea of network/constellation.

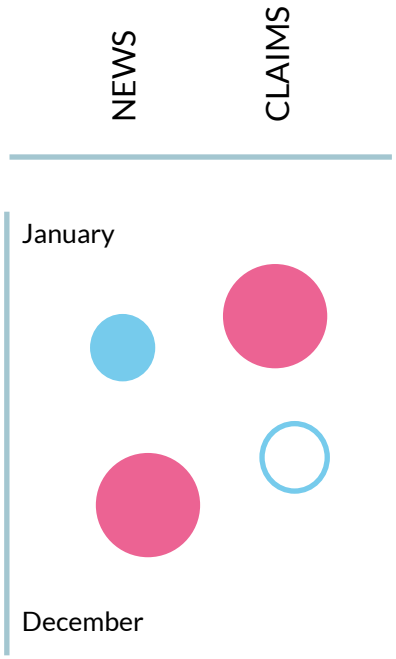
Secondly, we wanted to represent several axes in our data design. In addition to representing the data and its influences on Twitter, we wanted to create a timeline that would allow us to see the evolution of opinions as the research on COVID-19 progressed. Therefore, our timeline is represented vertically by the Y axis. The X axis, however, is divided into two categories, the News and the Claims.On this axis we wanted to contrast the news, which are information published by real newspapers, specialized magazines, with the claims, which are simple tweets talking about COVID-19. Note that news and claims can be true or false, whether intentional or not.

1. Use of data/

Mapping (annex 1)

The tweets are represented by luminous circles, to emphasize this network/constellation aspect. Finally, the size of these circles will represent their influence and it is the formula we have established that allows us to automatically define their size.

For example, a large blue circle represents a genuine influential tweet while a small red circle represents a less notable but still influential false information. Then, when the user selects a tweet, in addition to being displayed, it changes shape to stand out from the others and help the visitor during his experience.



1. Use of data/

Interactivity

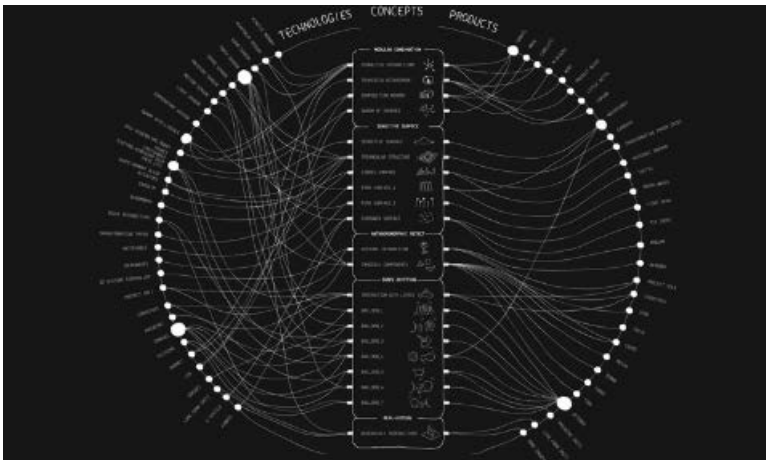
Regarding the interactivity of our project, we can find it on two different levels as for now. The first level is related to the user experience of our application. Indeed, during their navigation, the user will be able to interact with the graph itself. When hovering over the circles - i.e. the tweets - they will change shapes so that the user can identify where they are and more precisely which tweet they have selected.

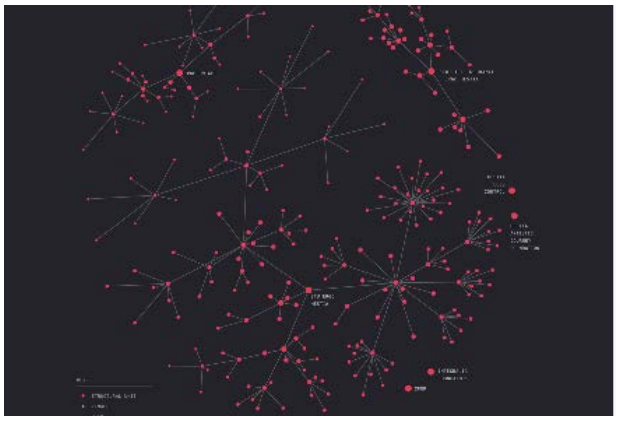
The second level of interaction appears with the menu we offer. This menu allows the users to play with the graph. As we saw earlier, one of our main objectives with this project is to show and allow the users to question the veracity of information, following our own vision of influence, hence the mathematical formula. However, we wanted the user to be able to make his own vision of the impact of these tweets. For this, we wanted to let the users play with the formula. They could play with the parameters to make them more or less influential on the sorting of the data.

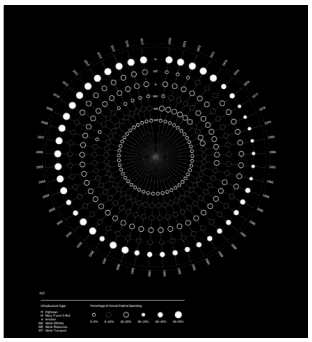
For example, we can imagine that for some people some parameters like likes or retweets are stronger factors to define the influence than the number of followers or following. In order to guide the user, we also thought of setting up a small help button that would explain the process and how the formula works.

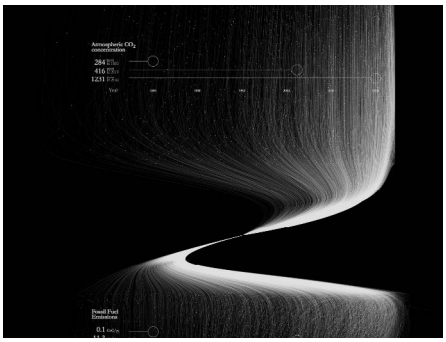
Chapter

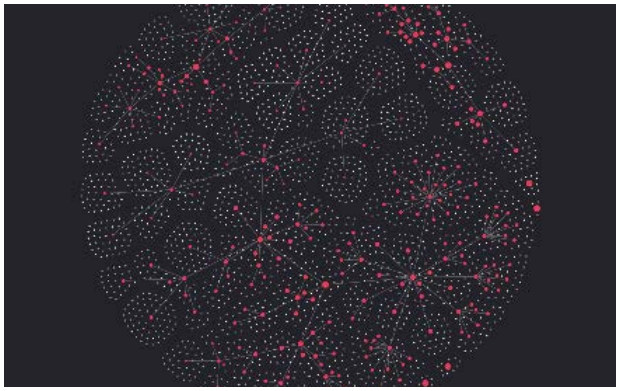
MOODBOARD

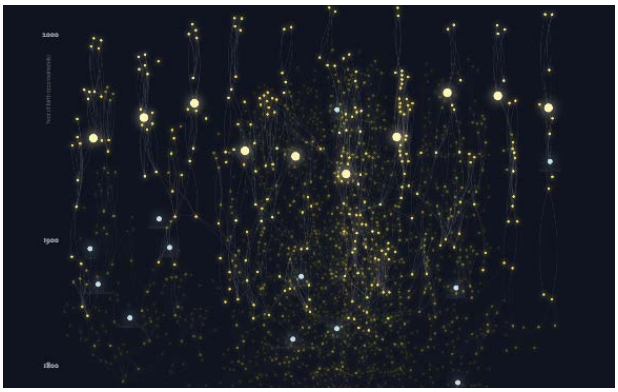


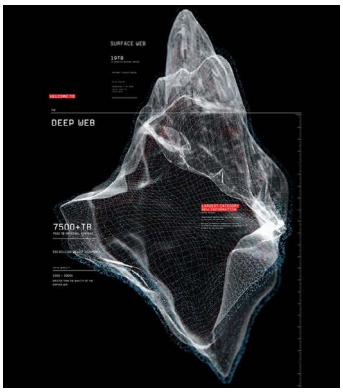


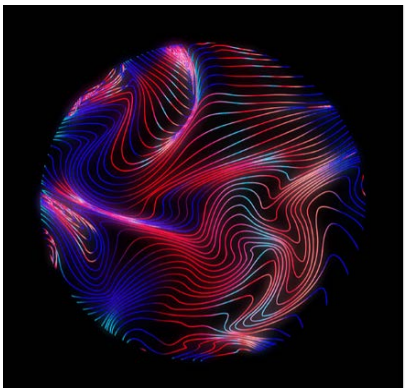








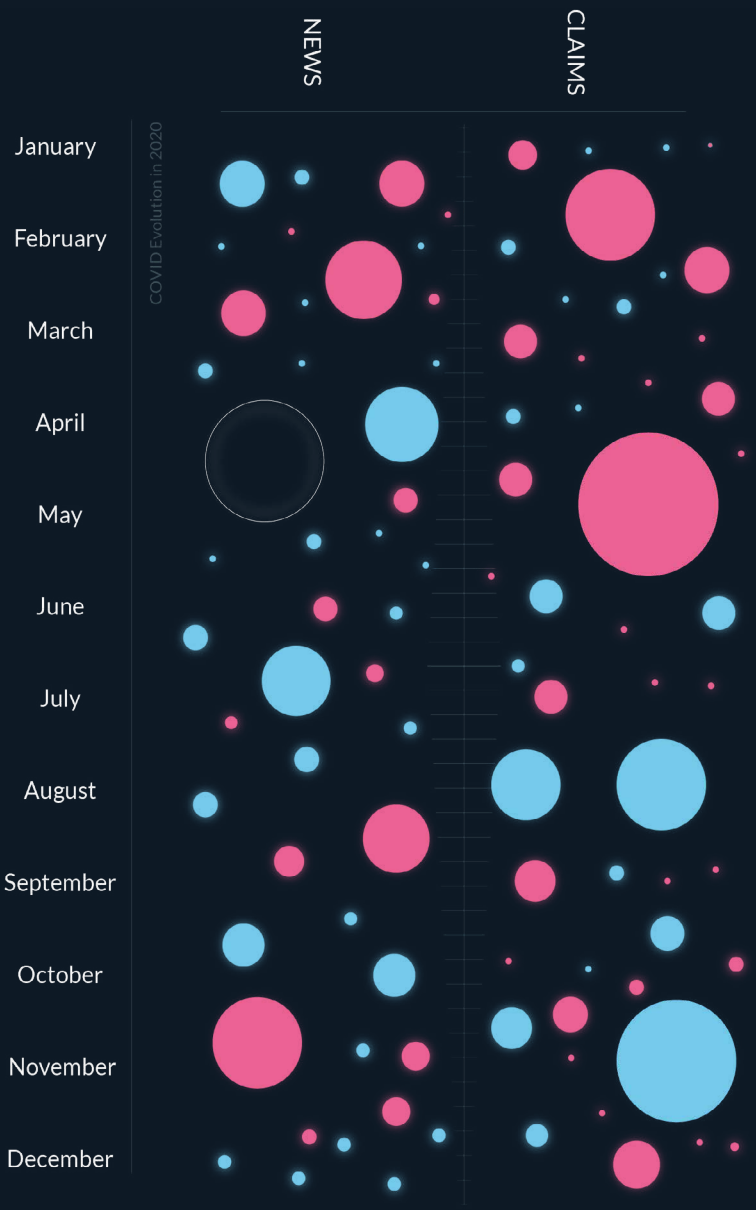




2. Final design/
Data visualisation and interactivity

To find and settle on a final design, we did several graphic researches.

Once our ideas were clear, we had to agree on the final format as well as the elements that should appear on our project. (annex 1)



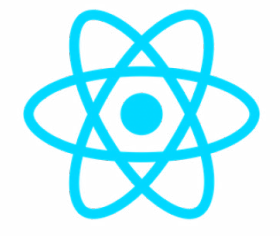
3. Implementation - Technology/

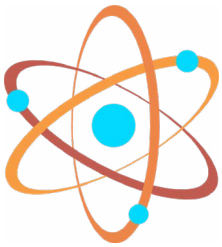
ReactJs

React-D3

Python

Pandas





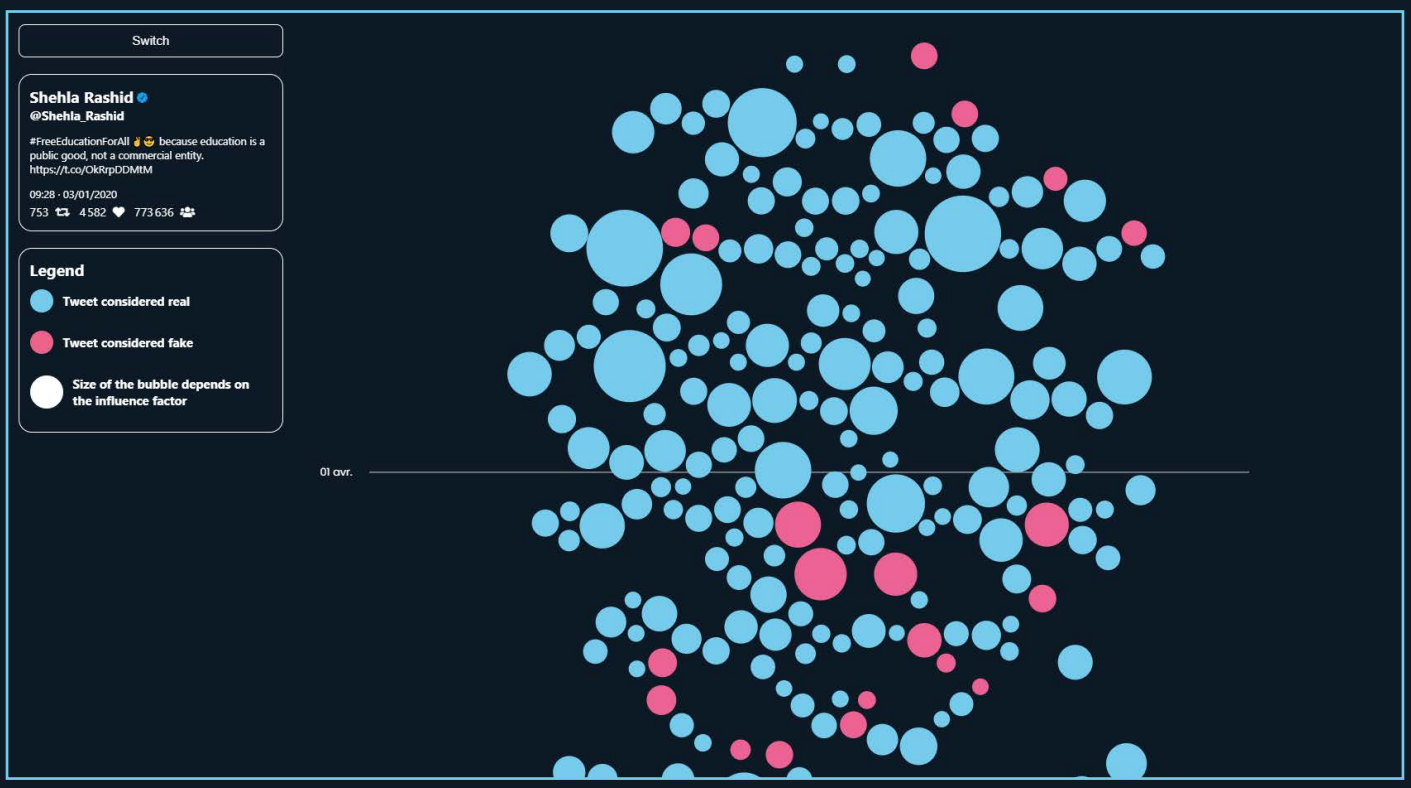
To do this project, we decided to use the ReactJS library. Indeed, in addition to having in our group people who have worked with this technology, the latter allowed us to use the D3-React library. This library, an extension of D3, allows us to generate graphics easily and in an optimal way for their use on the web. Moreover, ReactJS and D3 are standards in the world of web data visualization.

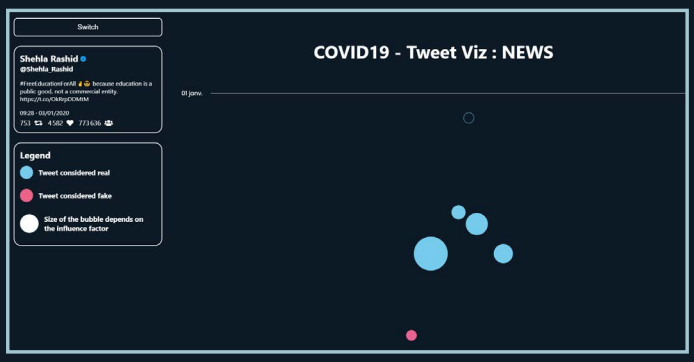
In addition to the more design-oriented web technologies, we also had to deal with the display of the data according to our dataset. This dataset presented several CSVs in which tweets were grouped (several million tweets). During our first tests we realized that loading all these files was very time consuming and complicated to set up. To solve this problem we decided to create a custom dataset, grouping relevant information defined according to our criteria (including the mathematical formula). To do this we decided to use Pandas, a Python library that allows us to perform operations on CSV files. In our case, the script we wrote allows us to merge the files, apply a pre-sorting as well as add new columns; all of this gathered in a single CSV file, our final dataset.

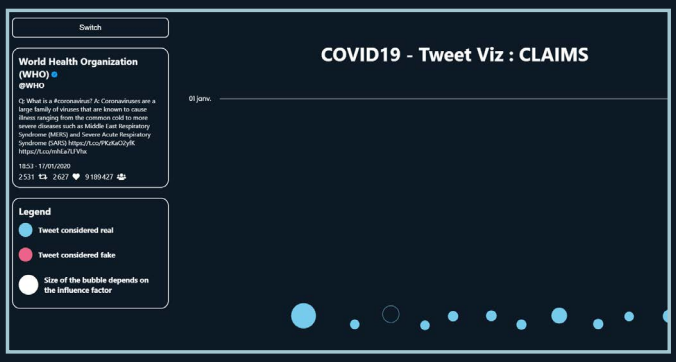
Unfortunately, due to a lack of time, we were not able to implement all the features of our model. In addition to some small graphics adjustments, glow effect for exemple, we would like to set up and finalize the interactive part, in particular with the help of financing. Indeed, we had the idea to create a menu (annex 1 model), allowing the user to play with the graph. Moreover, this interactive side would have allowed the users to understand our intentions by making our formula their own. It would also have allowed them to create their own vision of influence, by playing with the different parameters of the formula (followers, likes,...). We can note that despite this missing part, the user can interact with the graph by choosing a tweet to view it.

3. Implementation - Technology/

Final product







Voir le projet ici

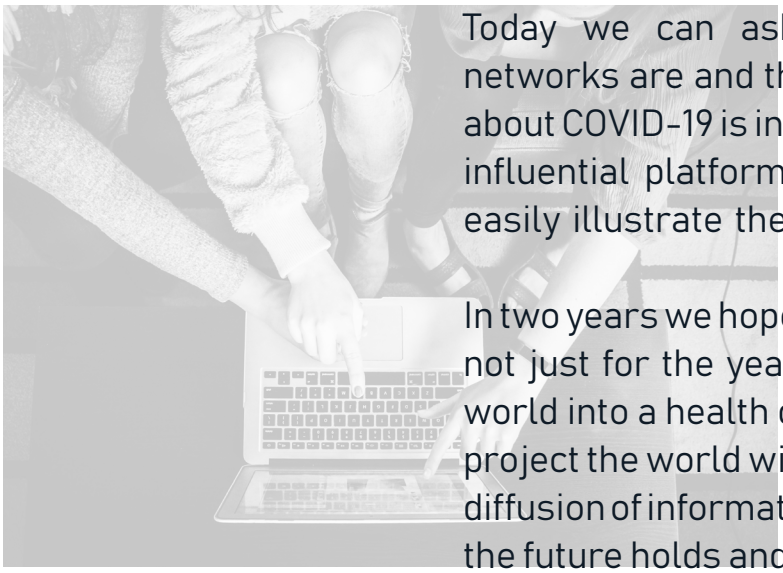
https://n0ls.github.io/COVID19-TweetViz/

Chapter

IMPACT

Impact/

Society and social networks



Always in accordance with the philosophy of MediaFutures, our project is embedded in a social dimension. **Covid-19 TWEETS VIZ** aims to raise questions about the impact of our words on social media. As we have seen previously, misinformation related to COVID-19 has caused confusion among people but also disruption and even deadly consequences. Understanding, detecting, interpreting and reducing the impact of misinformation can help these problems that directly impact our society. Today we can ask ourselves how influential social networks are and that is why the example of information about COVID-19 is interesting. As Twitter is one of the most influential platforms with its 199 million users, we can easily illustrate the potential impact of this information.

In two years we hope to be able to offer multi-year studies, not just for the year 2020. As COVID-19 has plunged the world into a health crisis, we want and hope that with our project the world will have a different vision regarding the diffusion of information on the internet. We don't know what the future holds and even though COVID-19 is almost over, we are not immune to reliving this situation in a few years.

Chapter

TEAM

LANDRODIE Nils

@n0ls

This project was really interesting overall. Having to both work on the **#technical** and design aspect is really challenging because the result really depends on both and every detail count. The theme and **#data** set we chose were also really interesting because they questioned the presence and **#impact** of fake news in our **#lives**.

#Technical part #React Dev #CSV #Python

29,1 k 24.3 k 102.7 k

LEFEVRE Erwann

@eight-rex

This **#project** was really interesting. We were each able to **#work** on our areas of interest to provide a satisfactory result.
Moreover, I was able to bring the knowledge I had acquired with the tutored project I did, **#Project 89**.

#Graphical part #Design #CSS

35,2 k 22,2 k 100,1 k

QUIRIN Cloé

@clozai

Working on a **#DataVisualization** project was completely new for me but it was a great experience. Searching for the best way to present the message we wanted to convey was really interesting. Also, incorporating a **#mathematical** aspect into our project proved to be quite challenging but it was interesting.

#Mathematics part #CSS

24,4 k 26,5 k 203,5 k

SCAVINNER Vincent

@vscav

This **#project** was a rewarding **#experience.**
Making our **#translatedData** and original idea become easier to grasp and understand through the use of technologies was a complex but essential **#process**.

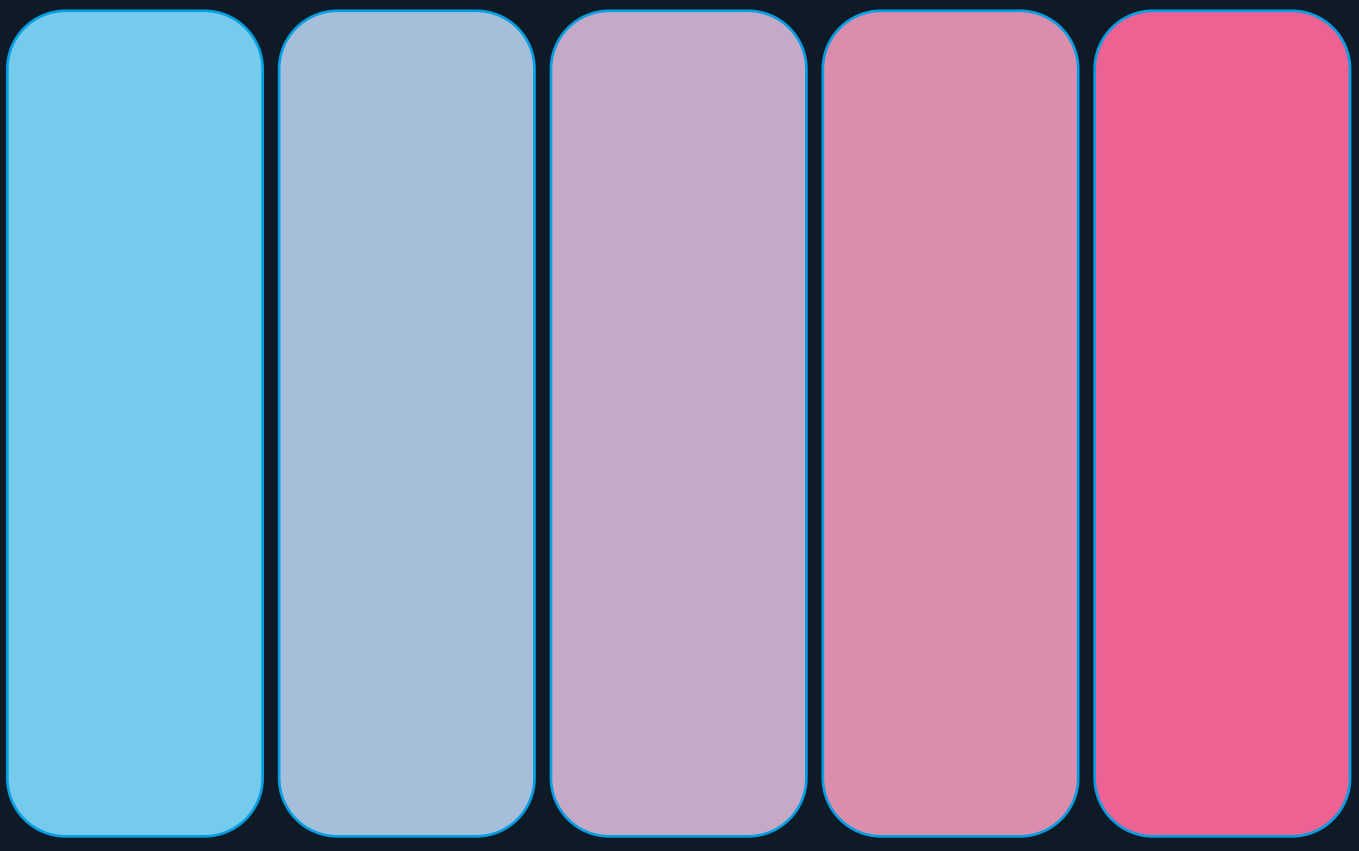
#Technical part #React Dev #Prog #CSV

44,8 k 36,2 k 113,9 k

Chapter

ANNEX 1
GRAPHIC GUIDELINES

C O L O R S



Real claiming

Fake claiming

Chapter

ANNEX 1
GRAPHIC GUIDELINES

F O N T

LATO BOLD

AaBbCc

Aa Bb Cc Dd Ee Ff Gg Hh
Ii Jj Kk Ll Mm Nn Oo Pp
Qq Rr Ss Tt Uu Vv Ww
Xx Yy Zz

0123456789

LATO REGULAR

AaBbCc

Aa Bb Cc Dd Ee Ff Gg Hh
Ii Jj Kk Ll Mm Nn Oo Pp
Qq Rr Ss Tt Uu Vv Ww
Xx Yy Zz

0123456789

LATO BOLD — TITRES

LATO REGULAR — CORPS DE TEXTE

Chapter

ANNEX 1
GRAPHIC GUIDELINES

IDEA

Chapter

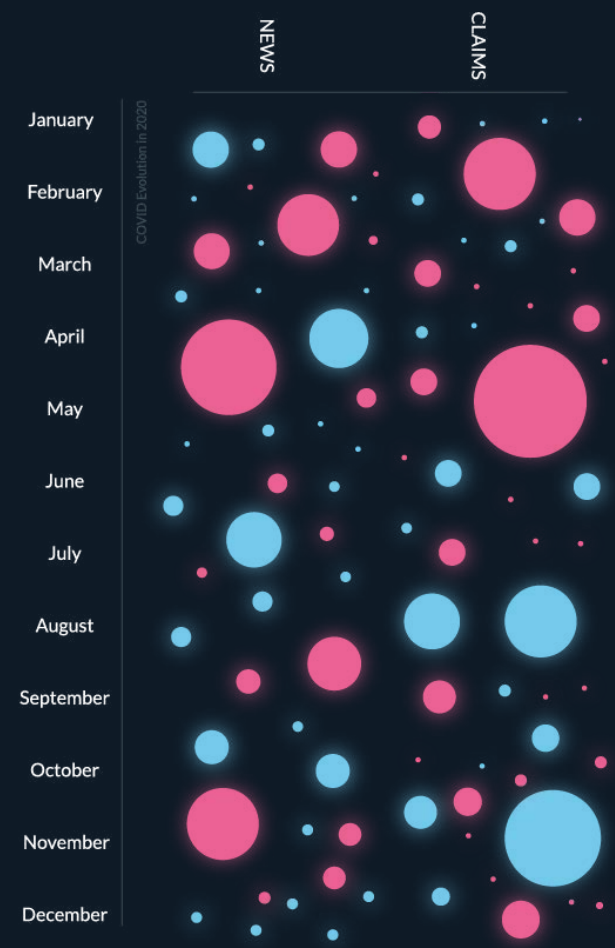
ANNEX 1
GRAPHIC GUIDELINES

IDEA

INTERACTIVITY PART

TWEET VISUALIZATION

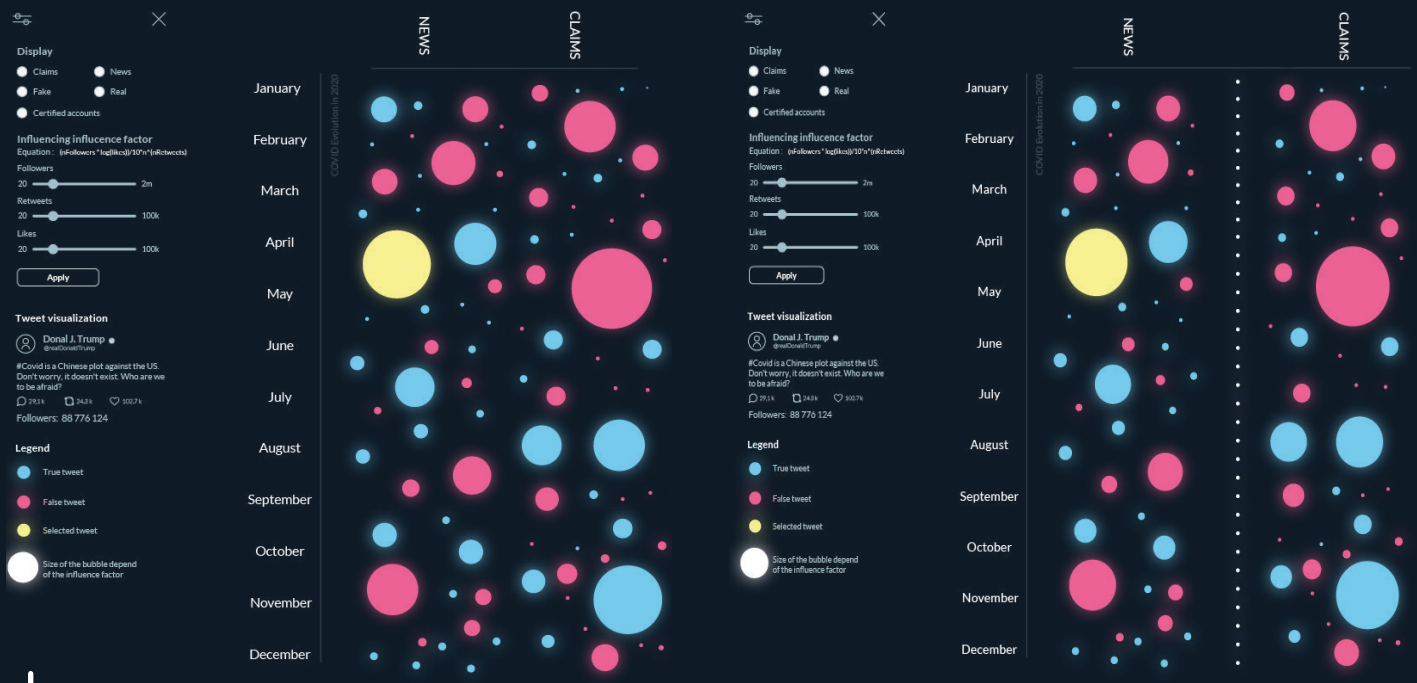
DATA VISUALIZATION



Chapter

ANNEX 1
GRAPHIC GUIDELINES

IMPROVEMENTS



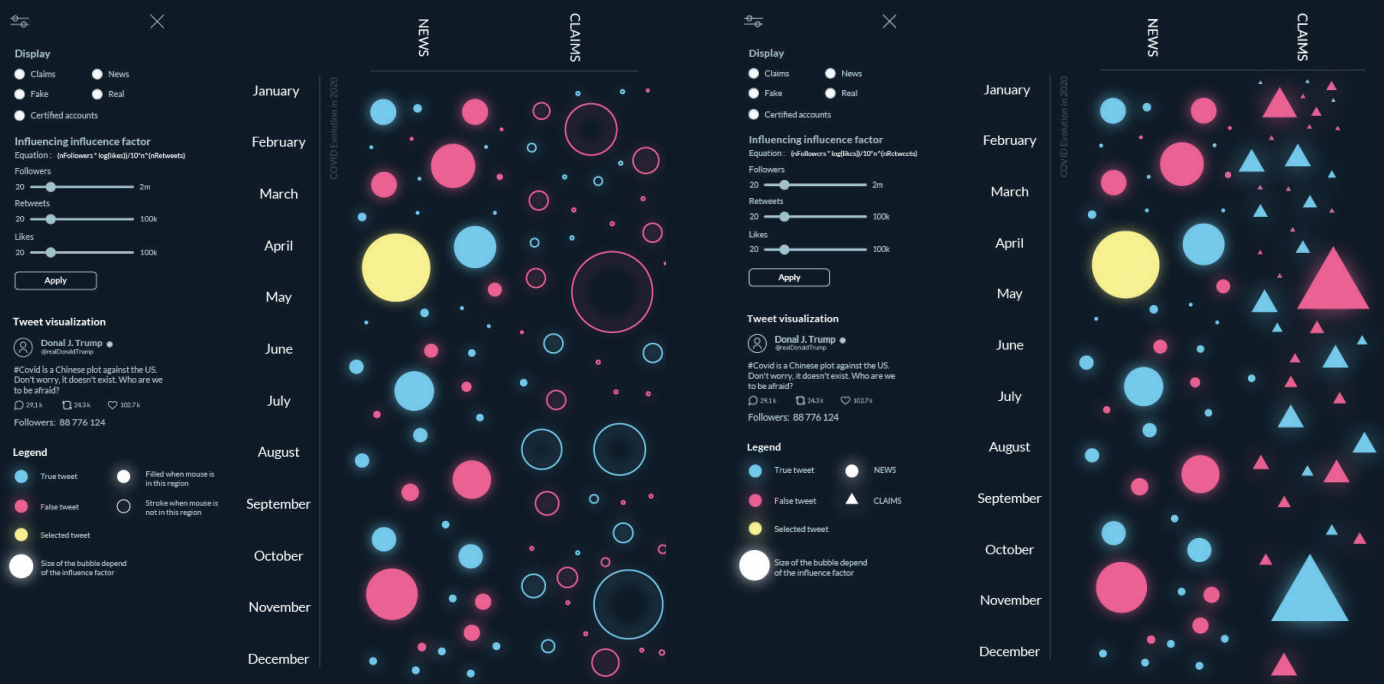
ADD LEGEND

AXIS SPLITER

Chapter

ANNEX 1
GRAPHIC GUIDELINES

IMPROVEMENTS

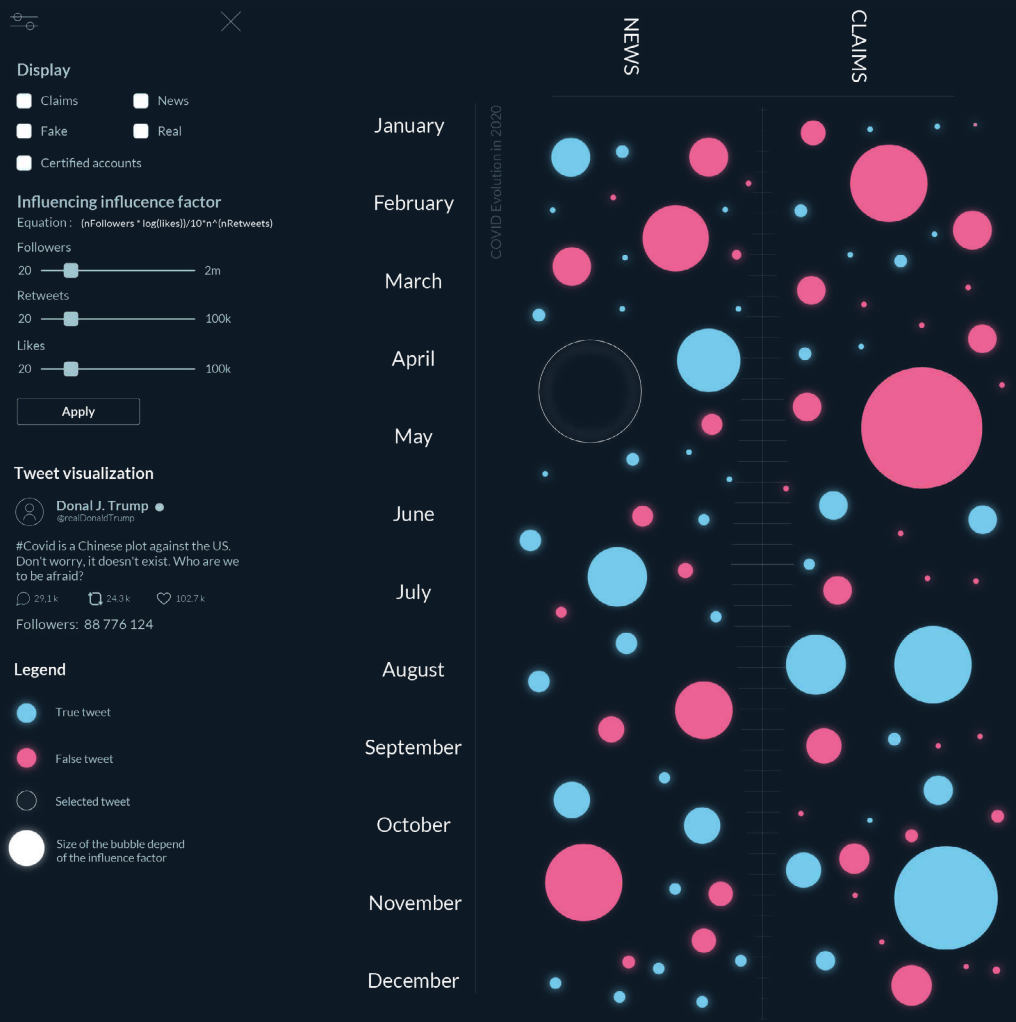


TRY NEW DESIGN

Chapter

ANNEX 1
GRAPHIC GUIDELINES

F I N A L



Chapter

ANNEX 2 - BENCHMARK

NOODY by David Bihanic and Thibault Jaillon/

Date: **2007**

Subject: **Design a device for graphically displaying search results via a browser.**

Data: **Answers following a search for information via a browser (Articles, Sites, images...)**

Data set: **More than 1000 sources, global newspapers (magazines and news sites, sorted by region) from all continents.**

Project **here**

In 2007, David Bihanic and Thibault Jaillon developed Noody 1.0, a visual search engine that reinvents the display and the user path. The information would no longer be hierarchical from top to bottom according to the relevance and precision of the search but by a visual representation system offering a better spatial hierarchy of metadata.

Questions such as, how to think of new methods of description according to space constraints? or how to propose a dynamic information retrieval system? how to prioritise information and how to represent it? are problems on which David Bihanic has worked.

Noody offers solutions based on an application that follows several semantic analysis methods. The results will be processed, as in a browser, according to their relevance to the topic searched. To begin, the application will define a zone of relevance which will be presented to the user by a white fuzzy zone. In this area, the results with the most useful information in relation to what the user is looking for will appear. In this area, the closer the information is to the epicentre, the more consistent the results will be with the browser search. By extension, other results will be pushed to the periphery of this relevance zone. This display will be presented directly to the user, even before the search.

Once the search is done, the application will analyse the keywords and bring out the information. This information will be displayed in the first place in the relevance zone but also everywhere on the screen. Indeed, as on a classic browser, Noody will bring out many different subjects and sites having a more or less close relationship with the searched subject. This is why the application will propose a visual result representing a graph of results on the whole screen allowing the user to move (with the mouse) in an almost infinite space. This will give the user the possibility to discover semantic fields outside the initial periphery.

About interactivity, the user will have several options. First of all, they will be able to carry out several searches via the dedicated zone. They will also be able to refine their search thanks to menus allowing them to choose a location, a date of publication, etc. With regard to mouse actions, the user can click on the semantic fields in order to divide them into several sub-fields presenting and specifying specific subjects as they are clicked. When the user hovers over the sub-fields, the topics and themes presented in the field will be displayed via an interactive drop-down menu. Finally, the user will be able to interact with the aesthetics of the graph by playing with parameters such as the gravity on the semantic fields, the friction, the force etc.

CONCLUSION

The data processing is very well thought out and could be similar to our concept. Indeed, in this example of Data Visualization, the data is processed according to its relevance but the user also has the possibility to use filters to specify his search. In our case this could be interesting, the influence of the veracity of the tweets that form our dataset, could be an option available to the user. We could also think of other options that would play on the layout and integration of our future design. Furthermore, the link between the display of the data and the interactivity of Noody makes the application both aesthetic and pleasant to use. Through the navigation and interactivity of the device, the user is immersed in a multitude of possibilities that make his navigation unique and pleasant.

ROYAL CONSTELLATIONS

Nadieh Bremer/

Date: **October 2016**

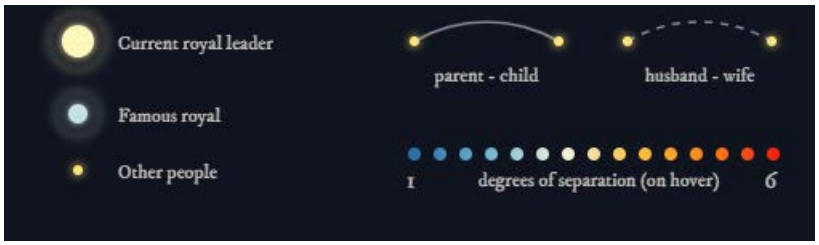
Subject: **Showcasing the connections in the European royal families for the past 1000 years**

Data: **all 10 of the current hereditary royal leaders of Europe and their ancestors**

Project **here**

While the project was originally supposed to be around the US elections, Nadieh Bremer felt that as a European, it might be more interesting to work on royal families and their connections. The first thing that came to her mind was about their bloodlines and how each family might be related to one other. While she knew that they were somehow related, she wanted to know how close they were of a relative.

The data is presented as different constellations. Each star represents a person. It is placed approximately on their year of birth in the vertical direction and to their closest relative who is a royal leader today in the horizontal direction. The years go from the 1000s to the 2000s. The interactivity in this data design is very important. At first, there seems to be a lot of information and the user might feel overwhelmed. However the user will very quickly understand that you can select the information you want to know. Indeed, all they need to do is hover over a star, and it will highlight only the star (i.e. person) and their relatives, up to a "6-degree of separation". The closest relatives have blue tint and it will graduate to red the higher the degree of relation is.



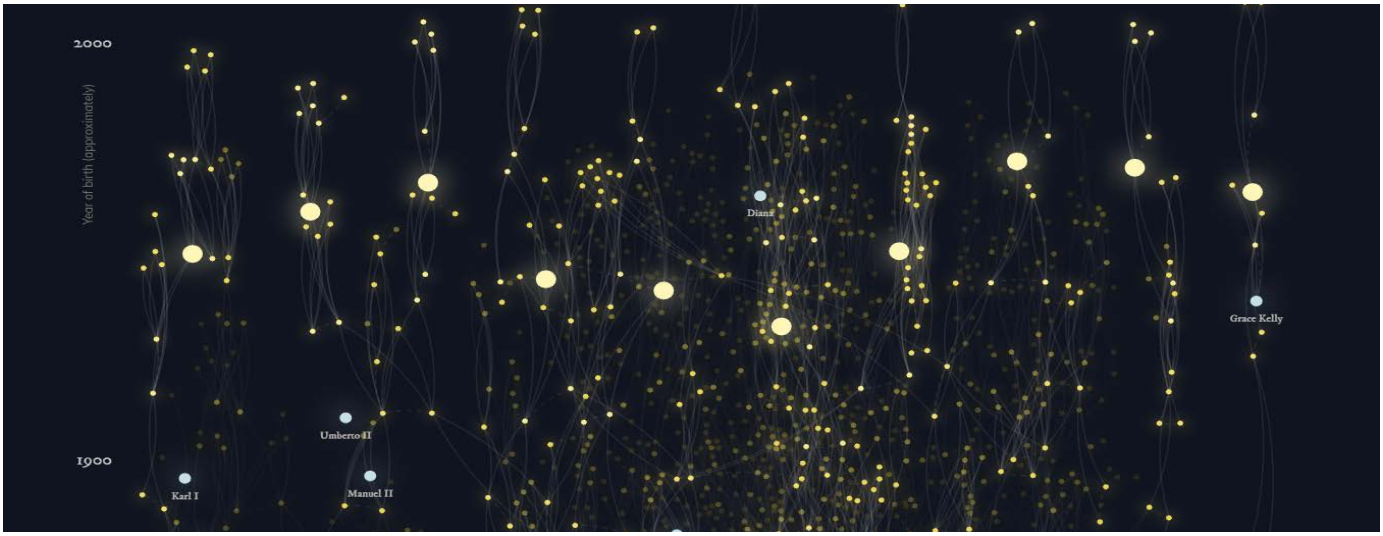
For a famous royal, a short description of the person will appear when hovering over the star. The user can also select two stars to see how they are connected to each other. With that the user can know the degree of separation between two people on the map.

CONCLUSION

The design of this project is very nice because it looks great (very simple but clever map) but is also very informative. The constellations are a great way to show the connections between each person but it also creates a very beautiful and interesting map to look at. The interactive aspect of it is also very interesting for the user because he can actually play with the data and select what he wants to see or analyze.

Nadieh Bremer managed to expose in a very elegant and easily understandable way the complex and intertwined relations of every european royal family.

This design could be really interesting for us to expose the relation between our data. We can also play with the size to showcase certain data. The chronological order is also an aspect that might be useful for our project to show the evolution of our data.



BLOOM's PROJECT by Dataveyes/

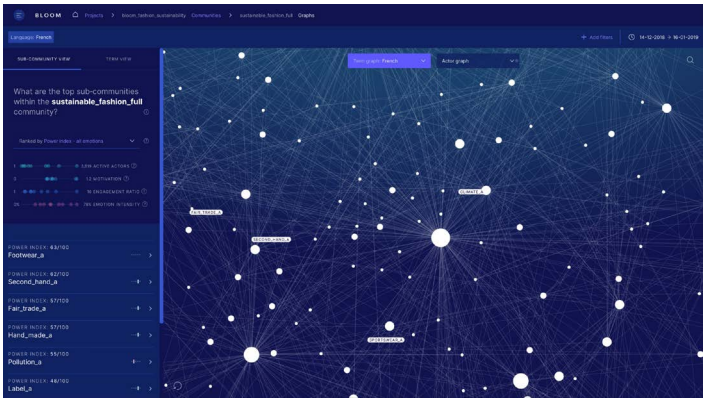
Date: **2019 - 2020**

Subject: **Design of a platform highlighting data to better understand it for the company Bloom**

Data: **The data is provided directly by the customer**

Project **here**

The Bloom Company works on a platform for deep analysis of network exchanges according to social and semantic links. Their company offers communication and marketing plans for clients so that they can expand on the web and anticipate possible strategies to put in place on the networks to increase their notoriety. To do this, they introduce advanced research, data processing algorithms to target their clients' core target audience, including consistent keyword research. However, Bloom's product is aimed at advanced users in these fields and that is why they called on Dataveyes to help them. Their mission was to set up a graphical interface to easily read and understand a data set.



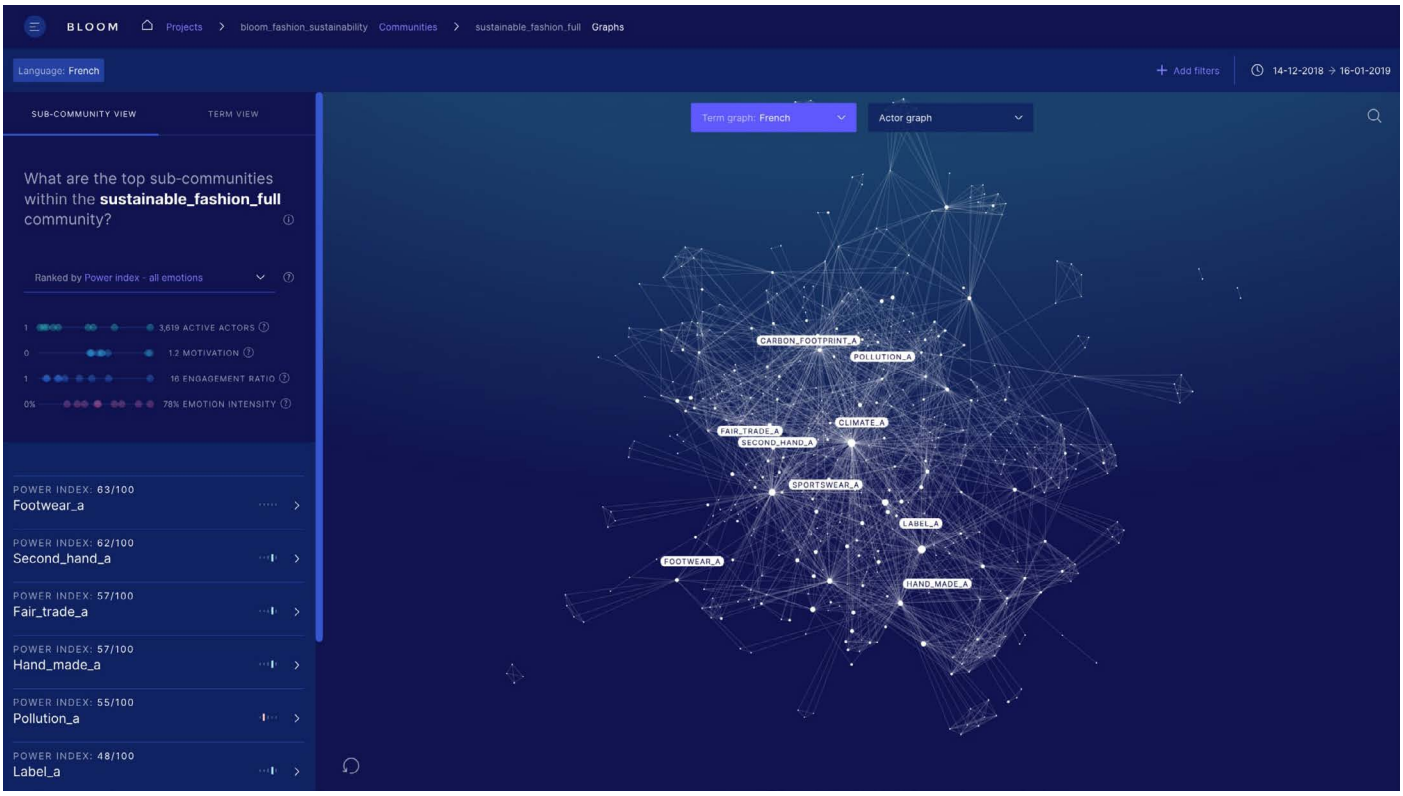
The data provided by Bloom is data extracted from public conversations between different users of social networks. Once retrieved, the data is analysed and processed by the company's programmes and algorithms. We can extract keywords, themes and trends from these conversations to enhance the data and make it usable.

In order to create an interface that presents the data to the user, Dataveyes made several proposals to its client. Their solution, the one that was retained, was to play on an ergonomic design, understandable while emphasizing the interactivity of the application. The data is represented in the form of graphs showing and proposing to the user a graph staging coherent conversations around the theme chosen by the latter.

The luminous points will correspond to the conversations and these will be linked together by physical links, in this case luminous lines, allowing all these data to be linked. In a second step, these data are hierarchized according to their theme, we can see the conversations including the searched theme as well as all the derivatives that we could find. In the manner of David Bihanic, Dataveyes here represents the results of a search on a browser. Moreover, the captions, titles and icons allow the user to navigate easily and to find information quickly. It is the interactivity of the application that allows users to answer their questions. In fact, while browsing, they can filter and refine their search via menus, zooms and an internal search system.

CONCLUSION

With a well thought-out, ergonomic design, with simple and clear titles, legends and logos, the application presented by Dataveyes is very interesting. It makes data, which is sometimes complicated or not accessible to everyone, understandable for non-expert clients of Bloom. Moreover, the interactivity offered by the site allows the user to play with the quantity of data proposed but also to be immersed in the heart of his research thanks to its design. For our project, it would be interesting to think about the layout of our elements and especially to think about the logic of our captions and titles. Indeed, we could take the example of this project in order to make the data accessible to all by using an elegant and well thought-out design.



Chapter

ANNEX 3 - ANALYSIS

Inoculating Against Fake News About COVID-19

https://www.frontiersin.org/articles/10.3389/fpsyg.2020.566790/full

Coronavirus misinformation, and how scientists can help to fight it

https://www.nature.com/articles/d41586-020-01834-3

Covid-19 : une uberisation des fake news

https://theconversation.com/covid-19-une-uberisation-des-fake-news-147375

« Covid-19 : une vague d'infodémie » (Désintox, ARTE)

https://youtu.be/9vgSOt9p60Y

Fake news in the age of covid-19

https://fbe.unimelb.edu.au/newsroom/fake-news-in-the-age-of-covid-19

There are two types of false information : intentionally and innocently spread. Either trying to "propagate chaos for the sake of political gains" or because of a "misconstrued understanding of the disease". Fake news works because we, as humans, are bad at detecting them. Once it is spread it is really difficult to correct it.