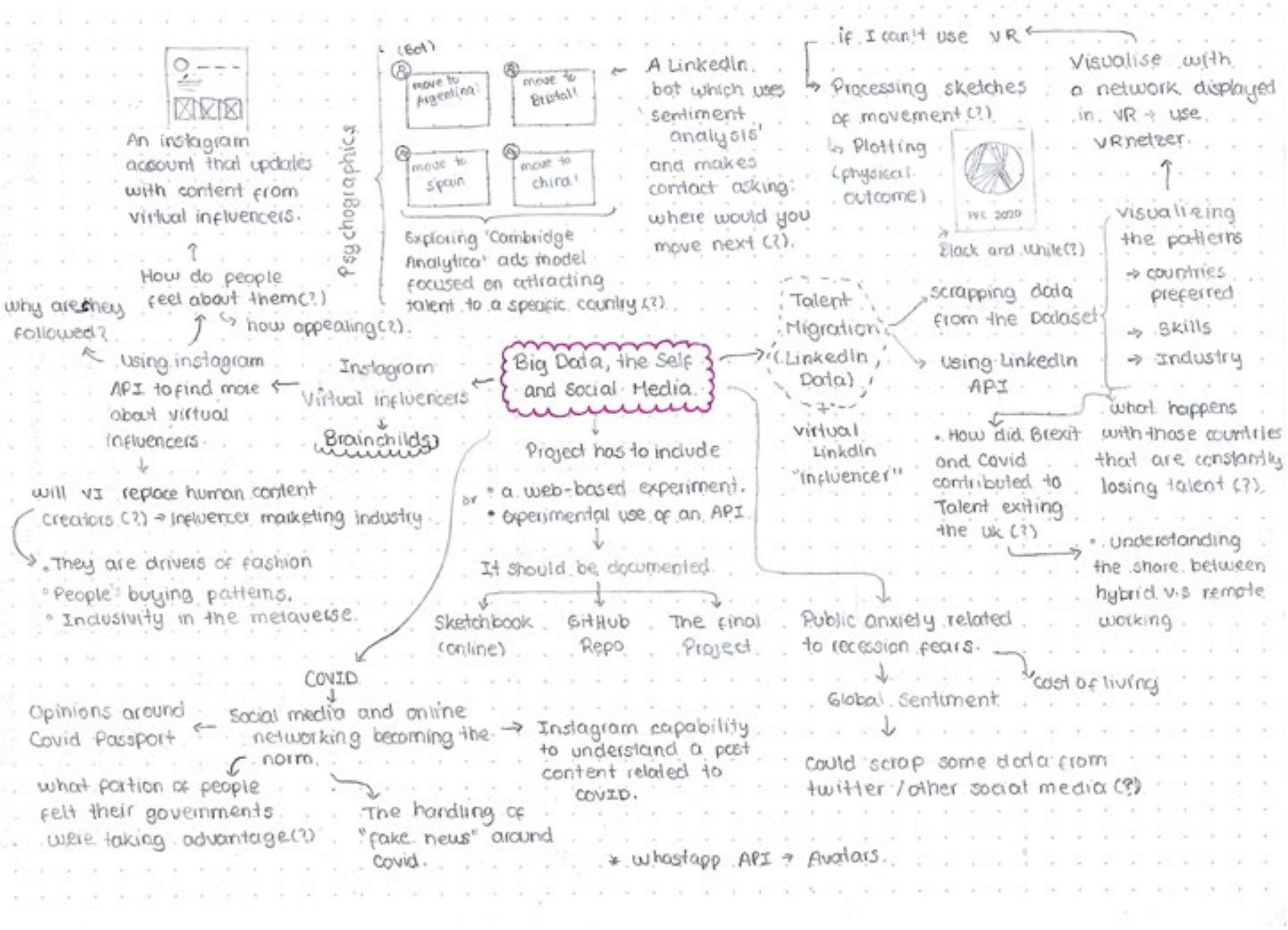
BIG DATA

Project Documentation


Big Data, the Self and Social Media
Instagram Virtual influencers
Brainchilds
An instagram account that updates with content from virtual influencers.
How do people feel about them(?)
how appealing(?).
why are they followed?
Using instagram API to find more about virtual influencers
will VI replace human content creators (?) → influencer marketing industry
They are drivers of fashion
People buying patterns.
Inclusivity in the metaverse.
Psychographics
move to Argentina
move to Bristol
move to spain
move to china
Exploring 'Cambridge Analytica' ads model focused on attracting talent to a specific country (?)
A LinkedIn bot which uses sentiment analysis and makes contact asking: where would you move next (?).
Talent Migration (LinkedIn Data)
virtual LinkedIn "influencer"
scrapping data from the Dataset
using LinkedIn API
if I can't use VR
Processing sketches of movement (?)
Plotting (physical outcome)
Black and white(?)
Visualise with a network displayed in VR → use VRnetzer
visualising the patterns
→ countries preferred
→ skills
→ industry
what happens with those countries that are constantly losing talent (?).
• How did Brexit and Covid contributed to Talent exiting the UK (?)
• understanding the share between hybrid v.s remote working
Project has to include
or
• a web-based experiment.
• Experimental use of an API
It should be documented
Sketchbook (online)
GitHub Repo
The final Project
COVID
Social media and online networking becoming the norm.
Opinions around Covid Passport
what portion of people felt their governments were taking advantage (?)
The handling of "fake news" around Covid.
Instagram capability to understand a post content related to COVID.
Public anxiety related to recession fears.
cost of living
Global Sentiment
Could scrap some data from twitter / other social media (?)
* whastapp API → Avatars.

• Should we rather find a balance between the rate of technological growth and the rate that we can develop suitable ethical guidelines.

Examples of datasets / Resources

• www.1000genomes.org
• International dialects of English Archive.
• Note: some algorithms don't rely on data anymore but they rather "learns" from itself. Such as a program learning how to play GO.
• www.superdatascience.com
• Data governance: British Academy talks on Robotics, AI and Society. (2018)
• UK Government's Data Science Ethical Framework
• World Bank Data
• European Union Open Data Portal
• The CIA World Factbook
•

The data science process

1) Identify the Question
   • what is the goal of my analysis?
   • what questions do I want to answer?
2) Prepare the Data
   • Source • Clean • Prepare the data for analysis.
   - Perform quality assurance verifications (?)
3) Analyse the Data
   - Build models
   - Perform data mining
   - Run text analytics (?)
4) Visualize the Data
   Translate complex insights into easy-to read visuals or animations
5) Present findings:
   Translate my findings into "comprehensible" language.

i with data, we have the advantage of deriving our insights from actual evidence. we are not changing the information to suit our ideas but we are rather formulating ideas in order to derive insights.

1) Identify the Question
   • Before we can prepare and analyse data, we must know what kind of data is that I need.
   • To find this data I need to understand my main question and main goal.
   • Such main question might need to be recalibrated in terms that the data will be able to understand.

Some key questions to make:
   - where is the data located (?)
   - Who is in charge of that data?
   - What would success look like for the project

Possible Project Ideas.

→ web crawler (different to web scraping)
• web scraping, data scraping or content scraping is when a bot downloads the content. This might target specific pages / websites.
A web crawler systematically browses the Internet with the purpose of web-indexing.

• A web crawler that continually collects and stores data and finally visualize it.
• An installation which is based on a large dataset which I have scrapped. (Physical manifestation of an online data source)
• A bot within a social media.
• A web browser extension (?)
• web crawling + VR Big data visualization
  Network Style visualization

2.1 Maslow's hierarchy of needs

Data science and physiology

At the bottom of Maslow's hierarchy are physiological factors, the basic needs for humans to simply survive. How can data help with those most basic requirements? How can it improve upon them?

Let's take the air we breathe as an example. Air pollution has been a major global cause for concern ever since the Industrial Revolution of the late 18th and early 19th centuries. We might automatically imagine smog to be a phenomenon of the past, as in the case of London in the 1950s, when coal emissions regularly covered the city, but smog continues to affect a large number of places around the world, from China to Brazil.

Any technologies that are designed to reduce pollution in affected cities are reliant on data: to improve the condition of the air, it must first be monitored.

Still feeling hesitant about the prospect of using AI in medicine?

Watson isn't the answer to all our problems, though. The machine's AI can still make mistakes. But the difference between machine doctors and human doctors is data, and as the technology to process growing quantities increases, so does the difference in ability between human and machine. After all, humans can absorb information from conferences, medical journals and articles, but we all have a finite capacity for storing knowledge. What's more, the knowledge that human doctors possess will largely be limited to their life experience. A machine doctor, on the other hand, can only get better the more data it is given. With instant access to data from other machines via the cloud, shared data can inform more accurate diagnoses and surgeries across the world. Thanks to exponential growth, these machines will have access to all manner of variations in the human body, leaving human knowledge flagging far behind.

Data science and belonging

After fulfilment of the second stage of Maslow's hierarchy (safety), the need for belonging within a social environment (family, friends, relationships) will follow. It states that humans need to be part of a community of people who share their interests and outlook on life. The perceived disconnect between technology and society has been a topic of much discussion in recent years. The internet is often criticized as contributing to an increasingly isolated existence where our every whim and need is catered for. As an outdoorsy person, I won't make any case in support of socializing in the digital over the physical world. However, I do believe that the relatively democratic accessibility that the internet affords to people all over the world at all hours of the day is to my mind a great asset to human existence and experience.

What's more, what makes social networks such as Facebook, Instagram and LinkedIn successful is not the usability of the platform – it's their *data*. A badly subscribed social network is unlikely to offer the same breadth of services as a well-subscribed network because social communication ultimately relies on relationships. If the data isn't there to connect us to the right information, whether that means human connections, images that appeal to us, or news stories on subjects in which we are interested, the social network will not be useful to us.

Data is helping to make our world much more interconnected, and it is not only aiding us in personal ventures like finding old school friends; it is also helping scholars and practitioners who are carrying out similar projects to find each other and partner up.

CASE STUDY Forging connections through LinkedIn

I love using LinkedIn – and I think that they have really applied their data to benefit both themselves and their users. A quick visit to the business network's 'People You May Know' tab will show you an inexhaustible list of recommendations for connections with LinkedIn's other users. Some of these might be people at your current workplace, but you may also notice people from your university, and even school friends, cropping up on the system as recommended connections. To do this, LinkedIn uses the data you post to your profile – background, experience, education, existing colleagues – and matches it with the profiles of others.

LinkedIn's technology has enabled thousands of people to rebuild connections with their past. And as these connections grow, so does the network's data, thereby generating yet more connections. Whenever you connect with another user, not only do you gain what they call a 'first-degree connection' but their linked colleagues become 'second-degree connections', thereby expanding your circle much further than may be apparent.

For LinkedIn, as with any other social media channels, all that is essential is input from its users. I have found numerous friends and ex-classmates on the site, many of whom have since gone into the same field as me and, thanks to data's ability to match us, this has opened up a new dialogue between old acquaintances. Knowing that I have access to friends and colleagues online builds a sense of community and maintains it long after we have moved on, whether that be from a city or a place of work, and I find this interconnectedness comforting.

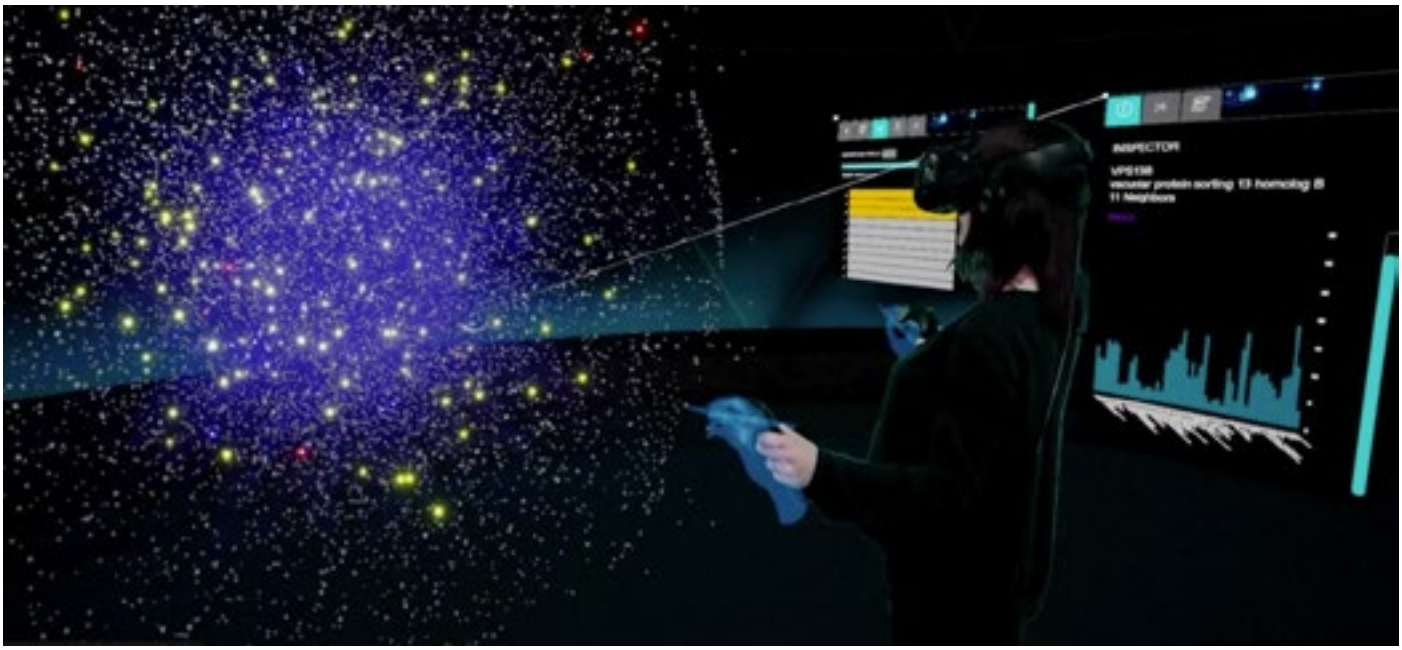
By connecting with others who share our interests, courses of study and location, LinkedIn can also give us a good insight into jobs that are relevant to us. When I was in the market for a new job, I started posting status updates to LinkedIn – the platform's data algorithms identified my needs according to key words that I had used, and this is how recruiters started to find me. What was even better was that since I was writing about subjects that interested me, LinkedIn's algorithms matched me to jobs that specifically required those branches of knowledge. It was even how this book's commissioning editor found me. How's that for social media channels' abilities to improve happiness?

VRNetzer

Attending the Imperial Lates helped me to gauge the different programs and technologies that practitioners are using. The main focus of the event was ‘Play’ and it presented games as a medium for learning.
In a conversation with one of the presenters, I got introduced to the tool VRNetzer, an Open Source project that allows the visualization of Networks.
In the experience you could see different clusters and zoom in and out to select which node you were interested in.

Figure. Imperial lates event banner

This made me remember a section of a book that presented a case study for Big Data visualisation through VR. Given the nature of the dataset I have found around Talent Migration around the world, I picture a world Network, or clusters that present the movement of people across the globe.

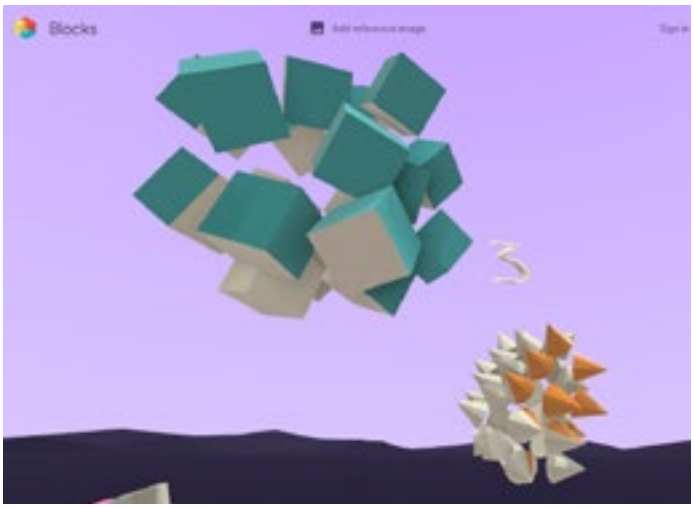


(2021) VRNetzer Autocore. menchelab. Available at: https://www.youtube.com/watch?v=Pd46211gc9U&t=5s (Accessed: March 22, 2023).

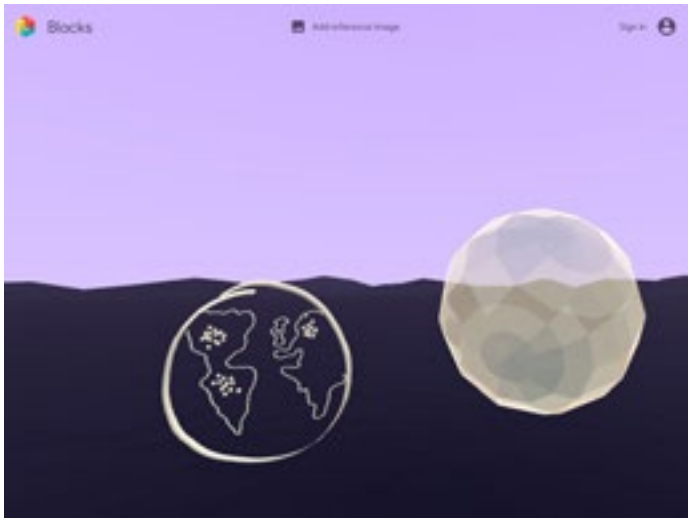
During the first setup of the VRNetzer environment, I presented some issues with the visualisation of the data. I could access the environment, but I couldn’t see the controllers or any preloaded data point. I used a VIVE index headset, which haven’t been tested by the developers of the project. So, I will be reviewing the executable file and load some data following the github instructions (https://github.com/menchelab/VRNetzer) The issue might also concern the headset itself, and I should test on the HTC VIVE.
The main point of testing this early is to understand if it would be a functional medium to present my data. I want to understand how much control I have on the platform.

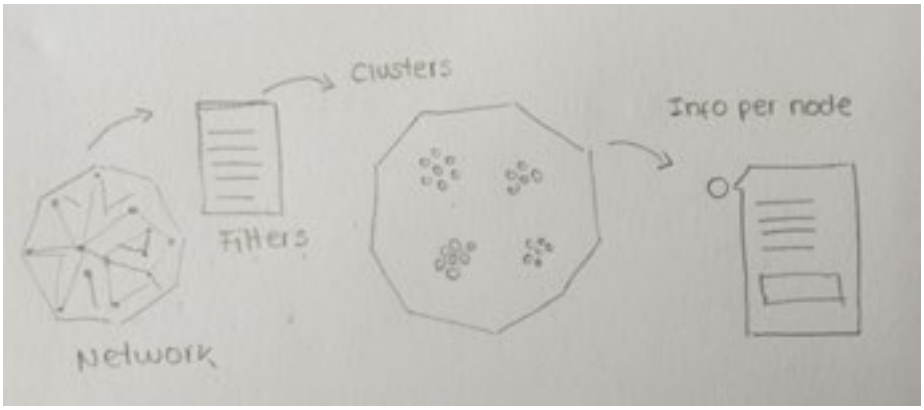
Mockups in Blocks-Google VR

I have also experimented with Blocks VR. The motive of these quick mockups was to interact with the space and understand how the clusters might look like. I tried different shapes and colours to differentiate the clusters.



Then I went into thinking how would those clusters look like within the earth. I clearly was thinking of it in 2D and then moved to a 3D sphere. I can picture the nodes within the sphere in groups that perhaps represent continents, and then those nodes moving from cluster to cluster according to changes in the data (triggered by the user).





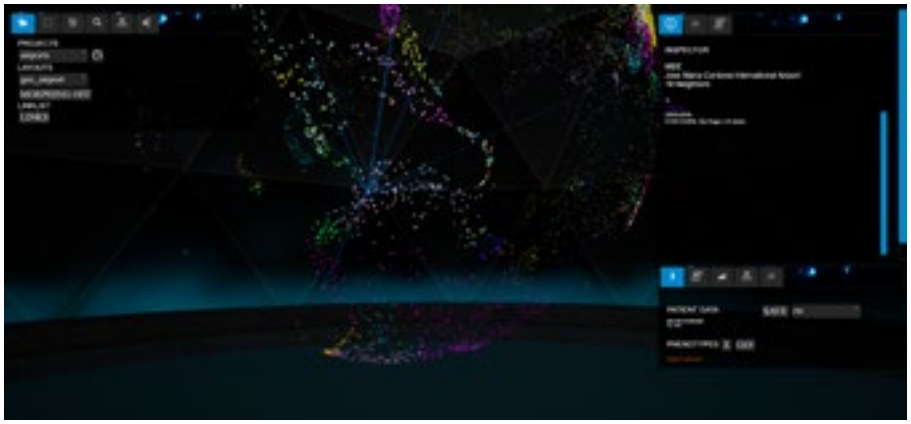
General functioning flow idea

Initial Network -> Filters by user -> New clusters -> Information per node

First look at the VRNetzer Analysis platform

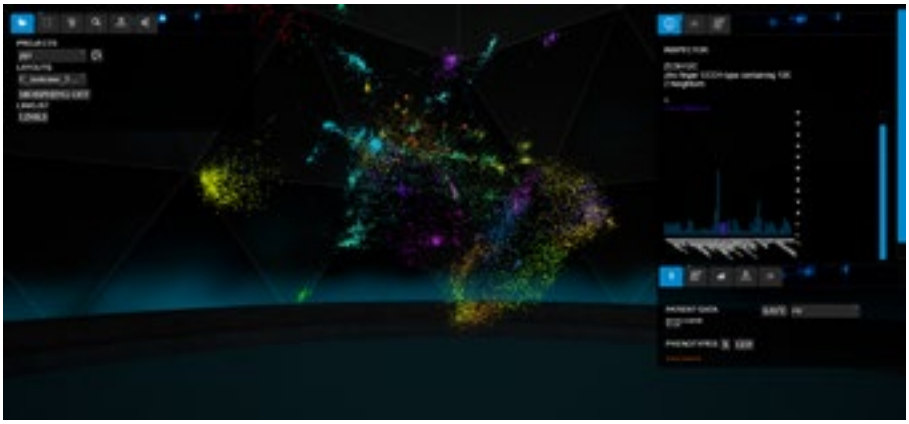
The platform uses a UI module, an Analytics module, and a VR module. To handle the data, it uses **MySQL workbench** from where we can create a new database and add new data through a **POST request** (generated in the UI module). As a first example, the platform shows a network of the world's airports. It does resemble the idea I had in mind, however, it could be expanded with interactions that would come from the LinkedIn dataset.

Nodes can show their immediate links



• Nodes can be clicked on, exposing visually some of the strongest links with other nodes (called neighbours).
• Clicking on the node also gives extra information about the node itself.

3D clusters. Exploring a node can give further information.



The nodes can be organized in different **layouts**. These layouts can be 2D or 3D and they can be selected by the user. The different layouts represent a change in the **CSV file**, and how data is positioned.

Challenges and Formative Assignment feedback

• Even though I had a formed project idea, it wasn't fully meeting the brief. I needed to include a 'web-based experiment' and I found that the amount of data I needed to 'embellish' the dataset I had found online was very challenging. The Linkedn API was also very restrictive with the information we could extract, and to GET relevant data, I had to fill in an application to become a '**LinkedIn Partner**'.

Furthermore, during the following classes, we got to explore '**sentiment analysis'** and how we could find the percentage of negativity or positivity of words. We also experimented with **posting to Twitter**, mainly because accessing the Twitter API to GET information had become a lot more restricted. This resulted in a new project idea that combines the different techniques we have learnt in class.

Those are the reasons why I have changed my initial idea, to a new project that still explores the subject of 'jobs' and 'talent'. However, I intend to explore how people are currently perceiving Artificial Intelligence in the workplace now, and in the future.

Link to Formative Assignmnet

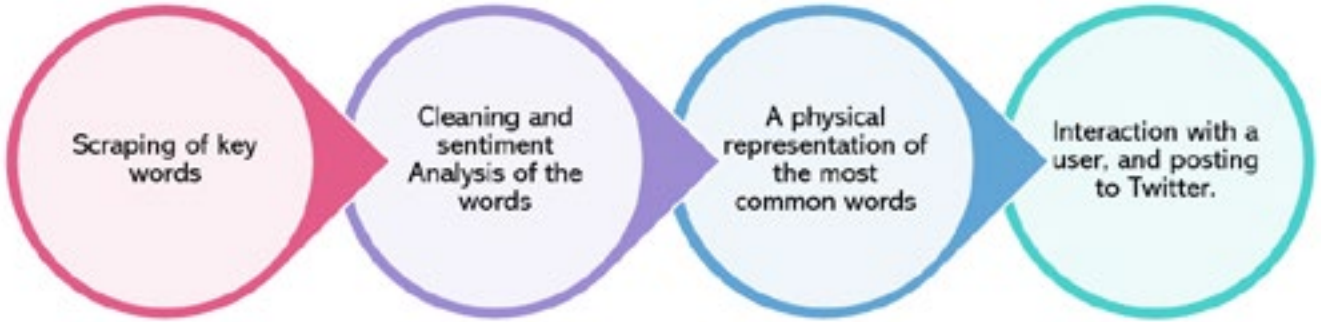
NEW PROJECT IDEA

How do people feel about AI, especially around the subject of job automation?

The recent release of Chat GPT has reached many users and has unveiled the power of AI and its disruptive nature. This large language model has caused many opinions and has sparked caution amongst many technologists.

The quick development of even more advanced models resulted in the publication of an 'Open Letter' which advocates for the pause of giant AI experiments for 6 months.
Such letter was signed by the likes of Elon Musk, Bart Selman, Yuval Harari, etc.

How should it work?



Scrapping of key words related to feelings:

Using the **'Pause Giant AI Experiments'** open letter as a starting point to extract key words. Then, navigate to hyperlinks present within the letter to keep extracting more words related to 'feelings'

Cleaning and Sentiment Analysis:

Counting the frequency of the words previously found and assessing how positive or negative they are.

Physical Representation:

Select the 8 most prominent words and accomodate:

- 2 words per candy in 4 designs:

perform 'word to vector' to decide which words should go together in one candy.

- Two colours: red and blue, making reference to the iconic Matrix scene.

- max 9 characters: using 'Lemmatization', to reduce the word.

Interaction with an user

Using a small gum ball machine as input for an user to express how the feel about AI, using the words that were previously printed in the chocolates.

By taking out a chocolate they will be also posting their opion on Twitter.

Design and Development

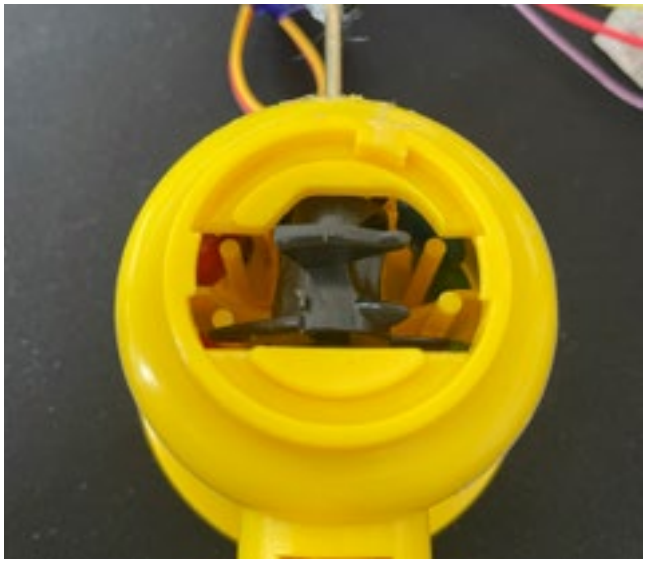
Why a Gum Ball machine ?

This small toy machine was randomly bought by my sister. I thought it was quite playful and also brought back memories of the uncertainty one can feel when turning the knob. You can see the sweets, but you can't choose which one you will get. I also like the idea of 'eating the words' that will be printed on the chocolates.

Testing the Gum Ball machine internal mechanism.

The machine is fully made of plastic and relies on user input. This meant that I could add movement through the back by connecting a servo motor.
The shape of the chocolate also matters, as some of them get stuck internally.

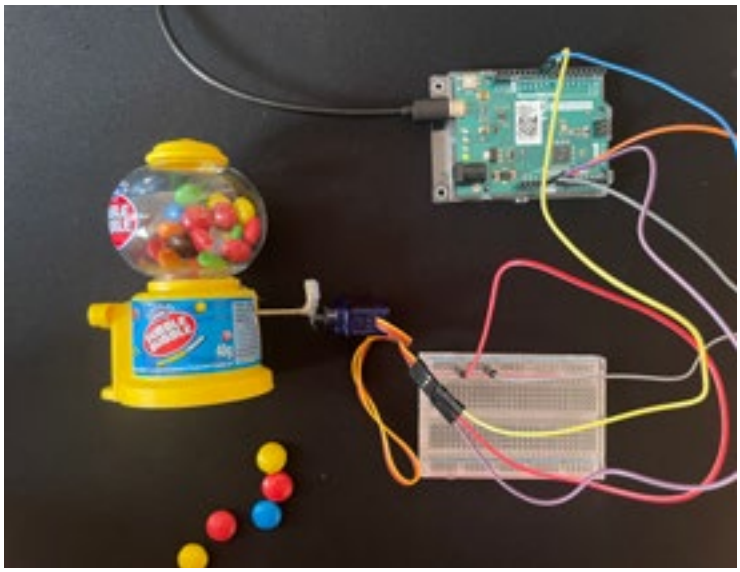
Watch video



Testing the Gum Ball machine with a 180 ° Servo.

One of the key aspects of using the machine is how the user will interact with it, and how such interaction allows for posting on Twitter. My first thought was to use a 180 servo to observe how well it would release the chocolates.

Watch video



Using the 180 ° Servo to post on Twitter

This idea came from the usage of the Tweepy library to post on Twitter. By creating rotation conditions with the Servo I could trigger Tweepy to create a new Tweet.
However, to be able to work within Python, I had to use pyfirmata and the Firmata library in Arduino. Firmata is a generic protocol for communicating with microcontrollers from software on a host.computer

```
import tweepy
import config
from pyfirmata import Arduino, SERVO, util
```

```
while True:
    x = input("input:")
    if x == "1":
        response = client.create_tweet(text='i love me')
        print(response)
        for i in range(0,180):
            rotateservo(pin, i)

        for i in range(180,1,-1):
            rotateservo(pin,i)
```

Testing the Gumball machine with a 360 ° Servo.

The continuous servo was a challenging component, since it works completely differently from a 180 servo. However, it does do a full spin and allows the sweets to fall faster, which at the same time reduces the number of sweets getting stuck.

Watch video

Checkpoint

• The **continuous servo** will be used to control the gumball machine. However, it needs to start and end at the same point. *(code improvement).*
• The servo motors will need to be held, as well as the gumball machine so that the only thing spinning is the knob *(physical structure improvement).*
• Twitter **can't accept** the same Tweet twice. Perhaps, a way to solve this, is allowing the user to generate a new input *(code improvement).*


Designs will be printed on all (2) colors.
Worried Anxious
Scared Frustrate
Excited Amazed
Ambitious Collab
designs
colors
Worried Anxious
Excited Amazed
Scared Frustrate
Ambitious Collab
Red
Dark Blue