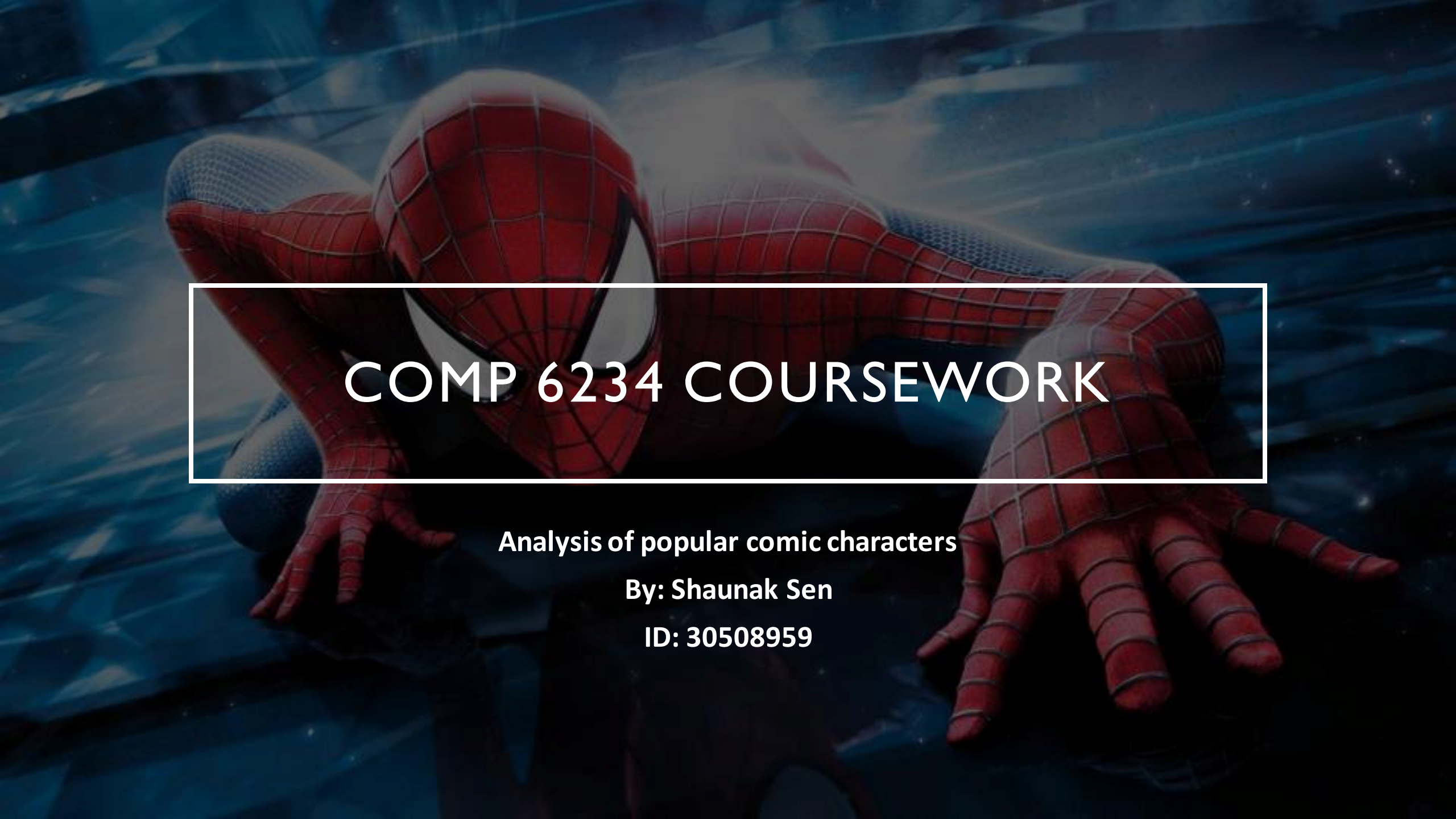

COMP 6234 COURSEWORK

Analysis of popular comic characters

By: Shaunak Sen

ID: 30508959

Data Story and questions to be answered

Almost everyone, whether it be a kid, or an adult has a favorite superhero! Comic characters are huge in number and popularity. Their popularity has further increased with the large number of movies being produced. I believe comic characters are highly relatable to a wide range of audience and it is also something which I will enjoy working with, so I have chosen it as the topic for my project.

I aim on performing analysis on data of popular comic characters (both Marvel and DC) and create visualizations and interactions that support and answer questions like the following:

- Do stronger characters in comics tend to be more intelligent?
- How are popular comic characters related to each other in terms of affiliations?
    - How often do good characters team up with bad characters?
- What are the most common and rare super powers that characters possess?
    - Which characters apart from Spiderman can sense danger?
- What factors influence the popularity of comic characters – strength, gender, character neutrality, other abilities
- How are popular characters demographically located?
    - Where are the strongest Marvel characters located?

Datasets to be used

I aim on performing analysis on the following datasets:

- https://github.com/fivethirtyeight/data/blob/master/comic-characters/dc-wikia-data.csv
- https://github.com/fivethirtyeight/data/blob/master/comic-characters/marvel-wikia-data.csv
- https://www.kaggle.com/claudiodavi/superhero-set#super_hero_powers.csv

The first two datasets follow a similar structure. They have been scraped from the following websites:

- http://marvel.wikia.com/Main_Page
- http://dc.wikia.com/wiki/Main_Page

The data was scraped in **August 2014** and has been split into two files according to which universe the character belongs to.

The third dataset was scraped from the website: https://www.superherodb.com/

This dataset was collected in **June 2017**

Column Description – First two datasets

Variable	Definition
page_id	The unique identifier for that character's page within the wikia
name	The name of the character
urlslug	The unique url within the wikia that takes you to the character
ID	The identity status of the character (Secret Identity, Public identity)
ALIGN	If the character is Good, Bad or Neutral
EYE	Eye color of the character
HAIR	Hair color of the character
SEX	Sex of the character (e.g. Male, Female, etc.)
GSM	If the character is a gender or sexual minority (e.g. Homosexual characters, bisexual characters)
ALIVE	If the character is alive or deceased
APPEARANCES	The number of appearances of the character in comic books (as of Sep. 2, 2014)
FIRST APPEARANCE	The month and year of the character's first appearance in a comic book, if available
YEAR	The year of the character's first appearance in a comic book, if available

Data Description – First Dataset

Column Name	Data Type
page_id	Integer
name	String
urlslug	String
ID	String (4 unique values, 2013 missing values)
ALIGN	String (3 unique values)
EYE	String
HAIR	String
SEX	String
GSM	String
ALIVE	String (2 unique values and 3 missing values)
APPEARANCES	Integer
FIRST APPEARANCE	String (to be converted to Date format, has 69 missing values)
YEAR	String (has 69 missing values)

First Dataset – Additional Information

- Size of dataset: **1.1 MB**
- Data dimensions: **6896 rows, 13 columns**
- Format of data file: .csv
- Missing values present: Yes
  - There are **2013** missing values in ID. ID might be an important feature in our analysis, so ID data needs to be scraped for these 2013 characters (if available)
  - ALIVE column has **3** missing values. On inspecting them we find that these rows do not correspond to any meaningful comic character. So we simply delete these 3 rows from our dataset
  - FIRST APPEARANCE and YEAR each have **69** missing values. It is observed that the characters for whom the YEAR is not known also have FIRST APPEARANCE as missing. This is expected.
- Source: http://dc.wikia.com/wiki/

Data Description – Second Dataset

Column Name	Data Type
page_id	Integer
name	String
urlslug	String
ID	String (4 unique values)
ALIGN	String (3 unique values)
EYE	String
HAIR	String
SEX	String
GSM	String
ALIVE	String (2 unique values and 3 missing values)
APPEARANCES	Integer
FIRST APPEARANCE	String (to be converted to Date format, has 815 missing values)
YEAR	String (has 815 missing values)

Second Dataset – Additional Information

- Size of dataset: **2.4 MB**
- Data dimensions: **16376 rows, 13 columns**
- Format of data file: .csv
- Missing values present: **Yes**
    - ALIVE column has **3** missing values. On inspecting them we find that these rows do not correspond to any meaningful comic character. So we simply delete these 3 rows from our dataset
    - FIRST APPEARANCE and YEAR each have **815** missing values. It is observed that the characters for whom the YEAR is not known also have FIRST APPEARANCE as missing. This is expected.
- Source: http://marvel.wikia.com/wiki/

Data Description – Third dataset

- This dataset has data for **668 unique** comic characters.
- For each character there is a list of possible super powers.
- There are **167** such superpowers. Thus there are **168** (name + 167 superpowers) columns and **668** rows in this dataset.
- Some of the superpowers are: Agility. Accelerated healing, Dimensional Awareness, Energy, Absorption etc.
- Each column corresponding to a super power for a specific superhero contains two possible values – **True** or **False**
- For e.g. Accelerated Healing for Ant Man is False implies that the character does not possess that super power
- There are **no missing values** in this dataset

Third dataset – Additional Information

Column Name	Data Type	Unique Values
hero_names	String	667 (all unique values)
[.. superpowers ..]	Boolean	True or False

- Size of dataset: **672.3 Kb**
- Data dimensions: **667 rows, 168 columns**
- Format of data file: .csv
- Missing values present: **No**
- Source: https://www.superherodb.com/

FEW PRELIMINARY VISUALIZATION IDEAS

- Question – Do stronger characters in comics tend to be more intelligent?
    - Data for strength level and intelligence needs to be scraped from the sources
- **Scatter plots** to visualize relationships between strength and intelligence
- **Preliminary Conclusion**: Strength and intelligence should have high correlation for comic characters. Characters like Batman, Hulk, Doctor Strange are all powerful and possess high intelligence

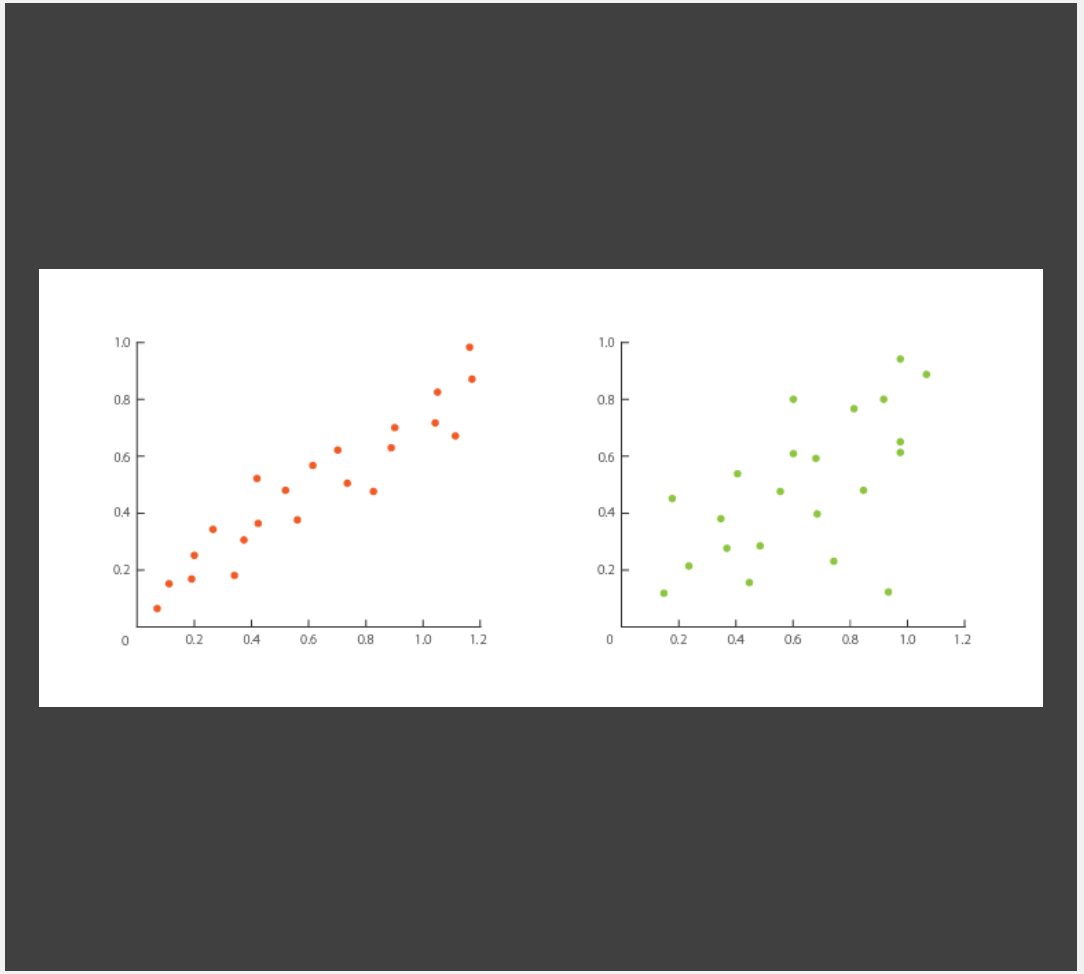

Image Source: https://datavizcatalogue.com/

- Question – What factors influence the popularity of comic characters – strength, gender, character neutrality, other abilities?
- Analysis of factors that affect popularity (number of appearances) by visualizing data by **scatter plots, bar graphs and correlation heat maps**
- Plotting a **bubble chart** to show influence of strength and intelligence on the popularity. Size of the bubbles will denote popularity
    - Here I have assumed that strength and intelligence are factors that influence popularity of a character

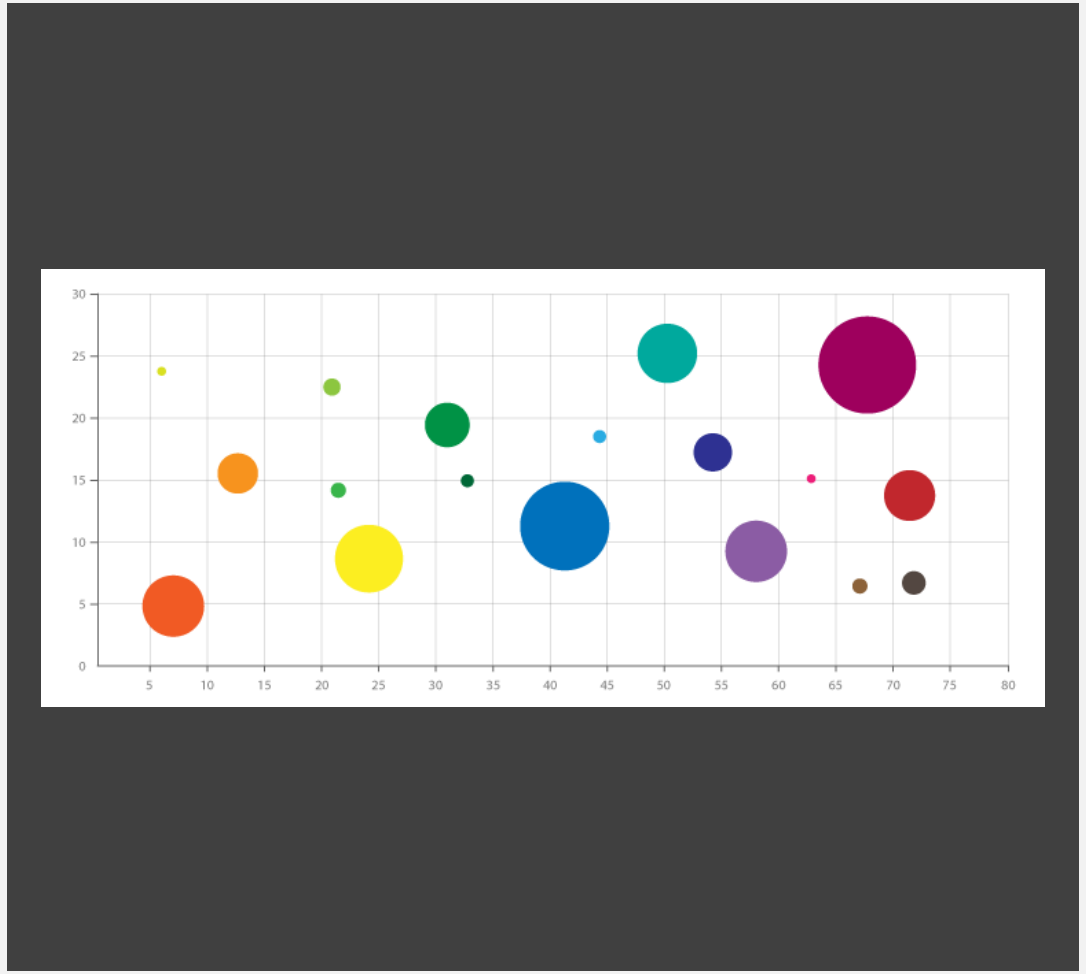

Image Source: https://datavizcatalogue.com/

- Question – How are popular characters demographically located?
  - Data for the location of base of operations needs to be scraped from the source
- Data can be visualized in form of a **dot map** (to visualize distribution of comic characters over a region) or as a **Choropleth Map** which will help answer questions like where are the strongest characters located ?
- **User interactions** can be promoted such that they can filter for specific regions on the map, or they can filter for specific attributes (like gender, strength, intelligence) to view the distribution of.

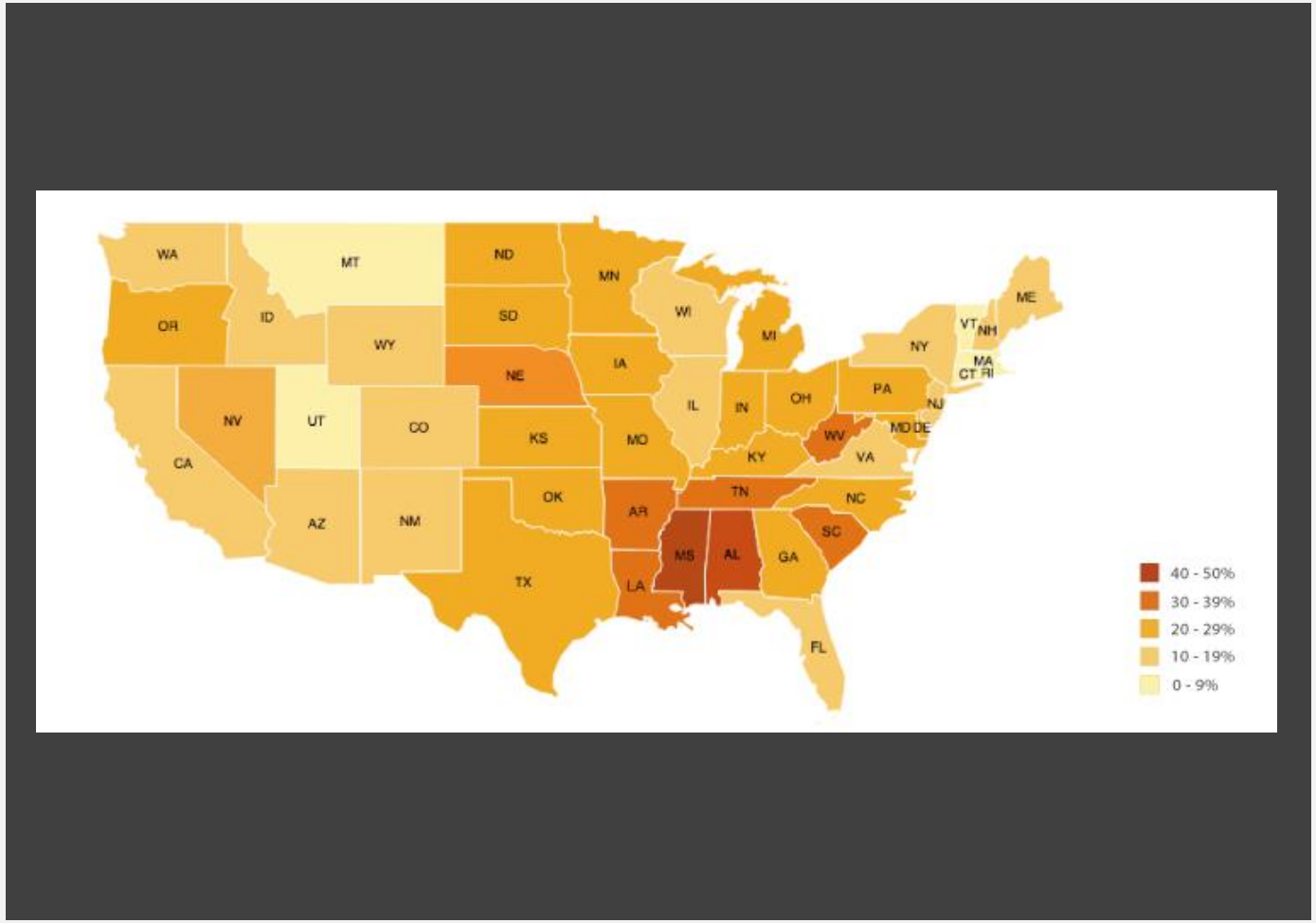

Image Source: https://datavizcatalogue.com/

- Question – How are popular comic characters related to each other in terms of affiliations?
- Data about affiliations of each character needs to be scraped from the source
- The relationship between various characters (according to common affiliations) can be shown through a **Network Diagram or Chord Diagram**
- Characters can be clustered based on their universe (Marvel or DC) or their character Alignment and interconnections between these groups can be visualized
- **Preliminary Conclusion**: Often good characters team up with negative characters to defeat a common enemy (e.g. Spiderman with Venom)

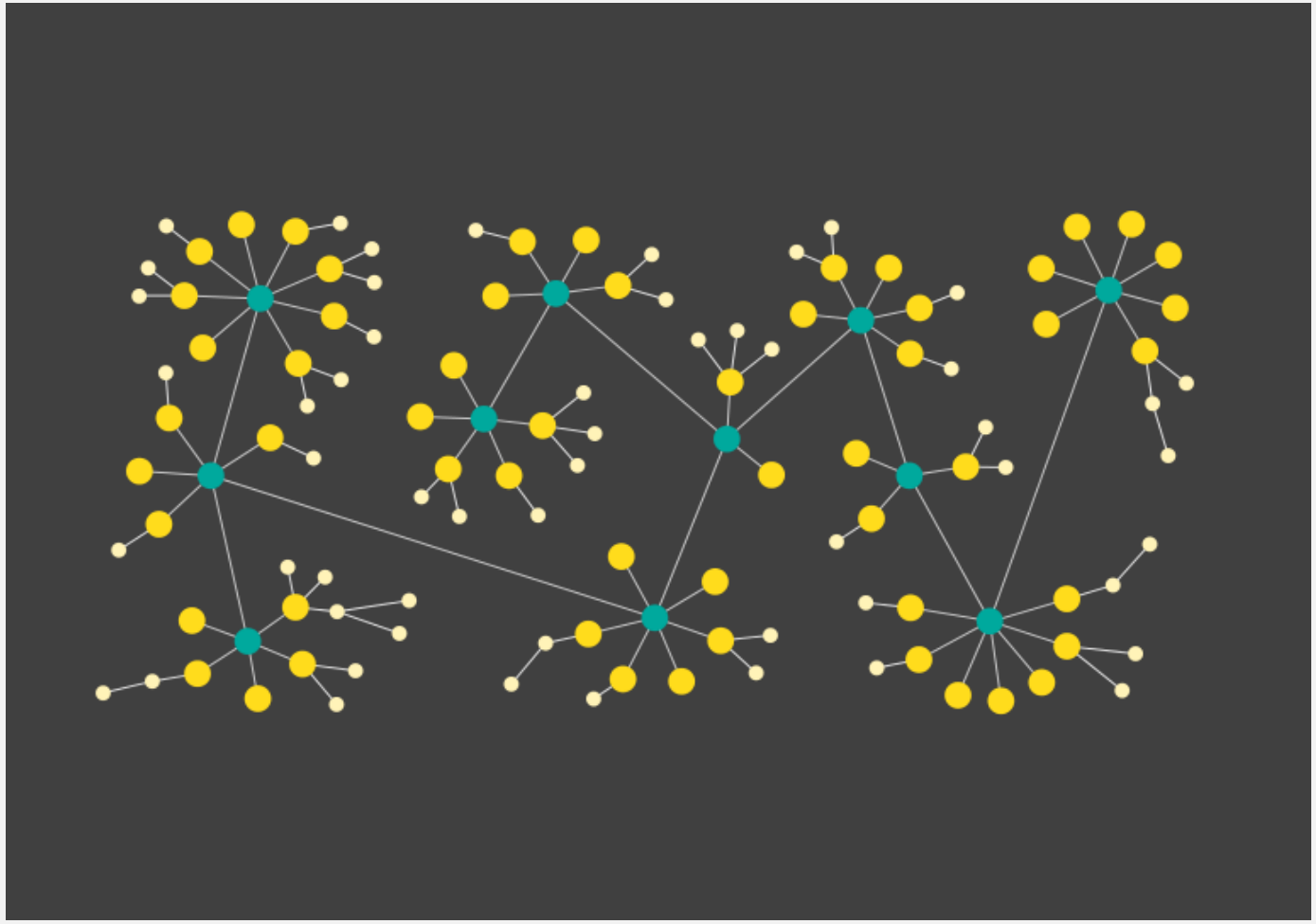

Image Source: https://datavizcatalogue.com/

THANK YOU

ss8n18@soton.ac.uk