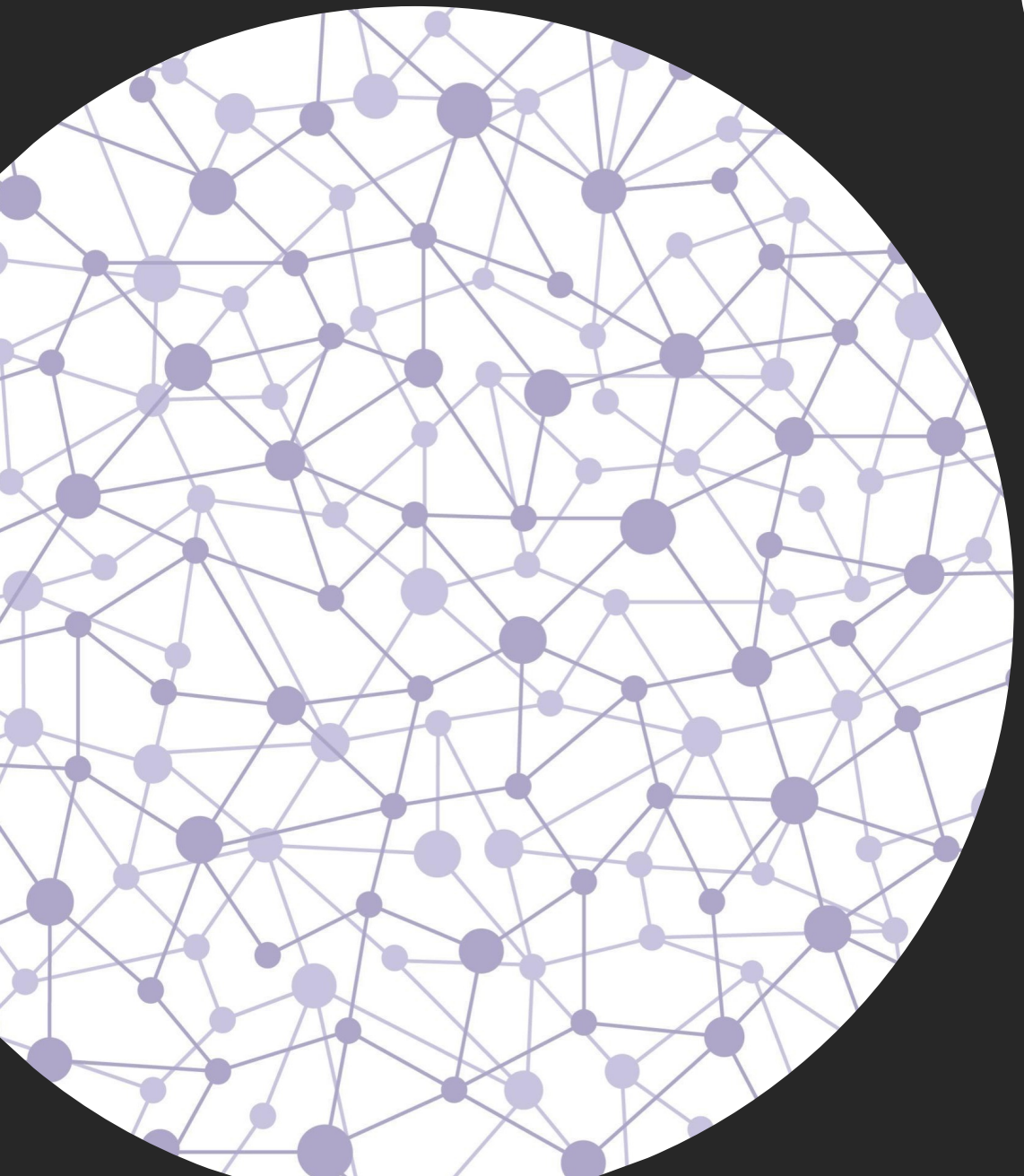
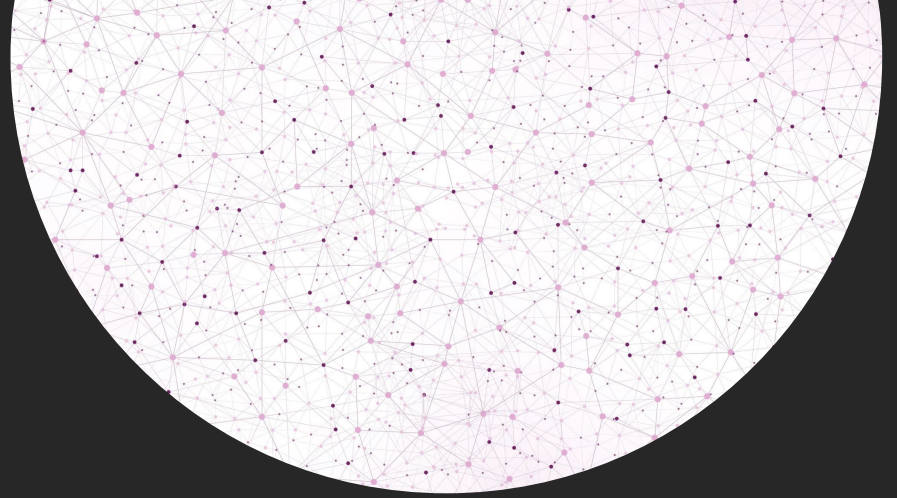




SNL Network/Collaboration Visualization

Tali Zacks

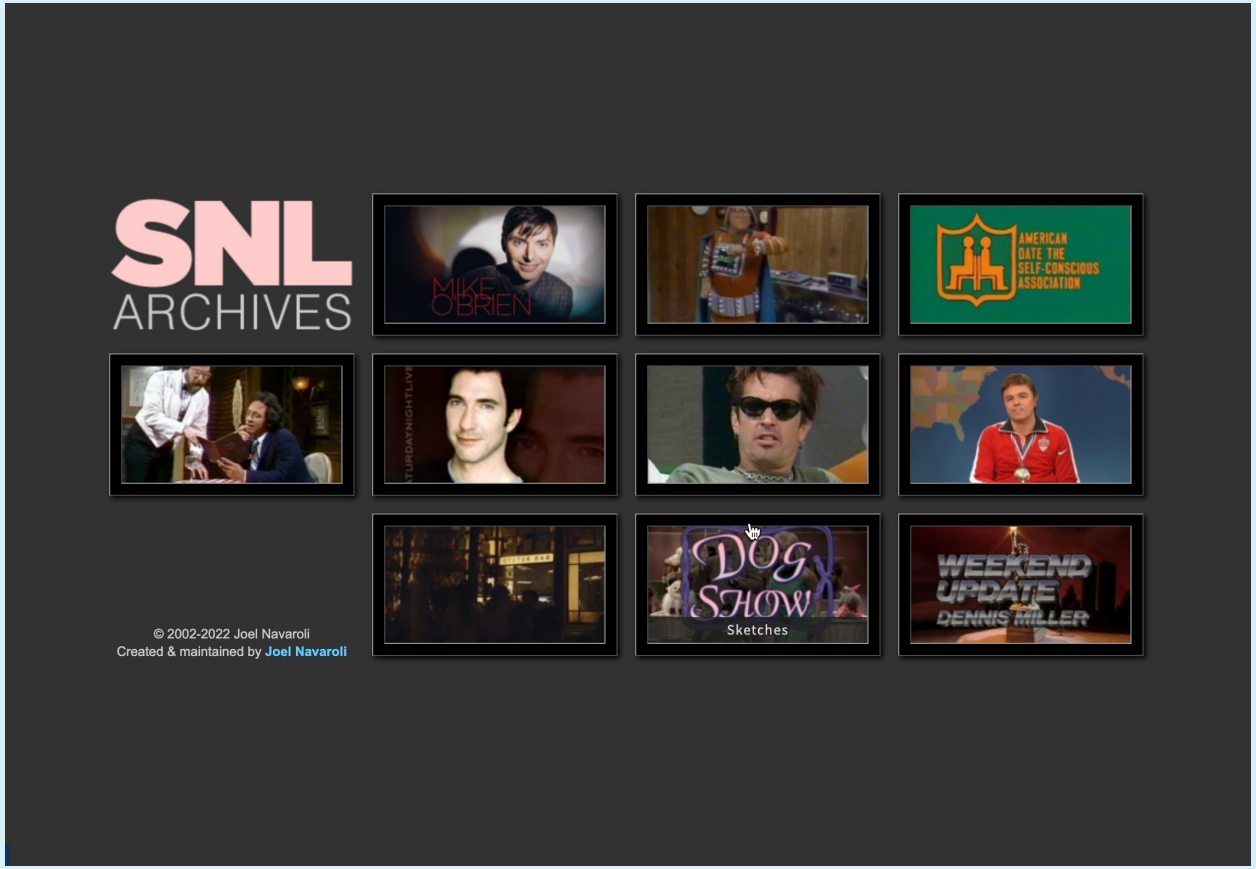
Idea Sparked

- I recently rewatched Brooklyn Nine-Nine, a show starring Andy Samberg (2005–2012), and noticed random appearances from a slew of SNL alumni, including (but **definitely** not limited to):
  - Fred Armisen (2002–2013) : Mlep(clay)nos (the clay is silent)
  - Maya Rudolph (2000–2007) : Karen Haas
  - Vanessa Bayer (2010–2017) : Debbie Fogle
- It made me think about all the collaborations that occur between SNL cast members and how frequently they include one another in their projects
  - Adam Sandler (1990–1995) movies featuring Rob Schneider (1990–1994) and David Spade (1990–1996)
  - Amy Poehler (2001–2008) and Tina Fey (1997–2006)
- I began digging into how I might be able to explore these connections

SNL Archives

- © 2002-2022 Joel Navaroli
  Created & maintained by **Joel Navaroli**

- GitHub and Kaggle Datasets created by scraping snlarchives.net
  - Hendrik Hilleckes (@hhllcks, hllcks@gmail.com, blog.hhllcks.de)
  - Colin Morris (http://colinmorris.github.io/)



SNL Archive Dataset

- **actors**
- appearances
- casts
- characters
- episodes
- hosts
- impressions
- seasons
- sketches
- tenure
- titles

	aid	url	type	gender
0	Kate McKinnon	/Cast/?KaMc	cast	female
1	Alex Moffat	/Cast/?AlMo	cast	male
2	Ego Nwodim	/Cast/?EgNw	cast	unknown
...	...	...	...	...
2304	Janis Ian	/Guests/?2	guest	female
2305	Valri Bromfield	/Guests/?5	guest	unknown

SNL Archive Dataset

- actors
- appearances
- **casts**
- characters
- episodes
- hosts
- impressions
- seasons
- sketches
- tenure
- titles

	aid	sid	featured	first_epid	last_epid	update_anchor	n_episodes	season_fraction
0	A. Whitney Brown	11	True	19860222.0	NaN	False	8	0.444444
1	A. Whitney Brown	12	True	NaN	NaN	False	20	1.000000
...	...	...	...	...	...	...	...	...
612	Fred Armisen	37	False	NaN	NaN	False	22	1.000000
613	Fred Armisen	38	False	NaN	NaN	False	21	1.000000

IMDb Datasets

- IMDb (Internet Movie Database) allows users to download HUGE, messy datasets
- I downloaded:
  - *names.basics.tsv (691.9 MB)*
  - *title.basics.tsv (757.3 MB)*
  - *title.pricipals.tsv (2.2 GB)*

  from https://datasets.imdbws.com/
- Licensing Information: Subsets of IMDb data are available for access to customers for personal and non-commercial use

names.basics.tsv

- nconst (string) - alphanumeric unique identifier of the name/person
- primaryName (string)– name by which the person is most often credited
- birthYear – in YYYY format
- deathYear – in YYYY format if applicable, else '\N'
- primaryProfession (array of strings)– the top-3 professions of the person
- knownForTitles (array of tconsts) – titles the person is known for

names.basics.tsv

	nconst	primaryName	birthYear	deathYear	primaryProfession	knownForTitles
0	nm0000001	Fred Astaire	1899	1987	soundtrack,actor,miscellaneous	tt0050419,tt0053137,tt0031983,tt0072308
1	nm0000002	Lauren Bacall	1924	2014	actress,soundtrack	tt0071877,tt0037382,tt0117057,tt0038355
2	nm0000003	Brigitte Bardot	1934	\N	actress,soundtrack,music_department	tt0054452,tt0056404,tt0057345,tt0049189
...	...	...	...	...	...	...
11554308	nm9993718	Aayush Nair	\N	\N	cinematographer	\N
11554309	nm9993719	Andre Hill	\N	\N	NaN	\N

title.basics.tsv

- tconst (string) - alphanumeric unique identifier of the title
- titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
- primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release
- originalTitle (string) - original title, in the original language
- isAdult (boolean) - 0: non-adult title; 1: adult title
- startYear (YYYY) – represents the release year of a title. In the case of TV Series, it is the series start year
- endYear (YYYY) – TV Series end year. '\N' for all other title types
- runtimeMinutes – primary runtime of the title, in minutes
- genres (string array) – includes up to three genres associated with the title

title.basics.tsv

	tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	end Year	runtime Minutes	genres
0	tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary,Short
1	tt0000002	short	Le clown et ses chiens	Le clown et ses chiens	0	1892	\N	5	Animation,Short
...	...	...	...	...	...	...	...	...	...
8854183	tt9916856	short	The Wind	The Wind	0	2015	\N	27	Short
8854184	tt9916880	tvEpisode	Horrid Henry Knows It All	Horrid Henry Knows It All	0	2014	\N	10	Adventure,Animation,Comedy

title.pricipals.tsv

- tconst (string) - alphanumeric unique identifier of the title
- ordering (integer) – a number to uniquely identify rows for a given titleId
- nconst (string) - alphanumeric unique identifier of the name/person
- category (string) - the category of job that person was in
- job (string) - the specific job title if applicable, else '\N'
- characters (string) - the name of the character played if applicable, else '\N'

title.pricipals.tsv

	tconst	ordering	nconst	category	job	characters
0	tt0000001	1	nm1588970	self	\N	["Self"]
1	tt0000001	2	nm0005690	director	\N	\N
2	tt0000001	3	nm0374658	cinematographer	director of photography	\N
...	...	...	...	...	...	...
49901290	tt9916880	4	nm10535738	actress	\N	["Horrid Henry"]
49901291	tt9916880	5	nm0996406	director	principal director	\N
49901292	tt9916880	6	nm1482639	writer	\N	\N
49901293	tt9916880	7	nm2586970	writer	books	\N
49901294	tt9916880	8	nm1594058	producer	producer	\N

Full Cast WikiTable

Table of *Saturday Night Live* cast members

Performer	Time on *SNL*	No. of seasons	Repertory Player	Featured Player	Middle Group	"Weekend Update" Anchor	Hosted	Best of...	Writer
Fred Armisen	2002–2013	11	✓	✓			✓		
Aristotle Athari	2021–present	1		✓					
Dan Aykroyd	1975–1979	4	✓			✓	✓	✓	✓
Peter Aykroyd	1980	1		✓					✓
Morwenna Banks	1995	1	✓						
Vanessa Bayer	2010–2017	7	✓	✓					
Jim Belushi	1983–1985	2	✓						✓
John Belushi	1975–1979	4	✓					✓	✓
Beck Bennett	2013–2021	8	✓	✓					
Jim Breuer	1995–1998	3	✓						
Paul Brittain	2010–2012	2		✓					
A. Whitney Brown	1986–1991	6		✓					✓
Aidy Bryant	2012–present	10	✓	✓					
Beth Cahill	1991–1992	1		✓					
Dana Carvey	1986–1993	7	✓				✓	✓	
Chevy Chase	1975–1976	2	✓			✓	✓	✓	✓
Michael Che	2014–present	8	✓	✓		✓			✓
Ellen Cleghorne	1991–1995	4	✓		✓				
George Coe	1975	1	✓						
Billy Crystal	1984–1985	1	✓			✓	✓		✓
Jane Curtin	1975–1980	5	✓			✓			
Joan Cusack	1985–1986	1	✓						
Pete Davidson	2014–present	8	✓	✓					
Tom Davis	1977–1980	3		✓					✓

- I then came across a Wikipedia page with a table featuring
  - *Every SNL cast member (all 159)*
  - *Their time on SNL*
  - *Number of seasons*
  - *And some columns with checkmarks denoting various roles on the show*
- I downloaded the table using an online WikiTable to csv converter
  - *Didn't keep the checkmark columns intact*
  - *I copied the table contents into Excel and used Find & Replace to change every checkmark into a "Yes"*

```
wiki_cast_excel = pd.read_csv('wiki_cast_excel.csv')
```

```
wiki_cast_excel.iloc[29]
```

```
Performer                  Brian Doyle-Murray
Time on SNL                    1980;1981–1982
No. of seasons                              2
Repertory Player                          Yes
Featured Player                           Yes
Middle Group                              NaN
"Weekend Update" Anchor                   Yes
Hosted                                    NaN
Writer                                    Yes
Name: 29, dtype: object
```

```
wiki_cast = wiki_cast_excel.fillna('No')
wiki_cast
```

	Performer	Time on SNL	No. of seasons	Repertory Player	Featured Player	Middle Group	"Weekend Update" Anchor	Hosted	Writer
0	Fred Armisen	2002–2013	11	Yes	Yes	No	No	Yes	No
1	Aristotle Athari	2021–present	1	No	Yes	No	No	No	No
2	Dan Aykroyd	1975–1979	4	Yes	No	No	Yes	Yes	Yes
3	Peter Aykroyd	1980	1	No	Yes	No	No	No	Yes
4	Morwenna Banks	1995	1	Yes	No	No	No	No	No
...	...	...	...	...	...	...	...	...	...
154	Casey Wilson	2008–2009	2	No	Yes	No	No	No	No
155	Fred Wolf	1996–1996	2	No	Yes	No	No	No	Yes
156	Bowen Yang	2019–present	3	Yes	Yes	No	No	No	Yes
157	Sasheer Zamata	2014–2017	4	Yes	Yes	No	No	No	No
158	Alan Zweibel	1980	1	No	Yes	No	No	No	Yes

159 rows × 9 columns

Formatting the Data

- I began by creating an IMDb names list using:

```
imdb_name_list = imdb_names_df.values.tolist()
```

- Then I took a subsection of the list which contained only entries with a name in the Wikipedia list of SNL actors:
- This pulled out all the entries with SNL cast member names but gave me upwards of 50 entries for some actors, all with separate 'knownForTitles'

```
cross_imdb_wiki_list = []
for person in imdb_name_list:
      if person[1] in wiki_cast['Performer'].values:
        cross_imdb_wiki_list.append(person)

def check_names(name):
    cross_name_list = (set([r[1] for r in cross_imdb_wiki_list]))
    result = False
    if name in cross_name_list:
        result = True
    return result
SNL_name_check = [r for r in cross_imdb_wiki_list if
check_names(r[1])]
```

Combining Title IDs

```
for i in SNL_name_check:
        for x in i:
         if str(x).startswith('tt'):
            titles = x.split(',')

        name = i[1]
        names_title_dict[name] += titles

for name, titles in names_title_dict.items():
        names_title_dict[name] = list(set(titles))
```

- Most IMDb entries have ‘knownForTitles’ separated by commas
- The position of those items within each dictionary is inconsistent:
  - {'actor,soundtrack,writer', 'nm0000195', '1950', '\\N', 'tt1748122,tt0128445,tt0362270,tt0335266', 'Bill Murray’}
  - {'nm0614853', '\\N', 'miscellaneous', 'Bill Murray', 'tt0097132’}
  - {'editor,editorial_department', '\\N', 'tt1732762,tt1210095', 'nm2966932', 'Bill Murray’}
- Rather than just 4 or 5 knownForTitles, some actors now have upwards of 50 title IDs associated with their name

(unnecessarily)

Combining Tables

- In order to combine the IMDb and Wikipedia DataFrames, I needed to set the indices to the actors' names in both

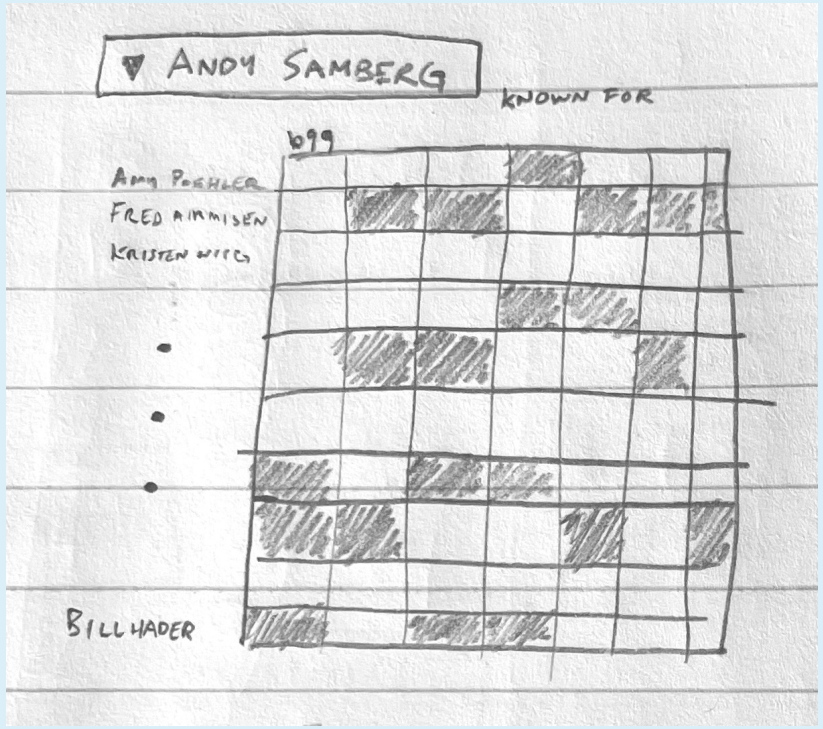
```
SNL_df = pd.DataFrame(SNL_cred_check, columns =
imdb_names.columns.tolist()).set_index('primaryName’)

wiki_cast = wiki_cast.set_index('Performer')

imdb_wiki_table = pd.concat([SNL_df, wiki_cast], axis=1, join="outer")
```

	nconst	birth Year	death Year	primary Profession	knownForTitles	Time on SNL	No. of seasons	Repertory Player	Featured Player	Middle Group	"Weekend Update" Anchor	Hosted	Writer
John Belushi	nm0000004	1949	1982	actor,soundtrack,writer	tt0078723,tt0072562,tt0077975,tt0080455	1975–1979	4	Yes	No	No	No	No	Yes
Jim Belushi	nm0000902	1954	\N	actor,music_department,producer	tt0097637,tt0072562,tt0117468,tt0095963	1983–1985	2	Yes	No	No	No	No	Yes
Dana Carvey	nm0001022	1955	\N	actor,soundtrack,writer	tt0072562,tt0108525,tt0105793,tt0295427	1986–1993	7	Yes	No	No	No	Yes	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Bowen Yang	NaN	NaN	NaN	NaN	NaN	2019–present	3	Yes	Yes	No	No	No	Yes
Sasheer Zamata	NaN	NaN	NaN	NaN	NaN	2014–2017	4	Yes	Yes	No	No	No	No

The Fun Part

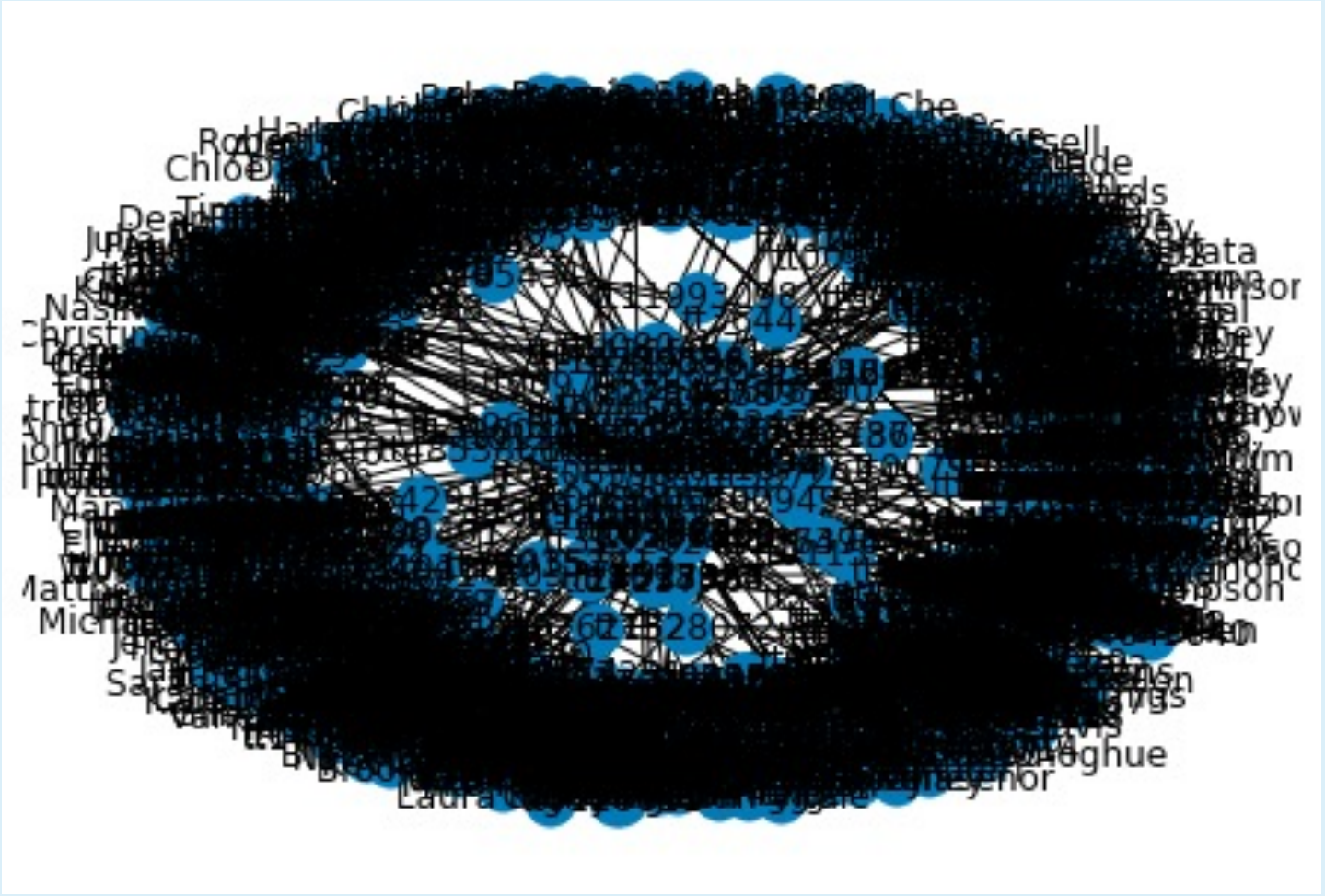


- Attempting to visualize the data introduced several problems to overcome
- The first being: How can I visualize collaboration frequency?
  - Network (node/edges) graph
  - **Heat maps**
- Eventual goal: Linked, interactive dashboard

← My initial vision

Attempting Network Graphs

- Using NetworkX and Plotly, I managed to create this horrific visual



Attempting Network Graphs

- Using pygraphviz, I managed to make these useless and impossible to read directed network graphs:
  - Undirected
  - Directed

Boolean Tables

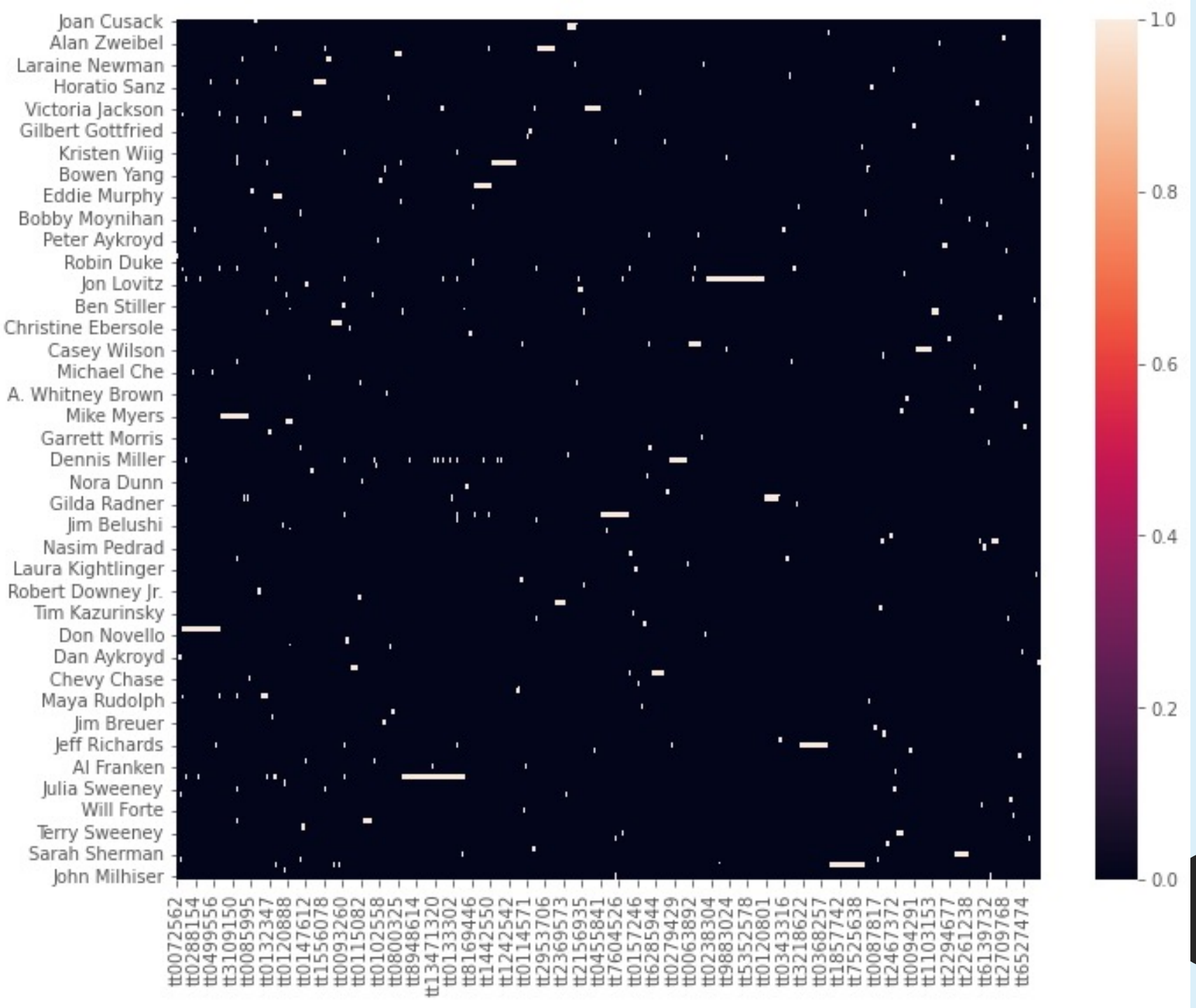
- At this point, I decided that a heat map would be my best solution
- I created the tables using a function to run through the titles vs names dictionary that I created and add a value depending on whether a title was present in both actor's dictionary values

```
values = list(set([x for y in same_title_dict.values() for x in y]))
data = {}
for key in same_title_dict.keys ():
    data[key] = [True if value in same_title_dict[key] else False for value in values]

boolean_table_name_cols = pd.DataFrame(data, index=values)
pd.DataFrame(boolean_table_name_cols)
```

	tt0072562	tt0077975	tt0078723	...	tt5117666	tt7242698	tt7778796
Joan Cusack	False	False	False	...	False	False	False
Phil Hartman	True	False	False	...	False	False	False
...	...	...	...	...	...	...	...
Brian Doyle-Murray	False	False	False	...	False	False	False
John Milhiser	True	False	False	...	False	False	False

“Full” Heat map



Reversed


tt0773756
tt3517306
tt0351810
tt1759761
tt0008582
tt0460619
tt8526284
tt0108122
tt0119528
tt0115279
tt2861424
tt15488832
tt5786442
tt0085995
tt3147486
tt4677934
tt5298646
tt0162349
tt0072265
tt5805752
tt2267998
tt0112230
tt0094291
tt0099674
tt2275913
tt0144168
tt0142646
tt0295476
tt1334555
tt0113492
tt0096626
tt3675186
tt11808942
tt2998406
tt4366830
tt2826134
tt0117108
tt11137268
tt3449744
tt10050772
tt0818620
tt2096673
tt0445934
tt1448500
tt5924202
tt1711525
tt0105986
tt0468999
tt1176740
tt0382382
tt0117652
tt14016930
tt0213790
tt0118632
tt11475900
tt8579674
tt0100419
tt0277370
tt0362270
tt0048860
tt0106179
tt8819940
tt5719700
tt1100051
tt0139239
tt0405336
tt0108709
tt0089822
tt0189246
tt5880050
tt0368107
tt0789775
tt0116778
tt3218622
tt1655460
tt2790152
tt3800692
tt8962130
tt0122905
tt1565981
tt8106790
tt0279429
tt7713450
tt0338616
tt2004334
tt0116695
tt6137590
tt0180052
tt0079687
tt4853866
tt2953706
tt0154574
tt4670472
tt0094834
tt0080455
tt0396362
tt0123965
tt2114006
tt5903450
tt6463510
tt10974848
tt0122718
tt1442550
tt0128445
tt0245686
tt0401759
tt0098987
tt11348726
tt0120863
tt0332379
tt0122933
tt6528132
tt0120801
tt2417108
tt0321473
tt5941780
tt7286456
tt6168110
tt0217061
tt3172618
tt8280446
tt2953902
tt0115367
tt0081518
tt0068120
tt7396544
John Belushi
Dan Aykroyd
Bill Murray
Mike Myers
Chevy Chase
Billy Crystal
Joan Cusack
Robert Downey Jr.
Chris Farley
Janeane Garofalo
Julia Louis-Dreyfus
Eddie Murphy
Jim Belushi
Dana Carvey
Adam Sandler
Christopher Guest
Anthony Michael Hall
Jon Lovitz
Jay Mohr
Randy Quaid
Chris Rock
Rob Schneider
Martin Short
Ben Stiller
Damon Wayans
Christine Ebersole
Will Ferrell
Jason Lee
Nora Dunn
Tim Meadows
Kevin Nealon
David Spade
Fred Armisen
Peter Aykroyd
Morwenna Banks
Jim Breuer
A. Whitney Brown
Beth Cahill
Ellen Cleghorne
George Coe
Pete Davidson
Tom Davis
Denny Dillon
Brian Doyle-Murray
Rachel Dratch
Robin Duke
Dean Edwards
Chris Elliott
Jimmy Fallon
Tina Fey
Will Forte
Al Franken
Ana Gasteyer
Gilbert Gottfried
Mary Gross
Bill Hader
Brad Hall
Rich Hall
Darrell Hammond
Phil Hartman
Jan Hooks
Yvonne Hudson
Melissa Hurst
Victoria Jackson
Leslie Jones
Chris Kattan
Tim Kazurinsky
Laura Kightlinger
Taran Killam
David Koechner
Cory Kreutzer
Matthew Laurence
Gail Matthius
Michael McKean
Mark McKinney
Kate McKinnon
Laurie Metcalf
Dennis Miller
Jeff Richards
Tracy Morgan
Garrett Morris
Laraine Newman
Don Novello
Mike O'Brien
Michael O'Donoghue
Cheri Oteri
Chris Parnell
Jason Sudeikis
Amy Poehler
Emily Prager
Colin Quinn
Gilda Radner
Tony Rosato
Jeff Richards
Ann Risley
Tim Robinson
Harry Shearer
Charles Rocket
Maya Rudolph
Horatio Sanz
Tom Schiller
Paul Shaffer
Molly Shannon
Sarah Silverman
Robert Smigel
Pamela Stephenson
Jason Sudeikis
Julia Sweeney
Terry Sweeney
Kenan Thompson
Danitra Vance
Dan Vitale
Patrick Weathers
Cas Wilson
Fred Wolf
Alan Zweibel
Seth Meyers
Lauren Holt
Michaela Watkins
Heidi Gardner
Sarah Sherman
Bobby Moynihan
Melissa Villaseñor
Kyle Mooney
Rob Riggle
Finesse Mitchell
Mikey Day
Andy Samberg
Paul Brittain
Nasim Pedrad
Chris Kelly
Abby Elliott
John Milhiser
Alex Moffat
Aristotle Athari
Jenny Slate
Beck Bennett
Vanessa Bayer
Sasheer Zamata
Kate Willett
Luke Null
Melissa Villaseñor
Michael Che
Jay Pharoah
Jon Rudnitsky
James Austin Johnson
Chloe Fineman
Ego Nwodim
Brooks Wheelan
Cecily Strong
Aidy Bryant
Chris Redd
Jim Downey
Bowen Yang
Andrew Dismukes
Siobhan Fallon
Punkie Johnson
1.0
0.8
0.6
0.4
0.2
0.0

Scaling Down

- I now knew how to create a Boolean table and binary heatmaps from said table
- Now I needed to figure out how to create a table for each actor which included only their titles
- First step: creating a dictionary for an actor's titles and the Boolean values for each actor based on the titles

```
data = {}
for actor in names_title_dict:
    values = names_title_dict[actor]
    headers = values[:]
    headers.insert(0,actor)
    actor_dict_2d = [headers]

    for key in names_title_dict:
        t_or_f = [1 if value in names_title_dict[key] else 0 for value in values]
        t_or_f.insert(0,key)
        actor_dict_2d.append(t_or_f)
    data[actor] = actor_dict_2d
```

Turning the dict into a Table

```
def create_table(i):
    table = pd.DataFrame(data[wiki_cast.iloc[i].name]).set_index(0)
    header = table.iloc[0]
    table.columns = header
    table = pd.DataFrame(table[1:], columns=header)
    return table
table = create_table(1)
```

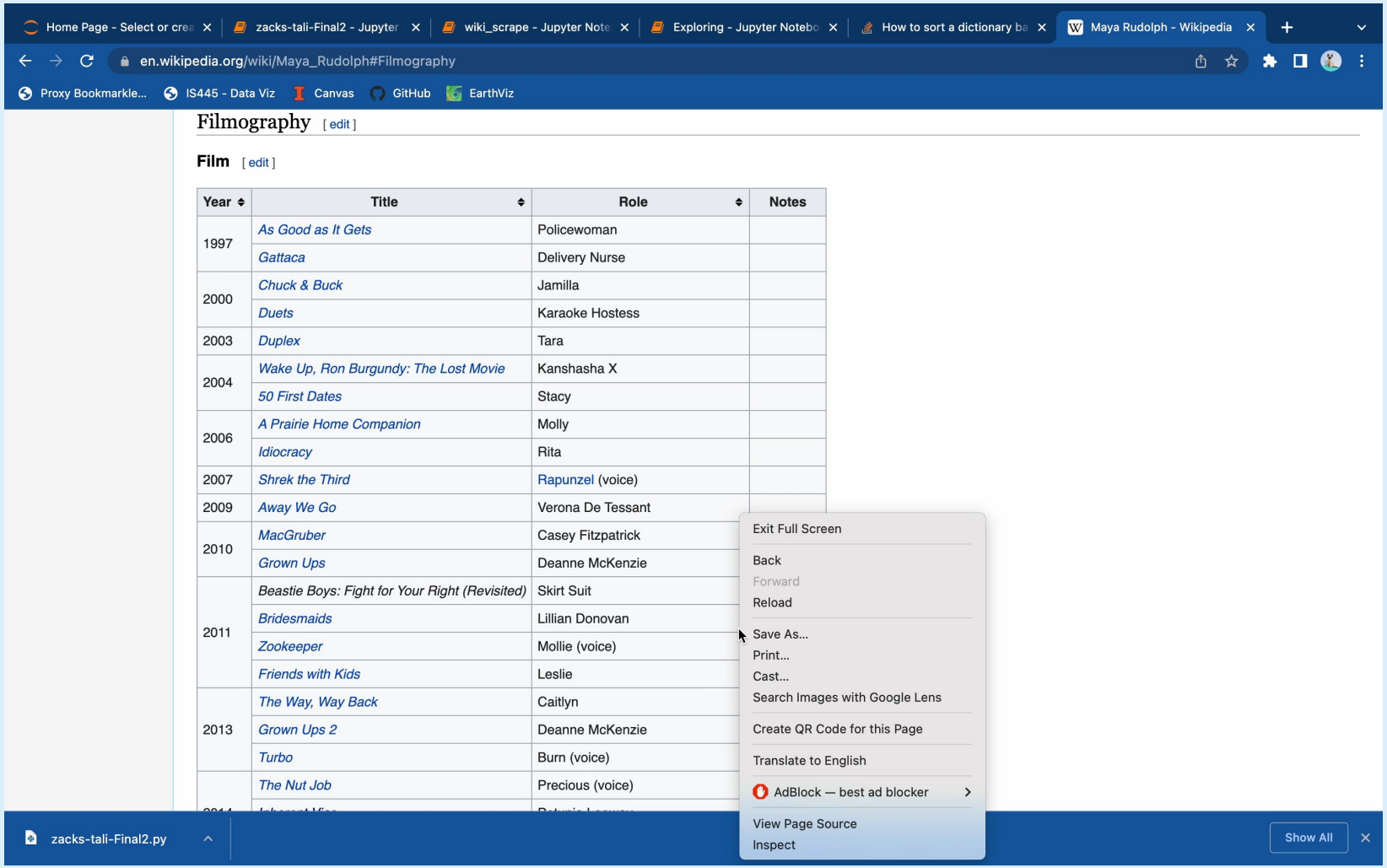
Prompt a Table

- I decided to create an interactive bar chart that would allow a user to select an actor's bar with a click
- Clicking on a bar prompts the creation of a table and, subsequently, a heat map (using bqplot)

Far from Complete Project Lists

- At this point, I have successfully created a bar chart which visualizes each actor's time at SNL in years and links to a heatmap, which compares the titles IMDb supplied as 'knownForTitles' for each actor
- This doesn't come nearly close enough to what I imagined
- I decided to scrape webpages for data for the first time… and then about 158 more times
    - I wanted to torture myself but also be certain that I was collecting the right data
- Actor Wikipedia pages often include a 'Filmography' section

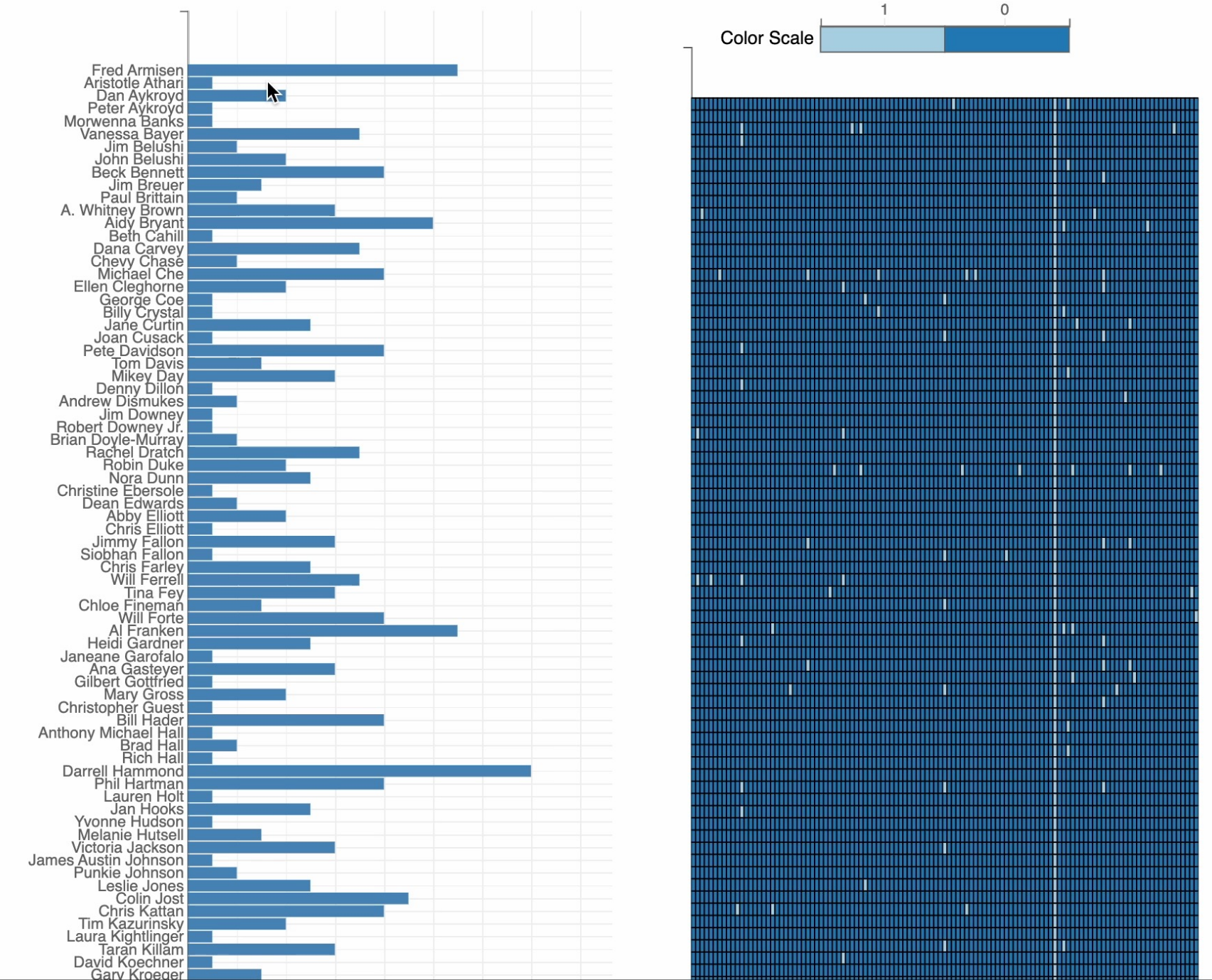
HTML 'Wiki Scraping'



SNL_Cast_Members_Full_Titles_dict

- The majority of cast member's Wikipedia pages had tables
    - Those that didn't required me to create the lists of projects by reading through their Wikipedia pages
- I compiled the contents of the film and tv lists for each actor and then appended it to SNL_Cast_Members_Full_Titles_dict as the value for the actor's name key
- Then I imported the wiki_scrape.py file into the notebook I was working in and redid ~the things~ I did to the original data in addition to:
    - Ensuring 'Saturday Night Live' was included in every value list
    - Reordering the dictionary to match the order of the bar chart

Adam Sandler: 5 seasons on SNL



Things I Learned

- Collecting and working with non-numerical data on a large scale from scratch (without a database to begin)
- Scraping the internet
- Setting up a workflow/schedule to hold myself accountable for due dates
- Importance of version control
    - It was nice to be able to go back in my file history when I thought I royally messed it up

What's Next?

- Other applications?
  - Compile a dictionary of any actor's projects… or really any person's projects… compile really any dictionary with value lists to see frequency of overlap in a heatmap
- I might add a histogram that displays the number of collaborations for each actor