Bulk Data Transfer

The SEC offers a bulk data download option for all of the data in both the Frame and Company Facts APIs. This data is updated nightly and can be accessed at http://www.sec.gov/Archives/edgar/daily-index/xbrl/companyfacts.zip (http://www.sec.gov/Archives/edgar/daily-index/xbrl/companyfacts.zip). While all available data is provided (~14.5k files), only those companies in the S&P 500 are needed for this project. Due to the number of files that need to be selected and copied, it's necessary to create a script to automatically select the correct files.

The `sp500_ciks.csv` data was sourced from the "S&P 500 component stocks" table on the List of S&P 500 companies (https://en.wikipedia.org/wiki/List_of_S%26P_500_companies) Wikipedia page.

In [1]:

```
import shutil
import pandas as pd
```

In [4]:

```
df = pd.read_csv('data/sp500_ciks.csv', dtype=str)
df.head()
```

Out[4]:

	Symbol	Security	SEC filings	GICS Sector	GICS Sub-Industry	Headquarters Location	Date first added	CIK	Founded
0	MMM	3M	reports	Industrials	Industrial Conglomerates	Saint Paul, Minnesota	1976-08-09	0000066740	1902
1	ABT	Abbott Laboratories	reports	Health Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	0000001800	1888
2	ABBV	AbbVie	reports	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	0001551152	2013 (1888)
3	ABMD	Abiomed	reports	Health Care	Health Care Equipment	Danvers, Massachusetts	2018-05-31	0000815094	1981
4	ACN	Accenture	reports	Information Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	0001467373	1989

In [12]:

```
src = 'E:/Downloads/sec_bulk_data'
dst = 'data/sec_bulk_data/'

for cik in df.CIK:
    try:
        shutil.copy(src + f'/CIK{cik}.json', dst)
    except FileNotFoundError as e:
        print('File not found for CIK:', cik, f'({df[df.CIK == cik].Security.values})')
```

```
File not found for CIK: 0001132979 (['First Republic Bank'])
```

I manually confirmed that the CIK for First Republic Bank is correct as shown above. However, the SEC's API shows that the data is missing when trying to access it through the website (https://data.sec.gov/api/xbrl/companyfacts/CIK0001132979.json (https://data.sec.gov/api/xbrl/companyfacts/CIK0001132979.json)). Therefore, it makes sense that the data would also be missing from the bulk data download. 499 / 500 companies were successfully found and copied.