Approach and Challenges

# Approach

## Google Search Query:

**Function:** google search (query, Num results=1000000)  
**Objective:** Perform a Google search with the specified query and fetch the HTML content of the search results page.  
**Details:** The function constructs a Google search URL using the query and the specified number of results. It then sends an HTTP GET request to fetch the HTML content.

## Extract URLs:

**Function:** extract\_urls(html)  
**Objective:** Parse the HTML content to extract URLs containing 'arcgis/rest/services'.  
**Details:** The function uses BeautifulSoup to parse the HTML and find all anchor (<a>) tags. It filters and extracts URLs that contain 'arcgis/rest/services'.

## Validate URLs:

**Function:** is\_valid\_url(url)  
**Objective:** Ensure extracted URLs are valid and contain the specified path.  
**Details:** The function checks if the URL starts with 'http' and contains 'arcgis/rest/services'.

## Save to CSV:

**Function:** save\_to\_csv (urls, filename='urls.csv')  
**Objective:** Save the cleaned and valid URLs to a CSV file.  
**Details:** The function creates a DataFrame from the extracted URLs, cleans them, filters out invalid URLs, and saves the result to a CSV file. It also prints a confirmation message when the file is saved successfully.

## Main Workflow:

1. Execute the Google search to get the HTML content.  
2. Extract the URLs from the HTML content.  
3. Save the cleaned and valid URLs to a CSV file and print a success message.

# Challenges Encountered

## Google Search Result Parsing:

**Challenge:** Google search results can be complex and might change frequently. Extracting URLs accurately required handling various formats and structures within the HTML.  
**Solution:** Used BeautifulSoup to parse the HTML and extract links, focusing on handling potential variations in the structure.

## Filtering and Cleaning URLs:

**Challenge:** Ensuring that only valid URLs containing 'arcgis/rest/services' are extracted and saved. Some URLs might have extra parameters or be malformed.  
**Solution:** Implemented a validation function to filter out invalid URLs and clean the extracted URLs before saving.

## Saving and Confirming File Creation:

**Challenge:** Confirming that the CSV file was saved successfully.  
**Solution:** Added a print statement to provide feedback on the success of the file-saving operation, ensuring users are informed.