Final Presentation

Objective

- Developing software requirements for examining funding activities, professional organizations, and publications to help identify key Ph.D. dissertation topics and postdoctoral research projects.
- Developing software to support each step of the process described above.

Past Interns

- Created a manual report
- Predicting new research results that will emerge in a field over the next five years
  - Topics: AI, Data Analytics, Cybersecurity, Quantum, and Modeling & Simulation
- Included organizations, journal names, funding sources, University and Labs, Researchers, etc

NSF Database

- Scrape data from the NSF Website and turn it into a data-frame.
- The first topic is AI.

In [120]:
```
nsf_awards_json_api = 'http://api.nsf.gov/services/v1/awards.json'
keywords = 'keyword=ai'

# Make the URL request
res = requests.get(nsf_awards_json_api + '?' + keywords)

# Get the awards data
json_response = res.json()["response"]
print(f'type(awards): {type(json_response)}')
awards = json_response["award"]

#print(awards)
```

```
type(awards): <class 'dict'>
```

In [85]:
```
#print(awards)
rdf = pd.json_normalize(awards)
rdf
```

:

	agency	awardeeCity	awardeeName	awardeeStateCode	fundsObligatedAmt	id	piFirstName	piLastName	publicAccessMandate
0	NSF	BALTIMORE	University of Maryland Baltimore County	MD	160000	2309760	Houbing	Song	1
1	NSF	ATLANTA	Georgia Tech Research Corporation	GA	7059902	2247790	Ashok	Goel	1
2	NSF	ATLANTA	Georgia Tech Research Corporation	GA	800000	2205152	Ali	Adibi	1
3	NSF	CORONA	New York Hall of Science	NY	74971	2230850	Dorothy	Bennett	1

In [86]:
```
rdf.shape
```

Final DataFrame

- Add additional columns for convenience
  - Date2 and NAME

ity	awardeeName	awardeeStateCode	fundsObligatedAmt	id	piFirstName	piLastName	publicAccessMandate	date	title	Date2	NAME
RE	University of Maryland Baltimore County	MD	160000	2309760	Houbing	Song	1	12/13/2022	Collaborative Research: CyberTraining: Pilot: ...	2022-12-13	Houbing Song
TA	Georgia Tech Research Corporation	GA	7059902	2247790	Ashok	Goel	1	10/27/2022	AI Institute for Adult Learning and Online Edu...	2022-10-27	Ashok Goel
TA	Georgia Tech Research Corporation	GA	800000	2205152	Ali	Adibi	1	09/13/2022	SCH: Intelligent Radiology Through Human-Machi...	2022-09-13	Ali Adibi
NA	New York Hall of Science	NY	74971	2230850	Dorothy	Bennett	1	08/31/2022	Conference on Human-Centered Approaches to Art...	2022-08-31	Dorothy Bennett
NE	University of Oregon Eugene	OR	299920	2225949	Lei	Jiao	1	08/30/2022	Collaborative Research: CNS Core: Small: Edge ...	2022-08-30	Lei Jiao

Funding overtime

- Aggregate the data to find the funding overtime.
- For example, 2022-08-18 had the least and 2022-10-27 had the most funding.

9	2022-05-20	1	10000.0
1	2022-02-08	1	19996.0
0	2022-01-28	1	20000.0
2	2022-02-16	1	20000.0
3	2022-02-28	1	42756.0
7	2022-04-07	1	74790.0
17	2022-08-31	1	74971.0
6	2022-03-17	1	105094.0
14	2022-08-26	1	150000.0
20	2022-12-13	1	160000.0
5	2022-03-15	1	175000.0
10	2022-05-26	4	263253.0
11	2022-07-07	1	299375.0
16	2022-08-30	2	299920.0
8	2022-05-03	1	397055.0
15	2022-08-29	1	600000.0
13	2022-08-22	1	790631.0

Out[92]:

	Date	Count	Total Funds Obligated Amount
0	2022-01-28	1	20000.0
1	2022-02-08	1	19996.0
2	2022-02-16	1	20000.0
3	2022-02-28	1	42756.0
4	2022-03-02	1	1739367.0
5	2022-03-15	1	175000.0
6	2022-03-17	1	105094.0
7	2022-04-07	1	74790.0
8	2022-05-03	1	397055.0
9	2022-05-20	1	10000.0
10	2022-05-26	4	263253.0
11	2022-07-07	1	299375.0
12	2022-08-18	1	0.0
13	2022-08-22	1	790631.0
14	2022-08-26	1	150000.0
15	2022-08-29	1	600000.0
16	2022-08-30	2	299920.0
17	2022-08-31	1	74971.0
18	2022-09-13	1	800000.0
19	2022-10-27	1	7059902.0
20	2022-12-13	1	160000.0

Funding based on institution

- Aggregate the data to find the amount of funding based on institutions.

Out[97]:

	Awardee Name	Count	Total Funds Obligated Amount
6	Embry-Riddle Aeronautical University	1	0.0
16	University of Chicago	1	10000.0
1	Baylor University	1	19996.0
18	University of Miami School of Medicine	1	20000.0
13	Rutgers University New Brunswick	1	20000.0
2	Cal Poly Pomona Foundation, Inc.	1	72610.0
11	Pennsylvania State Univ University Park	2	74790.0
9	New York Hall of Science	1	74971.0
3	California State University, Trustees	1	96812.0
15	University Enterprises Corporation at CSUSB	1	97668.0
0	Auburn University	1	105094.0
19	University of Notre Dame	1	150000.0
17	University of Maryland Baltimore County	1	160000.0
21	University of Southern California	1	175000.0
14	San Jose State University Foundation	1	263253.0
10	Northeastern University	1	299375.0
20	University of Oregon Eugene	1	299920.0
12	Purdue University	1	300000.0
7	Fayetteville State University	1	397055.0
5	Clemson University	1	600000.0
4	Carnegie-Mellon University	1	790631.0
22	University of Virginia Main Campus	1	1739367.0
8	Georgia Tech Research Corporation	2	7059902.0

Funding based on Professor

- Aggregate the data to find the amount of funding based on the faculty members.

	Name	Total Funds Obligated Amount
0	Ali Adibi	800000
1	Anh Nguyen	105094
2	Ashok Goel	7059902
3	Christopher Dancy	7479042756
4	Daniel Diaz	20000
5	Dorothy Bennett	74971
6	Filip Ilievski	175000
7	Frank Gomez	96812
8	Gabriel Granco	72610
9	Houbing Song	1600000
10	Jorge Ortiz	20000
11	Kelly Caine	600000
12	Lei Jiao	299920
13	Pablo Rivas	19996
14	Rebecca Willett	10000
15	Ronald Sandler	299375
16	Sambit Bhattacharya	397055
17	Stephen Baek	1739367
18	Tianqi Chen	790631
19	Xiaojun Lin	300000
20	Yongsuk Lee	150000
21	Yu Chen	263253
22	Yunfei Hou	97668

Topic Transformers

- Transformers is a subtopic of AI. The same steps were followed to make its data-frame.

:

	agency	awardeeCity	awardeeName	awardeeStateCode	fundsObligatedAmt	id	piFirstName	piLastName	publicAccessMandate	date	
0	NSF	KALAMAZOO	Western Michigan University	MI	199105	2138408	Pablo	Gomez	1	02/22/2022	Mul cosi appr
1	NSF	PRINCETON	Princeton University	NJ	194635	2203399	Niraj	Jha	1	01/28/2022	C Trai s
2	NSF	STATE COLLEGE	Solid State Ceramics, Inc.	PA	764749	1632476	Safakcan	Tuncdemir	1	09/22/2016	STTR Tem Cofirec
3	NSF	GAINESVILLE	University of Florida	FL	200000	1611048	Shuo	Wang	0	08/17/2016	High Fr Trai Windir
4	NSF	LINCOLN	University of Nebraska-Lincoln	NE	500000	1554497	Liyan	Qu	0	01/27/2016	C Ad Volta Magnet

The problem

- There are no overlap of projects between transformers and AI because the keywords “AI” and “Transformers” from the NSF database are extracted from the title not abstract.

```
common=df.merge(df_tran, how = 'inner' ,indicator=False)
common
```

agency	awardeeCity	awardeeName	awardeeStateCode	fundsObligatedAmt	id	piFirstName	piLastName	publicAccessMandate	date	title	Date2	NAME

```
#as shown above, 0 matches between AI and transformers, but transformers is a nested topic of AI
```

Keyword from Abstract

- When tried to extract keywords from the abstract, the following error were thrown.

```
Traceback (most recent call last):
  File "c:\work\git\tech_scout\tech_scout.py", line 33, in <module>
    if award['abstractText'] != None:
KeyError: 'abstractText'
```

Automated

- The keywords were used in a for loop to create an automated data-frame with the topics AI, cybersecurity, quantum, data analytics, and modeling & simulation.

_areas

leeStateCode	fundsObligatedAmt	id	piFirstName	piLastName	publicAccessMandate	date	title	Area	Topic
MD	160000	2309760	Houbing	Song	1	12/13/2022	Collaborative Research: CyberTraining: Pilot: ...	keyword=ai	ai
GA	7059902	2247790	Ashok	Goel	1	10/27/2022	AI Institute for Adult Learning and Online Edu...	keyword=ai	ai
GA	800000	2205152	Ali	Adibi	1	09/13/2022	SCH: Intelligent Radiology Through Human-Machi...	keyword=ai	ai
NY	74971	2230850	Dorothy	Bennett	1	08/31/2022	Conference on Human-Centered Approaches to Art...	keyword=ai	ai
OR	299920	2225949	Lei	Jiao	1	08/30/2022	Collaborative Research: CNS Core: Small: Edge ...	keyword=ai	ai
...	...	...	...	...	...	...	...	...	...
MD	300000	1708602	Mauro	Maggioni	0	04/28/2017	Statistical Learning for High-Dimensional Stoc...	keyword=modeling and simulation	modeling and simulation
VA	5000	1658908	John	Shortle	1	03/08/2017	Workshop: Towards an Ecosystem of Simulation M...	keyword=modeling and simulation	modeling and simulation
PA	346157	1727508	Wonpil	Im	0	01/20/2017	Bacterial Outer Membranes and	keyword=modeling and simulation	modeling and

PhD Students

- To achieve the Ph.D. students of the professors,
  - Use the google API and search engine

```python
import requests

# get the API KEY here: https://developers.google.com/custom-search/v1/overview
API_KEY = ""

# get your Search Engine ID on your CSE control panel
SEARCH_ENGINE_ID = ""

# the search query you want
query = "sean warnick byu"
# using the first page
page = 1
# constructing the URL
# doc: https://developers.google.com/custom-search/v1/using_rest
# calculating start, (page=2) => (start=11), (page=3) => (start=21)
start = (page - 1) * 10 + 1
url = f"https://www.googleapis.com/customsearch/v1?key={API_KEY}&cx={SEARCH_ENGINE_ID}&q={query}&start={start}"

# make the API request
data = requests.get(url, verify=False).json()

# get the result items
search_items = data.get("items")

# iterate over 10 results found
for i, search_item in enumerate(search_items, start=1):
    print('-----------')
    # Use the following to see other available keys
    #print(f'{search_item}, type: {type(search_item)}')
    print(f'Title: {search_item["title"]}')
    print(f'Link: {search_item["link"]}')
    print(f'Snippet: {search_item["snippet"]}')
    print(f'Pagemap: {search_item["pagemap"]}')
```

```
c:\work\git\tech_scout\.venv\lib\site-packages\urllib3\connectionpool.py:1045: InsecureRequestWarning: Unverified HTTPS request is being made to host '10.76.225.15'. Adding certificate verification is strongly advised. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warnings
  warnings.warn(
-----------
Title: Sean Warnick
Link: https://science.byu.edu/directory/sean-warnick
Snippet: Sean Warnick received the B.S.E. degree from Arizona State University in 1993, and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology ...
Pagemap: {'cse_thumbnail': [{'src': 'https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcQjEjCvkCgdp2lIDlTftTAuxJxr8FHlKejiFgSu9dPLblT9t5SvYEFP_IE', 'width': '225', 'height': '225'}], 'metatags': [{'og:image': 'https://brightspotcdn.byu.edu/dims4/default/c6657d4/2147483647/strip/true/crop/300x300+0+0/resize/1200x1200!/quality/90/?url=https%3A%2F%2Fbrigham-young-brightspot.s3.amazonaws.com%2F50%2Faf%2F83b0cc02535c86cb0356b105b3fb%2Fpersonphoto-cgi-1.jpeg', 'og:image:width': '1200', 'og:type': 'profile', 'twitter:card': 'summary_large_image', 'twitter:title': 'Sean Warnick', 'og:site_name': 'The College of Physical and Mathematical Sciences (CPMS)', 'og:image:url': 'https://brightspotcdn.byu.edu/dims4/default/c6657d4/2147483647/strip/true/crop/300x300+0+0/resize/1200x1200!/quality/90/?url=https%3A%2F%2Fbrigham-young-brightspot.s3.amazonaws.com%2F50%2Faf%2F83b0cc02535c86cb0356b105b3fb%2Fpersonphoto-cgi-1.jpeg', 'profile:username': 'Sean Warnick', 'og:ti
```

My Modification

- Go through each professor in the column
- Search up their name along with the institution and "Research group"
  - Georgia Tech Research Corporation Ali Adibi Research Group
- Store the result (links from the search) in a separate column
- Go through each link via a for loop and turn website text through spacy.

```
link_list = []
for x in df['Organization & Professor']:
    query = x + " Research Group"
    page = 1
    start = (page - 1) * 10 + 1
    url = f"https://www.googleapis.com/customsearch/v1?key={API_KEY}&cx={SEARCH_ENGINE_ID}&q={query}&start={start}"
    data = requests.get(url, verify=False).json()
    search_items = data.get("items")
    for i, search_item in enumerate(search_items, start=1):
        print(search_item["link"])
        link_list.append(search_item["link"])
```

The problem

- The google API and personal search engine is not always reliable.
- Could not finish the code.

```
link_list = []
for x in df['Organization & Professor']:
    query = x + " Research Group"
    page = 1
    start = (page - 1) * 10 + 1
    url = f"https://www.googleapis.com/customsearch/v1?key={API_KEY}&cx={SEARCH_ENGINE_ID}&q={query}&sta
    data = requests.get(url, verify=False).json()
    search_items = data.get("items")
    for i, search_item in enumerate(search_items, start=1):
        print(search_item["link"])
        link_list.append(search_item["link"])
```

```
/Users/far/Desktop/Personal Project/tech_scout2/tech_scout/env/lib/python3.10/site-packages/urllib3/connectionpool.
py:1045: InsecureRequestWarning: Unverified HTTPS request is being made to host 'www.googleapis.com'. Adding certif
icate verification is strongly advised. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warni
ngs
  warnings.warn(
```

```
TypeError                                 Traceback (most recent call last)
Cell In[118], line 9
      7 data = requests.get(url, verify=False).json()
      8 search_items = data.get("items")
----> 9 for i, search_item in enumerate(search_items, start=1):
     10     print(search_item["link"])
     11     #link_list.append(search_item["link"])

TypeError: 'NoneType' object is not iterable
```

```
query = "sean warnick byu"
# using the first page
page = 1
# constructing the URL
# doc: https://developers.google.com/custom-search/v1/using_rest
# calculating start, (page=2) => (start=11), (page=3) => (start=21)
start = (page - 1) * 10 + 1
url = f"https://www.googleapis.com/customsearch/v1?key={API_KEY}&cx={SEARCH_ENGINE_ID}&q={query}&start={start}"

# make the API request
data = requests.get(url, verify=False).json()

# get the result items
search_items = data.get("items")

# iterate over 10 results found
for i, search_item in enumerate(search_items, start=1):
    print('------------')
    # Use the following to see other available keys
    #print(f'{search_item}, type: {type(search_item)}')
    print(f'Title: {search_item["title"]}')
    print(f'Link: {search_item["link"]}')
    print(f'Snippet: {search_item["snippet"]}')
    print(f'Pagemap: {search_item["pagemap"]}')
```

```
/Users/far/Desktop/Personal Project/tech_scout2/tech_scout/env/lib/python3.10/site-packages/urllib3/connectionpool.
py:1045: InsecureRequestWarning: Unverified HTTPS request is being made to host 'www.googleapis.com'. Adding certif
icate verification is strongly advised. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warni
ngs
  warnings.warn(
```

```
TypeError                                 Traceback (most recent call last)
Cell In[7], line 18
     15 search_items = data.get("items")
     17 # iterate over 10 results found
---> 18 for i, search_item in enumerate(search_items, start=1):
     19     print('------------')
     20     # Use the following to see other available keys
     21     #print(f'{search_item}, type: {type(search_item)}')

TypeError: 'NoneType' object is not iterable
```

NIH Database

- The NIH database provides a lot more information.
- Some of the researchers in the columns (without titles) can be Ph.D. students.
  - Not all!

```
'''
{ "criteria": { advanced_text_search: { operator: "and", search_field: "projecttitle,terms",
"search_text": "brain disorder" } } }
'''


#params = { 'criteria': { 'advanced_text_search': { 'search_field': fields, 'search_text': searchstr } } }
params = { "criteria": { 'advanced_text_search': { 'operator': "and", 'search_field': "abstracts,projecttitle,terms"
"search_text": "data analytics" } } }

response = requests.post("https://api.reporter.nih.gov/v2/projects/search", json=params)

print(response.status_code)
print(response.text)
```

	profile_id	first_name	middle_name	last_name	is_contact_pi	full_name	title	project_title	award_amount	is_active	Organization
0	9841070	Gustavo		Glusman	True	Gustavo Glusman	PRINCIPAL SCIENTIST	DOCKET: accelerating knowledge extraction from...	609144	True	INSTITUTE FOR SYSTEMS BIOLOGY
1	15443354	Jennifer		Hadlock	False	Jennifer Hadlock	PRINCIPAL MEDICAL INFORMATICIST	DOCKET: accelerating knowledge extraction from...	609144	True	INSTITUTE FOR SYSTEMS BIOLOGY
2	8300690	ILYA		SHMULEVICH	False	ILYA SHMULEVICH	PROFESSOR	DOCKET: accelerating knowledge extraction from...	609144	True	INSTITUTE FOR SYSTEMS BIOLOGY
3	2089750	SERGIO	E	BARANZINI	False	SERGIO E BARANZINI		EVIDARA: Automated Evidential Support from Raw...	532891	True	INSTITUTE FOR SYSTEMS BIOLOGY
4	9454292	Sui		Huang	True	Sui Huang	PROFESSOR	EVIDARA: Automated Evidential Support from Raw...	532891	True	INSTITUTE FOR SYSTEMS BIOLOGY
5	10790241	Jonathan		Elmer	False	Jonathan Elmer		PREcision Care In Cardiac ArrEst - ICECAP (PRE...	1113712	True	STANFORD UNIVERSITY
6	6270502	KAREN	G	HIRSCH	True	KAREN G HIRSCH	ASSISTANT PROFESSOR	PREcision Care In Cardiac ArrEst - ICECAP (PRE...	1113712	True	STANFORD UNIVERSITY
7	14315038	Sung Eun	EUN	Choi	True	Sung Eun EUN Choi	INSTRUCTOR	Reducing oral health disparities in children u...	114113	True	HARVARD MEDICAL SCHOOL

Future Idea if The Google Api Works

- Run the website text of the links collected from the search through spacy.

```
In [1]: import spacy

        nlp = spacy.load("en_core_web_sm")

In [2]: text = """Ms. Farhana Uddin, an intern with the Department of Homeland Security, gave Steve great news
                  about the code she developed. She plans to share her developments with others at
                  the University of Maryland via Twitter."""

In [4]: d = nlp(text)
        print(d.ents)
        if d.ents:
            for entity in d.ents:
                print(f"Entity '{entity}' is of type '{entity.label_}'")

        (Farhana Uddin, the Department of Homeland Security, Steve, the University of Maryland, Twitter)
        Entity 'Farhana Uddin' is of type 'PERSON'
        Entity 'the Department of Homeland Security' is of type 'ORG'
        Entity 'Steve' is of type 'PERSON'
        Entity 'the University of Maryland' is of type 'ORG'
        Entity 'Twitter' is of type 'PRODUCT'
```

Manual Website Breakdown (not ideal)

- Manual breakdown of each website is possible (through the html tags)
- Not ideal
  - Every website is different
    - Automation not possible
  - Websites update
    - Previous code may not apply

Out[86]:

	Michael DeBuse	Neal Munson	Logan Nielsen	Alyssa Crezee	Caelen Miller	Kimley Morlant	Fritz-Carl Morlant	Joseph Mattson	Sean Gallacher	Koby Lewis	Tanner Day	Maxwell Hamilton	Michael King	Abram Aanderud
0	PhD	MS	MS	Research Assistant, Entry	Research Assistant, Entry	Research Assistant, Entry	Research Assistant, Entry	Research Assistant, Entry	Research Assistant, Entry	Research Assistant, Advanced	Research Assistant, Advanced	Research Assistant, Mid-level	Research Assistant, Leadership	Research Assistant, Entry

```
In [67]: URL = 'https://cs.byu.edu/department/directory/faculty-directory/sean-warnick/'
         page = requests.get(URL)

In [68]: page.status_code

Out[68]: 200

In [69]: page.headers

Out[69]: {'Date': 'Thu, 15 Dec 2022 08:00:52 GMT', 'Server': 'Apache/2.4.7 (Ubuntu)', 'X-Frame-Options': 'SAMEORIGIN', 'Vary
         ': 'Cookie,Accept-Encoding', 'Content-Encoding': 'gzip', 'Keep-Alive': 'timeout=5, max=200', 'Connection': 'Keep-Al
         ive', 'Transfer-Encoding': 'chunked', 'Content-Type': 'text/html; charset=utf-8'}

In [70]: page.text

Out[70]: ' \n<html lang="en">\n  <head>\n    <title>BYU Computer Science Department</title>\n    <meta charset="utf-8" />\n
         <meta http-equiv="x-ua-compatible" content="ie=edge" />\n    <meta name="viewport" content="width=device-width, ini
         tial-scale=1.0" />\n    <link rel="shortcut icon" href="/static/core/img/favicon.ico" />\n\n    <!-- <script async
         src="https://cdn.byu.edu/byu-theme-components/1.x.x/byu-theme-components.min.js"></script> -->\n    <!-- <link rel=
         "stylesheet" href="https://cdn.byu.edu/byu-theme-components/1.x.x/byu-theme-components.min.css" /> -->\n    <script
         \n      async\n      src="https://cdn.byu.edu/byu-theme-components/2.x.x/byu-theme-components.min.js"\n    ></scrip
         t>\n    <link\n      rel="stylesheet"\n      href="https://cdn.byu.edu/byu-theme-components/2.x.x/byu-theme-compone
         nts.min.css"\n      media="all"\n    />\n    <link\n      rel="stylesheet"\n      href="/static/core/css/main.css"\
         n      type="text/css"\n    />\n  </head>\n\n  <body class="body--flex">\n    <header>\n  \n\n\n\n<div class="navig
         ation">\n  \n\n<div id="wide-nav">\n  \n  <div id="wide-header">\n    <div id="wide-header-container">\n      \n\n<
         div class="site-title">\n    <a href="https://www.byu.edu/">\n        <img class="byu-logo-image" alt="BYU" src="/s
         tatic/core/img/byu-logo.svg">\n    </a>\n    <a class="cs-logo" href="/">\n        Computer Science\n    </a>\n</di
         v>\n      \n\n<div class="actions">\n    \n    <a class="header-login" href="/accounts/login/">Log In</a>\n    \n
         \n<a slot="actions" class="search-link" href="https://www.byu.edu/search-all">\n    <img class="search-icon" src="/
         static/core/img/search.svg" alt="Search">\n</a>\n<!-- <byu-search slot="search" action="navigate" action-target="ht
         tps://byu.edu/search/byu?keys=${search}" method="get"\n    placeholder="Search"></byu-search> -->\n</div>\n    </di
         v>\n  </div>\n  \n  <div class="nav-page-container">\n    <div class="nav-page-content">\n    \n      \n        <di
         v class="nav-dropdown">\n          \n          <button onfocus="onWideNavFocus(this)" class="nav-dropdown-tab">\n
         <text>Student Education</text>\n          </button>\n          \n          <div class="nav-dropdown-content" style=
```

Additional databases

- AFRL
  - Could not find an API
  - Research Areas: https://www.afrl.af.mil/About-Us/Fact-Sheets/Fact-Sheet-Display/Article/2282138/afosr-research-areas/
- ERDC
  - An API found but could not access it from my connection
  - Special authorization is required
  - https://www.erdc.usace.army.mil/Library/Electronic-Resources/
- ONR
  - Could not find an API
  - Research Areas: https://www.nrl.navy.mil/Our-Work/Areas-of-Research/
- DARPA
  - Could not find an API
  - Research Areas: https://www.darpa.mil/program/our-research/more
- IARPA
  - Could not find an API
  - Research Areas: https://www.iarpa.gov/research-programs

Additional databases

- DOE
  - API:
    - https://www.energy.gov/eere/buildings/application-programming-interface
    - Lacks relevant information
      - Project name, investigators, date, and funding
  - Research Areas: https://www.energy.gov/eere/education/research-topics
- **Department of Agriculture**
  - API:
    - https://pubag.nal.usda.gov/apidocs/
    - One of the most useful APIs!
    - Need a key to access it

USDA

- Example of the search results for quantum
- Just lacks the funding

Out[9]:

	date	nalt_all	subject	author_lastname	last_modified_date	language	title	startpage	usda_authored_publication	id
0	0000	[absorption, acetonitrile, adsorption, catalys...	[absorption, acetonitrile, adsorption, catalys...	[Wen, Yang, Zong, Ma, Wang, Lj]	2016-01-24T03:25:50.235Z	[English]	Photocatalytic H_2 production on hybrid catalys...	318	False	1002492
1	0000	[barium, calcium, carbon dioxide, carbonates, ...	[barium, calcium, carbon dioxide, carbonates, ...	[Larin, Rybakov, Zhidomirov, Mace, Laaksonen, ...	2016-01-23T03:54:52.463Z	[English]	Oxide clusters as source of the third oxygen a...	212	False	1002498
2	0000	[activation energy, active sites, aluminum, ca...	[activation energy, active sites, aluminum, ca...	[Mlinar, Zimmerman, Celik, Head-Gordon, Bell]	2016-01-23T03:53:37.033Z	[English]	Effects of Brønsted-acid site proximity on the...	65	False	1002837
3	0000	[electrons, hydrogen production, metals, oxida...	[electrons, hydrogen production, metals, oxida...	[Yang, Yan, Wang, Wen, Wang, Fan, Shi, Lj]	2016-01-23T03:53:04.276Z	[English]	Roles of cocatalysts in Pt–PdS/CdS with except...	151	False	1002901
4	0000	[alcohols, aluminum oxide, catalysts, ethers, ...	[alcohols, aluminum oxide, catalysts, ethers, ...	[Shou, Li, Ferrari, Sholl, Davis]	2016-01-23T03:51:06.602Z	[English]	Use of infrared spectroscopy and density funct...	150	False	1003123
5	0000	[autoxidation, electron transfer, irradiation,...	[autoxidation, electron transfer, irradiation,...	[Ronzani, Costanramone, Blanc, Benabbou, Beche...	2016-01-23T03:49:56.780Z	[English]	Visible-light photosensitized oxidation of α-t...	164	False	1003191
6	0000	[carbon, encapsulation, hydrogen production, i...	[carbon, encapsulation, hydrogen production, i...	[Peng, Zhang, Zeng, Li, Zhang, Lj]	2016-01-24T03:25:28.390Z	[English]	Carbon encapsulation strategy of Ni co-catalys...	156	False	1003207
7	0000	[catalysts, ligands, oxidation, photosensitizi...	[catalysts, ligands, oxidation, photosensitizi...	[Wang, Duan, Tong, Sun]	2016-01-23T03:50:22.732Z	[English]	Visible light-driven water oxidation catalyzed...	129	False	1003267
8	0000	[bioavailability, curcumin, formic acid, intra...	[bioavailability, curcumin, formic acid, intra...	[Li, Qiao, Li, He, Ye, Xiang, Lin, Guo]	2016-03-07T01:08:28.777Z	[English]	Metabolic and pharmacokinetic studies of curcu...	2751	False	1003712
9	0000	[acetonitrile, ammonium acetate, drugs, electr...	[acetonitrile, ammonium acetate, drugs, electr...	[ter Weijden, van den Broek, Ververs]	2016-03-05T01:53:56.349Z	[English]	Easy and fast LC–MS/MS determination of lidoca...	111	False	1003926
10	0000	[cadmium, computer software, databases, diesel...	[cadmium, computer software, databases, diesel...	[Şengül, Theis]	2015-04-15T02:34:42.132Z	[English]	An environmental impact assessment of quantum ...	21	False	1004747
11	0000	[annealing, ethanol, fluorescence, nanocomposi...	[annealing, ethanol, fluorescence, nanocomposi...	[Panigrahi, Bera, Basak]	2015-02-10T04:20:34.276Z	[English]	Ordered dispersion of ZnO quantum dots in SiO_2...	30	False	1006419
12	0000	[antibodies, biomedical materials, chemical co...	[antibodies, biomedical materials, chemical co...	[Zhang, Chen, Wang, Guo, Li, Shi]	2015-07-20T05:02:30.811Z	[English]	Preparation of highly fluorescent magnetic nan...	426	False	1006442

Out[9]:

_url	url	volume	publication_year	usda_funded_publication	issn	page	author_primary	doi
ad/?rm=	http://dx.doi.org/10.1016/j.jcat.2011.05.015	v. 281	2011	False	0021-9517	pp. 318-324	Fuyu Wen	10.1016/j.jcat.2011.05.015
ad/?rm=	http://dx.doi.org/10.1016/j.jcat.2011.05.002	v. 281	2011	False	0021-9517	pp. 212-221	Alexander V. Larin	10.1016/j.jcat.2011.05.002
ad/?rm=	http://dx.doi.org/10.1016/j.jcat.2012.01.002	v. 288	2012	False	0021-9517	pp. 65-73	Anton N. Mlinar	10.1016/j.jcat.2012.01.002
ad/?rm=	http://dx.doi.org/10.1016/j.jcat.2012.03.008	v. 290	2012	False	0021-9517	pp. 151-157	Jinhui Yang	10.1016/j.jcat.2012.03.008
ad/?rm=	http://dx.doi.org/10.1016/j.jcat.2012.12.011	v. 299	2013	False	0021-9517	pp. 150-161	Heng Shou	10.1016/j.jcat.2012.12.011
ad/?rm=	http://dx.doi.org/10.1016/j.jcat.2013.04.001	v. 303	2013	False	0021-9517	pp. 164-174	Filippo Ronzani	10.1016/j.jcat.2013.04.001
ad/?rm=	http://dx.doi.org/10.1016/j.jcat.2013.03.026	v. 303	2013	False	0021-9517	pp. 156-163	Tianyou Peng	10.1016/j.jcat.2013.03.026
ad/?rm=	http://dx.doi.org/10.1016/j.jcat.2013.06.023	v. 306	2013	False	0021-9517	pp. 129-132	Lei Wang	10.1016/j.jcat.2013.06.023
ad/?rm=	http://dx.doi.org/10.1016/j.jchromb.2011.07.042	v. 879	2011	False	1570-0232	pp. 2751-2758	Rui Li	10.1016/j.jchromb.2011.07.042
ad/?rm=	http://dx.doi.org/10.1016/j.jchromb.2011.11.030	v. 881-882	2012	False	1570-0232	pp. 111-114	E. ter Weijden	10.1016/j.jchromb.2011.11.030
ad/?rm=	http://dx.doi.org/10.1016/j.jclepro.2010.08.010	v. 19	2011	False	0959-6526	pp. 21-31	Hatice Şengül	10.1016/j.jclepro.2010.08.010
ad/?rm=	http://dx.doi.org/10.1016/j.jcis.2010.09.055	v. 353	2011	False	0021-9797	pp. 30-38	Shrabani Panigrahi	10.1016/j.jcis.2010.09.055
ad/?87...	http://dx.doi.org/10.1016/j.jcis.2010.09.084	v. 353	2011	False	0021-9797	pp. 426-432	Bingbo Zhang	10.1016/j.jcis.2010.09.084

Combined Topics

- Automated process
- Better for subtopics because the keywords (area) is being extracted from the abstract
- Can aggregate the data to show summaries

ear	usda_funded_publication	issn	page	author_primary	doi	pmcid_url	chorus_url	Area
011	False	0021-9517	pp. 318-324	Fuyu Wen	10.1016/j.jcat.2011.05.015	NaN	NaN	keyword=ai
011	False	0021-9517	pp. 212-221	Alexander V. Larin	10.1016/j.jcat.2011.05.002	NaN	NaN	keyword=ai
012	False	0021-9517	pp. 65-73	Anton N. Mlinar	10.1016/j.jcat.2012.01.002	NaN	NaN	keyword=ai
012	False	0021-9517	pp. 151-157	Jinhui Yang	10.1016/j.jcat.2012.03.008	NaN	NaN	keyword=ai
013	False	0021-9517	pp. 150-161	Heng Shou	10.1016/j.jcat.2012.12.011	NaN	NaN	keyword=ai
...	...	...	...	...	...	...	...	...
013	False	0168-3659	pp. 246-255	Fei Yan	10.1016/j.jconrel.2012.12.025	NaN	NaN	keyword=modeling and simulation
014	False	1067-4136	pp. 532-538	J. X. Liao	10.1134/S1067413614060083	NaN	NaN	keyword=modeling and simulation
011	False	0304-3894	pp. 1398-1404	N. Nasrallah	10.1016/j.jhazmat.2010.10.061	NaN	NaN	keyword=modeling and simulation

Merged Data (NSF + NIH)

- Combining the datasets are an option
  - This example is for data analytics
- NSF and NIH data-frames have **funding**
- Can combine the USDA data-frame with both or either for desired information

Organization	Name	Funding
AIRMETTLE, INC.	Donpaul Stephens	276000.0
Alabama State University	Carl Pettis	399976.0
BROAD INSTITUTE, INC.	Anne E. Carpenter	1955828.0
	BETH CIMINI	1136551.0
	Kevin William Eliceiri	1136551.0
CINCINNATI CHILDRENS HOSP MED CTR	Rhonda Szczesniak	406637.0
CYTODEL, INC.	PHILIP Arthur BAND	295365.0
Carnegie-Mellon University	Nicholas Nystrom	8914035.0
DARTMOUTH COLLEGE	Brian P Jackson	21074.0
EMORY UNIVERSITY	Alex Sox-Harris	677235.0
	Nader Nabile Massarweh	677235.0
FRED HUTCHINSON CANCER CENTER	Philip R Gafken	374776.0
Florida Atlantic University	Borko Furht	95751.0
Florida State University	Weikuan Yu	204767.0
Georgia Tech Research Corporation	Alan Porter	49992.0
	Haesun Park	3539333.0
HARVARD MEDICAL SCHOOL	Sung Eun EUN Choi	114113.0
HARVARD UNIVERSITY	Deanna Barch	726975.0
	Leah Helene Somerville	726975.0
ICAHN SCHOOL OF MEDICINE AT MOUNT SINAI	Li Shen	322926.0

Conclusion

- Finding Ph.D. students using the professors is difficult
  - The best approach till now is probably through the google API
  - Manually coding each website is not a good idea
- Gathering information about the organizations, institutions, investigators, funding, and dates is easily automated
- Different APIs show different projects in various areas
- It'll be interesting to see what other government agencies will show (the ones we did not have access to)

Future Suggestions

- Follow the pseudo-code for using the google API and spacey (included in the previous slides).
- Merge all the APIs from different sources to see whether there is any overlap between the projects.
- Follow up with the NSF API
  - Try to extract information from the abstract instead of the title