In [1]:

```
import numpy as np
import pandas as pd
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.ticker as mticker
from numpy.polynomial.polynomial import polyfit
import mplcursors
import squarify

from pandas import ExcelWriter
from pandas import ExcelFile
from IPython.display import FileLink, FileLinks
from datetime import datetime
import math
import xlsxwriter
import openpyxl
import io
import os

import tkinter as tk
from tkinter import simpledialog
import pixiedust

import plotly.express as px
import plotly.graph_objs as go
import chart_studio
import chart_studio.plotly as py
```

```
Pixiedust database opened successfully
```

Automating Collection Analysis Data Visualization in Jupyter Notebook:

What's Possible and Why Would You Do It

PAT LIENEMANN, ER ACCESS & DISCOVERY LIBRARIAN

LUWIS A.R.R. ANDRADI, CMT GA

NAT GUSTAFSON-SUNDELL, COLLECTIONS LIBRARIAN

EVAN RUSCH, INSTRUCTION LIBRARIAN

MINNESOTA STATE UNIVERSITY, MANKATO

MEMORIAL LIBRARY

JCA Production Line & Development Goals

1. Data Processing – Preparing the data and creating the StandardTitle
   i. **Python, JupyterNotebook**
   ii. MS Excel

2. Data Matching & Validation
   i. MS Excel
   ii. MS Access

Journal ID (JID)

3. Report Production
   i. MS Excel
   ii. **Python, Jupyter Notebook**

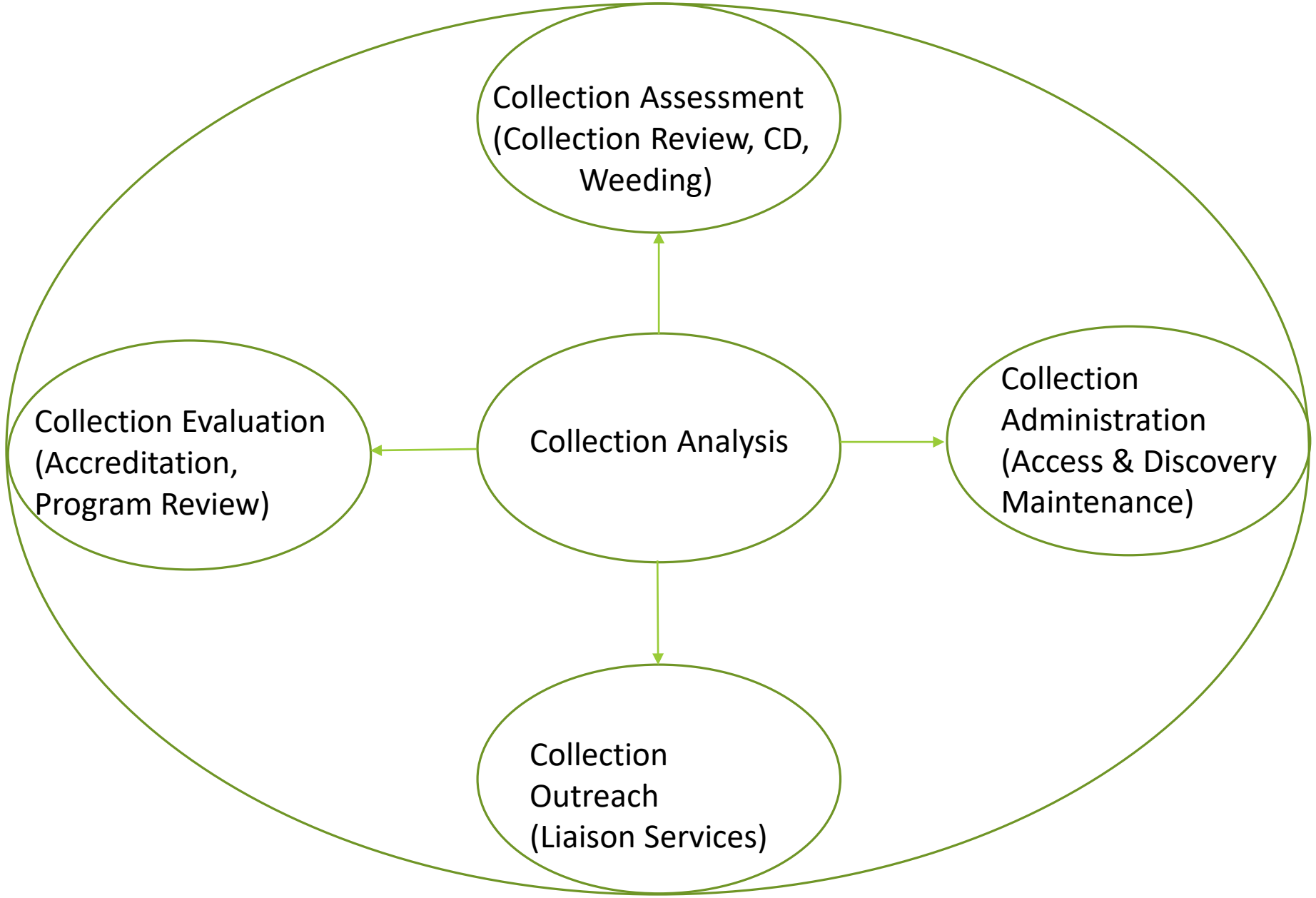
See the Video Introduction: https://link.mnsu.edu/85g

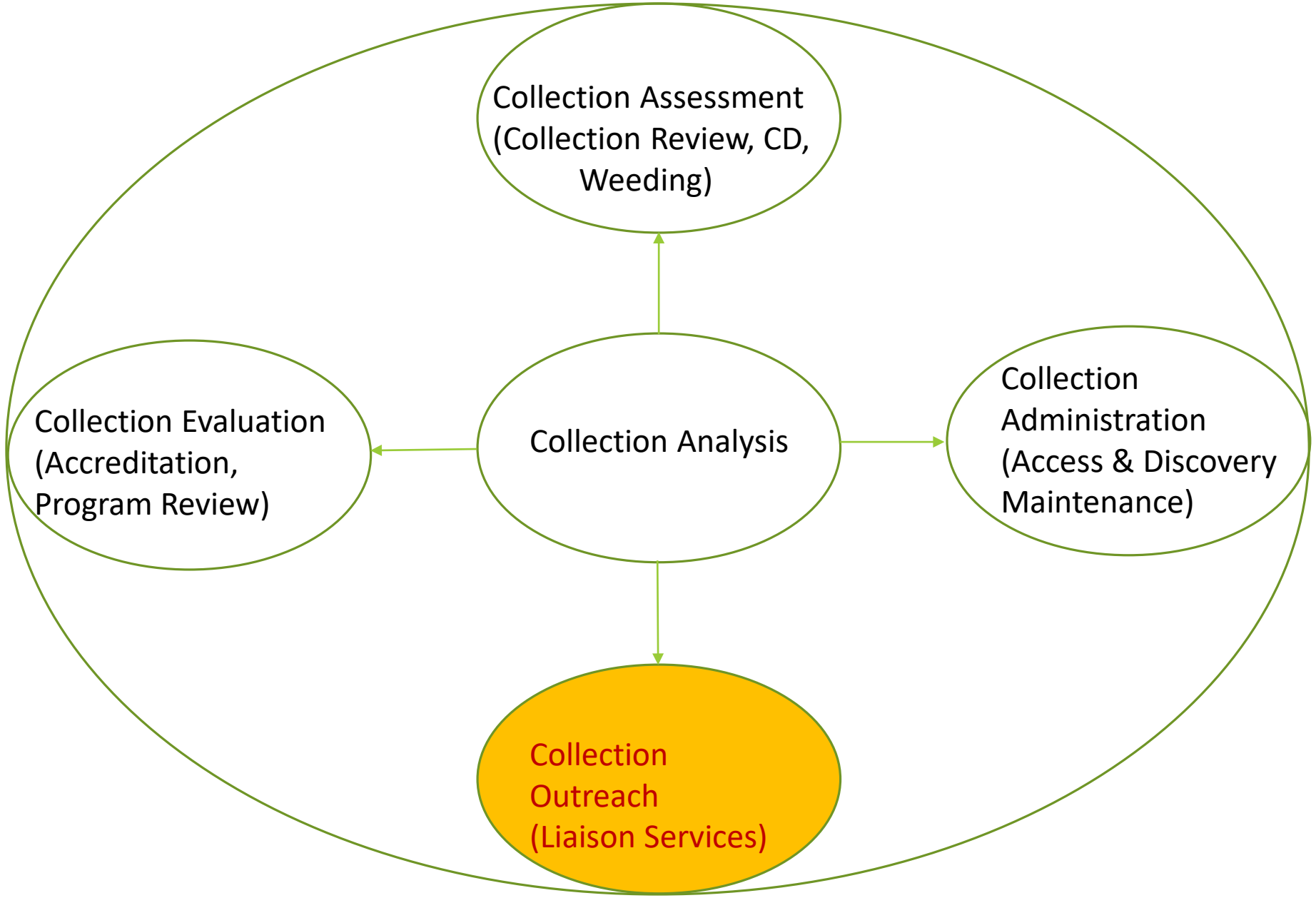
Minnesota State University, Mankato

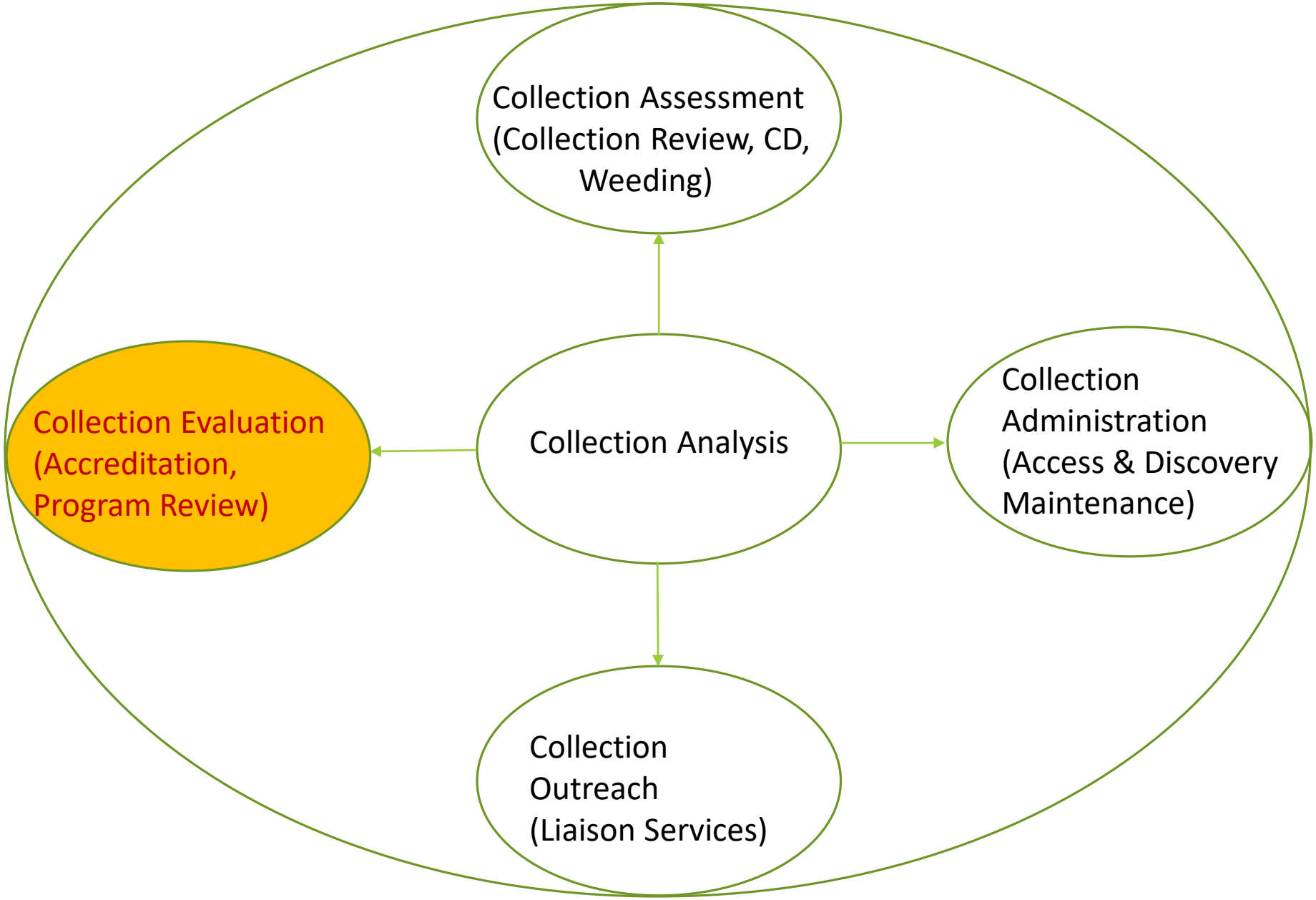
Library Services Collection Management Technology Lab

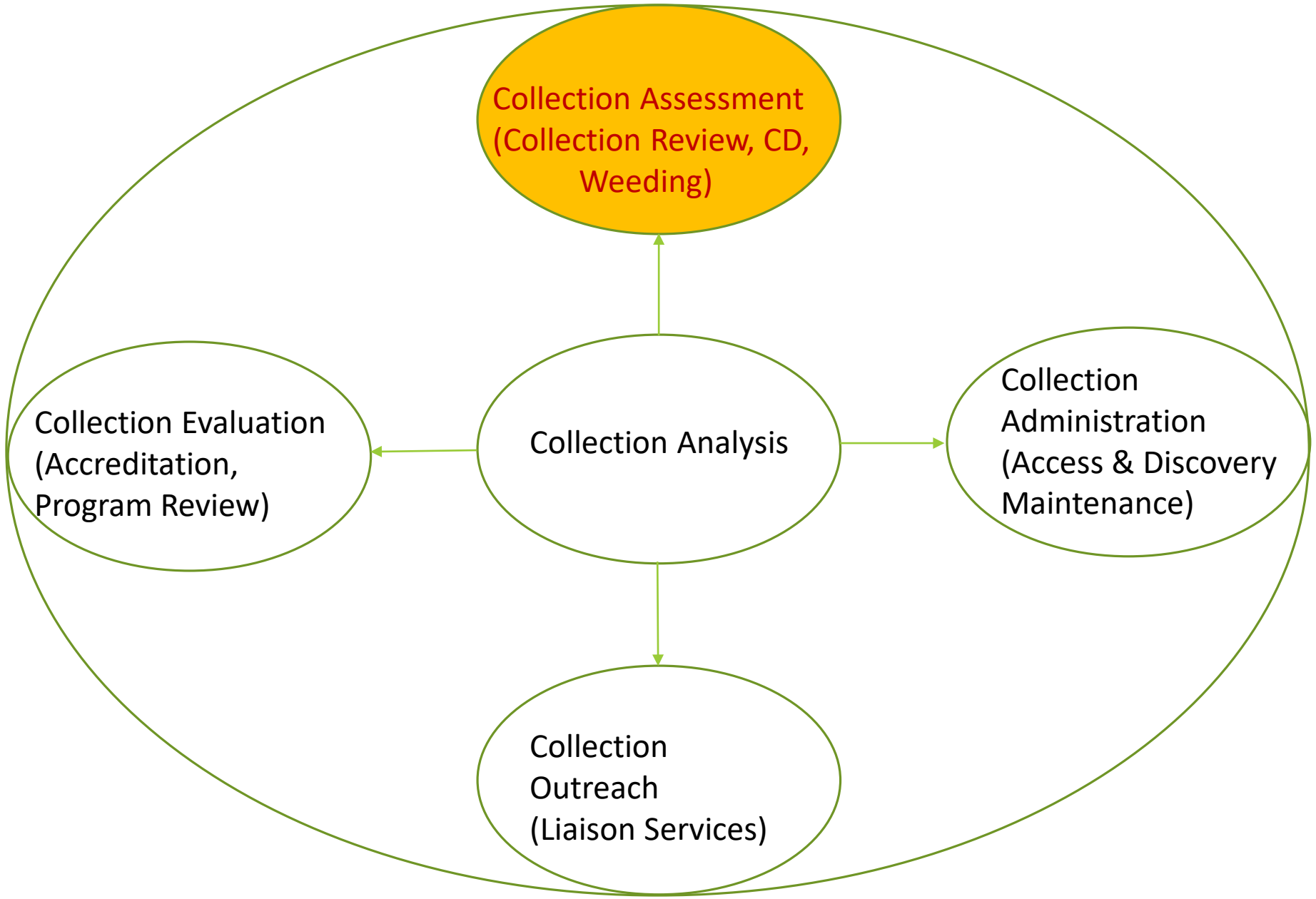
(A Sub-Group of the Journals Review Committee)

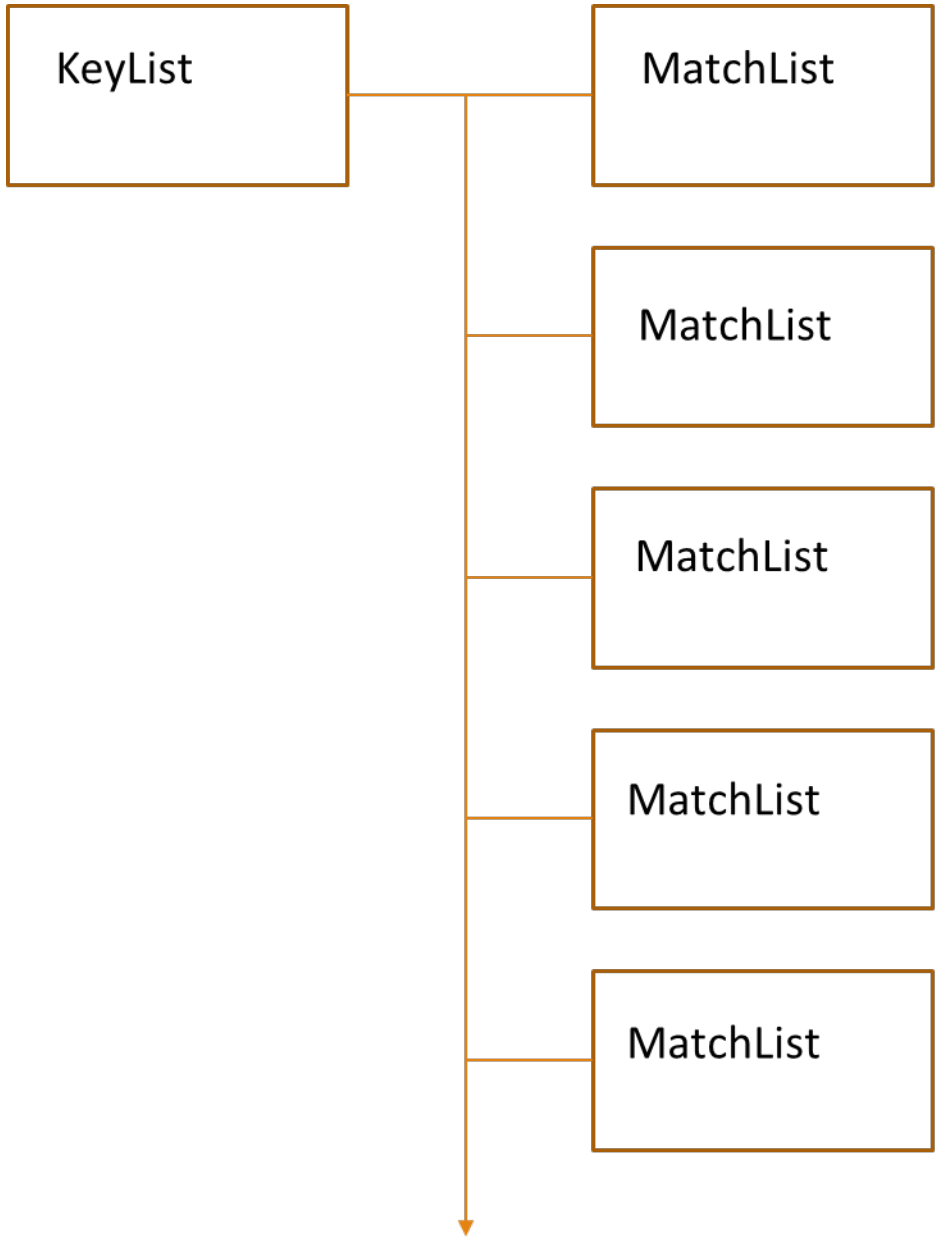
Luwis A.R.R. Andradi, CMT GA
Nat Gustafson-Sundell, Collections Librarian
Pat Lienemann, ER Access & Discovery Librarian
Jeff Rosamond, Technical Services Technician
Evan Rusch, Gov Docs & Instruction Librarian


Collection Assessment (Collection Review, CD, Weeding)
Collection Evaluation (Accreditation, Program Review)
Collection Analysis
Collection Administration (Access & Discovery Maintenance)
Collection Outreach (Liaison Services)


Collection Assessment (Collection Review, CD, Weeding)
Collection Evaluation (Accreditation, Program Review)
Collection Analysis
Collection Administration (Access & Discovery Maintenance)
Collection Outreach (Liaison Services)


Collection Assessment (Collection Review, CD, Weeding)
Collection Evaluation (Accreditation, Program Review)
Collection Analysis
Collection Administration (Access & Discovery Maintenance)
Collection Outreach (Liaison Services)


Collection Assessment (Collection Review, CD, Weeding)
Collection Evaluation (Accreditation, Program Review)
Collection Analysis
Collection Administration (Access & Discovery Maintenance)
Collection Outreach (Liaison Services)



KeyList:

- Scimago (Preferred!)
- Index
- Ulrich's List
- Academic Dept. Selected Titles
- Etc.

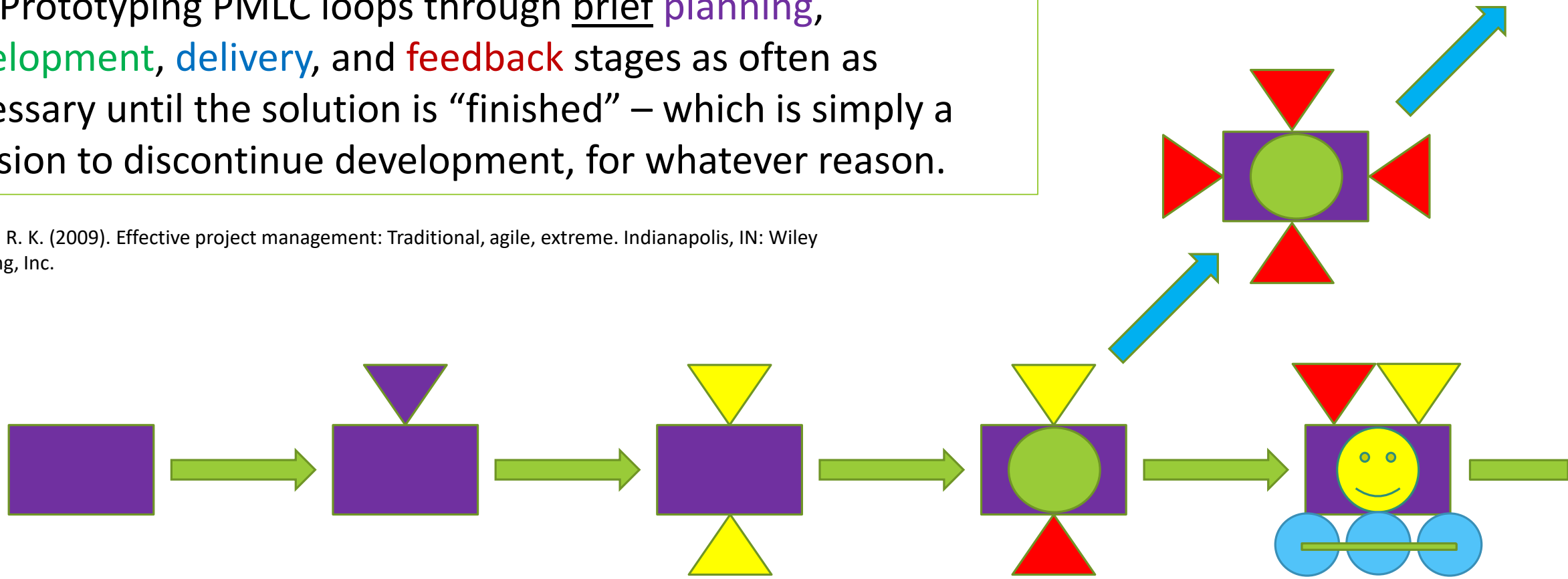
Data Sources (MatchLists):

- Serials Solutions (Holdings, Subject info)
- Aleph (POs, Payment History, Circs, Browses, ILL Loans)
- EbscoNet (Subscription info)
- COUNTER JR1, JR2, JR5 Reports (Usage Statistics)
- Vendors (Subscription info, Subject info)
- Scimago (Evaluative criteria, Subject info)
- Index (Subject info, Evaluative criteria, Coverage)
- Ulrich's List (Subject info, Journal info)
- Academic Dept. Selected Titles (Evaluative criteria)
- WMS
- Alma
- Etc.

Prototyping Project Management Life Cycle (PMLC)

The Prototyping PMLC loops through brief planning, development, delivery, and feedback stages as often as necessary until the solution is “finished” – which is simply a decision to discontinue development, for whatever reason.

Wysocki, R. K. (2009). Effective project management: Traditional, agile, extreme. Indianapolis, IN: Wiley Publishing, Inc.



Sample result for AUTHOR Query

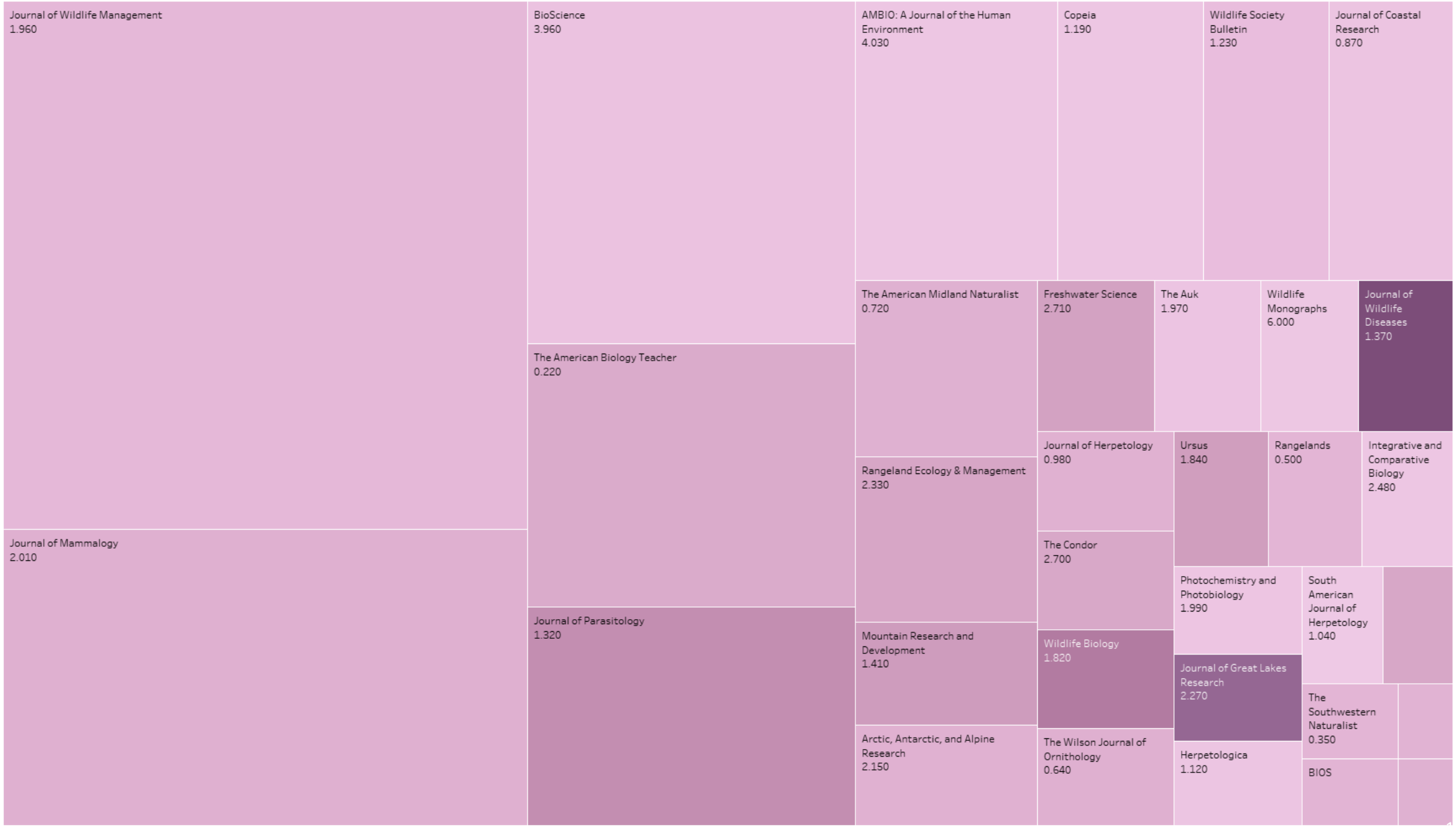
```
4   author_name, title_main ,doi_main,     (journaltitle)       reference_info          title_main,doi_main,author_name;
5   select * from reference_info
6
7   ## in the above query we have some null, hence counting it again removing the nulls
8   select t.author_name,t.title_main,t.DOI_main,sum(t.citcnt)  from  (
9   select author_name,title_main, doi_main,journaltitle,sum(journaltitle),case when length(journaltitle)=0 then 0 else 1 end as citcnt from reference_info) t group b
10  order by 1;
11  # adding ref_count in the query
12
13  select author_name,title_main, doi_main,ref_count,count(journaltitle),sum(case when length(journaltitle)=0 then 0 else 1 end ) Without_null from reference_info
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

author_name	title_main	doi_main	count(journaltitle)
DANIEL HOULIHAN	A Review of Behavioral Treatments used for LeschNyhan Syndrome	10.1177/0145445500242003	65
DANIEL HOULIHAN	An examination of response covariation in the behavioral treatment of identical twin boys with ...	10.1002/bin.2360090302	41
DANIEL HOULIHAN	Behavioral conceptualizations and treatments of Tourette's syndrome: A review and overview	10.1002/bin.2360080205	245
DANIEL HOULIHAN	Behavioral Manifestations of Adolescent School Relocation and Trauma	10.1300/j019v18n01_01	25
DANIEL HOULIHAN	Brief report: Identifying potential positive reinforcers in a residential treatment center for femal...	10.1002/bin.2360060206	25
DANIEL HOULIHAN	Brief report: Measuring selfefficacy with female adolescents who are conduct disordered: Valida...	10.1002/bin.2360060407	31
DANIEL HOULIHAN	EXPLORING THE REINFORCEMENT OF COMPLIANCE WITH 'DO' AND 'DON'T' REQUESTS AND TH...	10.2466/pr0.67.6.439-448	70
DANIEL HOULIHAN	Exploring the Reinforcement of Compliance with Do and Don'T Requests and the Side Effects: A...	10.2466/pr0.1990.67.2.439	35
DANIEL HOULIHAN	Predictors of peer helpfulness: Implications for youth in residential treatment	10.1002/bin.2360070106	81
DANIEL HOULIHAN	Recognizing and treating Rett syndrome in schools	10.1177/0143034311403058	139
DANIEL HOULIHAN	Relationship Satisfaction, Sexual Satisfaction, and Sexual Problems in Sexsomnia	10.1080/19317610903510489	35
DANIEL HOULIHAN	The Use of In Vivo Desensitization for the Treatment of a Specific Phobia of Earthworms	10.1177/1534650107300863	45
DANIEL HOULIHAN	Using sociometric measures to predict help seeking behaviors of youth in a positive peer culture ...	10.1002/bin.2360090203	29

Package	Subscription Usage Package-Level, JR1, 5 Yr Mean	Subscription Usage Package-Level, JR1, 3 Yr Mean	Subscription Usage Pacakge-Level, JR1, 5 Yr Volatility (range/mean)	Subscription Usage Pacakge-Level, JR1, 3 Yr Volatility (range/mean)	Usage % HTML, 2013	Usage % HTML, 2014	Usage % HTML, 2015	Usage % HTML, 2016	Usage % HTML, 2017	Usage % HTML, Trend
	420	363	0.65	0.12	15%	17%	19%	15%	19%	
	1348	1264	0.33	0.10	17%	10%	10%	15%	25%	
	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	194	241	1.06	0.71	0%	1%	12%	7%	4%	
	631	804	1.25	0.73	43%	21%	24%	31%	34%	
	742	669	0.58	0.09	44%	58%	51%	53%	59%	
	236	219	0.47	0.50	51%	75%	71%	76%	58%	
	840	821	0.52	0.15	41%	50%	46%	50%	47%	
	1333	1797	1.31	0.48	8%	9%	41%	11%	10%	
	59,287	65,605	0.35	0.04	56%	0%	58%	65%	63%	
	2,750	3,849	1.26	0.33	48%	66%	69%	69%	70%	
	10	5	2.79	0.40	0%	0%	0%	0%	0%	N/A
	654	656	0.79	0.60	20%	20%	28%	59%	61%	
	1,268	1,535	0.91	0.53	3%	9%	7%	12%	52%	
	574	696	0.73	0.27	16%	14%	20%	40%	27%	

Collection Breakdown by Journals: 2013-2017 Usage, Subscription Usage Percent, and 2017 CiteScore



JCA Production Line & Development Goals

1. Data Processing – Preparing the data and creating the StandardTitle
    i. MySQL
    ii. MS Excel
    iii. Open Refine (tested)

2. Data Matching & Validation
    i. MS Excel
    ii. MS Access

3. Report Production
    i. MS Excel
    ii. Tableau

JCA Production Line & Development Goals

1. Data Processing – Preparing the data and creating the StandardTitle
   i. **Python, Jupyter Notebook**
   ii. MS Excel

2. Data Matching & Validation
   i. MS Excel
   ii. MS Access

3. Report Production
   i. MS Excel
   ii. **Python, Jupyter Notebook**

JCA Production Line & Development Goals

1. Data Processing – Preparing the data and creating the StandardTitle
   i. **Python, Jupyter Notebook**
   ii. MS Excel

2. Data Matching & Validation
   i. MS Excel
   ii. MS Access

Journal ID (JID)

3. Report Production
   i. MS Excel
   ii. **Python, Jupyter Notebook**

Data Category	Current Version
CiteScore (from Scopus)	2015-2017
Click-Throughs (Alma)	*Development Only*
Cost Data (from vendor journal lists + Alma)	2020
Demand Data (majors & hours)	*Development Only*
ER Holdings (Alma)	Run 1/21/2020
Faculty citations and publications (from Cross Ref)	*Prototype Only*
Interlibrary Loans (Aleph)	2017
Interlibrary Loans (Alma & OCLC)	*Development Only*
JR1 (article downloads in a year)	2018 (back to 2013)
JR2 (article turnaways)	2018 (back to 2013)
JR5 (article downloads in a year per publication year)	2018
Local Data (comments)	*Prototype Only*
Print Ser Browses and Loans (Alma)	2019
Print Ser Holdings (Alma)	2019
Scimago (from Scopus)	2018

Data Category	Current Version
CiteScore (from Scopus)	2015-2017
Click-Throughs (Alma)	*Development Only*
Cost Data (from vendor journal lists + Alma)	2020
Demand Data (majors & hours)	*Development Only*
ER Holdings (Alma)	Run 1/21/2020
Faculty citations and publications (from Cross Ref)	*Prototype Only*
Interlibrary Loans (Aleph)	2017
Interlibrary Loans (Alma & OCLC)	*Development Only*
JR1 (article downloads in a year)	2018 (back to 2013)
JR2 (article turnaways)	2018 (back to 2013)
JR5 (article downloads in a year per publication year)	2018
Local Data (comments)	*Prototype Only*
Print Ser Browses and Loans (Alma)	2019
Print Ser Holdings (Alma)	2019
Scimago (from Scopus)	2018

1. **Liaison Journal Collection Analysis (LJCA) Report**
   - Supports Collection Outreach and Collection Evaluation
   - One report for each department or program (ideally)

2. **Scimago Master Blaster (SciMB) Report**
   - Supports Collection Outreach, Evaluation, and Assessment
   - Matches the "universe" of journals to our collection
   - Allows us to analyze collection strengths and gaps from a high-level vantage
   - Allows us to 'slice' the data at the subject level, and retain an overall context

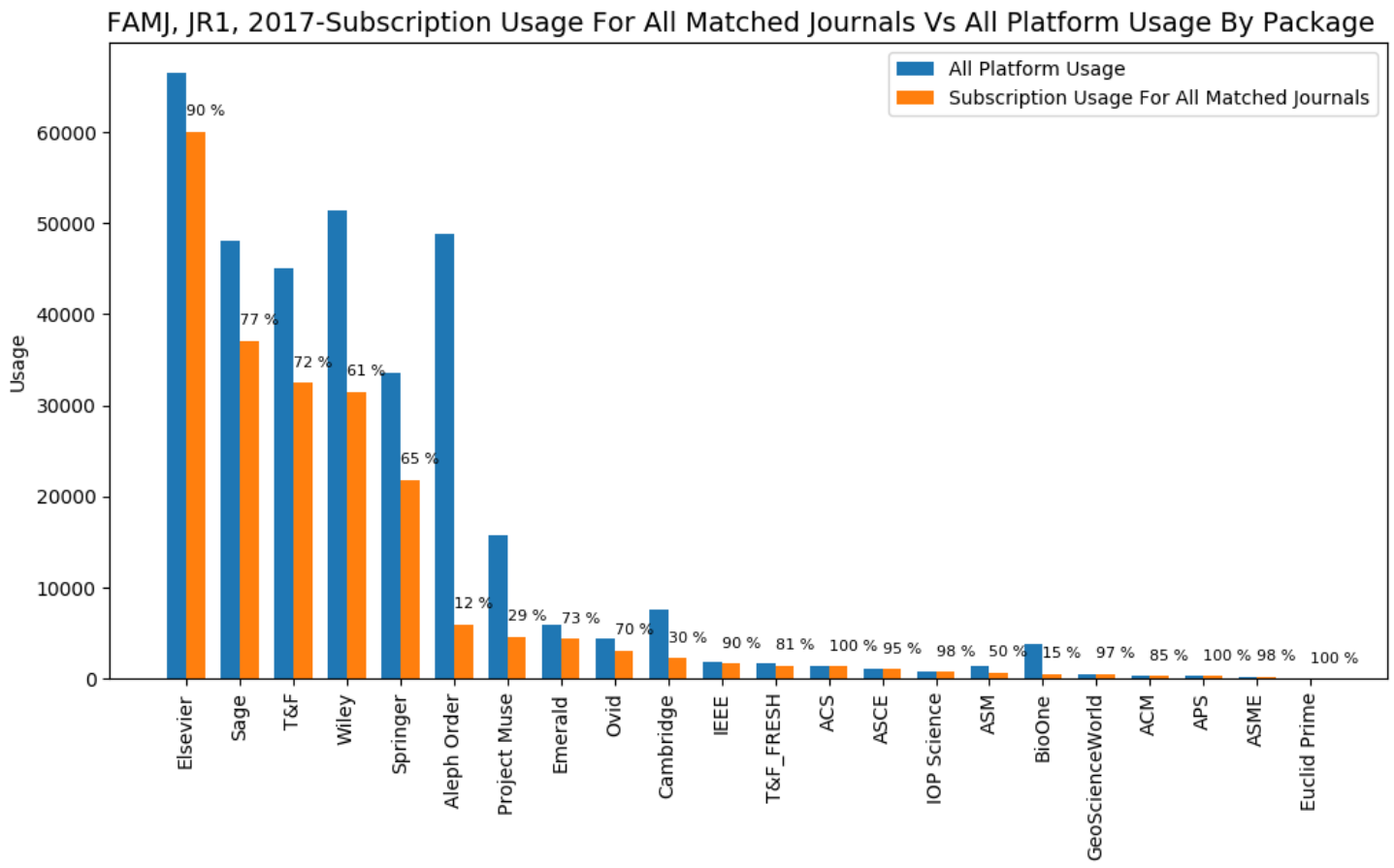
3. **Collection Review (CR) Report**
   - Supports Collection Assessment
   - Matches all journal and journal package subscriptions to all relevant data
   - Allows us to see which subscriptions are actionable

4. **Package Level Analysis Report (PLAR)**
   - Supports Collection Assessment
   - Rolls up data from the CR Report, and adds extra package level variables

See recorded demonstration of SciMB

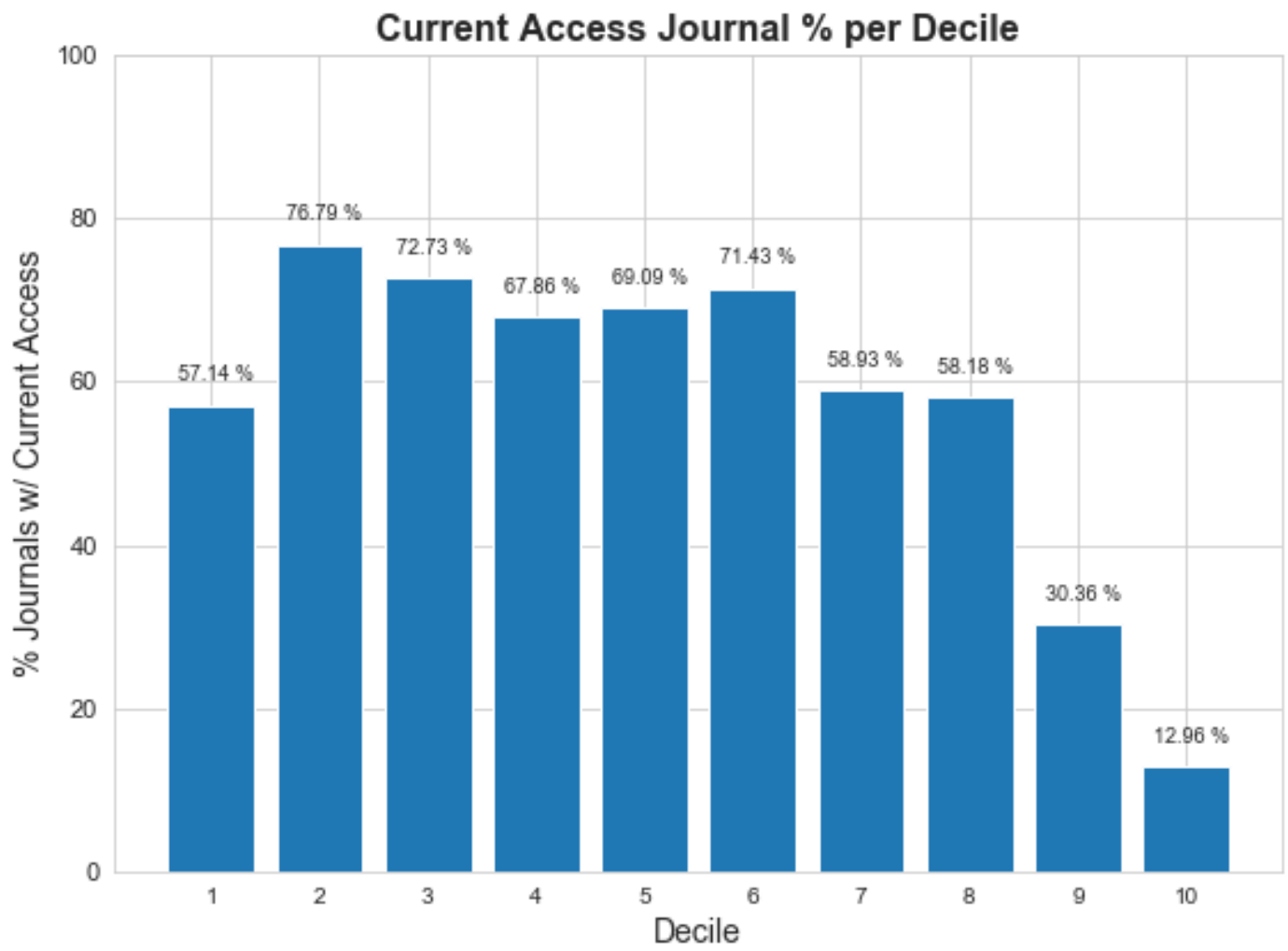
This slide is a placeholder for a ‘live’ demonstration of the SciMB report data, version AY2020, Fall.

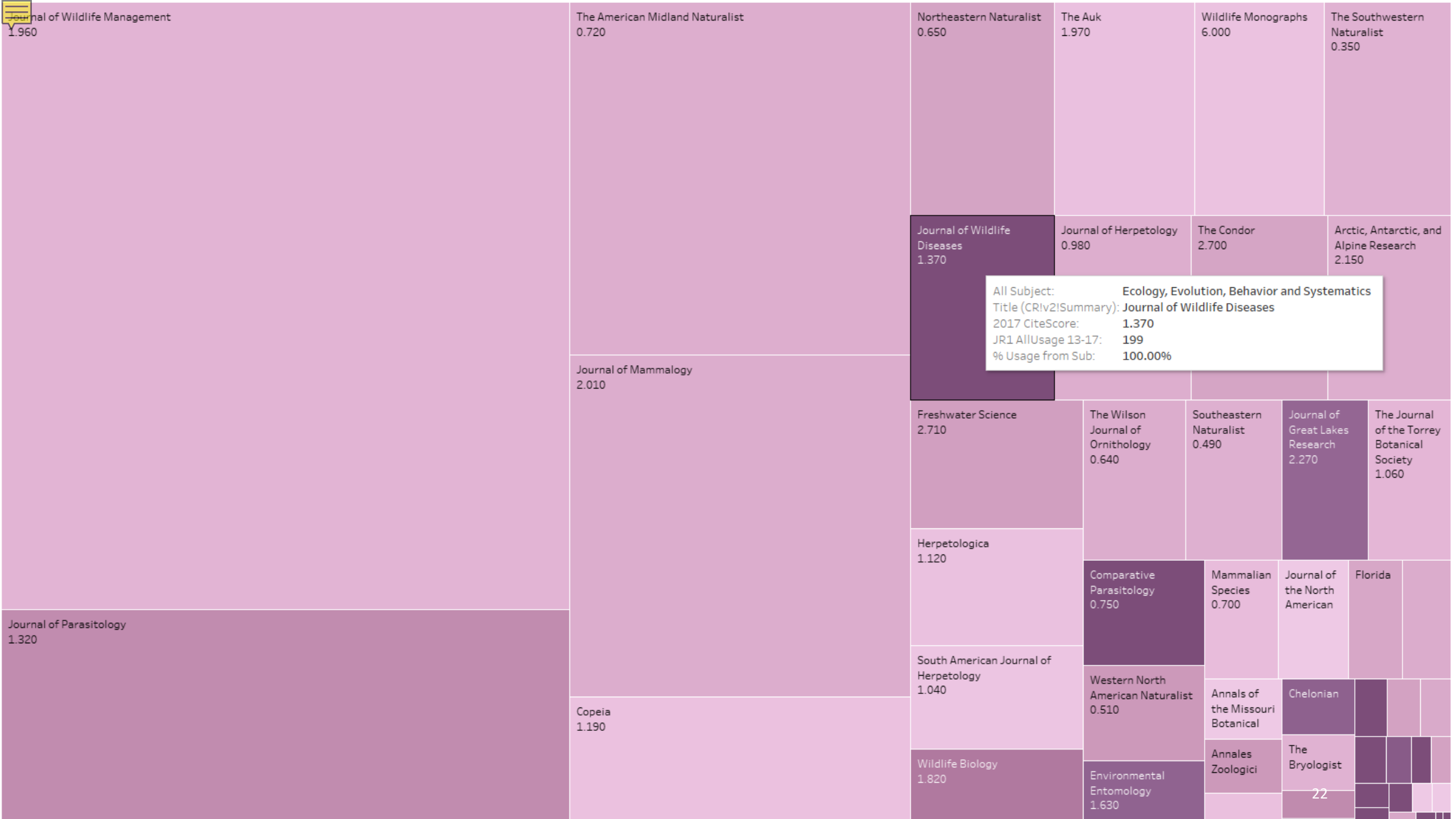
Power of Visualizations



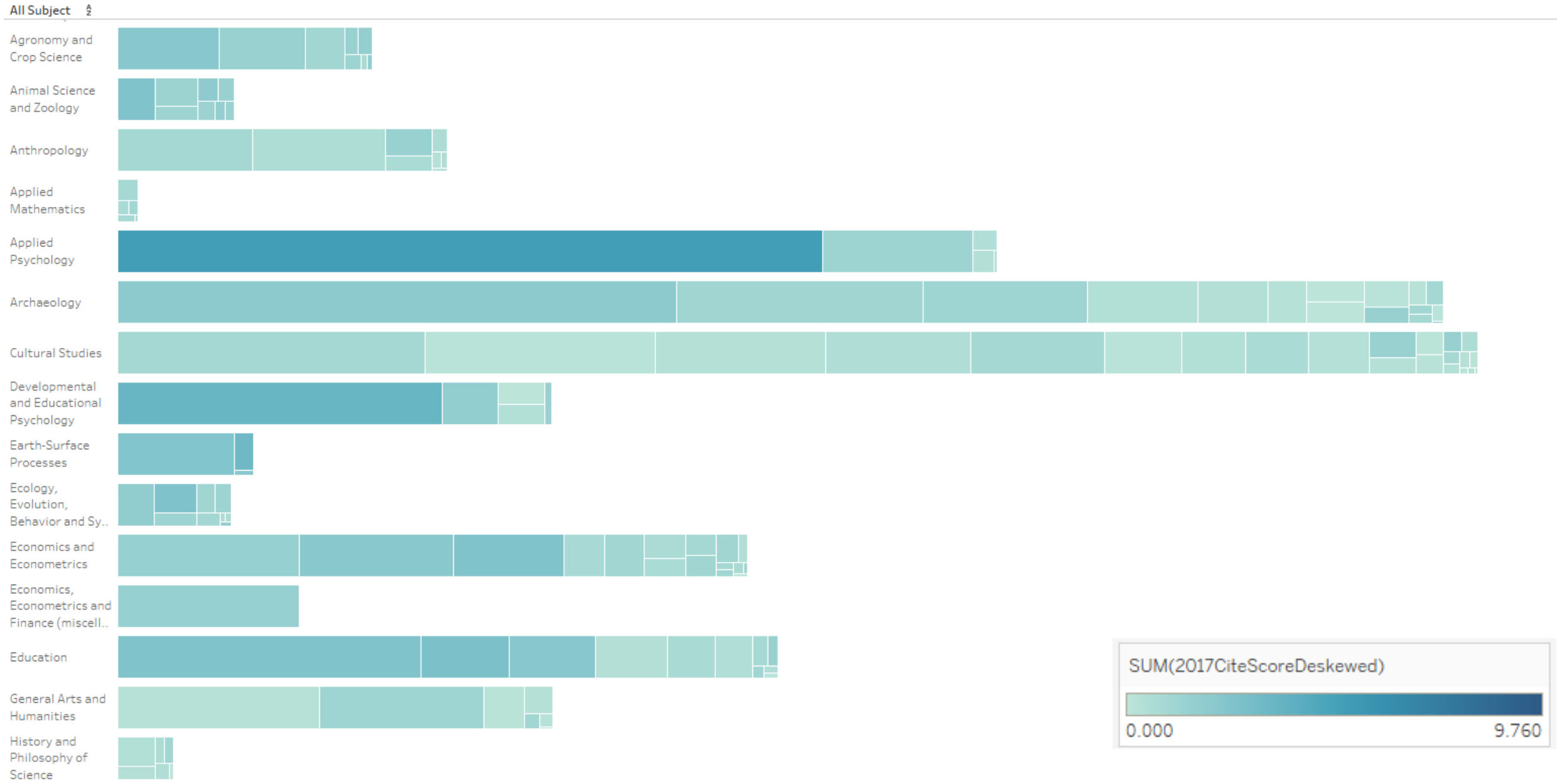
- Journal Collection Evaluation and Support
- Data Driven Decisions

MS Excel Visualizations




Journal of Wildlife Management
1.960
The American Midland Naturalist
0.720
Northeastern Naturalist
0.650
The Auk
1.970
Wildlife Monographs
6.000
The Southwestern Naturalist
0.350
Journal of Wildlife Diseases
1.370
Journal of Herpetology
0.980
The Condor
2.700
Arctic, Antarctic, and Alpine Research
2.150
All Subject: Ecology, Evolution, Behavior and Systematics
Title (CR!v2!Summary): Journal of Wildlife Diseases
2017 CiteScore: 1.370
JR1 AllUsage 13-17: 199
% Usage from Sub: 100.00%
Journal of Mammalogy
2.010
Freshwater Science
2.710
The Wilson Journal of Ornithology
0.640
Southeastern Naturalist
0.490
Journal of Great Lakes Research
2.270
The Journal of the Torrey Botanical Society
1.060
Herpetologica
1.120
Comparative Parasitology
0.750
Mammalian Species
0.700
Journal of the North American
Florida
Journal of Parasitology
1.320
South American Journal of Herpetology
1.040
Western North American Naturalist
0.510
Annals of the Missouri Botanical
Chelonian
Copeia
1.190
Annales Zoologici
The Bryologist
Wildlife Biology
1.820
Environmental Entomology
1.630

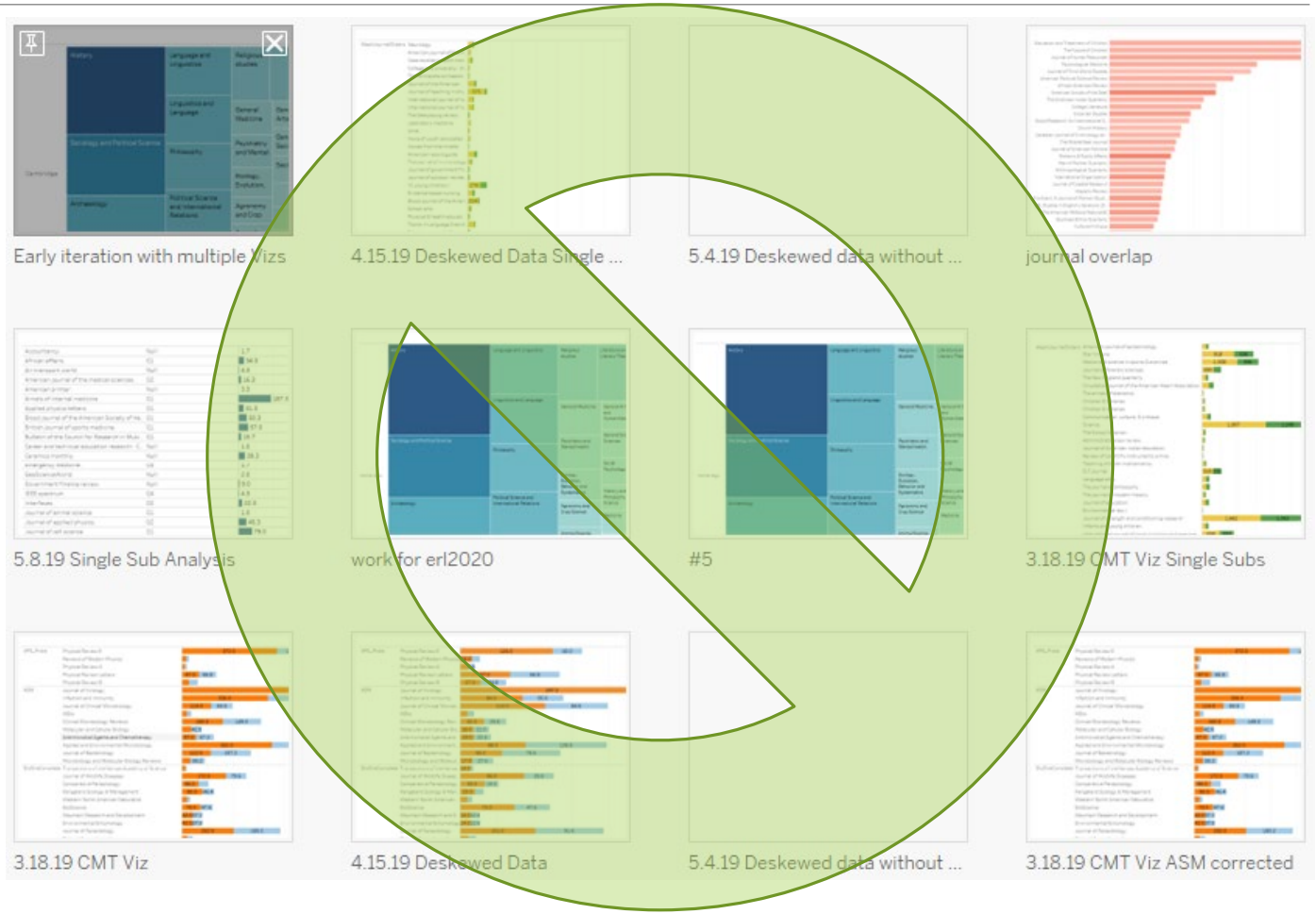
Single Collection Breakdown by Subjects, Journal Title, 2013-2017 Usage, and 2017 CiteScore



Why we have rejected Tableau

(for the most part)

- Not sharable
- Boutique Approach
- Not Scalable



New Approach Needed

21 Core Visualizations

+ Powerful, Dynamic Features

+ Automation

= Python & Jupyter Notebook

LJCA

Journal Collection Analysis database application (JCA db)

Produces a variety of reports

Liaison Journal Collection Analysis (LJCA) report.

130+ Data variables

21 Visualizations

Finished Reports

More than 70 subject area and report has to be developed for each subject area.

The "Jupyter Implementation"(JI)

Design a prototype

Replace a previous version of manually implemented data visualizations

Jupyter Notebook is a web-based interactive computational environment.

Benefit

The JI automates the production of data visualizations.

The JI enables reproducible results.

The JI functions as a teaching tool.

Some Jupyter Notebook Codes

```
In [ ]: import numpy as np
        import pandas as pd
        import matplotlib
        %matplotlib inline
        import matplotlib.pyplot as plt
        import seaborn as sns
        import matplotlib.ticker as mticker
        from numpy.polynomial.polynomial import polyfit
        import mplcursors
        import squarify

        from pandas import ExcelWriter
        from pandas import ExcelFile
        from IPython.display import FileLink, FileLinks
        from datetime import datetime
        import math
        import xlsxwriter
        import openpyxl
        import io
        import os

        import tkinter as tk
        from tkinter import simpledialog
        import pixiedust

        import plotly.express as px
        import plotly.graph_objs as go
        import chart_studio
        import chart_studio.plotly as py
```

Packages and Modules

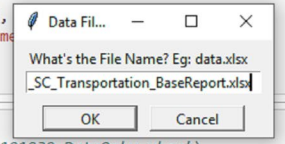
```
In [*]: # Newest- Scim_SA_Immun&Micro_BaseReport.xlsx

        # """Get User input- Edited_Scim_SC_SocioPoliSci_2018_BaseReport_SummaryOnly.xlsx  old"""
        #Scim_SC_Transportation_BaseReport.xlsx

        ROOT = tk.Tk()

        ROOT.withdraw()
        # the input dialog
        DataFileName = simpledialog.askstring(title="Data File Name",
                                              prompt="What's the File Name

        DataFileName
```




```
n [10]: # Read excel file
        #OrginalExcelFile = pd.read_excel('Data/SciMB_AY20_v1_FINAL_191030_DataOnly.xlsx')
        OrginalExcelFile = pd.read_excel(DataFileName)
        OrginalExcelFile.head()
```

ut[10]:

	JID	TITLE	ISSN1	ISSN2	SOURCEID	RANK	TYPE	SJR	SJR_BEST_QUARTILE	H_INDEX	...	CITESCORE_IMPRINT	CITESCORE_PUBLISHERC
0	JID	Title	ISSN1	ISSN2	Sourceid	Rank	Type	SJR	SJR Best Quartile	H index	...	CiteScore Imprint	CiteScore Publish
1	327	Analytic Methods in Accident Research	2213-6657	NaN	21100261712	1	journal	4.662	Q1	23	...	Elsevier	N
2	596	Journal of Travel Research	0047-2875	NaN	14813	2	journal	3.176	Q1	114	...	SAGE	Uni
3	684	Tourism Management	0261-5177	NaN	16547	3	journal	2.924	Q1	159	...	Elsevier	Unite

Importing Excel File

Discussion

JI could create the initial 21 visualizations successfully.

Able to export all the visualizations and tables into an Excel file.

JI will eventually be shared with librarians across the country.

Report Accomplish three goals:

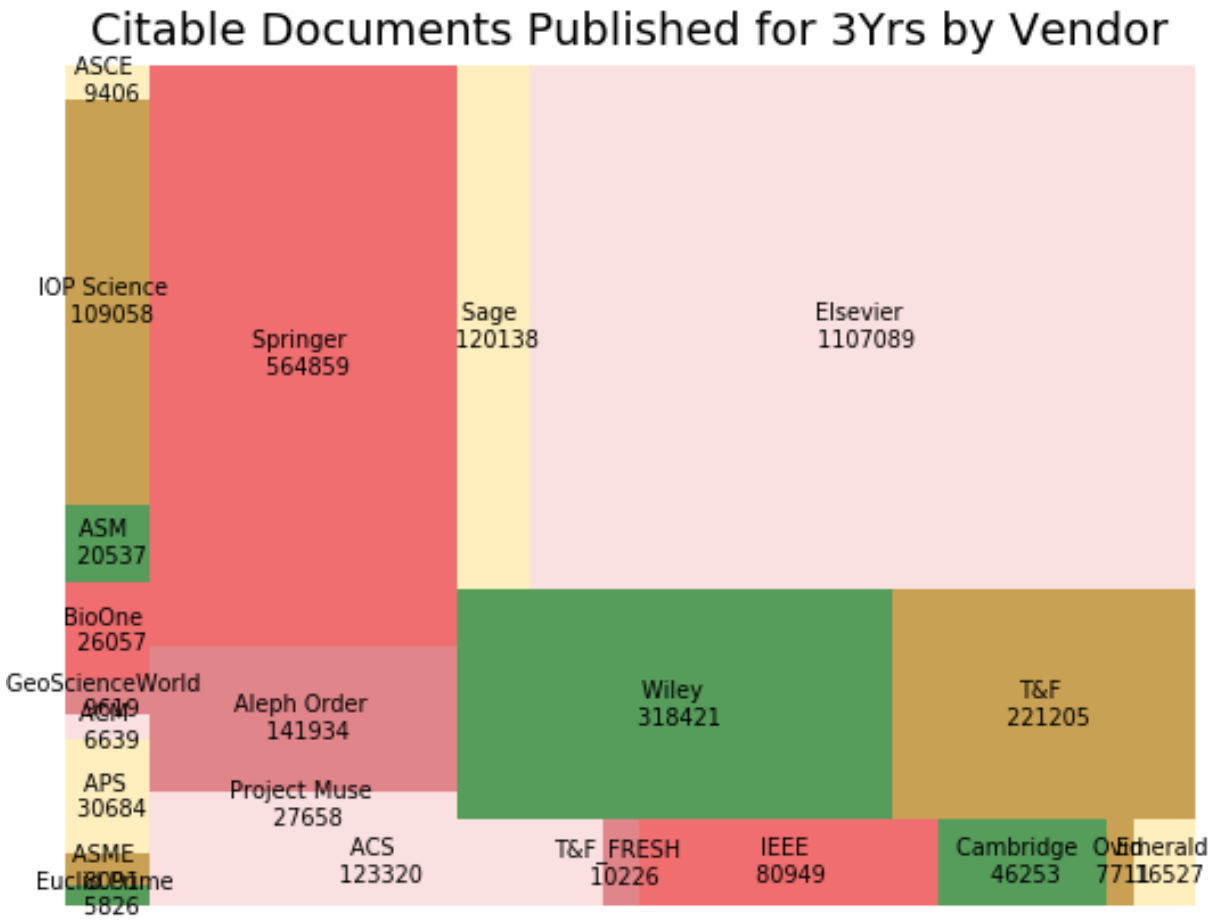
- Automation
- Reproducibility
- User Education

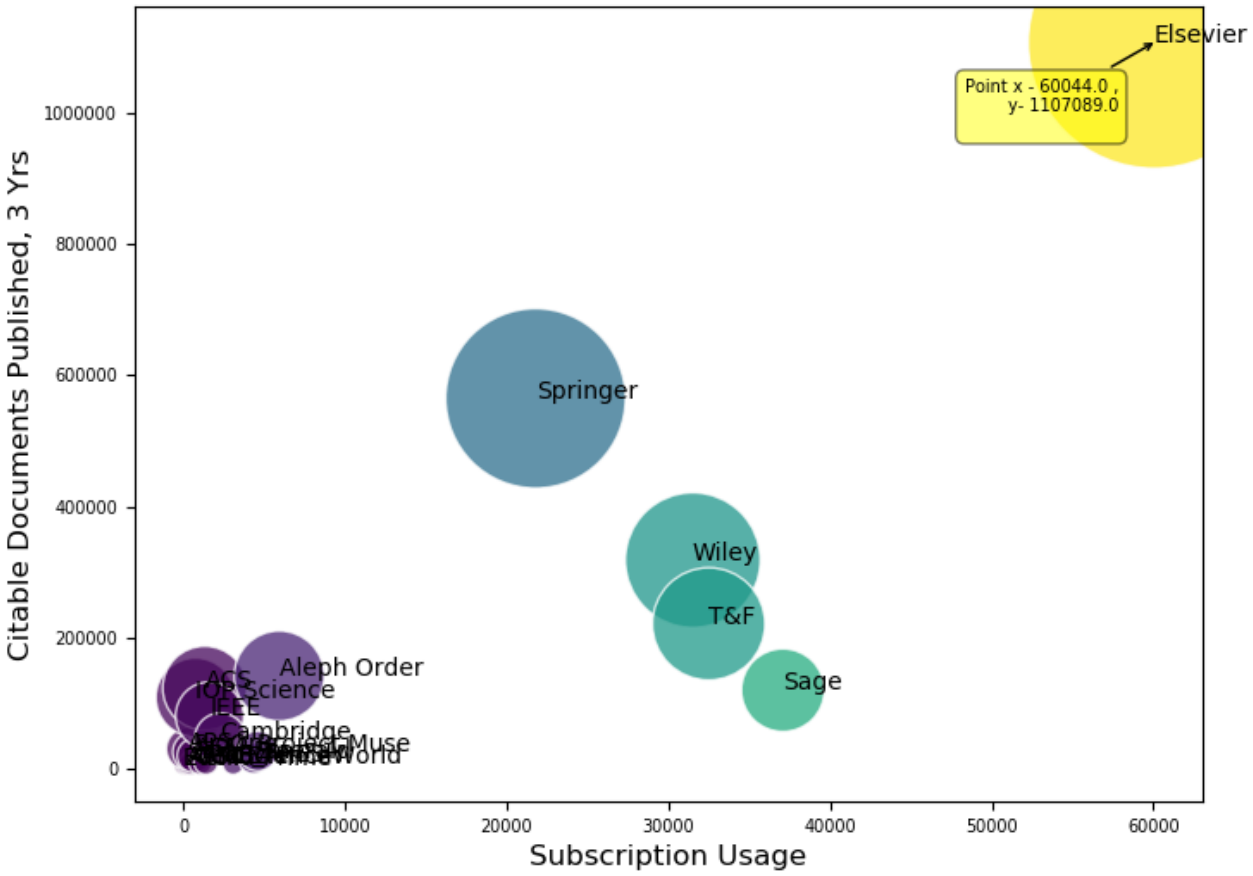
Development of Additional Reports.

Finding some other powerful python libraries to create visualizations.

Guiding collection decisions for next calendar year.

Jupyter Notebook Demonstration


Citable Documents Published for 3Yrs by Vendor
ASCE 9406
IOP Science 109058
Springer 564859
Sage 120138
Elsevier 1107089
ASM 20537
BioOne 26057
GeoScienceWorld
Aleph Order 141934
Wiley 318421
T&F 221205
APS 30684
Project Muse 27658
ASME
ACS 123320
T&F
FRESH 10226
IEEE 80949
Cambridge 46253
Emerald 16527
5826


Elsevier
Point x - 60044.0 ,
y- 1107089.0
Citable Documents Published, 3 Yrs
Subscription Usage
Springer
Wiley
T&F
Sage
Aleph Order
ACS
IOP Science
IEEE
Cambridge
Project Muse
World
0
200000
400000
600000
800000
1000000
0
10000
20000
30000
40000
50000
60000

Pixiedust

```
# Importing pixiedust library
import pixiedust
# check avilable Sample Data
pixiedust.sampleData()
# Set Sample data to 6- home prices
home_df = pixiedust.sampleData(6)
# Check the first few rows in dataset
home_df.head()
#home_df.loc[0:10,'LOCATION':'YEAR BUILT']
display(home_df)
```

Options Share

Schema

Table

Search table

Showing 100 of 4369 rows

SOLD DATE	PROPERTY TYPE	ADDRESS	CITY	STATE	ZIP	PRICE	BEDS	BATHS
nan	Single Family Residential	1262 Main St	Cotuit	MA	2635.0	1175000	5.0	3.5
November-9-2017	Single Family Residential	57 Philip St	Medfield	MA	2052.0	1650000	5.0	6.5
September-29-2017	Single Family Residential	20 Middle St	Hingham	MA	2043.0	2437500	5.0	5.5
February-21-2018	Single Family Residential	95 Racing Beach Ave	Falmouth	MA	2540.0	1780000	4.0	3.0
March-31-2017	Single Family Residential	56 Lorena Rd	Winchester	MA	1890.0	1775000	5.0	5.0
December-21-2017	Single Family Residential	997 Memorial Dr	Cambridge	MA	2138.0	2500000	4.0	3.5
June-22-2017	Single Family Residential	29 Nickerson Rd	Lexington	MA	2421.0	1650000	4.0	3.5

Table View Options

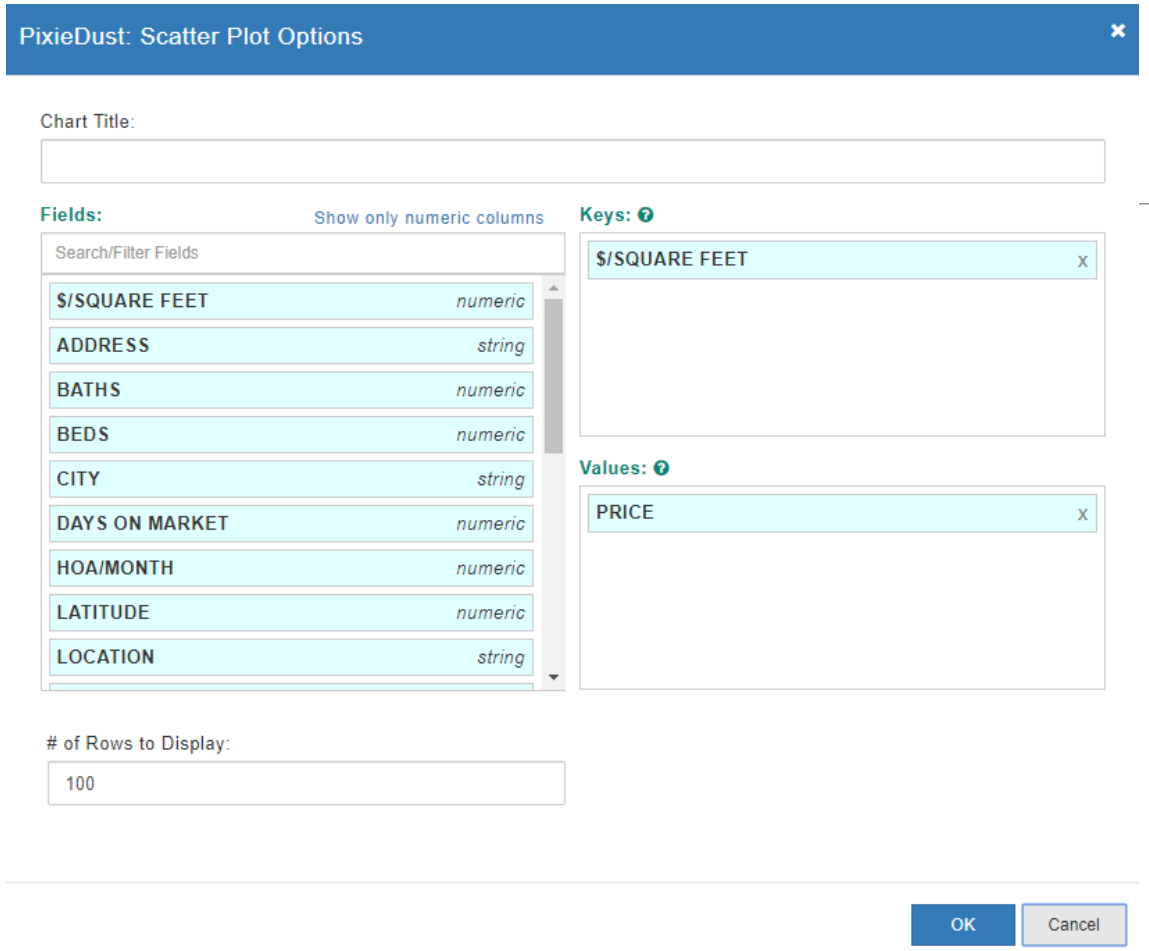
Hide Schema

Hide Search

Hide Row Count

Show Rows:

All


PixieDust: Scatter Plot Options
Chart Title:
Fields:
Show only numeric columns
Search/Filter Fields
$/SQUARE FEET numeric
ADDRESS string
BATHS numeric
BEDS numeric
CITY string
DAYS ON MARKET numeric
HOA/MONTH numeric
LATITUDE numeric
LOCATION string
Keys:
$/SQUARE FEET
Values:
PRICE
# of Rows to Display:
100
OK
Cancel


display(home_df)
Options
Share
Renderer:
matplotlib
House Price Compare to Area
$/SQUARE FEET
Scatter Plot Options
Chart Size: 100
D3 Rendering (mpld3)
PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
250
500
750
1000
1250
1500
1750
2000
$/SQUARE FEET

PixieDust: Map View Options

Chart Title:

Fields: Show only numeric columns

Search/Filter Fields

$/SQUARE FEET	*numeric*
ADDRESS	*string*
BATHS	*numeric*
BEDS	*numeric*
CITY	*string*
DAYS ON MARKET	*numeric*
HOA/MONTH	*numeric*
LATITUDE	*numeric*
LOCATION	*string*

Keys:

- LATITUDE x
- LONGITUDE x

Values:

- PRICE x

# of Rows to Display: 100

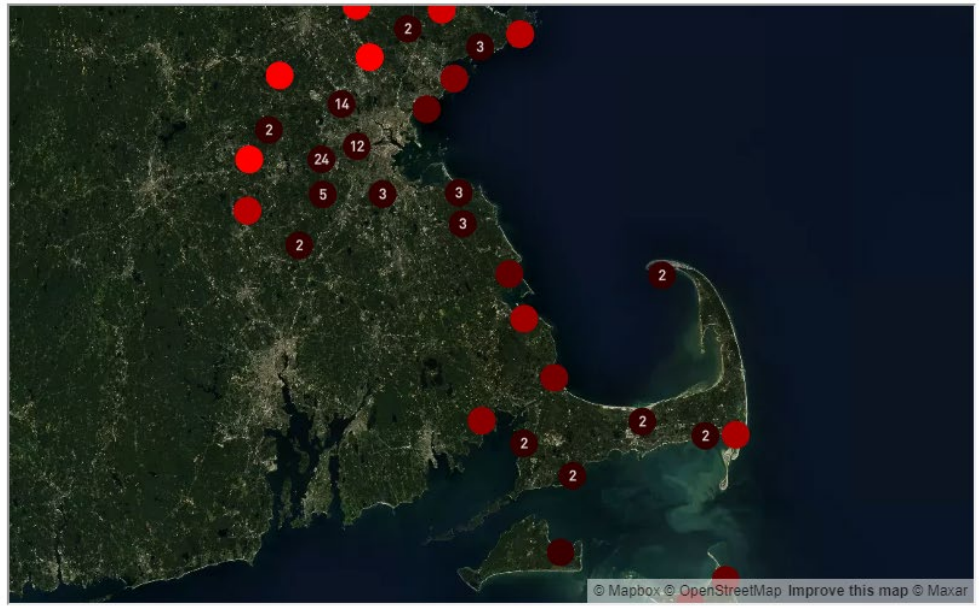
Mapbox Access Token: pk.eyJ1IjoibWFwYm94IiwiYSI6ImNpejY4M29iazA2Z2gycXA4

Custom Base Color:

Secondary Custom Base Color:

```
display(home_df)
```

Options | Share | Renderer: mapbox



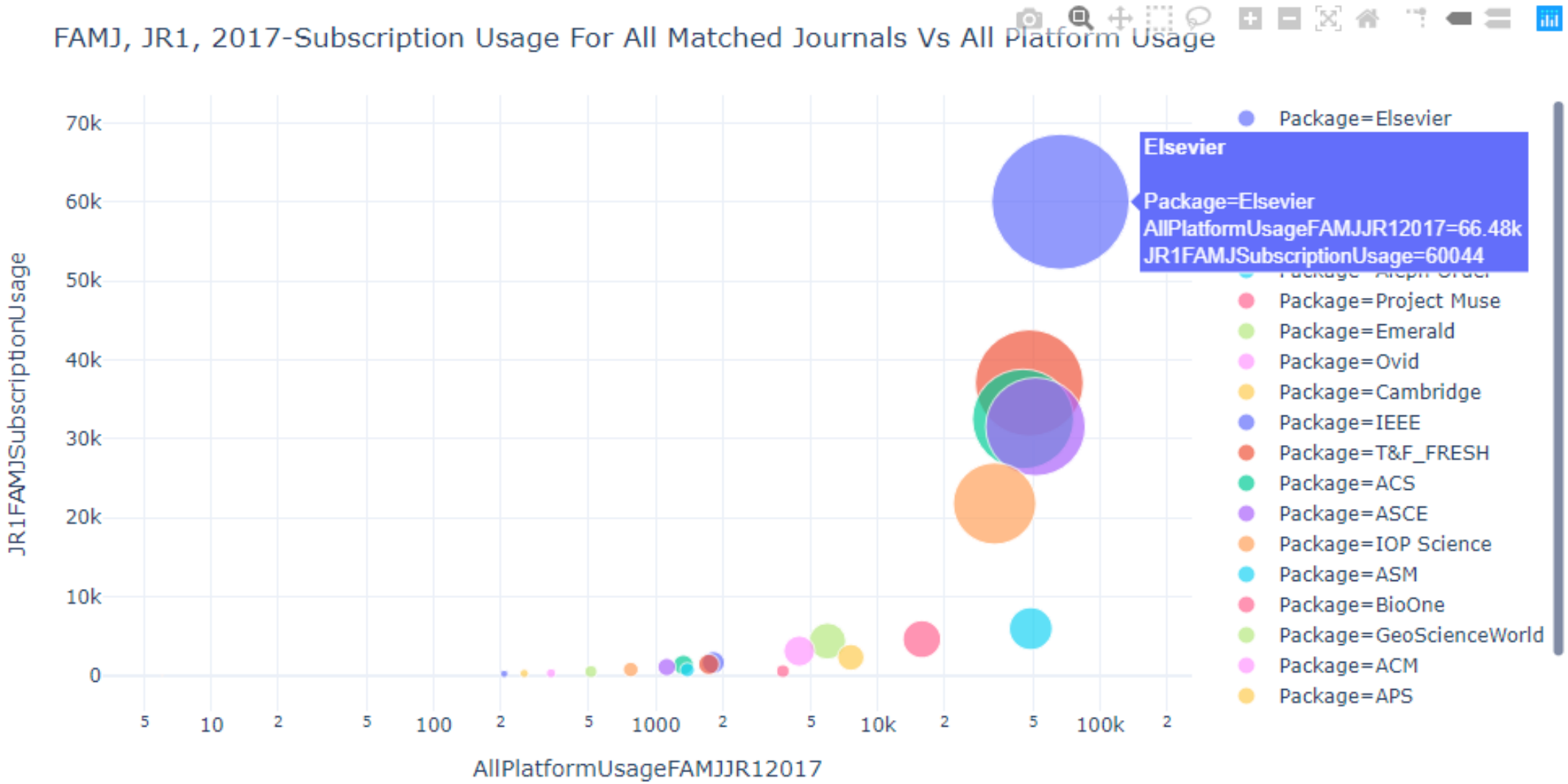
Map View Options

- Show legend
- Map Size: 90
- Opacity: 80
- Style: choropleth-cluster
- Color Ramp: Custom Monochrom
- basemap: satellite-v9
- Number of Bins: 5

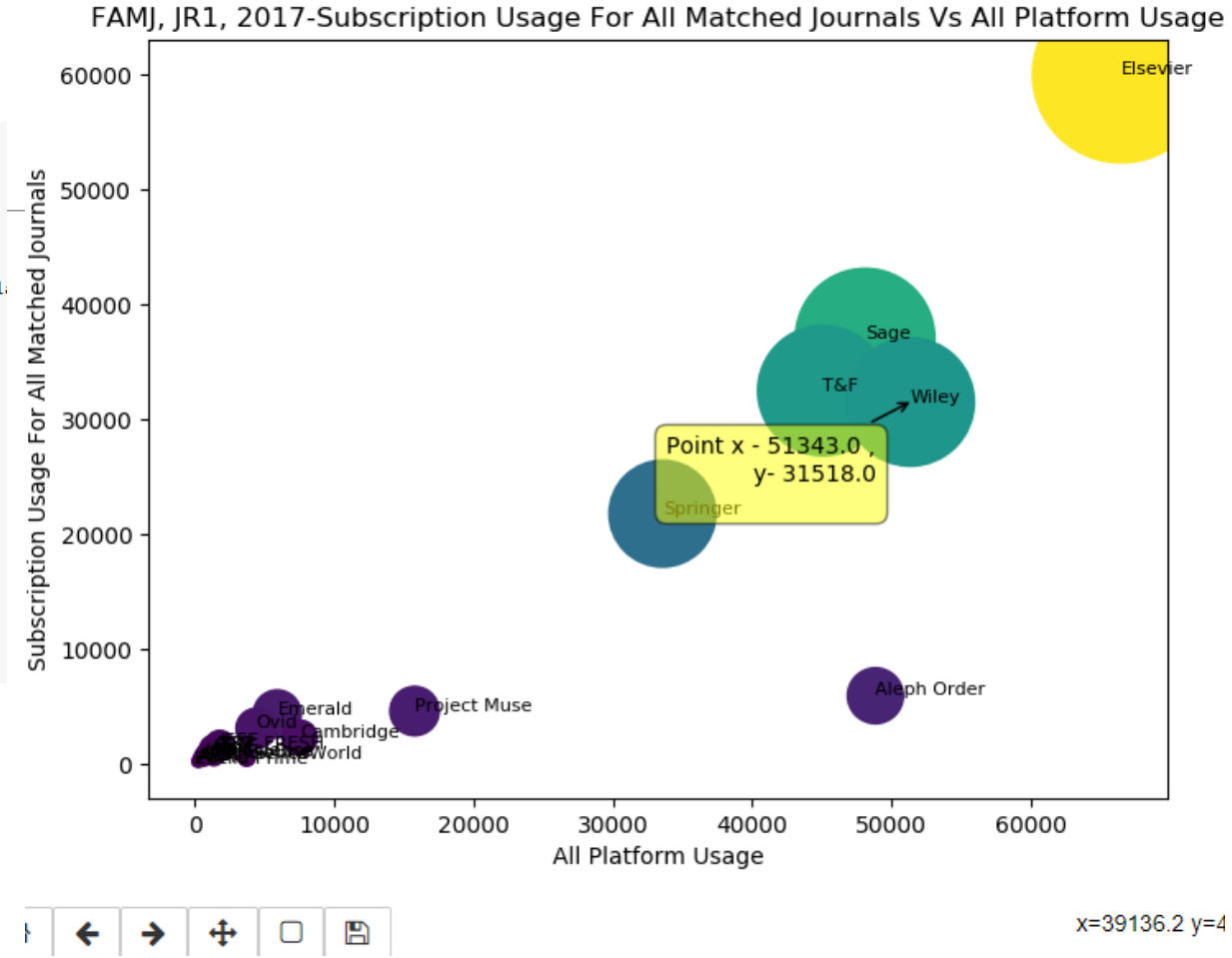
Bin Options

```python
import plotly.express as px

fig = px.scatter(SubUsageVsAllPlatformUse2017, x="AllPlatformUsageFAMJJR12017", y="JR1FAMJSubscriptionUsage",
                 size="JR1FAMJSubscriptionUsage", color="Package",hover_name="Package", log_x=True, size_max=60)
fig.update_layout(template='plotly_white',title='FAMJ, JR1, 2017-Subscription Usage For All Matched Journals Vs All Platform U
fig.show()
```

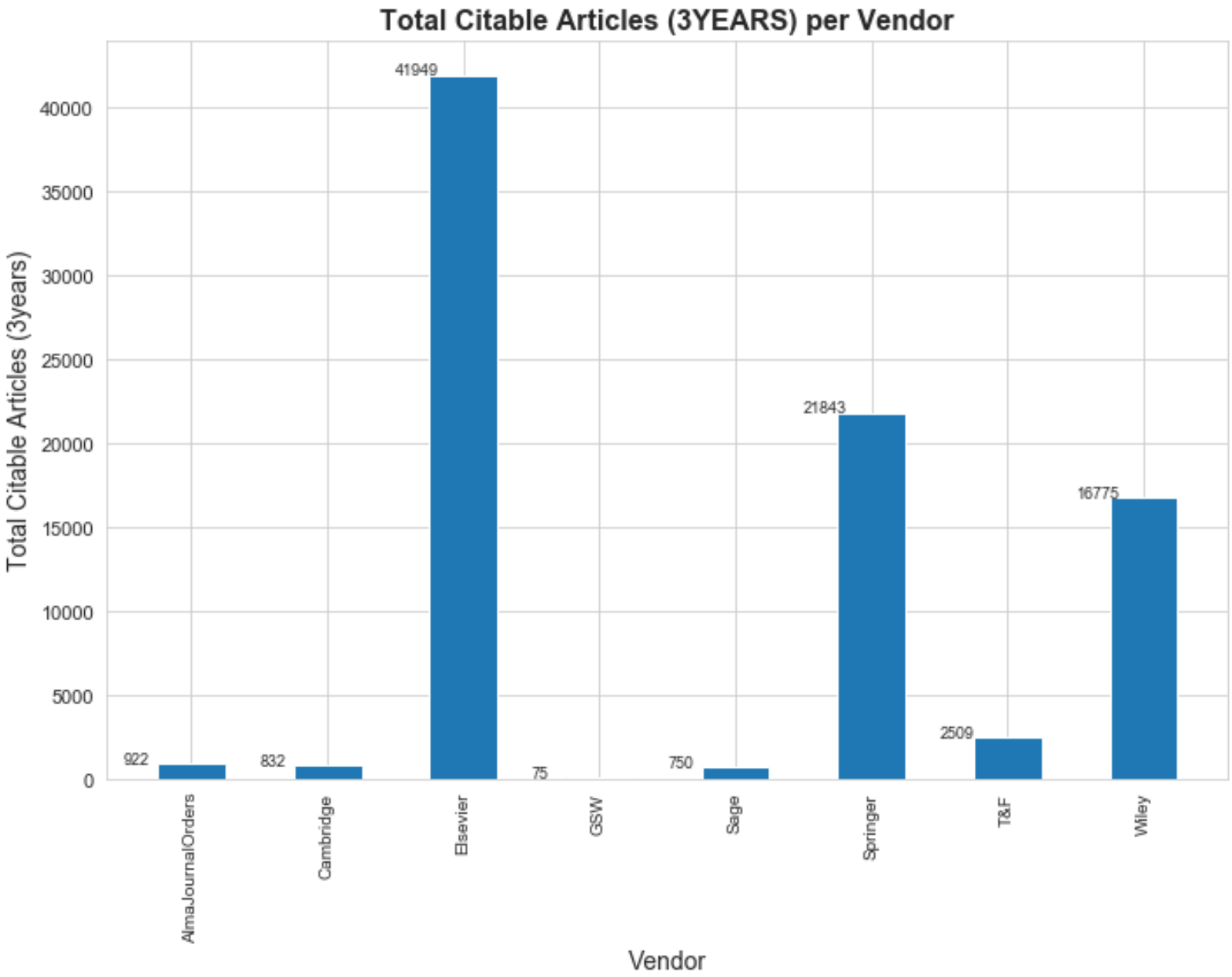


```
%matplotlib notebook
#calling it a second time may prevent some graphics errors
%matplotlib notebook
import matplotlib.pyplot as plt
fig , ax = plt.subplots(figsize=(8,6))
a = SubUsageVsAllPlatformUse2017['JR1FAMJSubscriptionUsage']/1000
# Create a scatter plot (Use s - make the size of each vendors usage)
plt.scatter('AllPlatformUsageFAMJJR12017', 'JR1FAMJSubscriptionUsage',s= x , c='JR1FAMJSubscriptionUsage', data=SubUsageVsAllPl
plt.title("FAMJ, JR1, 2017-Subscription Usage For All Matched Journals Vs All Platform Usage",fontsize=12)
plt.xlabel('All Platform Usage',fontsize=10)
plt.ylabel('Subscription Usage For All Matched Journals',fontsize=10)
list1 = list(SubUsageVsAllPlatformUse2017['Package'])
i = 0;
for row in SubUsageVsAllPlatformUse2017.itertuples():
    h = list1[i]
    i=i+1
    h = str(h)
    c = row.JR1FAMJSubscriptionUsage
    d = row.AllPlatformUsageFAMJJR12017
    ax.text(d,c,s = h, size = 8)
crs = mplcursors.cursor(ax,hover=True)
crs.connect("add", lambda sel: sel.annotation.set_text(
    'Point x - {} ,\n'
          'y- {}\n'.format(sel.target[0], sel.target[1])))
plt.show()
```



Bringing it Back Together

- Portable AND Sharable
- Impact of automation
- Potential for further automation



data.xlsx

Visualization demonstration