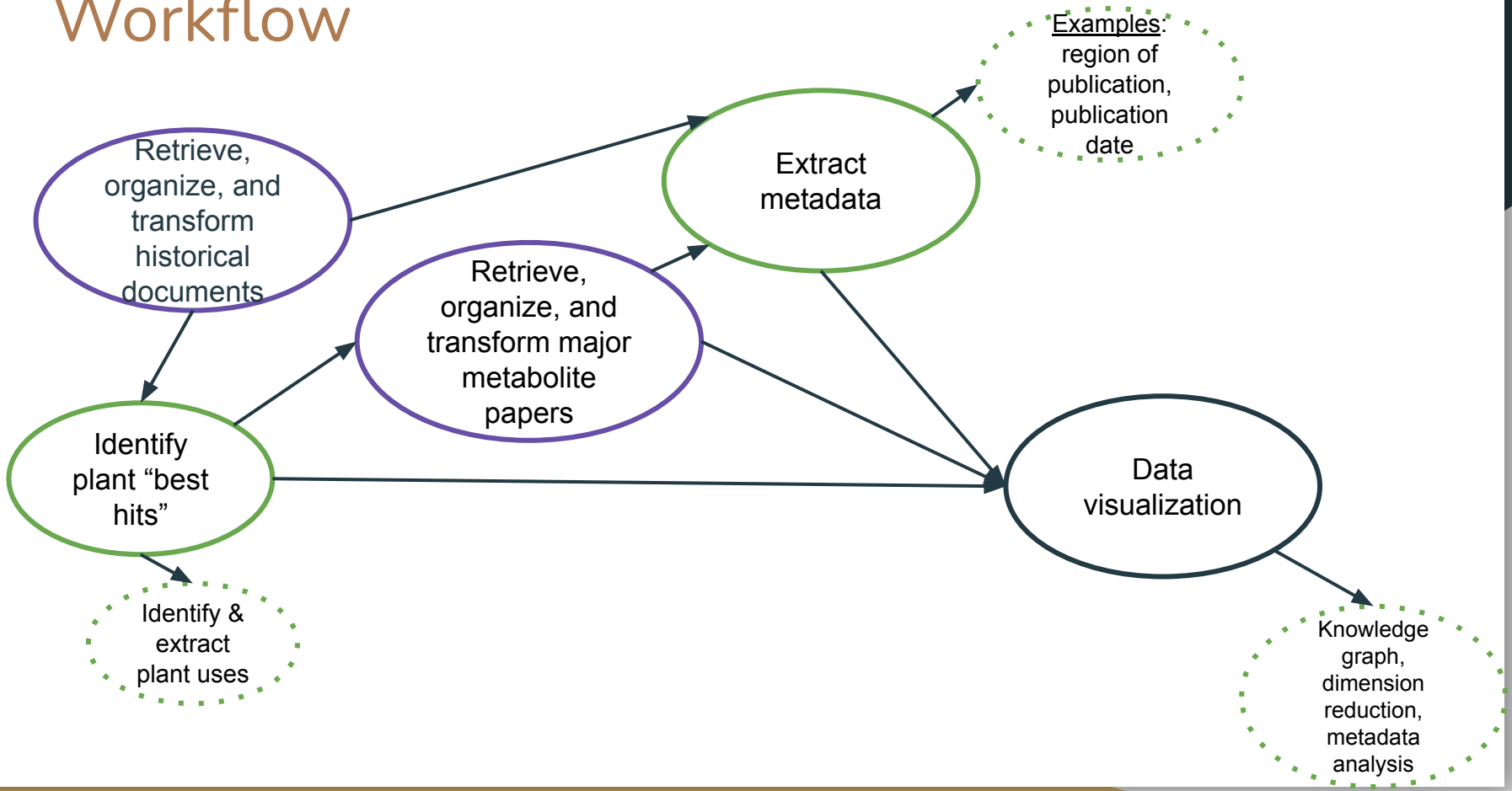
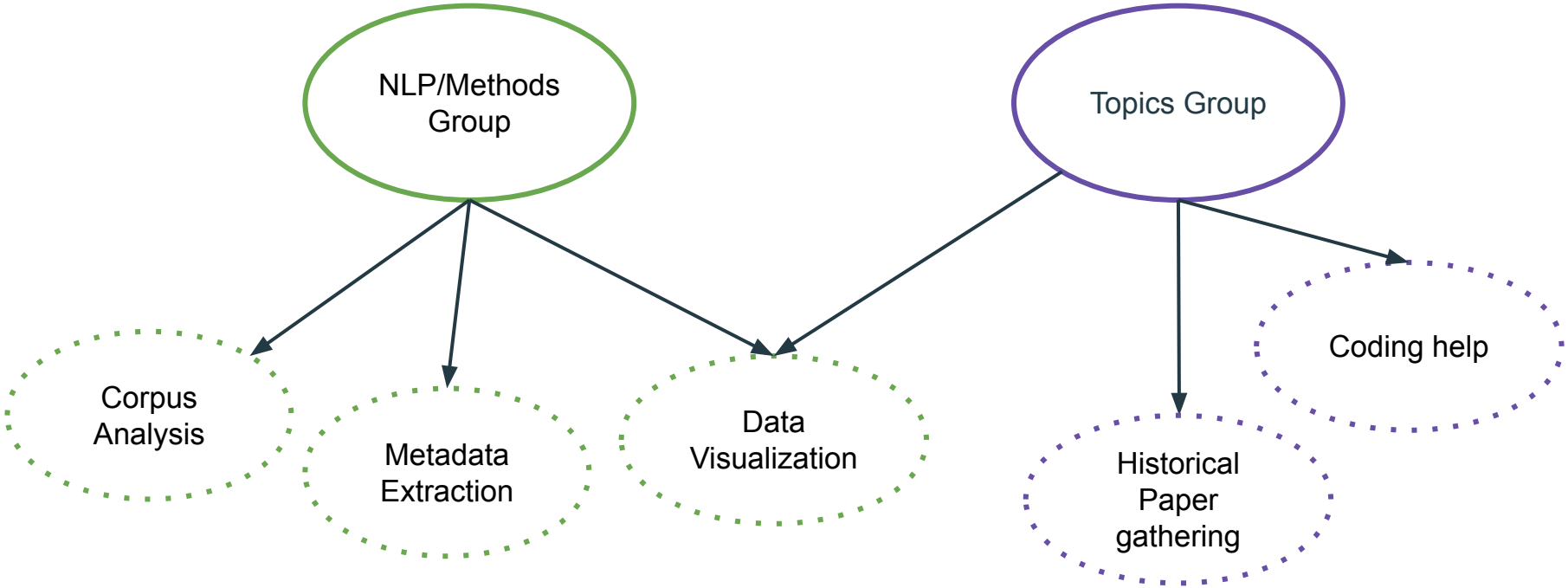
Project Update
11.21.2023

Workflow


Examples: region of publication, publication date
Retrieve, organize, and transform historical documents
Extract metadata
Retrieve, organize, and transform major metabolite papers
Identify plant “best hits”
Data visualization
Identify & extract plant uses
Knowledge graph, dimension reduction, metadata analysis

Subgroups


NLP/Methods Group
Topics Group
Corpus Analysis
Metadata Extraction
Data Visualization
Historical Paper gathering
Coding help

Hypothesis:

What were heavily documented plants in historical texts,

What purpose (if any) did these plants serve,

And can we find the metabolites that align with those purposes?

Hypothesis:

Topics group main focus for now

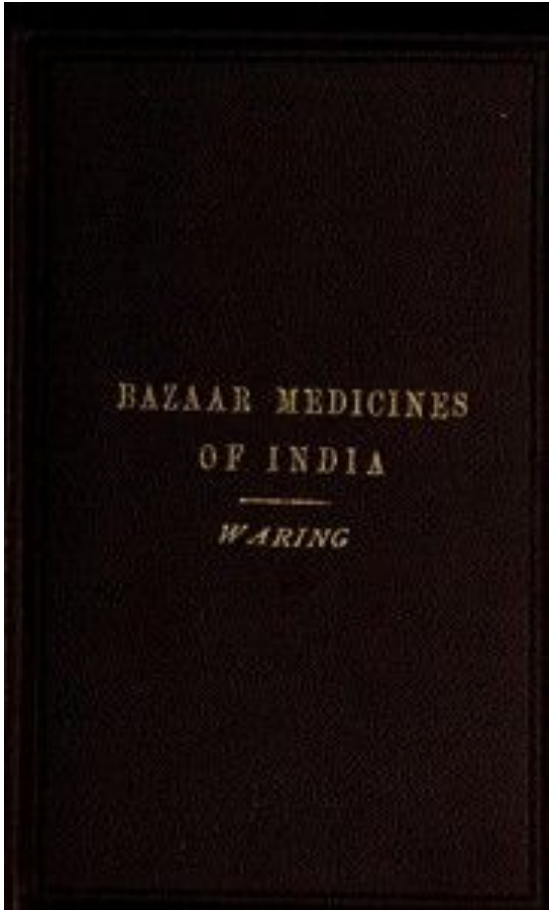
What were heavily documented plants in historical texts,

What purpose (if any) did these plants serve,

And can we find the metabolites that align with those purposes?

Our “historical paper” criteria

- Before 1900s
- Location the text focuses on
  - Continent-wise
- About plants in anyway.
- Preferably in English/ english translation
- Either a PDF or .txt format available

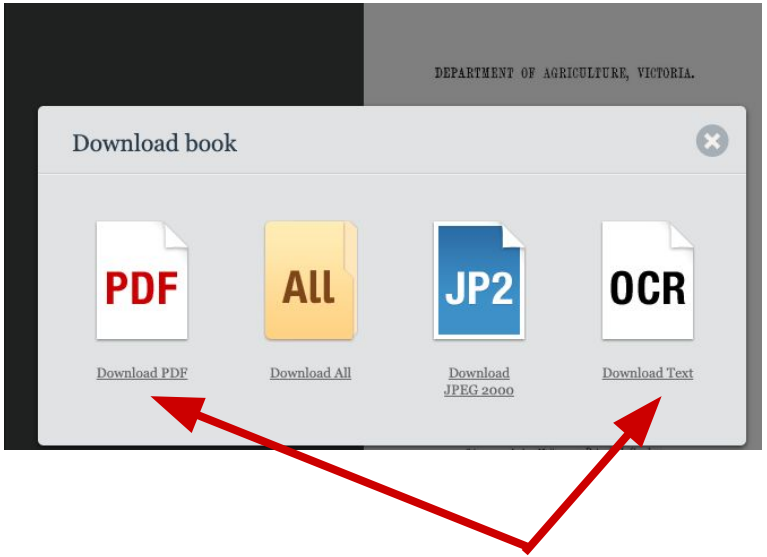


Sources we got papers from:

- Gutenberg project
  - Created a bot to gather all the “botany” books
- Biodiversity Heritage Library
- Google Books
- University libraries
- AI given sources
  - ChatGPT gave us leads on books
- Internet Archive

Gathering and Passing of Historical Documents

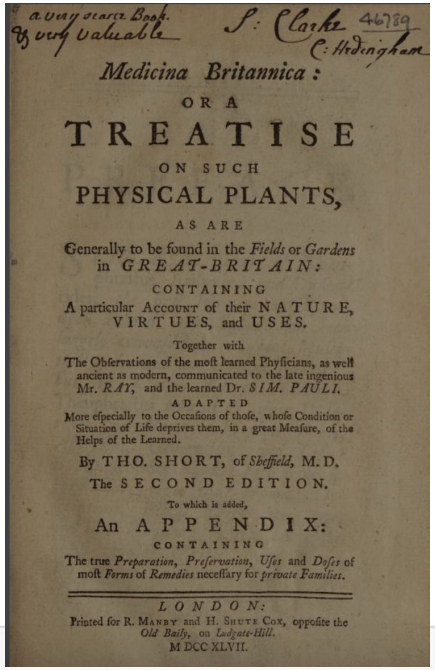
- Group effort in finding all the historical documents and placing them in a google sheets with links and as much metadata as we could find.
- Checked all gathered to make sure they fit criteria
- Downloaded and placed into the google drive to pass off to other group



On-going problems / projects

- Some books/articles were not in english
    - Working on translations so we can include these books
- Translation pdf → txt was challenging in some cases

On-going problems / projects



a very scarce Book. & very valuable S. Clarke 46789 C. Hedingham

Medicina Britannica:

OR A

TREATISE

ON SUCH

PHYSICAL PLANTS,

AS ARE

Generally to be found in the *Fields* or *Gardens* in *GREAT-BRITAIN:*

CONTAINING

A particular Account of their NATURE, VIRTUES, and USES.

Together with

The Obfervations of the moft learned Phyficians, as well ancient as modern, communicated to the late ingenious Mr. *RAY*, and the learned Dr. *SIM. PAULI.*

ADAPTED

More efpecially to the Occafions of thofe, whofe Condition or Situation of Life deprives them, in a great Meafure, of the Helps of the Learned.

By THO. SHORT, of *Sheffield*, M. D.

The SECOND EDITION.

To which is added,

An APPENDIX:

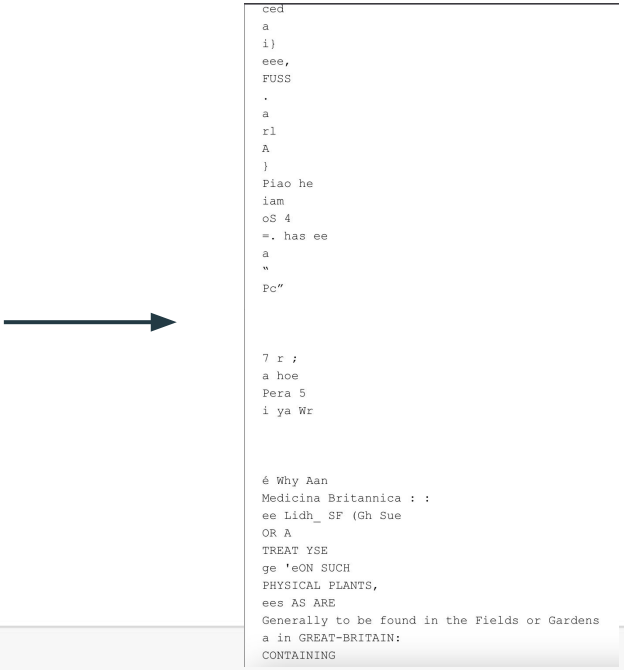
CONTAINING

The true *Preparation*, *Preservation*, *Ufes* and *Doſes* of moft *Forms* of *Remedies* neceſſary for *private Families.*

LONDON:

Printed for R. MANBY and H. SHUTE COX, oppoſite the *Old Bath*, on *Ludgate-Hill.*

M DCC XLVII.



```
ced
a
i}
eee,
FUSS
.
a
rl
A
}
Piao he
iam
oS 4
=. has ee
a
“
Pc”



7 r ;
a hoe
Pera 5
i ya Wr



é Why Aan
Medicina Britannica : :
ee Lidh_ SF (Gh Sue
OR A
TREAT YSE
ge 'eON SUCH
PHYSICAL PLANTS,
ees AS ARE
Generally to be found in the Fields or Gardens
a in GREAT-BRITAIN:
CONTAINING
```

```
In [16]: df.iloc[1,1]
```

```
Out[16]: '\' \n7 \n1 \n’ \n\' \ni, \n; \n7 \n5 \n’ \n& \n’ \n= \nai \nbo \npf: \nced \na \ni} \neee, \nFUSS \n. \na \nrl \nA
         \n} \nPiao he \niam \noS 4 \n=. has ee \na \n“ \nPc” \n\n\n\n7 r ; \na hoe \nPera 5 \ni ya Wr \n\n\n\né Why Aan \nM
         edicina Britannica : : \nee Lidh_ SF (Gh Sue \nOR A \nTREAT YSE \nge \'eON SUCH \nPHYSICAL PLANTS, \nees AS ARE \nG
         enerally to be found in the Fields or Gardens \na in GREAT-BRITAIN: \nCONTAINING \nA particular Account of their N
         A T UR EF, \nA ¢V RT US, and IS ES, \nTogether with \nThe Obfervations of the moft learned Phyficians, as welt \nan
         cient as modern, communicated to the late ingenious \nMr. RAY, and the learned Dr. SIM. PAULI. \nADAP:T ED \nMore e
```

Question:

Michael's group working on this

What were heavily documented plants in historical texts,

What purpose (if any) did these plants serve,

And can we find the metabolites that align with those purposes?

The “Goal”

- Collect as many plant metabolite papers as possible from modern databases.
- Extract metadata on the papers within the corpus.
- Refine the corpus to only include relevant hits.
  - Removing stop words
  - Removing punctuation

Approach in R

```r
library(rcrossref)
library(dplyr)
library(purrr)
#Dataframe creation of metadata----
fetch_papers <- function(keywords, num_results = 500) {
  query <- paste(keywords, collapse = " ")
  works <- cr_works(query = query, filter = c(type = "journal-article"), limit = num_results)

  if (!is.null(works) && !is.null(works$data) && nrow(works$data) > 0) {
    papers <- works$data %>%
      mutate(title = map_chr(title, ~as.character(.x[1])))
    return(papers)
  } else {
    print("No papers retrieved.")
    return(NULL)
  }
}

# Keywords for the search
keywords <- c("metabolite profile", "plants")

# Fetch papers based on keywords
papers <- fetch_papers(keywords)
```

Makes a dataframe of papers and associated metadata

```
> colnames(papers)
 [1] "alternative.id"      "container.title"     "created"             "deposited"           "published.online"    "doi"                 "indexed"
 [8] "issn"                "issue"               "issued"              "member"              "page"                "prefix"              "publisher"
[15] "score"               "source"              "reference.count"     "references.count"    "is.referenced.by.count" "subject"          "title"
[22] "type"                "url"                 "volume"              "abstract"            "language"            "short.container.title" "author"
[29] "link"                "license"             "reference"           "funder"              "published.print"     "update.policy"       "assertion"
[36] "archive"             "subtitle"            "update_to"
```

Corpus Collection in Python

```python
import requests
from Bio import Entrez

def fetch_papers_by_year(keywords, start_year, end_year, retmax=10, api_key=None):
    base_url = "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/"
    all_ids_by_year = {}

    #Loop through by year and collect paper IDS through the api key
    for year in range(start_year, end_year + 1):
        for retstart in range(0, 9999, retmax):
            search_url = f"{base_url}esearch.fcgi?db=pubmed&term={'%20'.join(keywords)}&retmax={retmax}&retstart={retstart}&

            if api_key:
                search_url += f"&api_key={api_key}"

            search_response = requests.get(search_url)
            search_data = search_response.json()

            # Check for errors in the response
            if "ERROR" in search_data["esearchresult"]:
                print(search_data["esearchresult"]["ERROR"])
                break

            current_ids = search_data["esearchresult"]["idlist"]
            all_ids_by_year.setdefault(year, []).extend(current_ids)

            # If fewer IDs than retmax were returned, we've reached the end and can stop fetching
            if len(current_ids) < retmax:
                break

    print("IDs retrieved for each year:")
    for year, ids in all_ids_by_year.items():
        print(f"{year}: {len(ids)} IDs")

    return all_ids_by_year
```

Loops by year to bypass the max api collection limit

Setting up for running the fetching function

Sets year / keywords for loop

```python
# Example usage
api_key = '2a712e0d47f3436ef738ec764ade7b1bee09'
start_year = 2000
end_year = 2023
keywords = ["metabolite", "plant"]
papers_by_year = fetch_papers_by_year(keywords, start_year, end_year, retmax=10, api_key=api_key)
```

Adjust year and search words

Python Metadata Collection

```
# Now you can loop through the papers_by_year dictionary and fetch metadata for each year
#for year, pubmed_ids in papers_by_year.items():
    #papers_metadata = fetch_metadata(pubmed_ids)

    #if papers_metadata:
        #print(f"Year: {year}")
        #for pubmed_id, data in papers_metadata.items():
            #print(f"PubMed ID: {pubmed_id}")
            #print(f"Abstract: {data.get('Abstract', 'N/A')}")
            #print(f"Funding: {', '.join(data.get('Funding', ['N/A']))}")
            #print(f"Citations: {data.get('Citations', 'N/A')}")
            #print("\n")
```

This is where I left off.

- Metadata extraction works for Abstract, Title, PubmedID % Publisher
- Citations & Funding are not extracting

Important Pieces of Metadata:

- Reference count
- Referenced by
- Region of publication

Next Step:

- **Assemble corpus by extracting abstracts from NCBI scraping**

Data Visualization

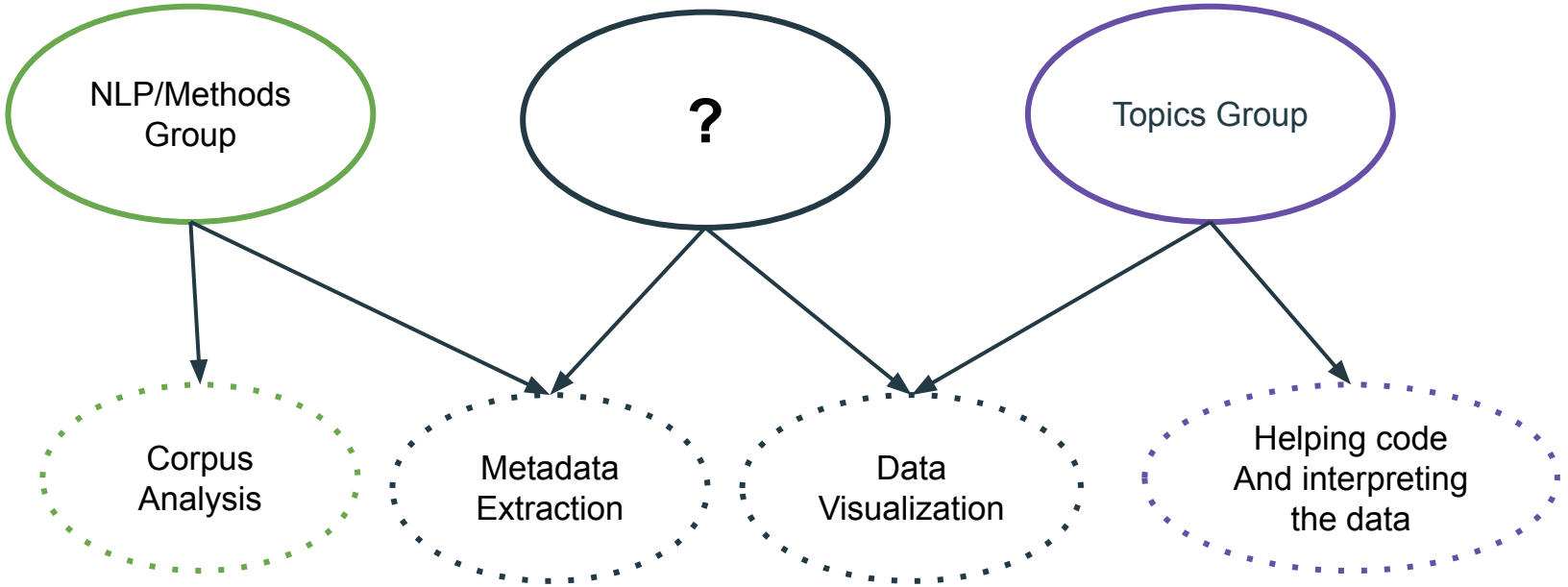
			8 Continents (let me update zone of countries again)													
	Group of function	Total Number of publication	Asia		Africa		North America		South America		Antarctica		Europe		Australia/Oceania	
			2000-2010	2011-2020	2000-2010	2011-2020	2000-2010	2011-2020	2000-2010	2011-2020	2000-2010	2011-2020	2000-2010	2011-2020	2000-2010	2011-2020
1	Plant survial and alive (ex.hibit herbivor eating, promote growth)															
2	Pharmaco(drug-like) (ex. Cocian, selerarinone)															
3	Traditional herbal (ex. eugenol (clove oil)															
4	Food															
5																

Data Visualization

	Class of metabolite		list	Plants family	Function
1	Alkaloids				
		1	Pyrrolizidine alkaloid		defense mechanism against insect herbivores
		2	Tropane alkaloids	Solanaceae	act as anticholinergic effect on central nervouse system
		3	Mescaline		
2	alkylamides				
3	Amines				
4	Carbohydrates and organic acids				
5	Cyanogenic glycosides				
6	Flavonoids and Tannins				
7	Glucosinolates				
8	Lectins, peptides and polypeptides				
9	Non-protein amino acids (NPAAs)				
10	Phenylpropanoids, lignins, coumarins and lignans				
11	Polyacetylenes, fatty acids and waxes				
12	Polyketides				
13	Steroids and saponins				
14	Terpenes				

What comes next?

What's Next: Workflow



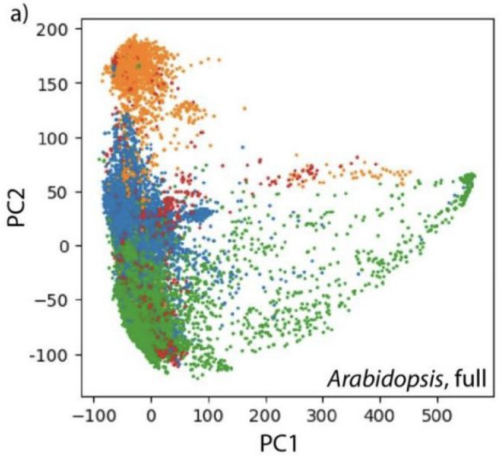
What's Next: Hypothesis

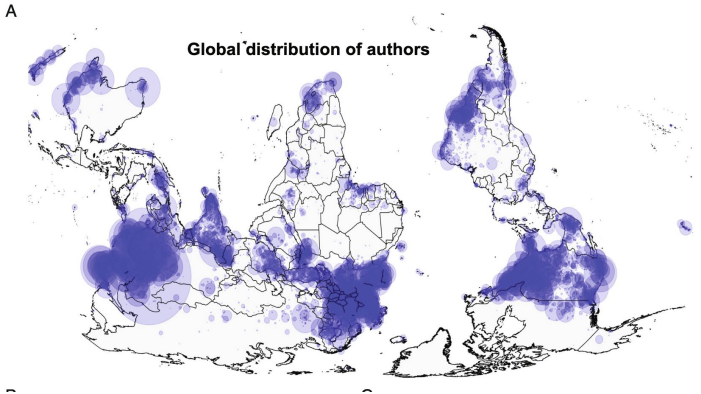
- Do we still like this hypothesis?
- Are there more specific questions we want to answer?
- What kinds of questions do we want to ask of our corpus?
    - What keywords should we use to create subgroups or clusters?

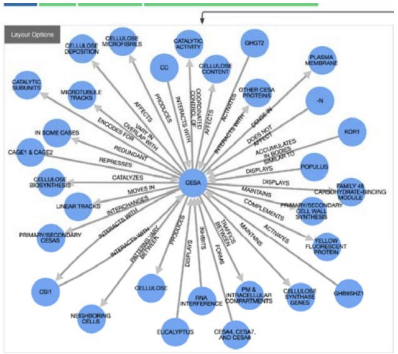
What's Next: Open Questions

- How do we pull whole papers, not just abstracts, to extract metabolite profiles?
- How do we interface between historical names for things and modern day names for things?
    - Find sources for common names of plants
        - Possibly IPNI or SciName Finder but need to check how comprehensive they are with a list of plants
- How often do nonsensical characters show up in the historical texts?

What's Next: Data Visualization







Preliminary Data Visualization Ideas

Need to see the data before finalizing but:

- World map comparing what purposes each region had for studying plants (little pie charts on continents similar to the authorship paper).
  - Medicinal
  - Agricultural
  - Religious
  - Etc.
- Word clouds of metabolites that show up often
- Heatmap of metabolites or common plans or common uses?
- World map for visualizing the plants used on each continent to treat similar diseases. And if possible, include the present metabolites and these plants.