The Data Incubator | Capstone proposal
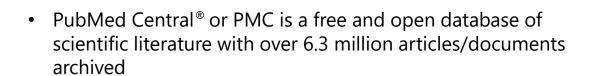
Cluster and Visualize PMC search results

Arpan S. Neupane

Motivation

1

Form a research question
"I wonder how the immune responses in the lung function?"

2

- PubMed Central® or PMC is a free and open database of scientific literature with over 6.3 million articles/documents archived
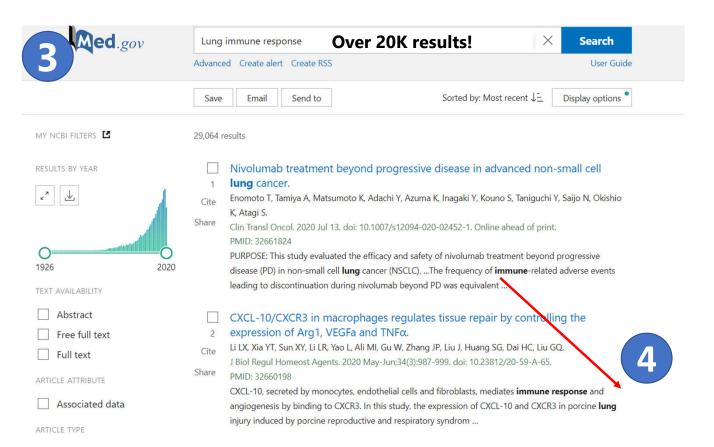
- The FIRST place most students or researchers start their query

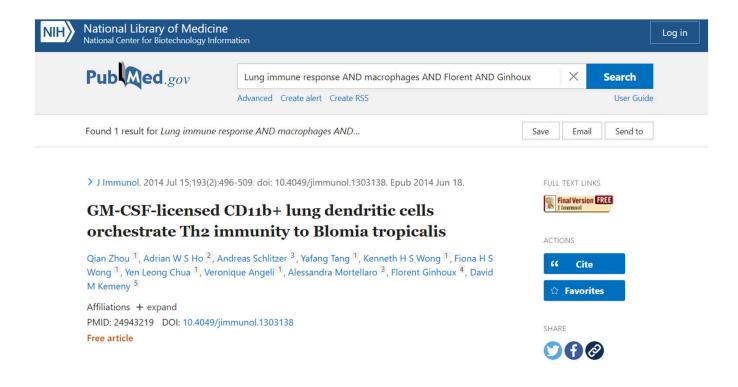- Most likely, the FIRST google result will also take you to PMC

Time to open lots of TABS!!!

Current solutions

- Narrowing the search query!
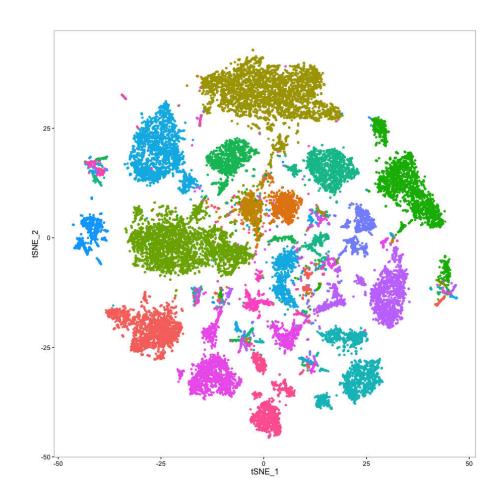- May not actually give you want you want!

Why is it an issue worth tackling?

- Time lost reading one article after another to identify "similar" ones

- Missed opportunities at identifying articles with similar findings if results are in different pages. Page 1 vs. Page 10 vs. Page 200

- Financial repercussion of potentially (and unnecessarily) repeating an experiment/study, wasting valuable resources!
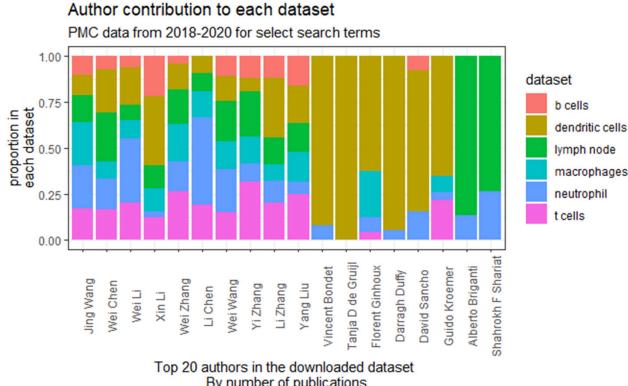
Proposed solution

- Cluster results from PMC based on topics

- Visualize

- Each "island" would house information for articles that may provide similar information

- Overall act as a first pass filter

Satijalab.org

A little bit about the data and Initial exploration

- Utilized **R** and **RISmed** to gather ~25K individual PMC records

- Query terms: b cells, dendritic cells, macrophages, lymph node, neutrophils, t cells

- Published: 2018-2020

Author contribution to each dataset

PMC data from 2018-2020 for select search terms
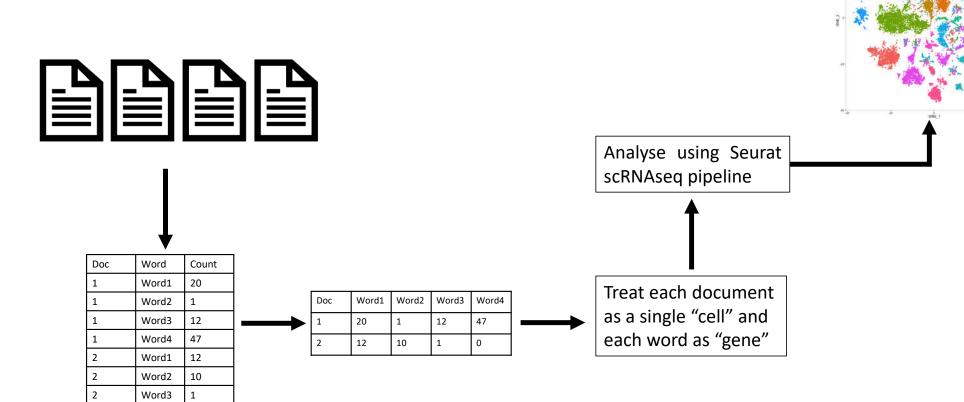
Top 20 authors in the downloaded dataset
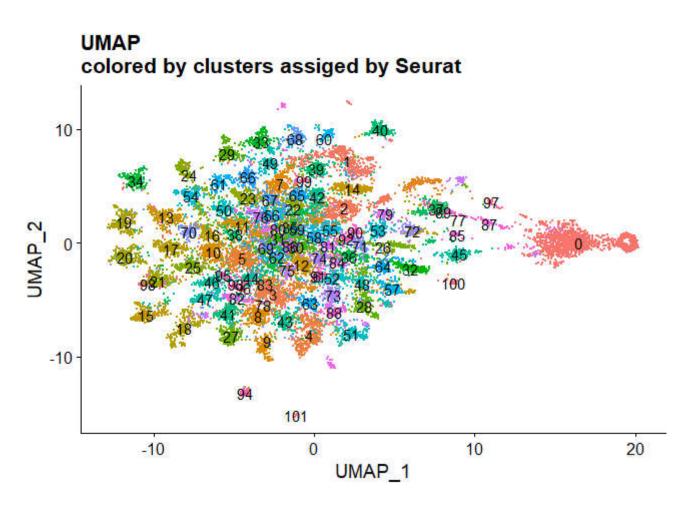By number of publications
Authorname:Last First

Data (~5000) pertaining to search terms (dataset) were downloaded from PMC on 2020-06-23.

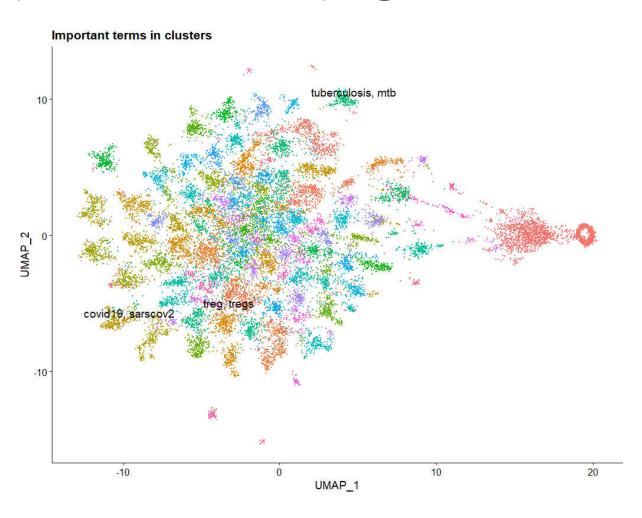Does not fully reflect all publications from 2018-2020

Analysis

Doc	Word	Count
1	Word1	20
1	Word2	1
1	Word3	12
1	Word4	47
2	Word1	12
2	Word2	10
2	Word3	1

Doc	Word1	Word2	Word3	Word4
1	20	1	12	47
2	12	10	1	0

Treat each document as a single "cell" and each word as "gene"

Analyse using Seurat scRNAseq pipeline
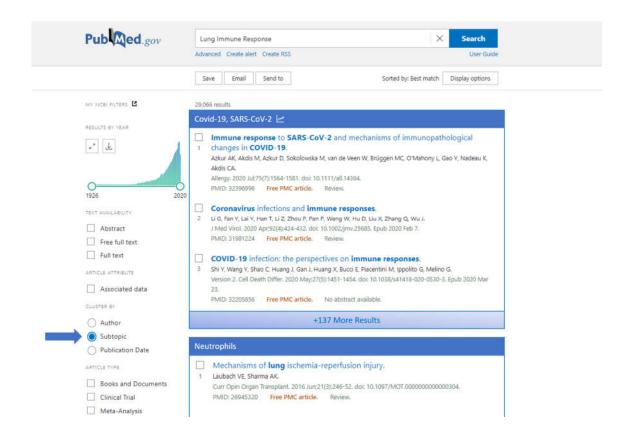
Clustering all documents

Why did they cluster? Identifying topics

Important terms in clusters

Not limited to "visualization" only

Who would this benefit?

- Students and Researchers

- Companies looking to market/cater to specific type of research

- Public! Visual patterns are much more interesting than simply reading texts

☺ Thank you