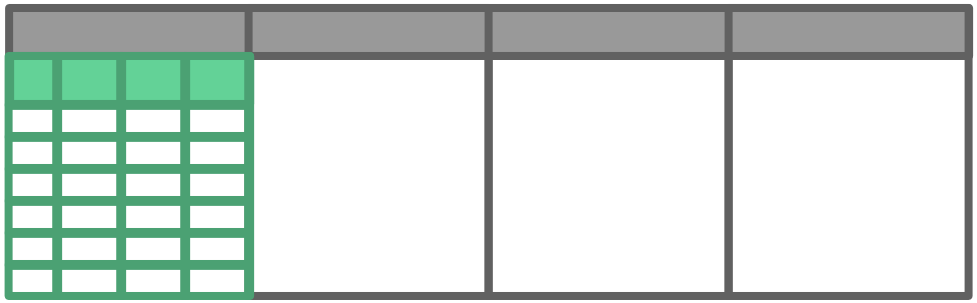
Scholarly Data Recommender

Andrew Yeh

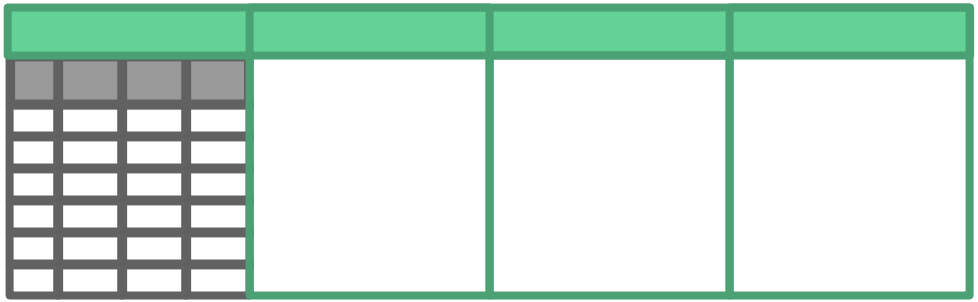
Normal Data vs. Scholarly Data

Normal data is what we usually work with.

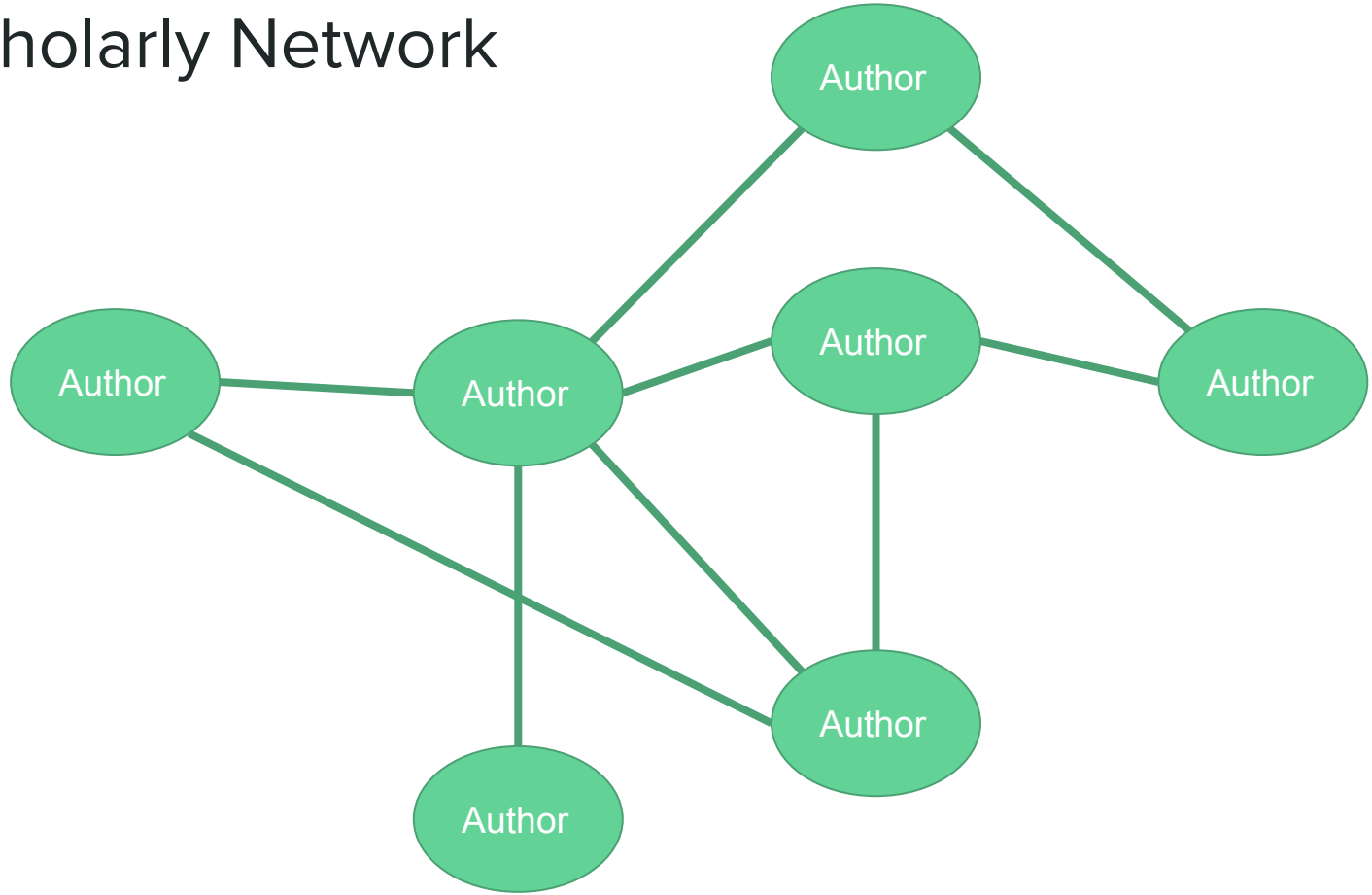


Normal Data vs. Scholarly Data

“Scholarly” data is the metadata of the project, such as the author, publication, citations, etc.



Scholarly Network



Data Source

National Center for Biotechnology Information

Database on papers relating to:

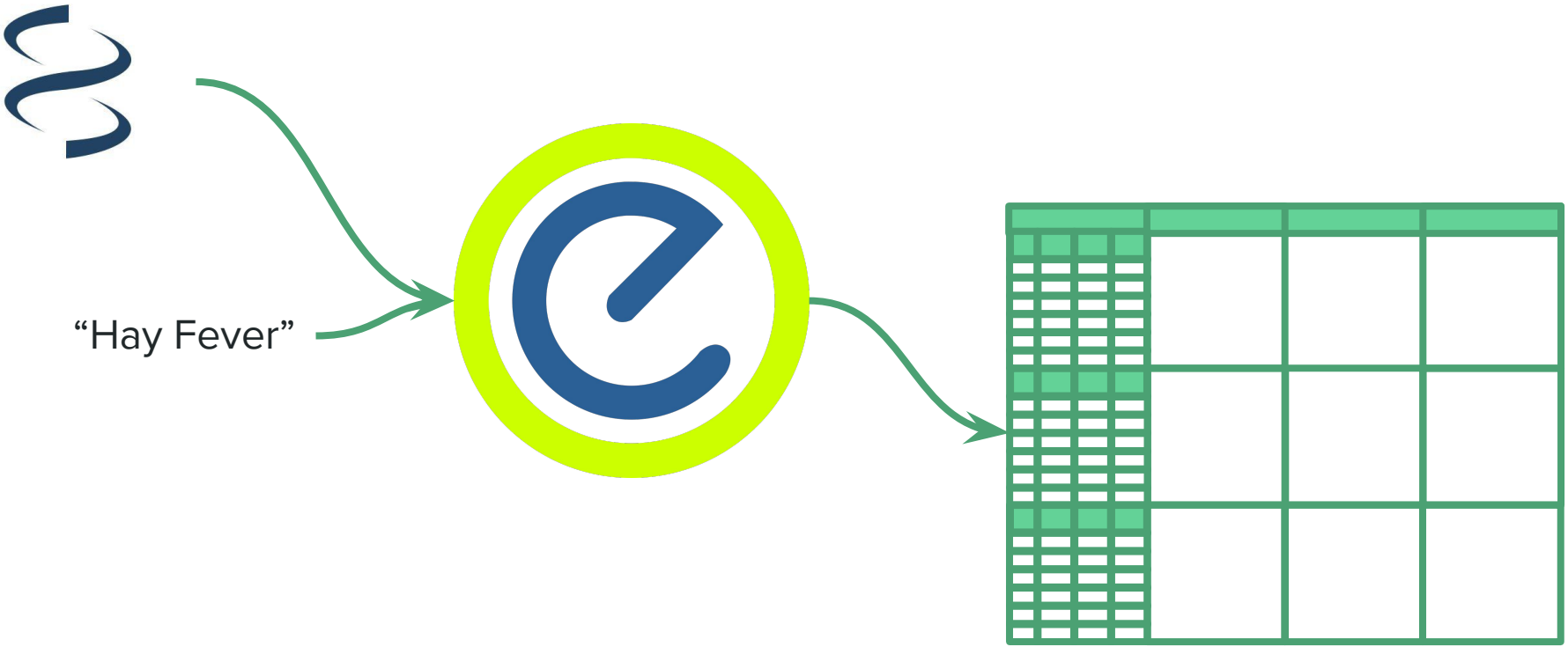
- Biotechnology
- Biomedicine
- Bioinformatics

Data Extraction

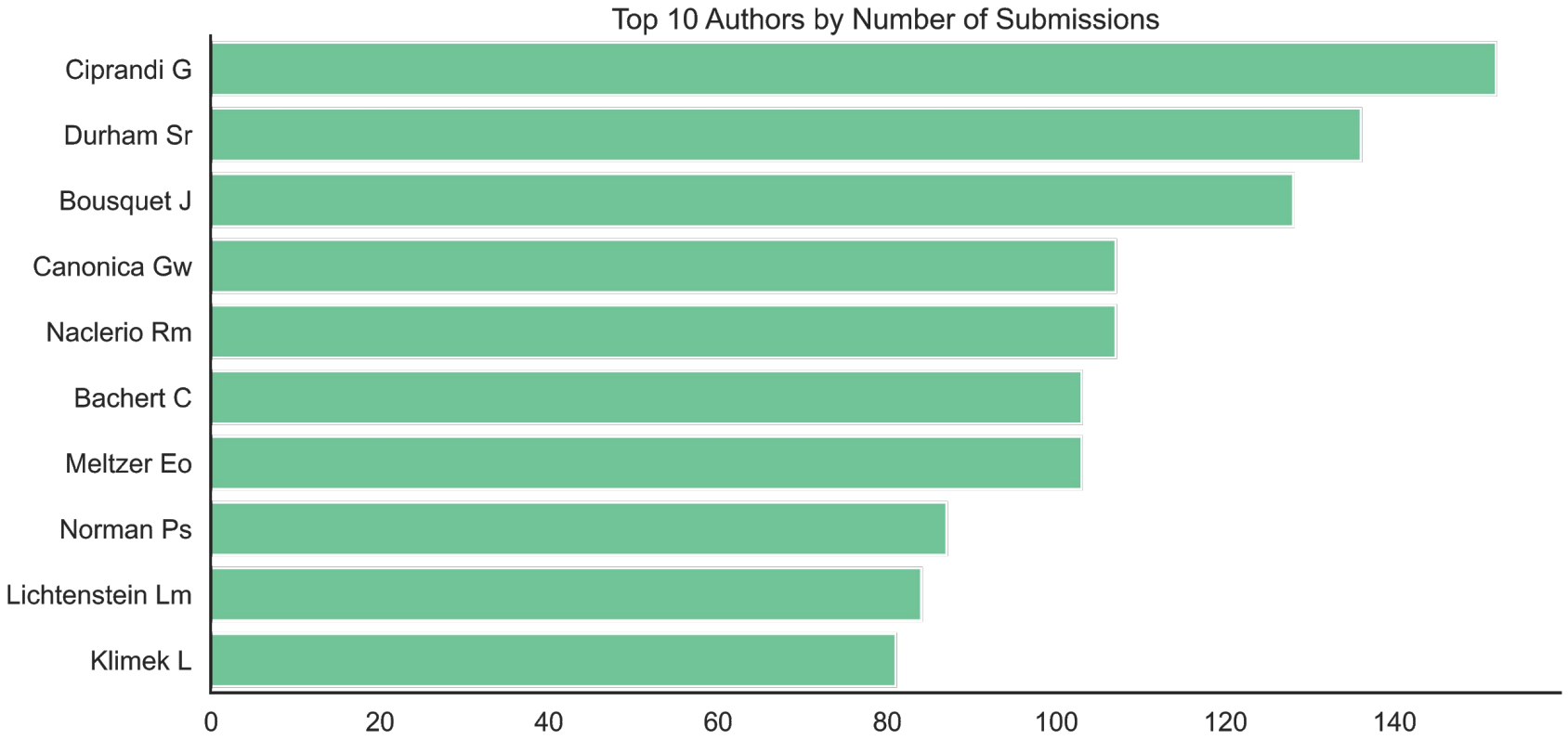
Entrez

NCBI’s primary text search and retrieval system

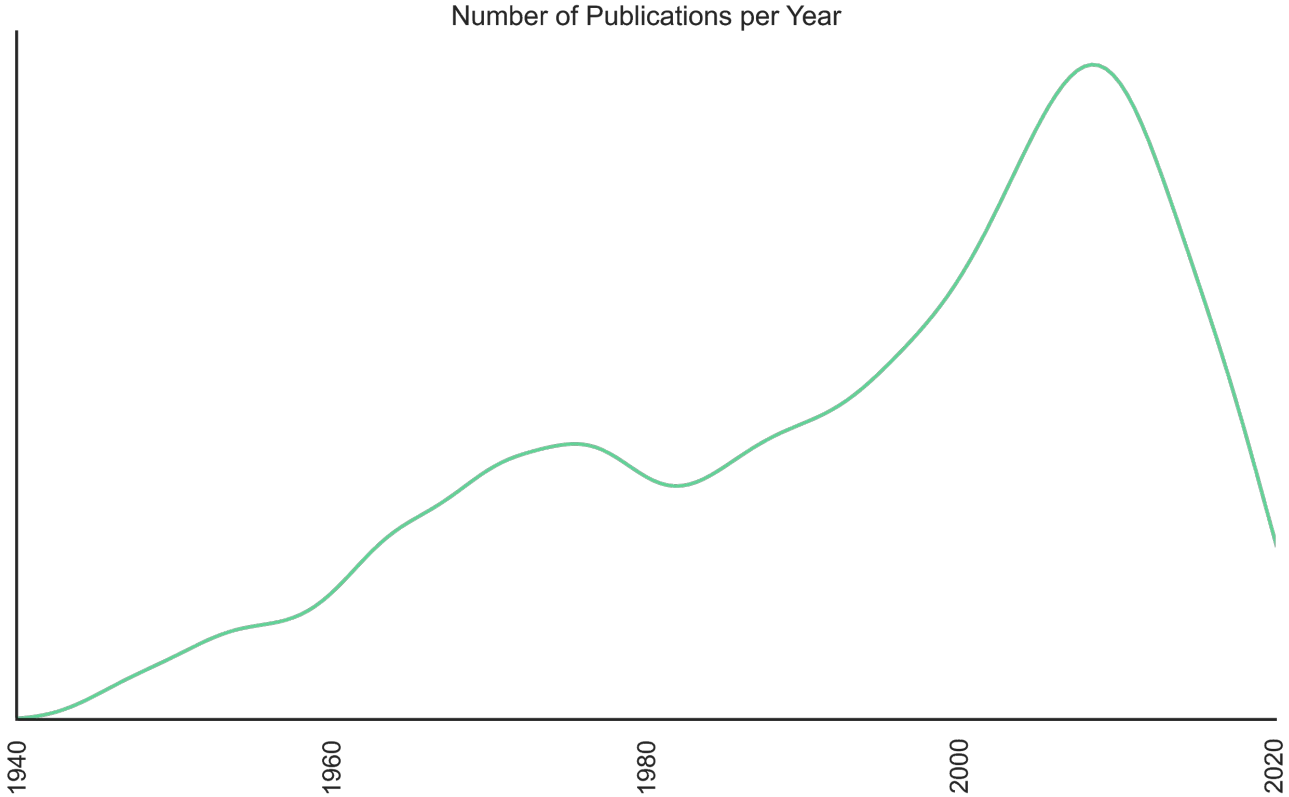
Data Scraping Process



Exploratory Data Analysis

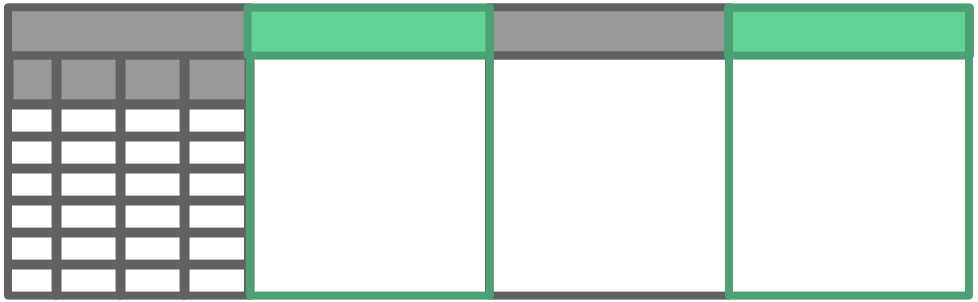


Exploratory Data Analysis



Data Preprocessing

Not all metadata is needed



Recommending Significant Articles

What makes a document significant?

- ‘Significant’ articles are ones that made an impact in the academic community
- Highly referenced documents are the building blocks of scholarly networks

Recommending Significant Articles

Problem: Not all documents have reference numbers

Solution: Utilize NLP and a logistic regression classifier to predict if a given document could be considered ‘significant’ given its key terms

- Precision is our sole focus
- Model ended up with a precision score of 0.768

Choosing Significant Articles

MetaData	Values
Title	Allergic Rhinitis and its Impact on Asthma
Authors	J Bousquet, P Van Cauwenberge, N Khaltaev, J Denburg, ...
Journal Published	Allergy
Location	Denmark
Language	English
Year Published	2008

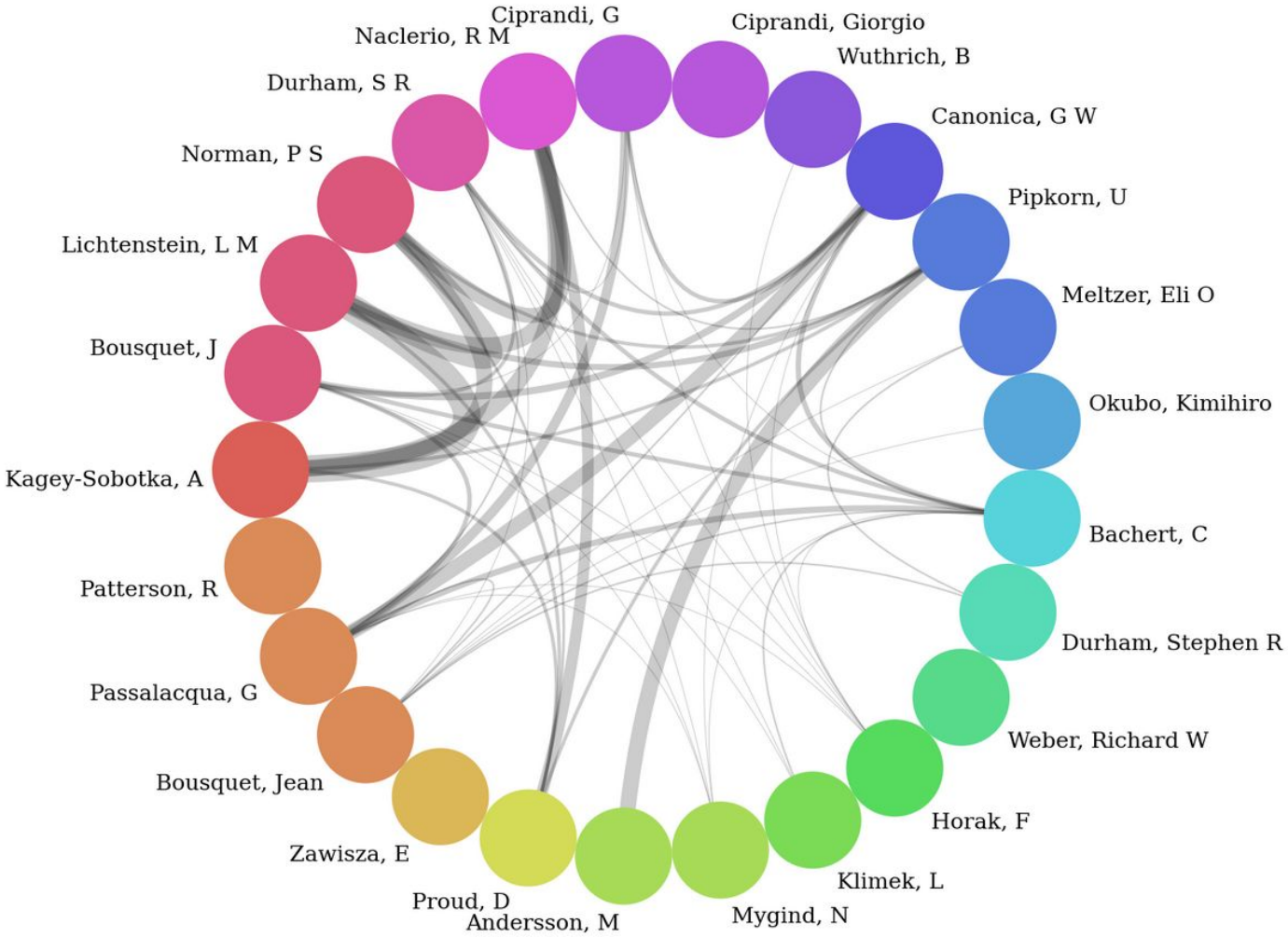
Further Suggestions

Utilize cosine similarities to perform content-based recommendations

	Titles
Recommendation 1	Local allergic rhinitis: concept, clinical manifestations, and diagnostic approach.
Recommendation 2	Managing hay fever during the exam period.
Recommendation 3	Prevalence of asthma, respiratory symptoms and allergic disorders among adolescents in the province of Aquila
Recommendation 4	Otitis media and eustachian tube dysfunction

Explanatory Data Analysis

In scholarly data, exploratory data analysis is for initial data visualization while explanatory data analysis is for data visualization of the results.


Ciprandi, G
Ciprandi, Giorgio
Naclerio, R M
Wuthrich, B
Durham, S R
Canonica, G W
Norman, P S
Pipkorn, U
Lichtenstein, L M
Meltzer, Eli O
Bousquet, J
Okubo, Kimihiro
Kagey-Sobotka, A
Bachert, C
Patterson, R
Durham, Stephen R
Passalacqua, G
Weber, Richard W
Bousquet, Jean
Horak, F
Zawisza, E
Klimek, L
Proud, D
Mygind, N
Andersson, M


J Investig Allergol Clin Immunol
Allergy
Nurs Times
J Allergy Clin Immunol
Epidemiol Prev
Sastre J
Mullol J
Davila I
Bartra J
Valero A
Passalacqua G
Durham Sr
Bachert C
Canonica Gw
Bousquet J
Pearce L
Potterton D
Frankland Aw
Watson Rt
Scadding Gk
Norman Ps
Lichtenstein Lm
Valenta R
Melzer Eo
Mattei A
Angelone Am
Di Stefano R
Sbarbati M
Cialfi D

Future Steps and Post-MVP Goals

1. More visuals, and learn about network plots
2. Implementing using the abstract in the NLP model as well

Contacts - Andrew Yeh

LinkedIn:

https://www.linkedin.com/in/andrew-yeh-3a3649117/

GitHub:

https://github.com/awyeh64/