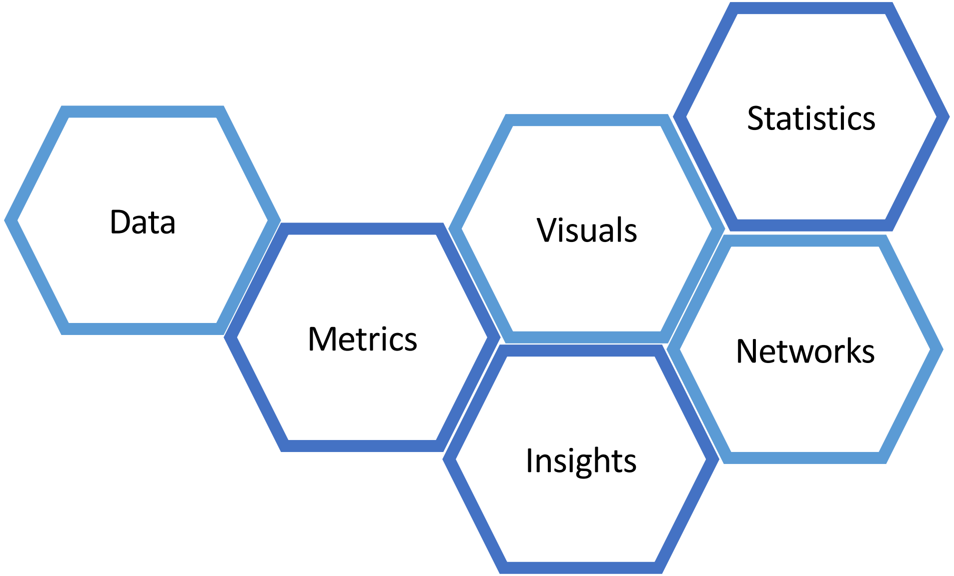
Citation Networks and Impact Analysis

National Library of Medicine

Isaac Robinson (CDF)
Harvard University, Mathematics, 2022

Special Thanks:
Anna Calcagno (NLM)
Kenneth Wilkins (NIDDK)



Data

Metrics

Visuals

Insights

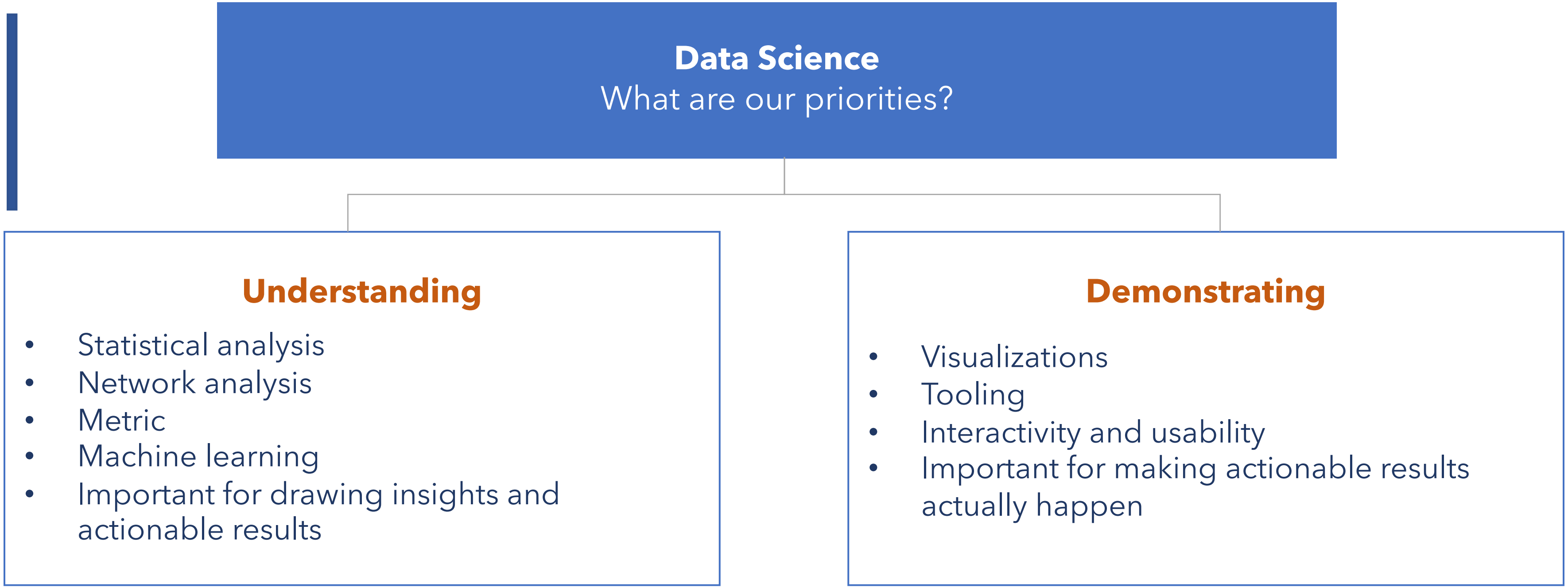
Statistics

Networks

NIH National Institutes of Health

PROJECT OBJECTIVES

1. Aggregate public and private data sources to provide wholistic and selectively accessible information
2. Create multiplex citation graph pipeline for future research
3. Suggest new impact metrics for funding institutions with non-traditional risk profiles
4. Package the above into easily digestible formats for technical and non-technical users



Data Science
What are our priorities?

Understanding

- Statistical analysis
- Network analysis
- Metric
- Machine learning
- Important for drawing insights and actionable results

Demonstrating

- Visualizations
- Tooling
- Interactivity and usability
- Important for making actionable results actually happen

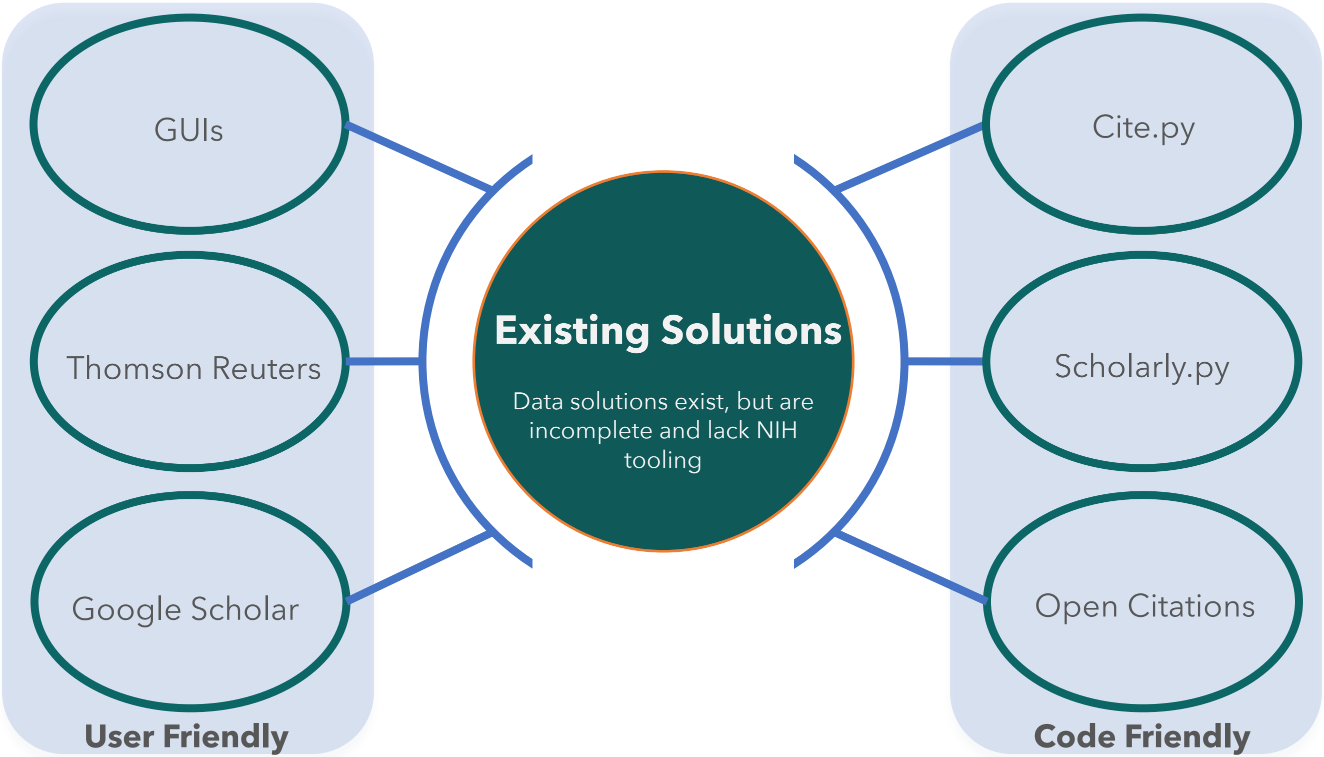
Key Takeaway

Data analysis can be a bit obtuse to those outside of a project, but those people are essential for generating insights.

NIH National Institutes of Health

Step 1: Getting the Data

EXISTING OPTIONS



What to do :/

Introducing iCiteNLM

1. A library for curating all the data you need to build citation graphs
2. Accesses multiple pipelines including Google Scholar, Open Citations, and more.
3. Provides general tooling as well as NIH-specific functions, such as PMID to DOI conversions
4. Lots of room to expand and include NIH-specific resources like InCite

Introducing iCiteNLM

Curate

A library for curating all the data you need to build citation graphs

Private-Public

Accesses multiple open access pipelines including Google Scholar and Open Citations, as well as support for private options

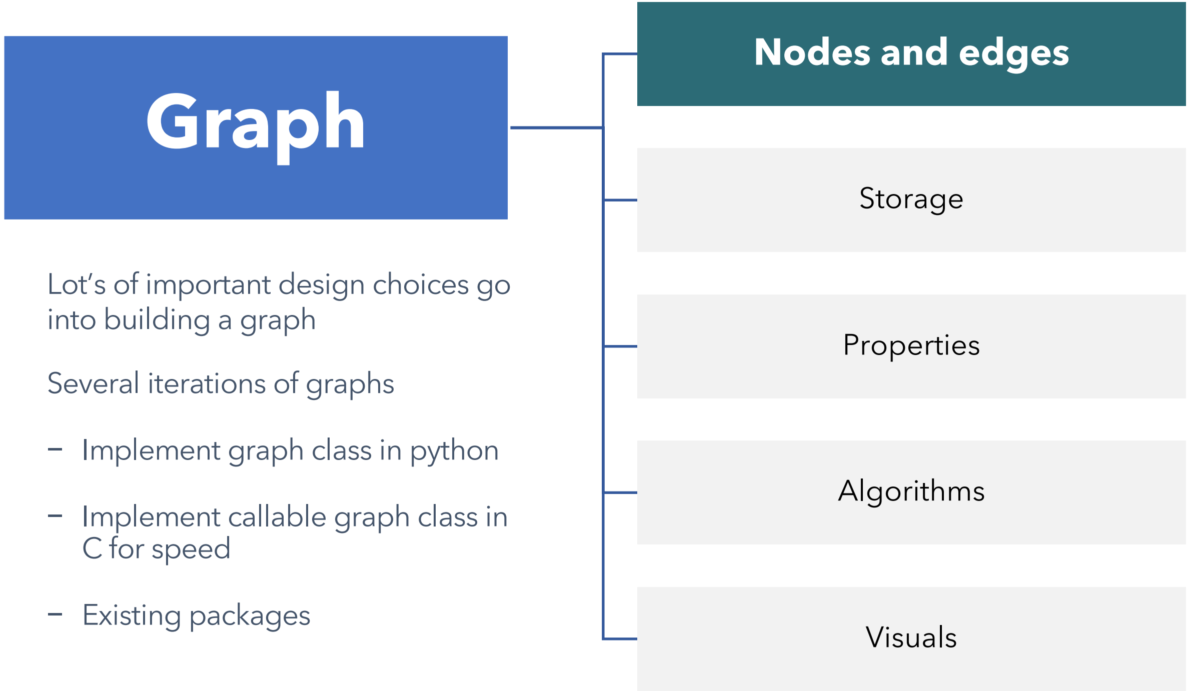
By NIH, for NIH

Provides general tooling as well as NIH-specific functions, such as PMID to DOI conversions

Now add it to the graph!

Step 2: Create the Graph

Graph Basics



Graph

Nodes and edges

Storage

Properties

Algorithms

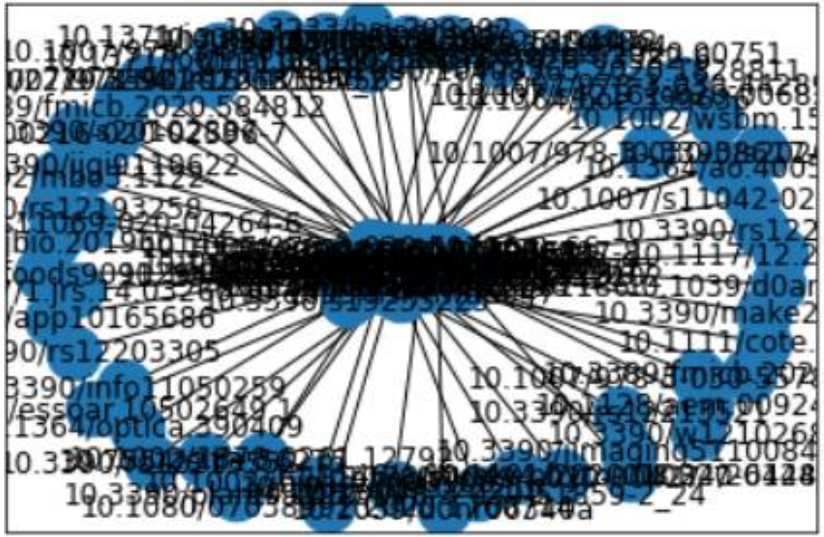
Visuals

- Lot’s of important design choices go into building a graph
- Several iterations of graphs
  - Implement graph class in python
  - Implement callable graph class in C for speed
  - Existing packages

Visualizing the Graph

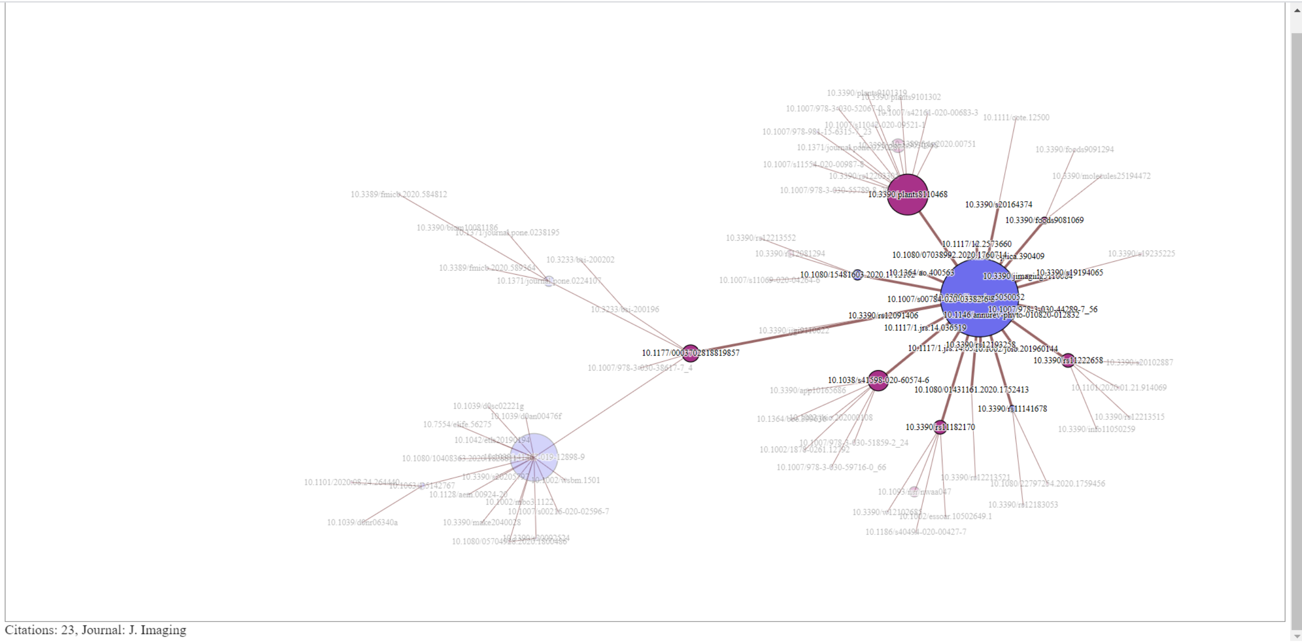
Pre-Built Solutions

- Most graphs used for **data science** are only displayed for exploratory purposes
- The accepted solution is matplotlib
- Hard to understand, only manipulatable using code, limited controllability
- Poor node positioning

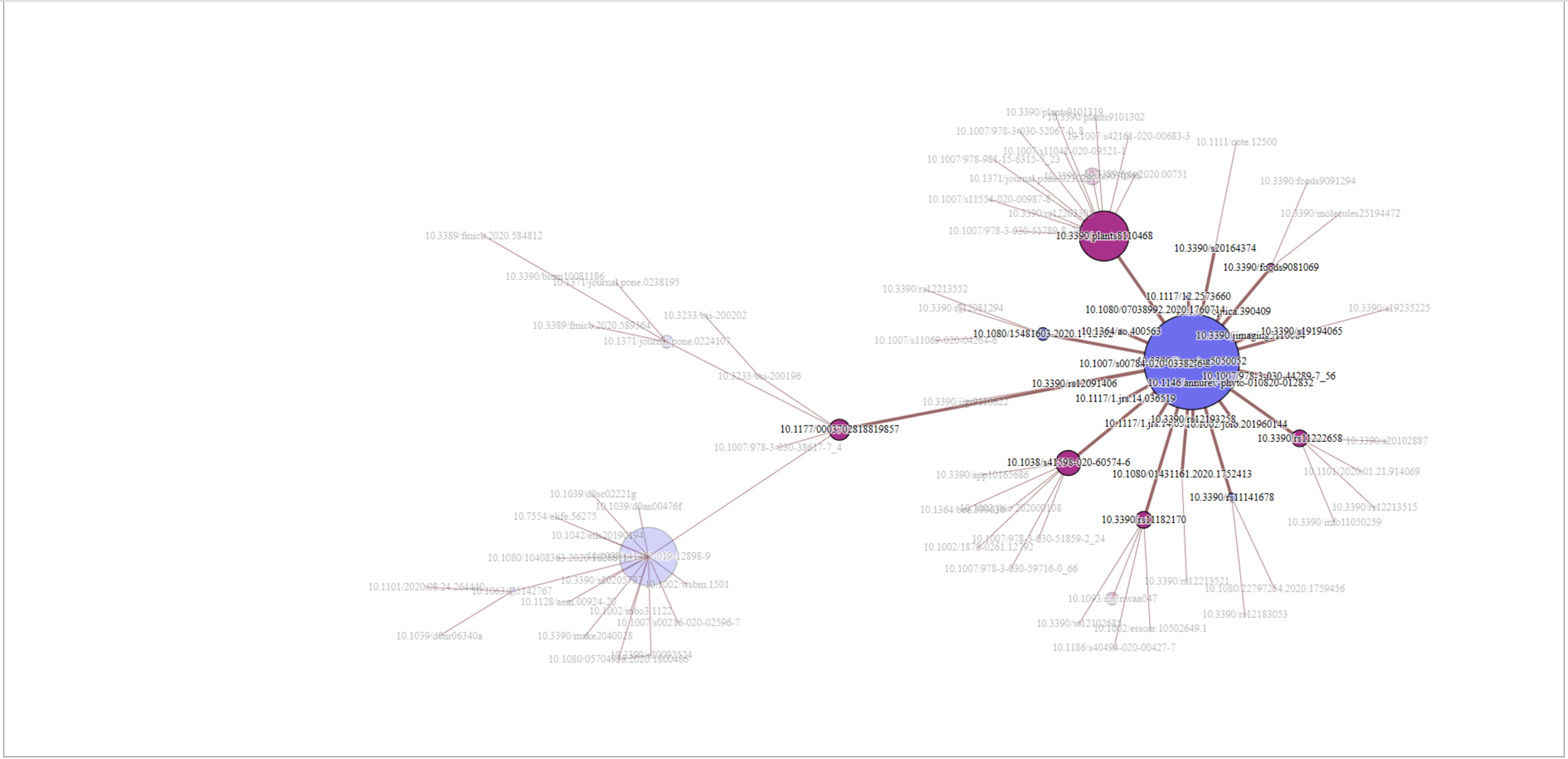


My Solution: Web-Server, HMTL, D3

- Looked for examples of the prettiest visuals and then found what they used
    - NYT + D3
    - Fell in love with the library
- A DOM-manipulator, making it easy to launch (as easy as launching a website)
- Any desired features can be added, or embed as part of a dash



Citations: 23, Journal: J. Imaging



Citations: 23, Journal: J. Imaging

NIH National Institutes of Health

Step 3: Metrics

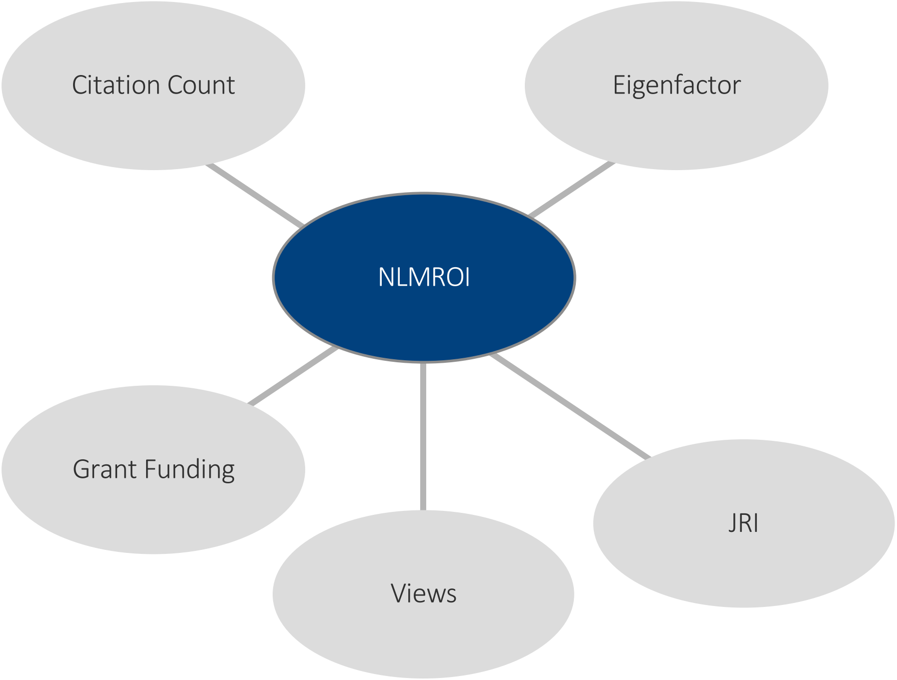
Introducing NLMROI

1. Citation count
2. Eigenfactor
3. JRI
4. Views and Web2.0 Metrics
5. Grant funding

The risk profile for governmental organizations necessitates a different metric

Existing Metrics

1. Citation count
2. Eigenfactor
3. JRI
4. Views and Web2.0 Metrics
5. Grant funding



CIVIC DIGITAL FELLOWSHIP

Introducing NLMROI

$$NLMROI = \frac{\sum C_i E_{c_i} + \sum \sqrt{C_j} E_{cj} + \sum \sqrt[3]{C_k} E_{c_k}}{\sum Funding}$$

- Government-specific metric
- Consistent with existing literature and metrics including Eigenfactor, JRI, Impact Factor etc.
- Helps account for differences in research areas (still best to compare within one area)

Now add it to the graph!

Step 4: Usability

Outputs

1. Graph written in D3 and hosted locally means it is easily web-publishable
    1. Open-source/publicly available options of iCiteNLM makes data privacy less of an issue and means anyone can access and see this data
2. For those interested, a host of algorithms were implemented on the graph including clustering, k-cliques, shortest-path, centrality, and more
    1. These can all be accessed through the python module
3. Excel file has been formatted for easy analysis

Next Steps

Only the beginning

1. Any metric you can imagine can be added to the graph
    1. Parallel graphs for co-authorship
    2. Weighing for small vs large teams
2. Include funding information for 2nd, 3rd degree nodes
    1. Validate metric using studies and experts
3. Host graph on data servers and expand to higher degrees
4. Any other graph analysis technique you would like!
    1. Causal inference
    2. Dominating sets
    3. Network flow

CIVIC DIGITAL FELLOWSHIP

Other Tail Ends

1. Topic modeling with Latent Dirichlet Allocation
2. NER with Bert-based machine learning models
3. Kleinberg burst-detection
4. Excel files

CIVIC
DIGITAL
FELLOWSHIP

Thank you to everyone for this opportunity!

Please reach out if you have questions/comments, or if you think any of these techniques could be useful in other areas!

Email: isaac_robinson@college.harvard.edu

Sources

Leveraging Citation Networks to Visualize Scholarly Influence Over Time
J. Portenoy, J. Hullman, J.D. West (2016)
arXiv: 1611.07135

Static ranking of scholarly papers using article-level Eigenfactor (ALEF)
I. Wesley-Smith, C. T. Bergstrom, and J. D. West (2016)
The 9th ACM International Conference on Web Search and Data Mining (WSDM)

CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature
C Chen
Journal of the American Society for information Science …, 2006

Author-Level Eigenfactor Metrics: Evaluating the Influence of Authors Institutions and Countries Within the SSRN Community.
J.D. West, M.C. Jensen, R.J. Dandrea, G.J. Gordon, and C.T. Bergstrom, (2013)
Journal of the American Society of Information Science and Technology 64: 787-801

Journal impact evaluation: a webometric perspective
Thelwall, M.
Scientometrics 92, 429–441 (2012)

Sources and Papers

Eigenfactor: ranking and mapping scientific knowledge
J.D. West, (2010)
PhD Dissertation. University of Washington
Big Macs and Eigenfactor Scores: Don't Let Correlation Coefficients Fool You
J.D. West, T.C. Bergstrom, C.T. Bergstrom, (2010)
Journal of the American Society for Information Science & Technology. 61(9): 1800-1807

How to improve the use of metrics: Learn from Game Theory
J.D. West, (2010)
Nature 465:871-872 [pdf]
Network Analysis and Indicators
Ding Y., Rousseau R., Wolfram D.

Causal Inference for Social Network Data
Elizabeth L. Ogburn Department of Biostatistics

Searching for intellectual turning points: Progressive knowledge domain visualization
C Chen
Proceedings of the National Academy of Sciences, 2004

Sources and Papers

Impact factor: a valid measure of journal quality?
Saha, Somnath et al.
Journal of the Medical Library Association : JMLA vol. 91,1 (2003): 42-6.

Response to "Big Macs and Eigenfactor Scores: The Correlation Conundrum"
J.D. West, T.C. Bergstrom, C.T. Bergstrom, (2010)
Journal of the American Society for Information Science & Technology 61:2592

Dominating Sets in Social Network Graphs
Laura L. Kelleher, Margaret B. Cozzens,
Mathematical Social Sciences,
Volume 16, Issue 3, 1988