Development of an Automated Workflow for Analysis of Data Availability Statements in NIEHS-Sponsored Publications

National Institute of Environmental Health Sciences (NIEHS):
Program Analysis Branch

Kristi Pettibone — PhD

Christie Drew — PhD

Chris Duncan - PhD

CIVIC DIGITAL FELLOWSHIP

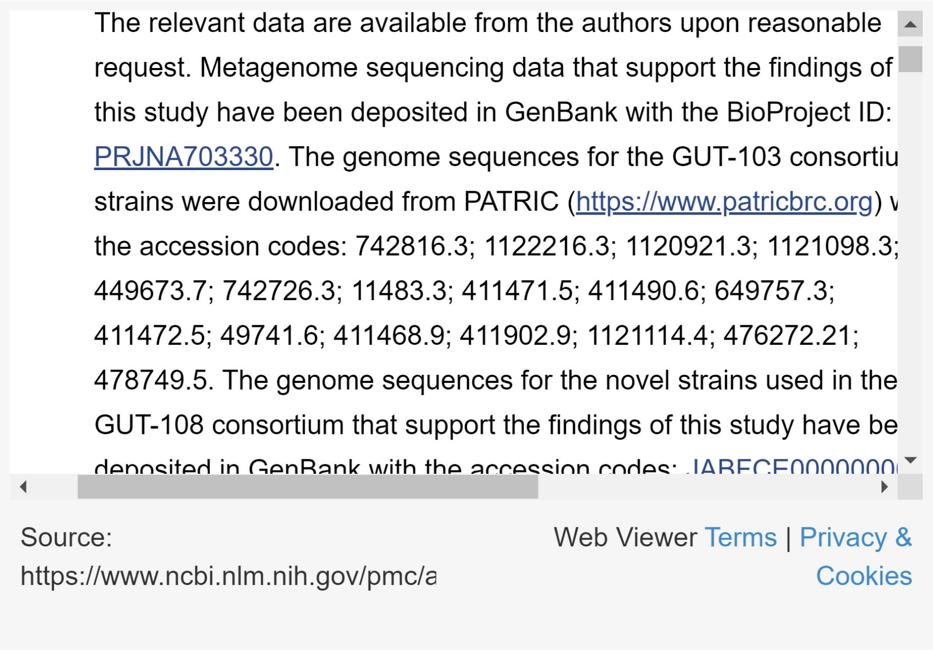
NIH

Kashyap Sreeram

Duke University

Bioinformatic Systems in Neurogenetics
and Mental Health Policy

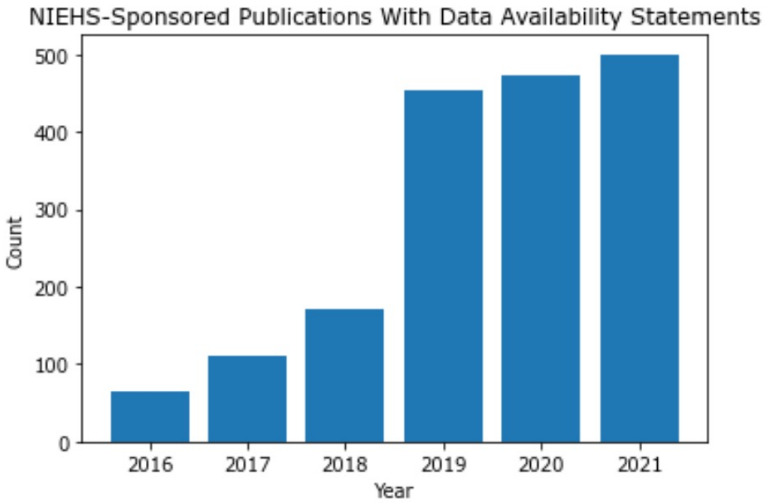
DATA AVAILABILITY STATEMENTS



The relevant data are available from the authors upon reasonable request. Metagenome sequencing data that support the findings of this study have been deposited in GenBank with the BioProject ID: PRJNA703330. The genome sequences for the GUT-103 consortiu strains were downloaded from PATRIC (https://www.patricbrc.org) v the accession codes: 742816.3; 1122216.3; 1120921.3; 1121098.3; 449673.7; 742726.3; 11483.3; 411471.5; 411490.6; 649757.3; 411472.5; 49741.6; 411468.9; 411902.9; 1121114.4; 476272.21; 478749.5. The genome sequences for the novel strains used in the GUT-108 consortium that support the findings of this study have be deposited in GenBank with the accession codes: JABFCE0000000

Source: https://www.ncbi.nlm.nih.gov/pmc/a

epresentation of Data Availability through Data Availability Statements (DAS)



Number of Publications with DAS in 2021 is 5x the number in 2016 → Merits an analysis of proper data availability

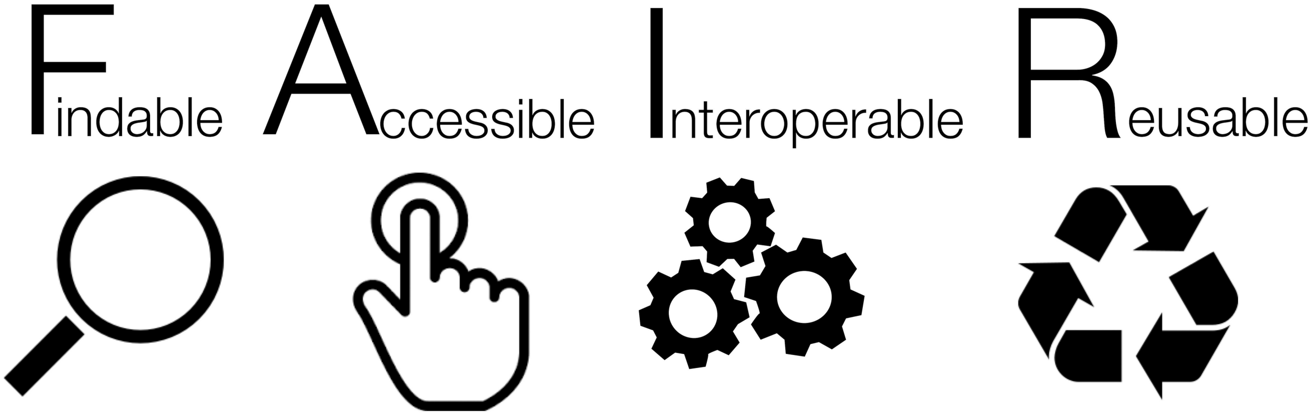
FAIR PRINCIPLES OF DATA SHARING

F – Are the Metadata and data easy to find for humans and computers?

A – Can the data be properly accessed using a standardized protocol?

I – Do the data use broadly applicable language and/or integrate with other data?

R – Are the data able to be replicated/optimized for reuse in different settings?



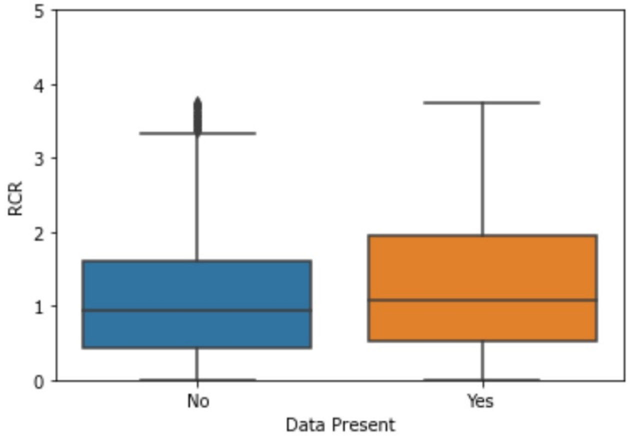
CIVIC DIGITAL FELLOWSHIP

KEY QUESTIONS

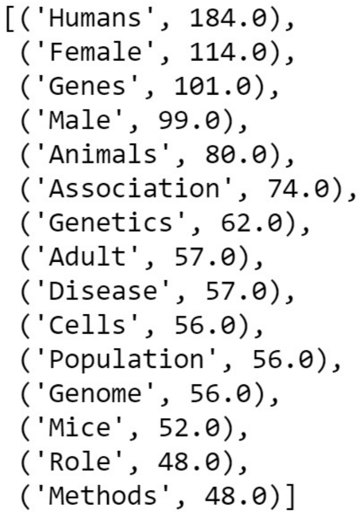
- **<u>Higher-Level Question:</u>** What do authors say in their Data Availability Statements?
- **<u>Secondary Questions:</u>**
  - Do DAS provide insight into the translational applicability of a publication?
  - How can we automate a workflow to extract important metadata information from publication DAS?
    - Do DASs need to be standardized?
  - What specific repositories or datasets are being accessed?
  - How are data being cited/can we quantify data reuse?
  - Do NIEHS-sponsored publications follow the FAIR principles?

OVERVIEW OF NIEHS PUBLICATIONS

- **Technologies Used:** Python (Pandas, Numpy, sk-learn)
- **Variables of Interest:** Relative Citation Ratio (RCR), MeSH Terms Extracted



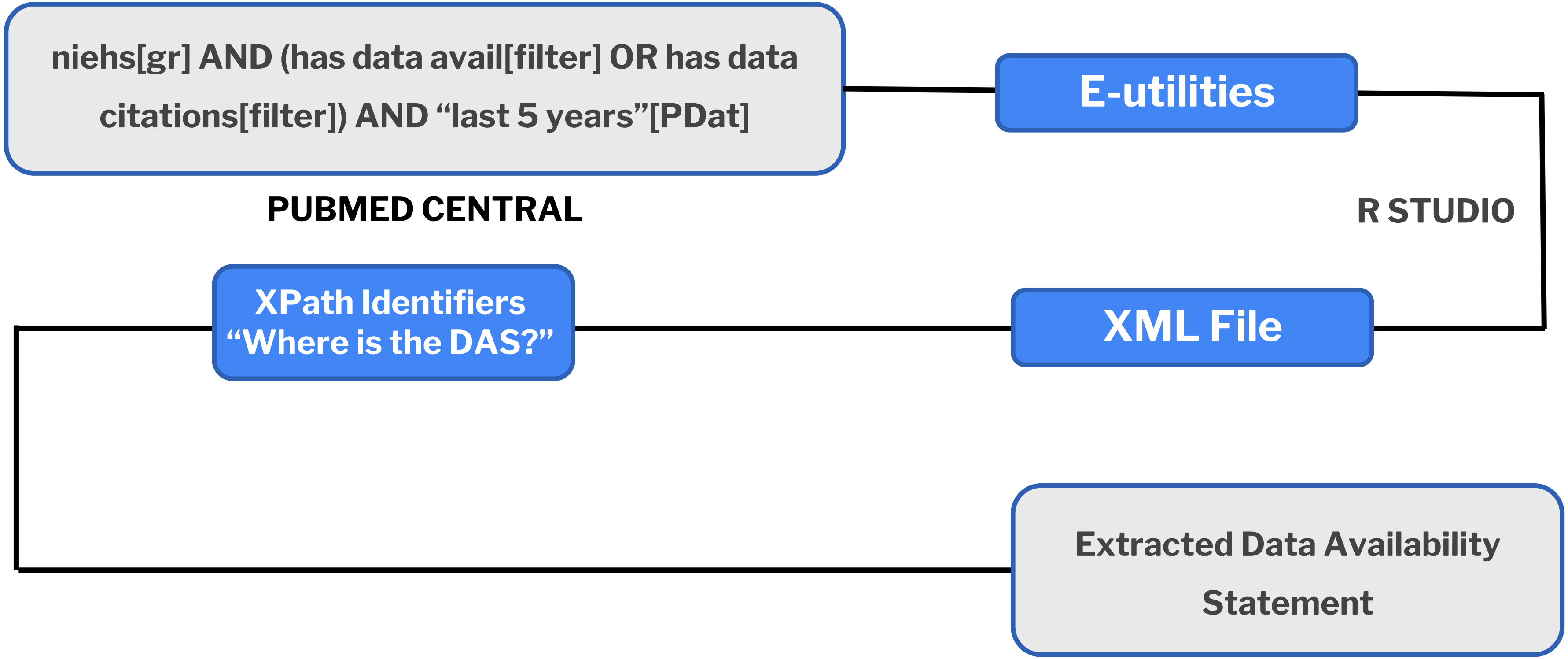
Publications with data present have higher RCR, on average



```
[('Humans', 184.0),
 ('Female', 114.0),
 ('Genes', 101.0),
 ('Male', 99.0),
 ('Animals', 80.0),
 ('Association', 74.0),
 ('Genetics', 62.0),
 ('Adult', 57.0),
 ('Disease', 57.0),
 ('Cells', 56.0),
 ('Population', 56.0),
 ('Genome', 56.0),
 ('Mice', 52.0),
 ('Role', 48.0),
 ('Methods', 48.0)]
```

Of the publications that share their data, human/genomic studies have the highest instances of data sharing

TEXT SCRAPING METHODOLOGY



niehs[gr] AND (has data avail[filter] OR has data citations[filter]) AND “last 5 years”[PDat]

PUBMED CENTRAL

E-utilities

R STUDIO

XML File

XPath Identifiers
“Where is the DAS?”

Extracted Data Availability Statement

TEXT-SCRAPING RESULTS

15,511

Publications Funded by NIEHS

2,971

?

Unique Publications with Possible DAS

1,539

✓

Unique Publications with Actual Data/DAS

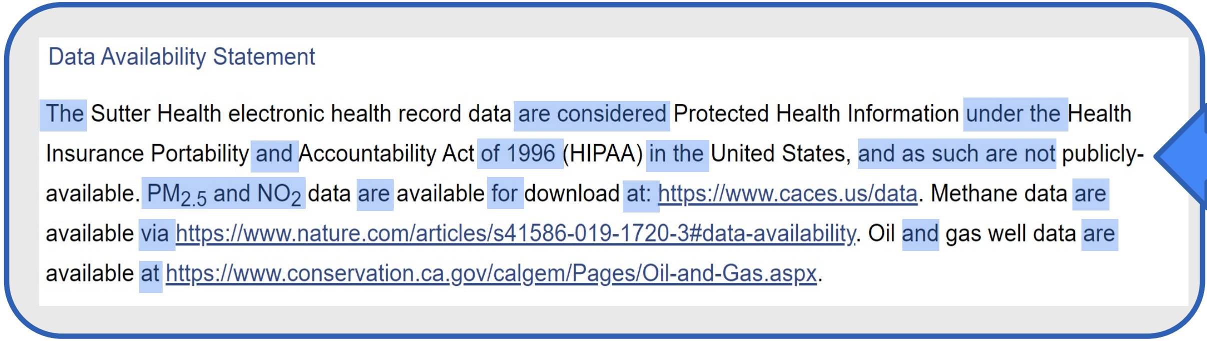
DATA SHARING STATEMENT CONTENT

Questions:

- Can we categorize data sharing into different types?
- What specific repositories are being utilized for data re-use?
- To what extent does NLP capture important aspects of a publication's DAS?

Technologies Used:

- Python NLTK NLP library – tokenization of DAS

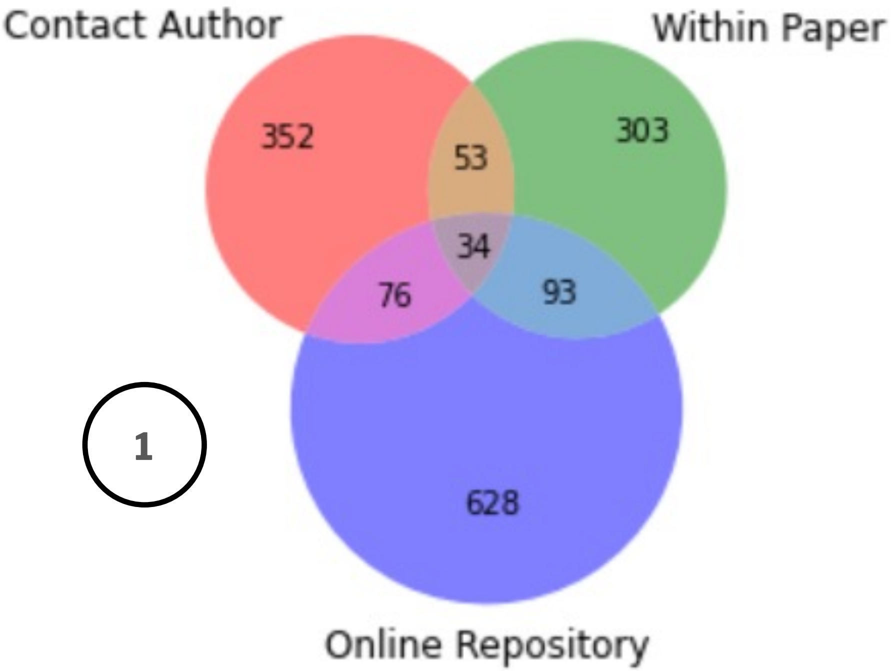


Data Availability Statement

The Sutter Health electronic health record data are considered Protected Health Information under the Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the United States, and as such are not publicly-available. $PM_{2.5}$ and NO_2 data are available for download at: https://www.caces.us/data. Methane data are available via https://www.nature.com/articles/s41586-019-1720-3#data-availability. Oil and gas well data are available at https://www.conservation.ca.gov/calgem/Pages/Oil-and-Gas.aspx.

- Removal of "Background Noise"
- Removal of words with numerical characters

RECURRING TRENDS IN DAS - CLASSIFICATION



1

40.8% → 🗂

📄 ← 19.7%

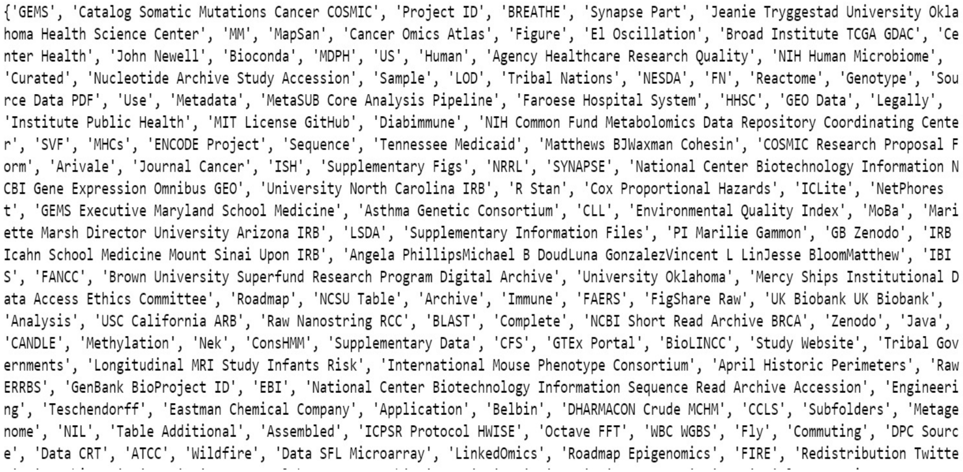
22.9% → ✉

16.6% of Statements are involved in more than one type of data sharing

OVERVIEW OF NIEHS PUBLICATIONS

- Part-of-Speech Tagging to extract repository names
- Background Noise (non-repositories) manually removed

NCBI Gene Expression Omnibus	**188**
GitHub	**95**
The Cancer Genome Atlas Program (TCGA)	**63**
ENCODE	**32**
NCBI Sequence Read Archive (SRA)	**30**



```
{'GEMS', 'Catalog Somatic Mutations Cancer COSMIC', 'Project ID', 'BREATHE', 'Synapse Part', 'Jeanie Tryggestad University Okla
homa Health Science Center', 'MM', 'MapSan', 'Cancer Omics Atlas', 'Figure', 'El Oscillation', 'Broad Institute TCGA GDAC', 'Ce
nter Health', 'John Newell', 'Bioconda', 'MDPH', 'US', 'Human', 'Agency Healthcare Research Quality', 'NIH Human Microbiome',
'Curated', 'Nucleotide Archive Study Accession', 'Sample', 'LOD', 'Tribal Nations', 'NESDA', 'FN', 'Reactome', 'Genotype', 'Sou
rce Data PDF', 'Use', 'Metadata', 'MetaSUB Core Analysis Pipeline', 'Faroese Hospital System', 'HHSC', 'GEO Data', 'Legally',
'Institute Public Health', 'MIT License GitHub', 'Diabimmune', 'NIH Common Fund Metabolomics Data Repository Coordinating Cente
r', 'SVF', 'MHCs', 'ENCODE Project', 'Sequence', 'Tennessee Medicaid', 'Matthews BJWaxman Cohesin', 'COSMIC Research Proposal F
orm', 'Arivale', 'Journal Cancer', 'ISH', 'Supplementary Figs', 'NRRL', 'SYNAPSE', 'National Center Biotechnology Information N
CBI Gene Expression Omnibus GEO', 'University North Carolina IRB', 'R Stan', 'Cox Proportional Hazards', 'ICLite', 'NetPhores
t', 'GEMS Executive Maryland School Medicine', 'Asthma Genetic Consortium', 'CLL', 'Environmental Quality Index', 'MoBa', 'Mari
ette Marsh Director University Arizona IRB', 'LSDA', 'Supplementary Information Files', 'PI Marilie Gammon', 'GB Zenodo', 'IRB
Icahn School Medicine Mount Sinai Upon IRB', 'Angela PhillipsMichael B DoudLuna GonzalezVincent L LinJesse BloomMatthew', 'IBI
S', 'FANCC', 'Brown University Superfund Research Program Digital Archive', 'University Oklahoma', 'Mercy Ships Institutional D
ata Access Ethics Committee', 'Roadmap', 'NCSU Table', 'Archive', 'Immune', 'FAERS', 'FigShare Raw', 'UK Biobank UK Biobank',
'Analysis', 'USC California ARB', 'Raw Nanostring RCC', 'BLAST', 'Complete', 'NCBI Short Read Archive BRCA', 'Zenodo', 'Java',
'CANDLE', 'Methylation', 'Nek', 'ConsHMM', 'Supplementary Data', 'CFS', 'GTEx Portal', 'BioLINCC', 'Study Website', 'Tribal Gov
ernments', 'Longitudinal MRI Study Infants Risk', 'International Mouse Phenotype Consortium', 'April Historic Perimeters', 'Raw
ERRBS', 'GenBank BioProject ID', 'EBI', 'National Center Biotechnology Information Sequence Read Archive Accession', 'Engineeri
ng', 'Teschendorff', 'Eastman Chemical Company', 'Application', 'Belbin', 'DHARMACON Crude MCHM', 'CCLS', 'Subfolders', 'Metage
nome', 'NIL', 'Table Additional', 'Assembled', 'ICPSR Protocol HWISE', 'Octave FFT', 'WBC WGBS', 'Fly', 'Commuting', 'DPC Sourc
e', 'Data CRT', 'ATCC', 'Wildfire', 'Data SFL Microarray', 'LinkedOmics', 'Roadmap Epigenomics', 'FIRE', 'Redistribution Twitte
```

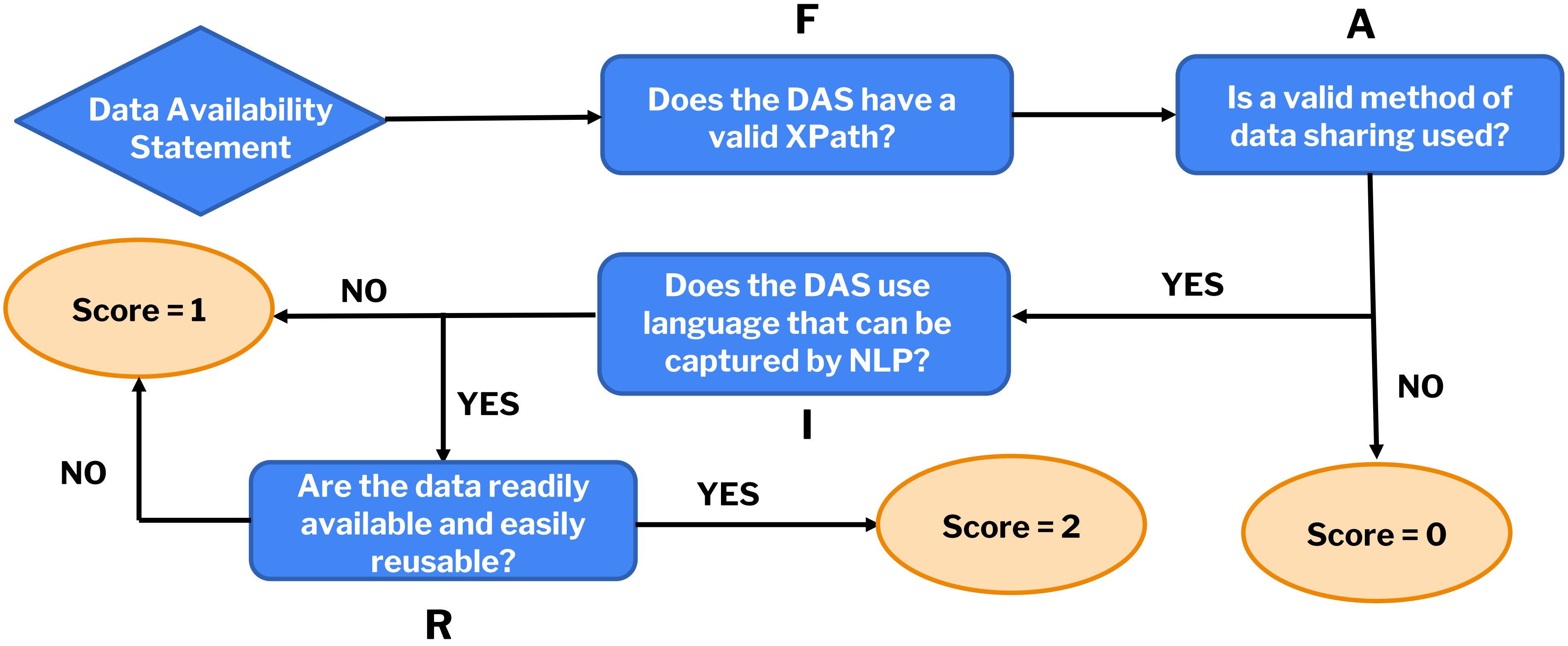
Over 972 Unique Repositories

FAIR DATA SHARING STANDARDS

Overarching Question:

Do DASs show evidence that publications follow the FAIR data sharing principles? Can we use this to quantify the degree to which DAS are successful?

“SCORE-BASED” FAIR CLASSIFICATION



F

A

Data Availability Statement

Does the DAS have a valid XPath?

Is a valid method of data sharing used?

YES

NO

Score = 0

Does the DAS use language that can be captured by NLP?

I

NO

Score = 1

YES

Are the data readily available and easily reusable?

R

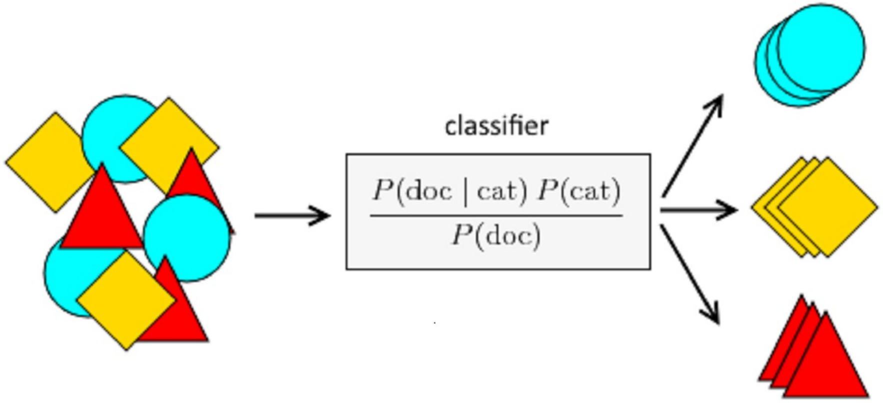
NO

YES

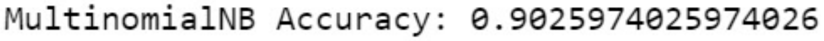
Score = 2

RECURRING TRENDS IN DAS – SENTIMENT ANALYSIS

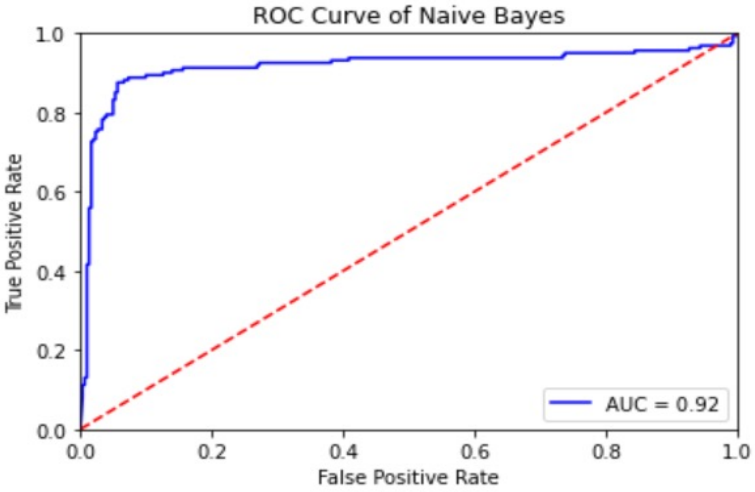
Naïve Bayes Classifier – classification algorithm based on conditional probability that assumes independent features



$$\frac{P(\text{doc} \mid \text{cat})\, P(\text{cat})}{P(\text{doc})}$$



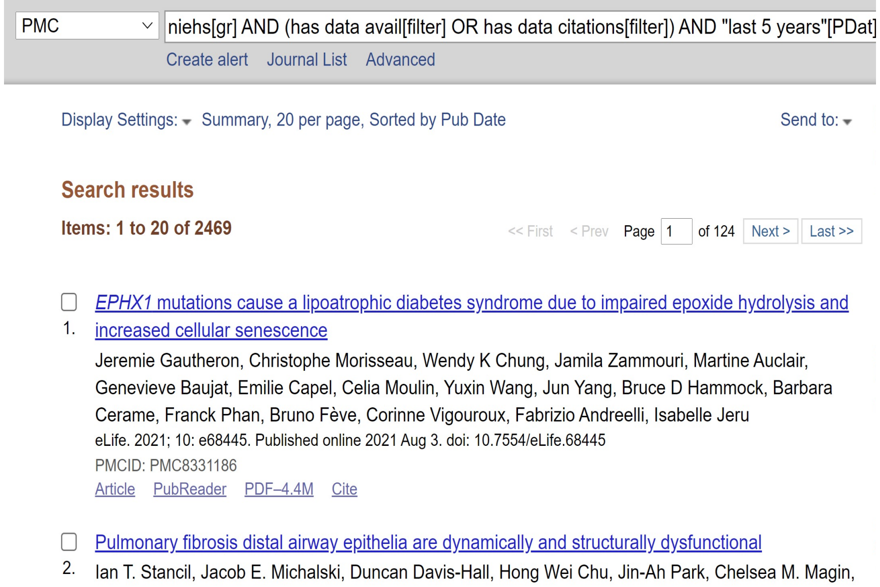
```
MultinomialNB Accuracy: 0.9025974025974026
```



Conclusion: With an accuracy score of 90.26% and AUC of 92%, the data can be categorized distinctly based on the FAIR data sharing principles. NIEHS-supported DAS success can be quantified and future publications can be validated using this model.

FUTURE DIRECTION

- Use publication metadata to further establish translational applicability of publications
- Implement a PMC-integrated software tool that:
  - Automatically standardizes DASs
  - Extracts important metadata from publications and DASs
  - Validates use of FAIR principles for data sharing



niehs[gr] AND (has data avail[filter] OR has data citations[filter]) AND "last 5 years"[PDat]

Summary, 20 per page, Sorted by Pub Date

Search results

Items: 1 to 20 of 2469

Page 1 of 124

□ 1. *EPHX1* mutations cause a lipoatrophic diabetes syndrome due to impaired epoxide hydrolysis and increased cellular senescence
Jeremie Gautheron, Christophe Morisseau, Wendy K Chung, Jamila Zammouri, Martine Auclair, Genevieve Baujat, Emilie Capel, Celia Moulin, Yuxin Wang, Jun Yang, Bruce D Hammock, Barbara Cerame, Franck Phan, Bruno Fève, Corinne Vigouroux, Fabrizio Andreelli, Isabelle Jeru
eLife. 2021; 10: e68445. Published online 2021 Aug 3. doi: 10.7554/eLife.68445
PMCID: PMC8331186
Article PubReader PDF–4.4M Cite

□ 2. Pulmonary fibrosis distal airway epithelia are dynamically and structurally dysfunctional
Ian T. Stancil, Jacob E. Michalski, Duncan Davis-Hall, Hong Wei Chu, Jin-Ah Park, Chelsea M. Magin,

CIVIC DIGITAL FELLOWSHIP

SPECIAL THANKS

Special Thanks to:

- **NIEHS: Program Analysis Branch**
  - Kristi Pettibone
  - Christie Drew
  - Steven Tuyishime
  - Nidhi Gera

- **NIEHS: Genes, Environment, and Health Branch**
  - Chris Duncan

- **NIEHS: Office of Data Science**
  - Trey Saddler

- **National Institutes of Health**
  - Jessica Mazerik
  - Allissa Dillman
  - Jacqueline Cattell

- **Coding It Forward**
  - Rachel Dodell & Ariana Soto