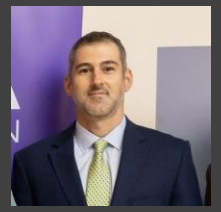
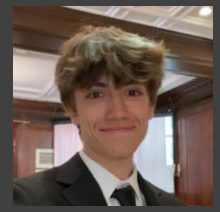
HARVARD DATA SCIENCE REVIEW

Discovering Data Sets in Unstructured Corpora: Discovering Use and Identifying New Opportunities

Nick Pallotta, John M Locklear, Xiangyu Ren, Victor Robila, and Adel Alaeddini



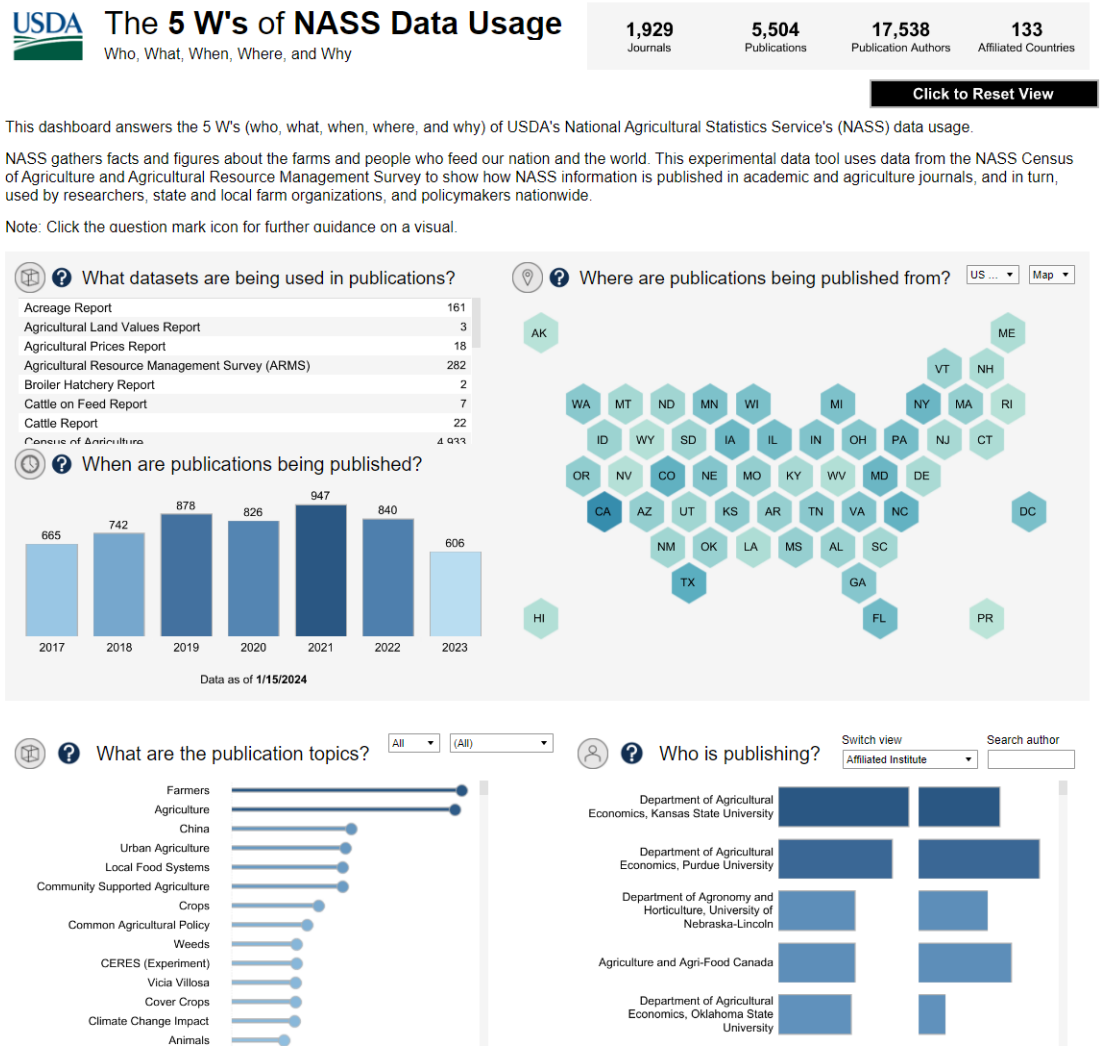

National Agricultural Statistics Service (NASS)

…..Timely, Accurate, and <u>Useful</u> Statistics in Service to Agriculture – NASS Mission Statement

- 400+ reports per year; crops, livestock, environmental, and economic data, etc…

NASS utilized the democratizing data work to create a dashboard called the [5W's of NASS Data Usage]

- 5W's = Who, what, where, when, why of NASS data usage
- Peer reviewed research articles; along with topics, authors, institutions, etc…



Response Rate Impact?

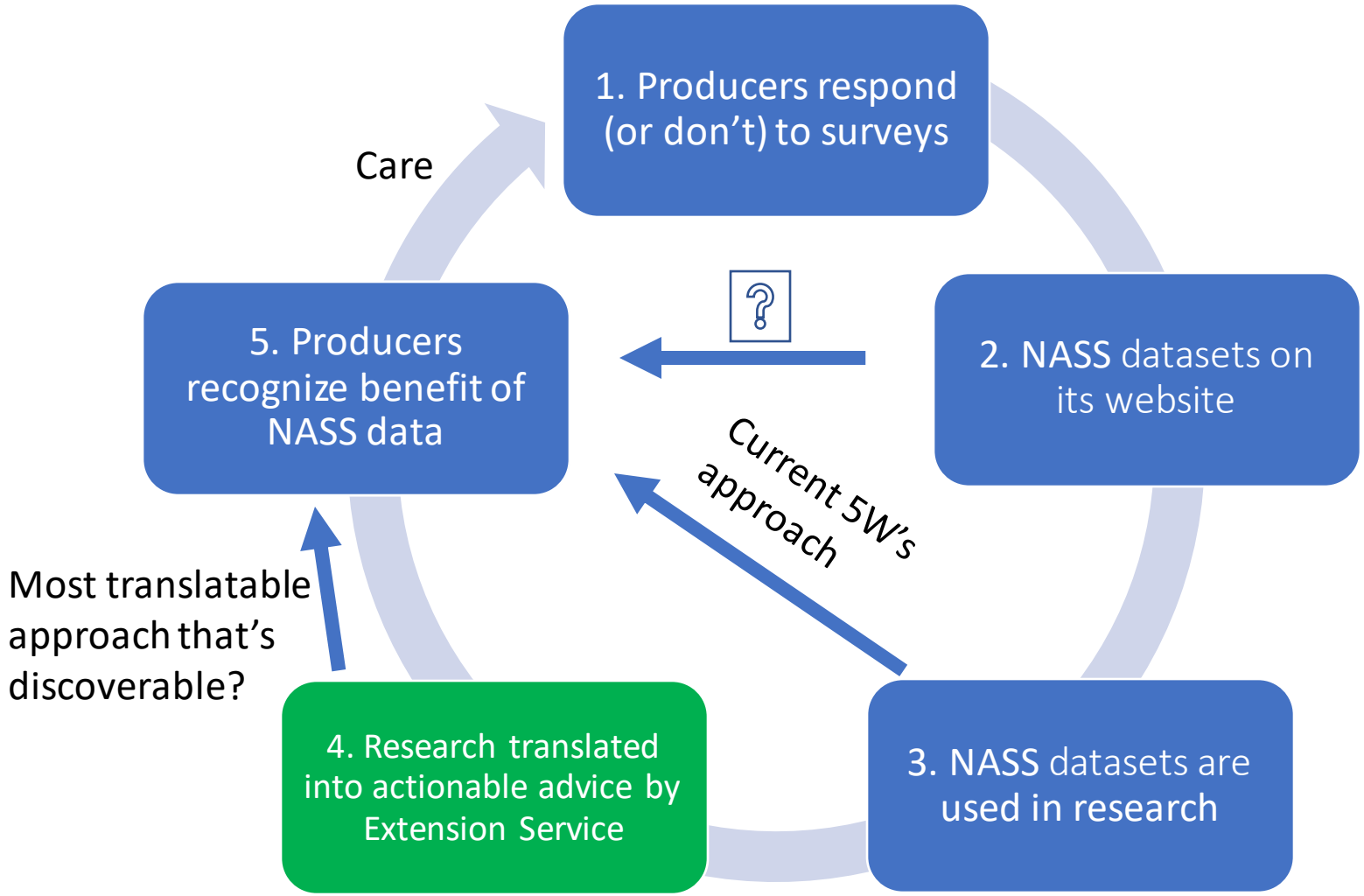
Survey Response Rates: *"What is this data going to be used for?"*

- *2023 experiment, if farmers are emailed the dashboard showing research usages, will there be a higher response rate?*
- *Non-statistically significant increase in response rates*

Ranking of importance – Why do you not respond to NASS Surveys?		
Social media web survey	State statistician interviews with farmers	Focus groups
1. Time (burden)	1. Time (burden)	1. Do not see benefit (reciprocity)
2. Do not see benefit (reciprocity)	2. Do not see benefit (reciprocity)	2. Time (burden)
3. Privacy concerns (authority)	3. Privacy concerns (authority)	3. Questionnaires are redundant
4. Questionnaires not relevant to farmers or ranchers	4. Accuracy	4. Privacy concerns (authority)
5. Questionnaires not very user friendly	5. Other	5. Other

NASS Response Rate Research Team

Value Recognition Circle



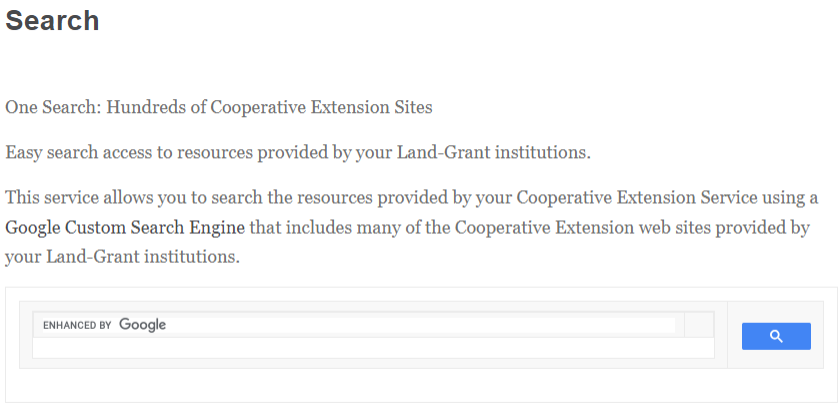
Cooperative Extension Service

EXTENSION FOUNDATION

"Extension provides non-formal education and learning activities …. to farmers and other residents of rural communities and urban areas…..taking knowledge gained through research and education and bringing it directly to the people to create positive changes."

- Service provided through the more than 100 land grant institutions
  - Decentralized, content hosted on individual university-maintained websites

- Search service existed that hit 158 URLs
- Those URLs would be the "unstructured corpora"



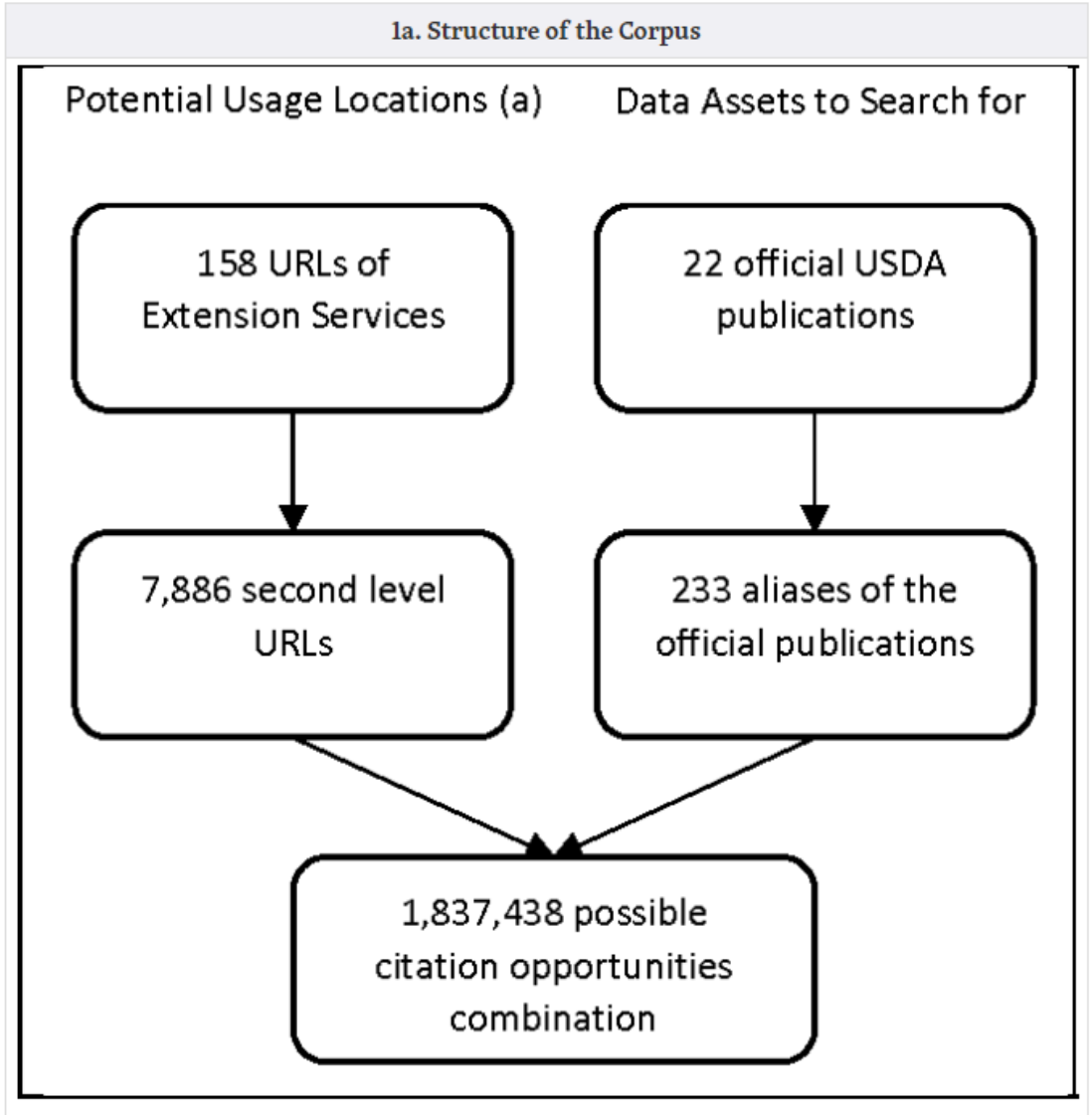
What Datasets Are We Going to Look For?

HARVARD DATA SCIENCE REVIEW

- Twenty-two reports + Aliases
- Other names
- Frequency can be included
- Generic report names
- No DOI's (yet)

Report Name	Alias 1
Census of Agriculture	Agricultural Census
Cattle	January Cattle
Farm Labor	Labor
Crop Production	Crop
Farm Production Expenditures	ARMS

Identified Datasets and Data Reports



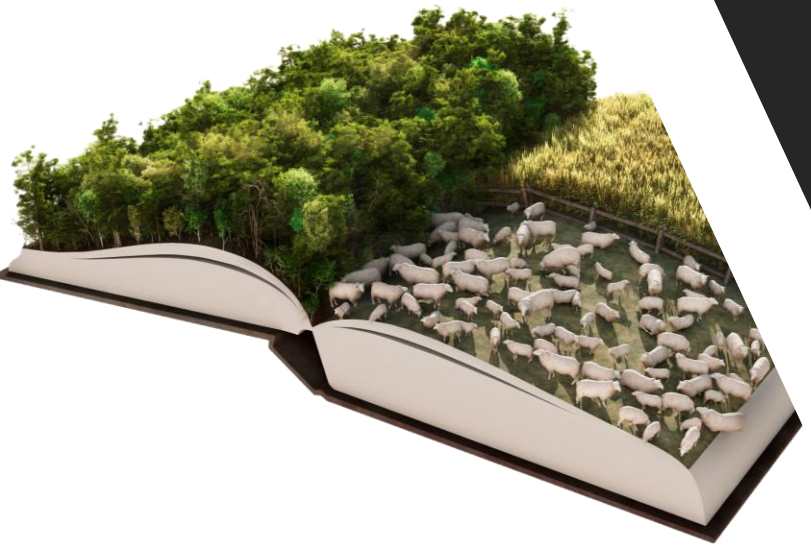
JSON files from Extension Foundation

USDA NIFA Backed Extension Bot:

- Extension Foundation Team created a GenAI backed Bot
  - John M. Locklear
  - Dr. David Warren (OSU)
- Created machine readable JSON files to feed the bot
- Our team used the JSON files to look for dataset usages

Universities Contributing

- Oklahoma State University (1,540)
- Oregon State University (5,203)
- University of Georgia (874)
- University of Florida Publications (6,402)
- University of Florida Blogs(10,354)
- University of Kentucky (324)
- University of Tennessee(2,602)
- LSU (1917)
- Penn St. (4,740) *In development*
- University of California – Integrated Pest Management (1,675)
- Ask Extension Knowledgebase (313,443)



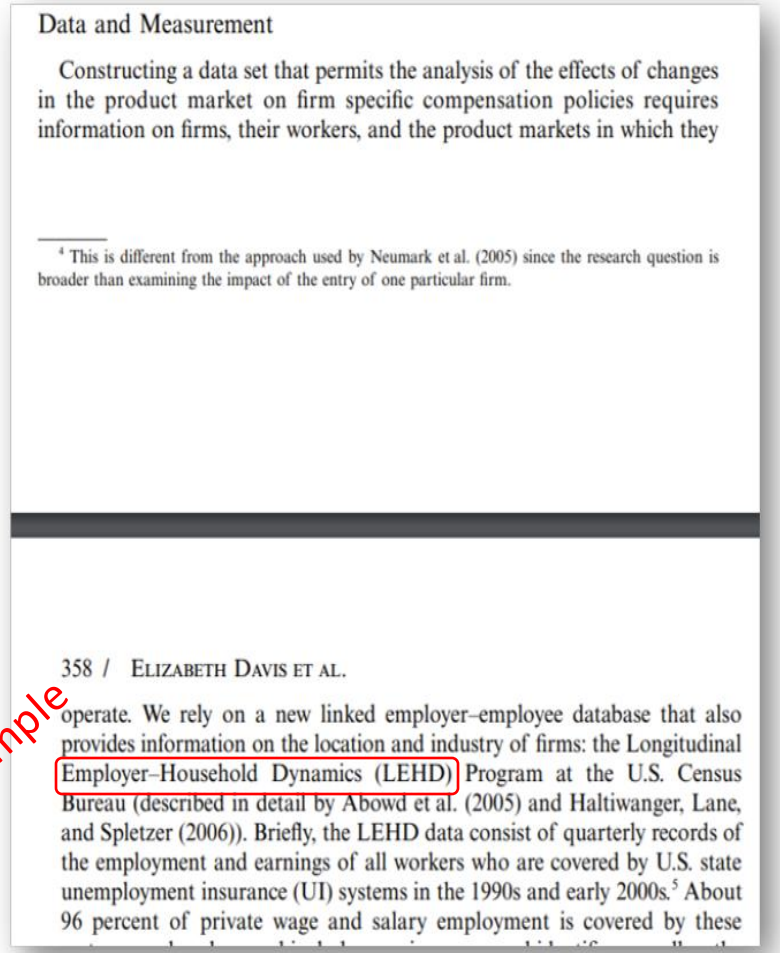
Methodology

Dr. Adel Alaeddini

Dataset extraction using Kaggle ML Models

HDSR HARVARD DATA SCIENCE REVIEW

- Kaggle Competition Finalist
- Transformer-Enhanced Heuristic Search
- Training **the transformer classifier** by whether a string refers to a dataset/survey/report.
  - Using **the Schwartz-Hearst algorithm** to identify LONG-NAME (Acronym) pattern candidate dataset using the pre-trained classifier prediction model and minimum document frequency.
  - Adding **Aliases of USDA datasets** to train the pre-trained classifier prediction model.
- Findings:
  - 15 datasets from the 3,000 Oregon State articles.
  - 20 datasets from 1,500 Oklahoma State articles


Data and Measurement

Constructing a data set that permits the analysis of the effects of changes in the product market on firm specific compensation policies requires information on firms, their workers, and the product markets in which they

[4] This is different from the approach used by Neumark et al. (2005) since the research question is broader than examining the impact of the entry of one particular firm.

358 / Elizabeth Davis et al.

operate. We rely on a new linked employer–employee database that also provides information on the location and industry of firms: the Longitudinal Employer–Household Dynamics (LEHD) Program at the U.S. Census Bureau (described in detail by Abowd et al. (2005) and Haltiwanger, Lane, and Spletzer (2006)). Briefly, the LEHD data consist of quarterly records of the employment and earnings of all workers who are covered by U.S. state unemployment insurance (UI) systems in the 1990s and early 2000s.[5] About 96 percent of private wage and salary employment is covered by these

Example

Dataset extraction looking for hyperlinks

HDSR
HARVARD DATA SCIENCE REVIEW

- Algorithmic web scraper applied to both the Oregon State and the Oklahoma State JSON corpora
  - consisted of the hyperlinks found in the articles that were part of the original corpora.
  - Were these hyperlinks to datasets?

- Oregon State (72,336 sublinks)
  - 14 links to data assets were found
  - 4 of which do not reference specific data (e.g, https://quickstats.nass.usda.gov/).

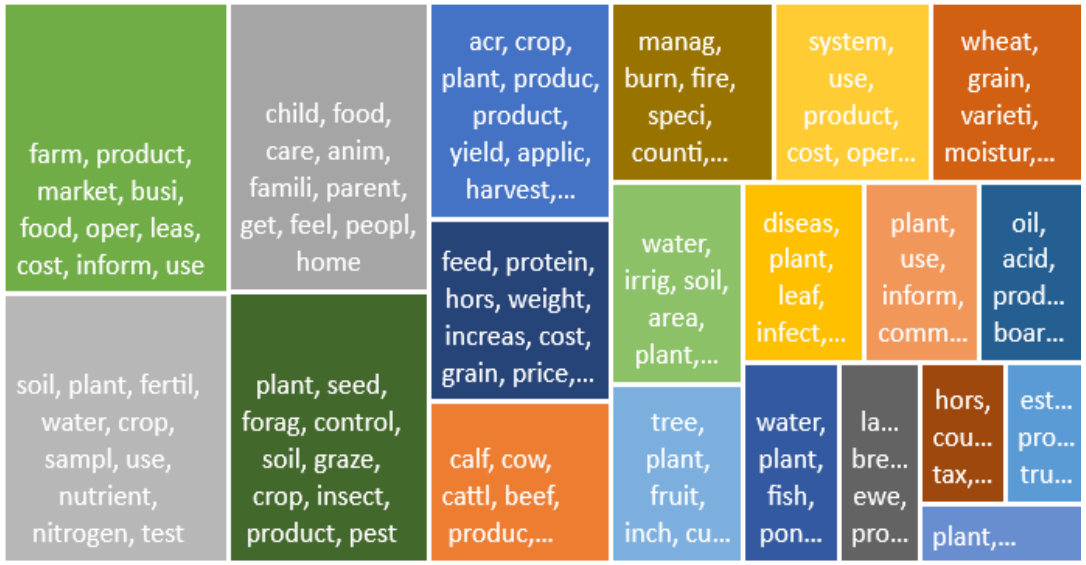
- Oklahoma State (69,845 sublinks)
  - 8 Quick Stats links of National Agricultural Statistics Service
  - 3 Links to the Rural-Urban Continuum Codes dataset (another USDA dataset)

Bag of Words Text Analysis

- Latent Dirichlet allocation (LDA) multiscore method
  - Gensim package :
  - Examples of main topics covered in Oklahoma State reports
  - Topic number:20 & Words in each topic: 10

The main topics covered in Oregon State University reports

The main topics covered in Oklahoma State reports



Machine Learning for Estimation of Missed Opportunities

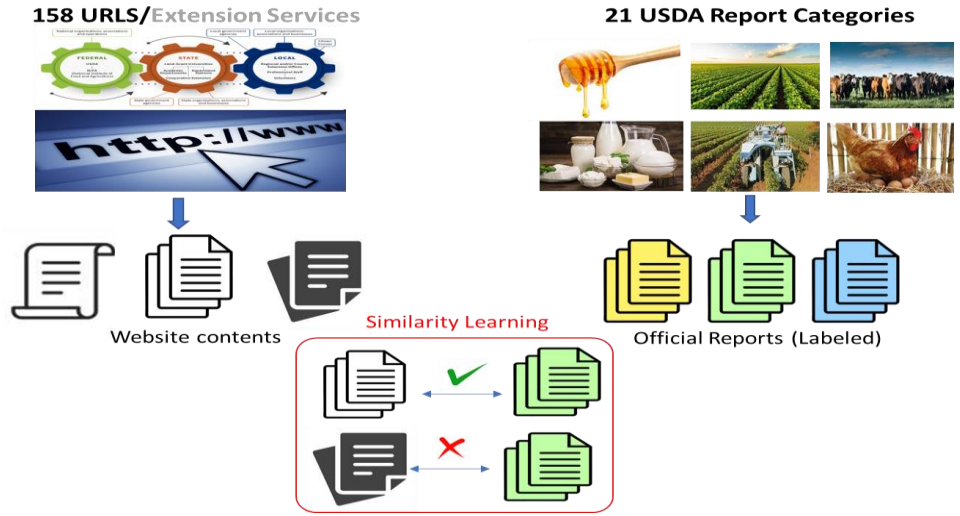
HARVARD DATA SCIENCE REVIEW

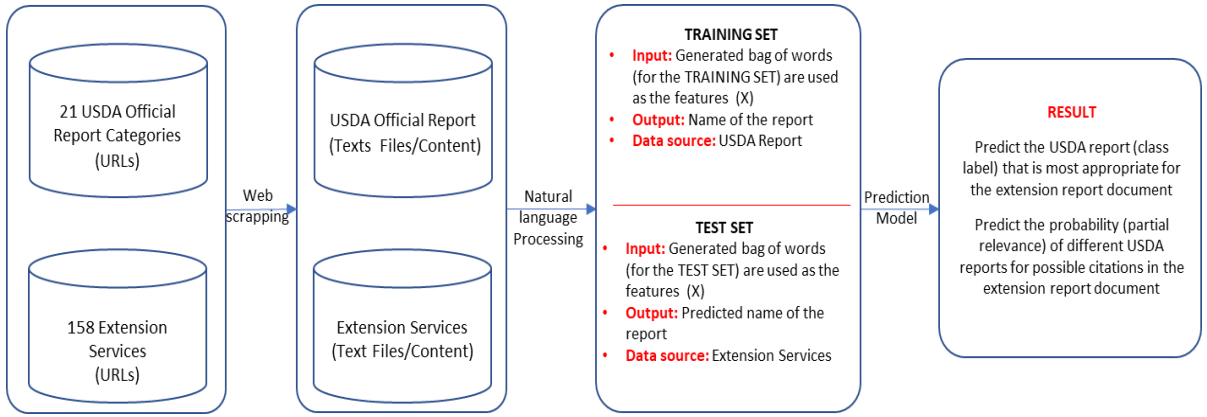
Basic Idea: Utilize bag of words as features to train a classifier to estimate the probability of an extension services document (158 URLS) requiring a citation from the USDA Reports

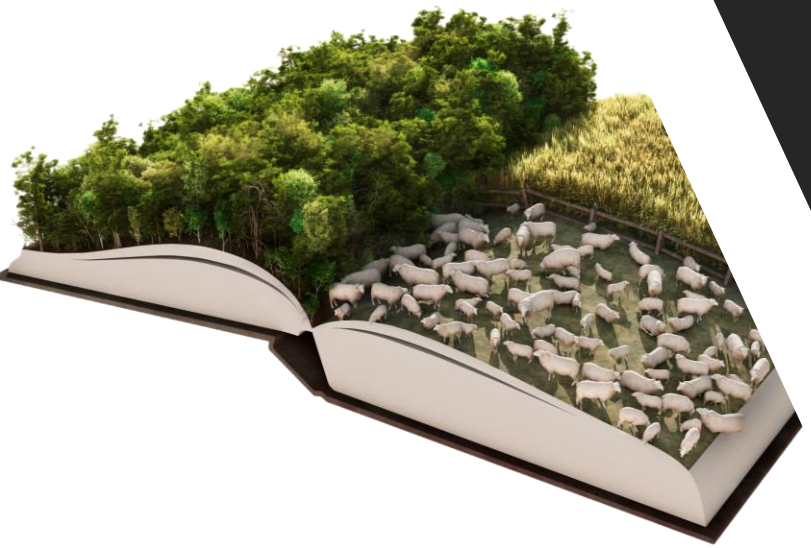
Feature Extraction: Bag of word analysis with n=35 top frequent words; No ordering used for simplicity

Model: Feed forward neural network with 1 hidden layer (70 nodes)

Result: 90% accuracy in assigning correct label (appropriate referencing)







Wrap-up and Next Steps

Nick Pallotta
nick.pallotta@usda.gov

Next Steps for NASS

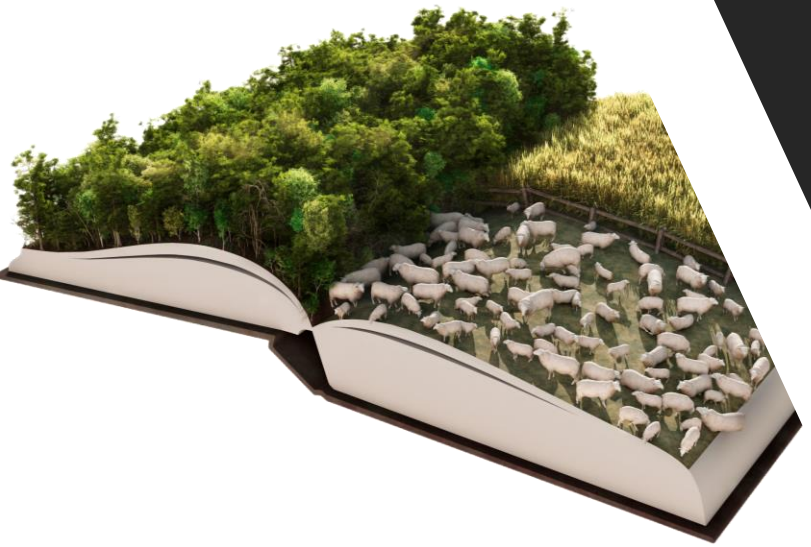
HDSR HARVARD DATA SCIENCE REVIEW

- **To Summarize:** lots of work for what we found

- Met with Extension Service content creators.
  - How can we reduce the burden on data users to cite our data?
    - DOI's (will allow us to take advantage of new services)
    - Citation import capabilities

DataCite Usage Tracker

- Larger corpus of machine-readable files
- Or, a centralized Extension corpus?



Questions