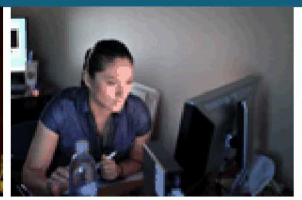
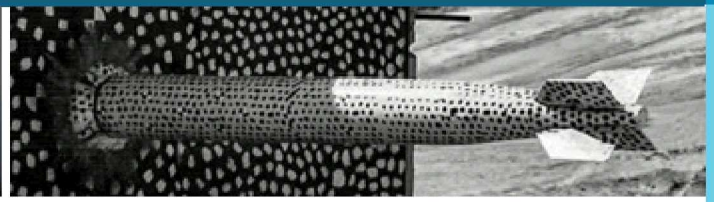
SAND2020-6630C

TOPIC MODELING WITH NATURAL LANGUAGE PROCESSING FOR IDENTIFICATION OF NUCLEAR PROLIFERATION-RELEVANT SCIENTIFIC AND TECHNICAL PUBLICATIONS





INMM Annual Meeting 2020

Presented by Jon Bisila

Team: Jon Bisila, Zoe Gastelum, Danny Dunlavy, Craig Ulmer

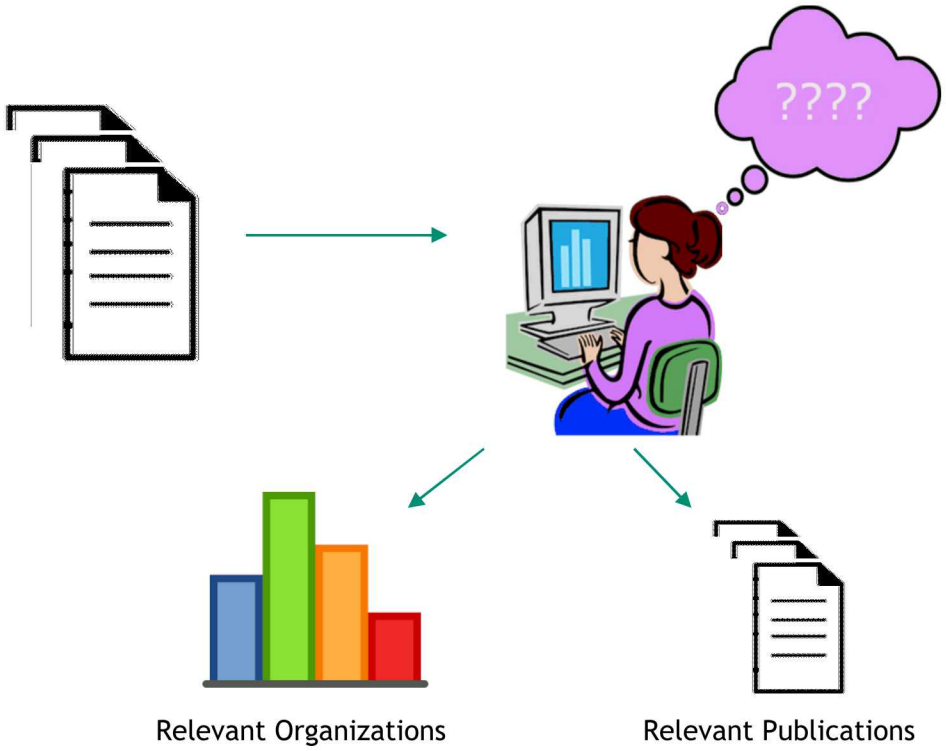
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Depa
Energy's National Nuclear Sec
Administration under contract DE-N

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

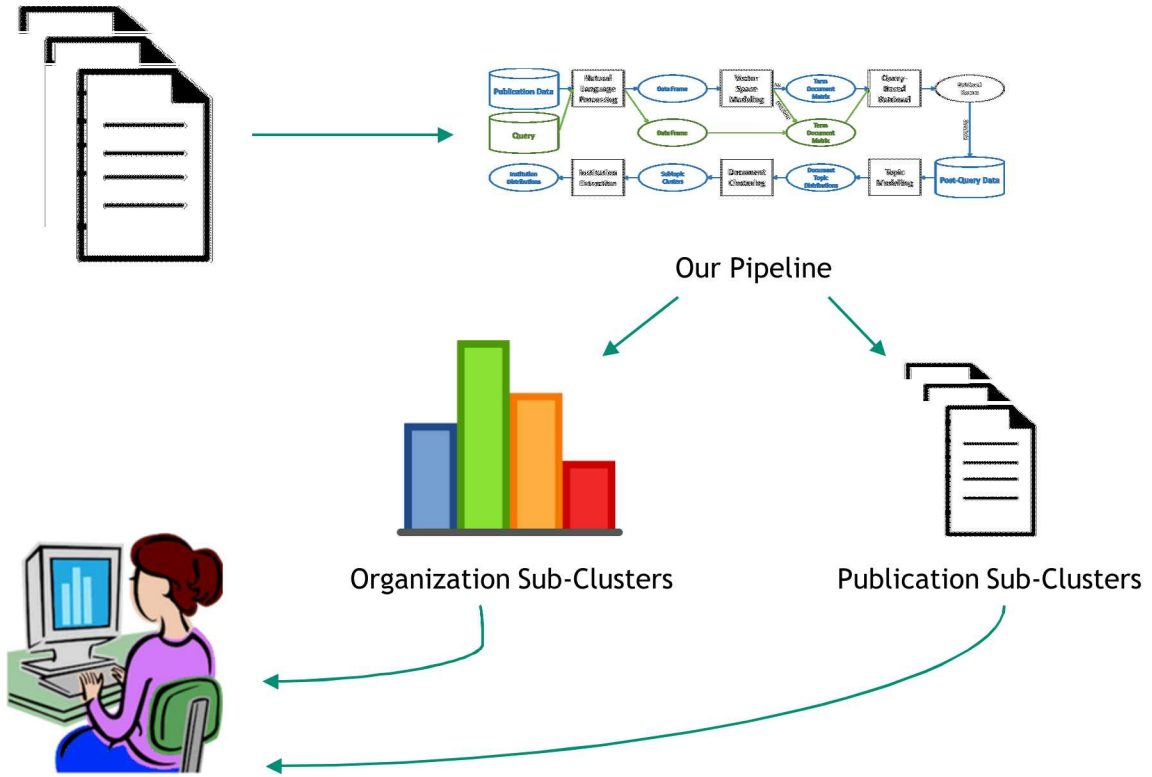
Efficiently identifying relevant scientific publications remains a challenge

How can we help analysts more efficiently find what they are looking for?

Current: Manual Triage

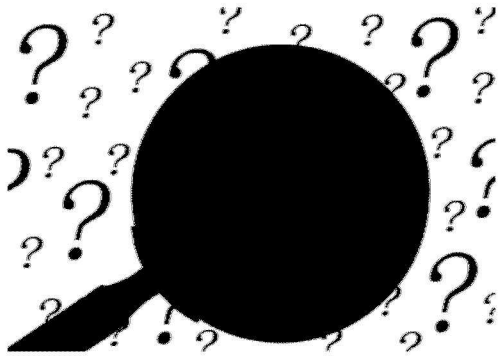


Proposed: Assisted Triage

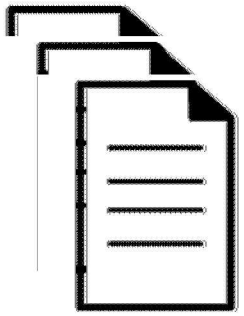


Challenges

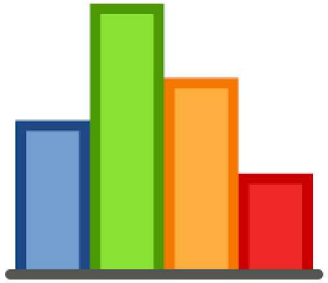
- Search strings used by analysts to find relevant information have limitations
- Sifting out relevant documents still requires manual triage
- Efficiently identifying organizations involved can be difficult



Effective Search Strings



Relevant Publications



Relevant Organizations

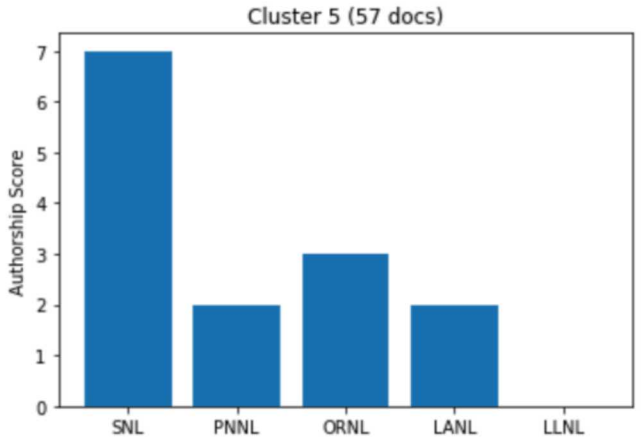
Our Approach

- Build from existing work to augment existing analyst search process
- Utilize topic modelling and document clustering to assist triage
- Present distributions of organizations in a more user-friendly manner

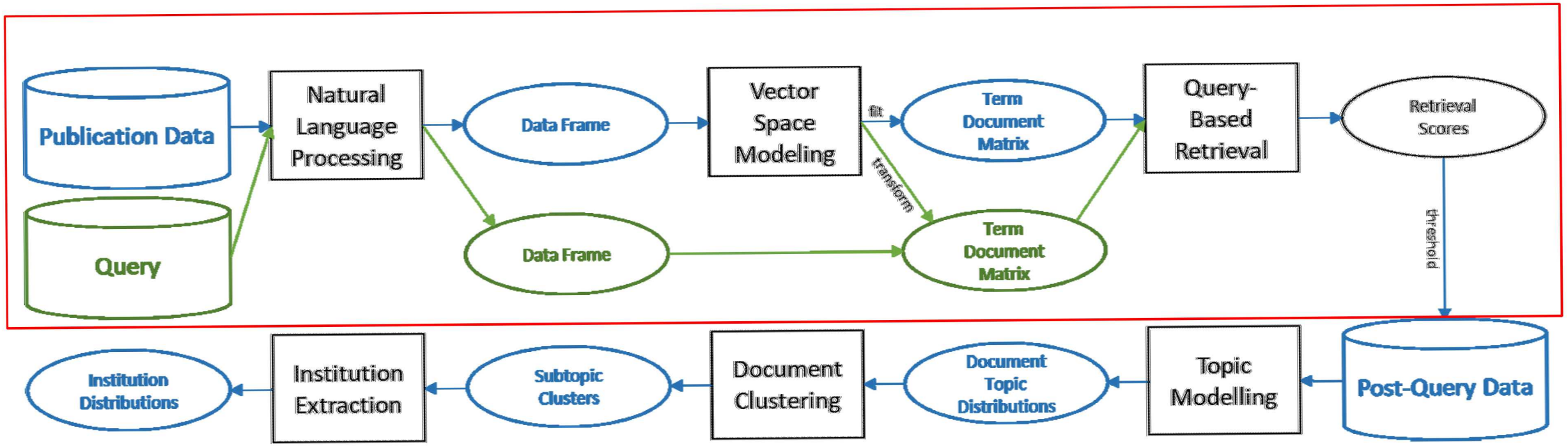
Topic Modelling

```
0.058*"compon" + 0.047*"contain" + 0.041*"polym" + 0.035*"solvent"
0.042*"layer" + 0.039*"shell" + 0.031*"flow" + 0.028*"electrod"
0.026*"modul" + 0.023*"optic" + 0.022*"system" + 0.018*"devic"
```

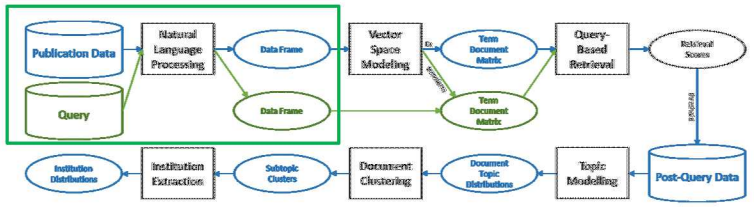
Distribution of Organizations



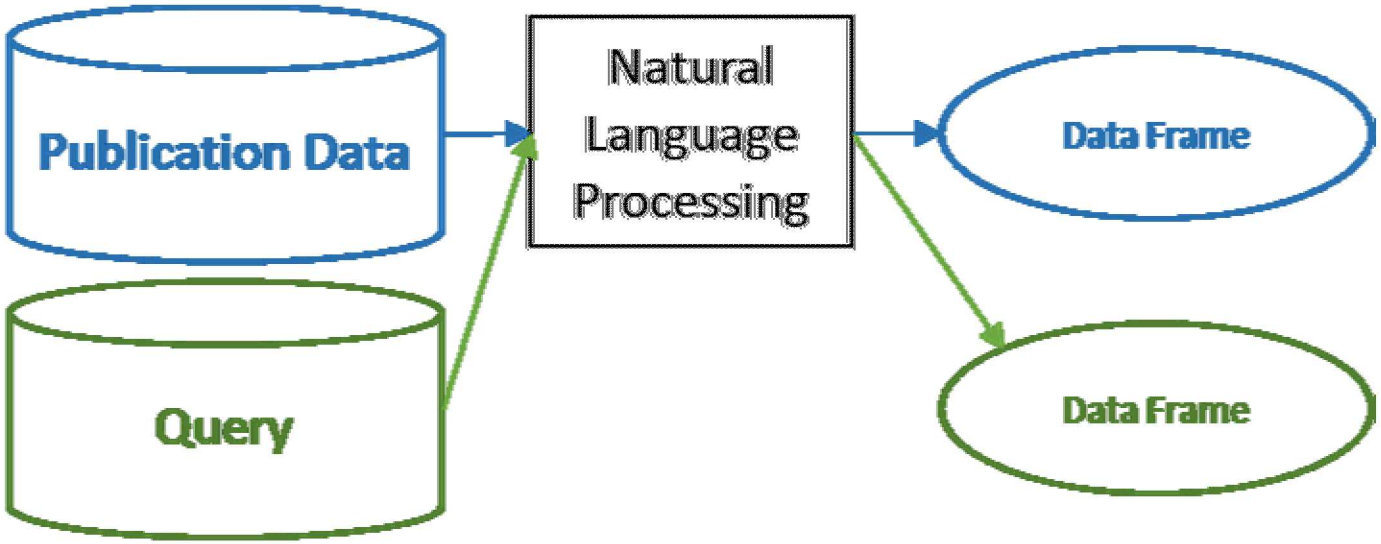
Workflow


Publication Data
Query
Natural Language Processing
Data Frame
Data Frame
Vector Space Modeling
fit
transform
Term Document Matrix
Term Document Matrix
Query-Based Retrieval
Retrieval Scores
threshold
Post-Query Data
Topic Modelling
Document Topic Distributions
Document Clustering
Subtopic Clusters
Institution Extraction
Institution Distributions

Methods: Natural Language Processing

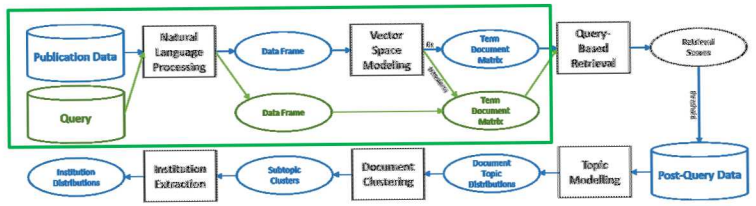


- Example transformations* of Publication Data:
  - Using word roots instead of words: *proliferate*, *proliferation*, *proliferates*… --> *proliferat*
  - Remove words that do not help discriminate topics: *a, an,* "the", "in", etc.

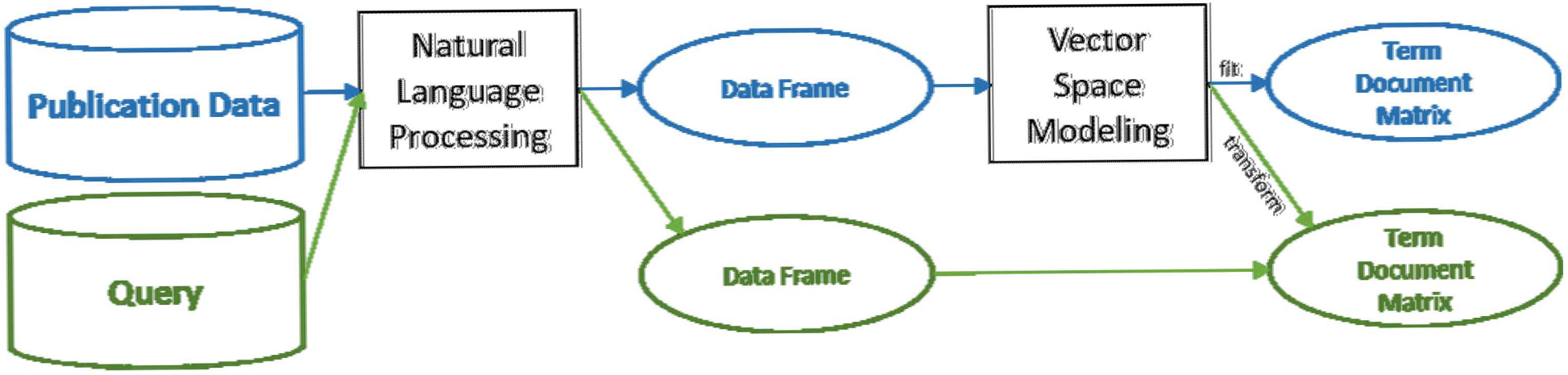


* Note: Details regarding all natural language processing (NLP) tasks performed are in the full paper

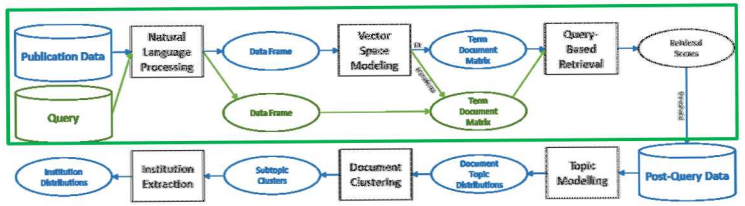
Methods: Vector Space Modeling



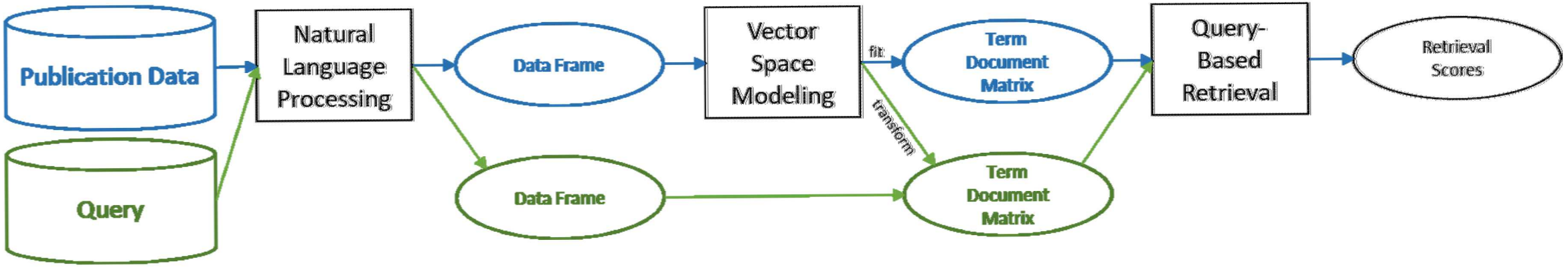
- Vector Space Models (VSMs) provide another interpretation of the data
- *Term-document matrix:* columns represent terms, rows represent documents



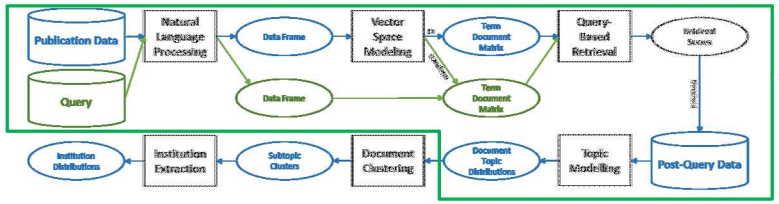
Methods: Query-Based Information Retrieval



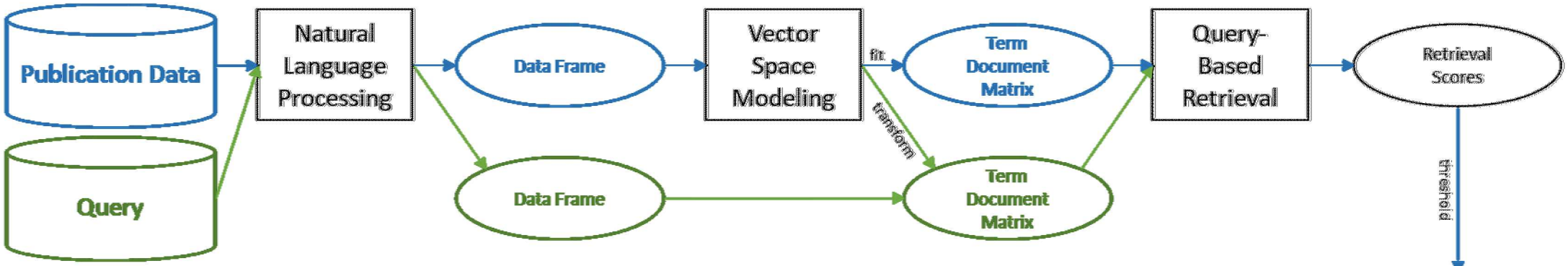
- Simulate Boolean search efficiently using single matrix-vector product
- Provide a score for each document in relation to the query



Methods: Topic Modeling



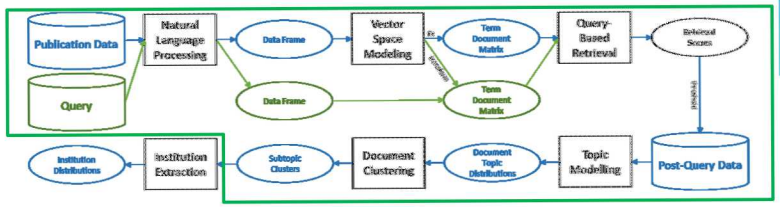
- Build topic models using Latent Dirichlet Allocation (LDA)
- Documents represented as distributions of topics
- Topics composed of distributions of terms



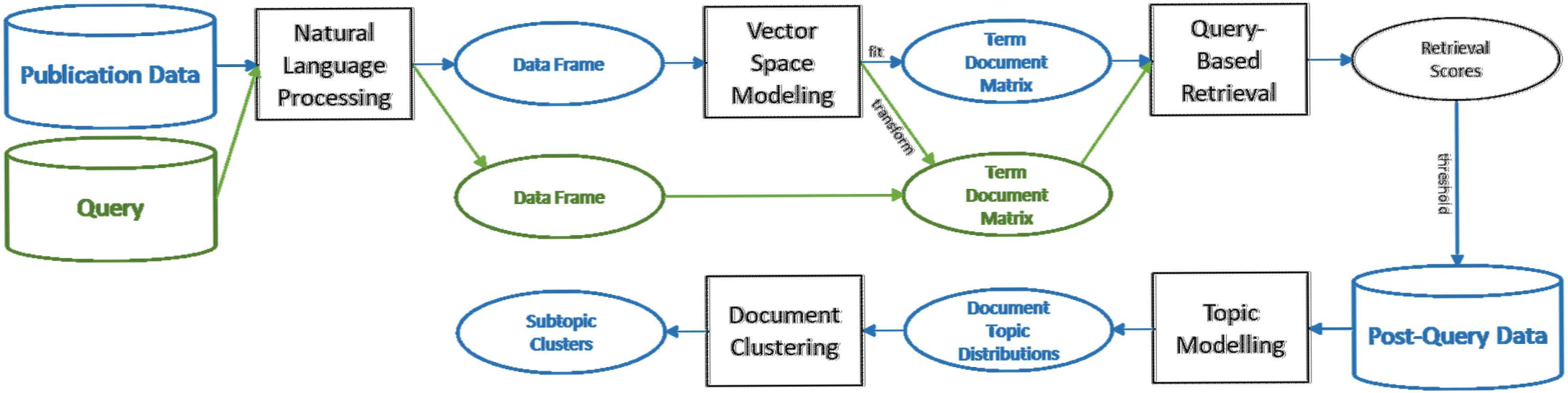
Topics and Weighted Terms

```
0.058*"compon" + 0.047*"contain" + 0.041*"polym" + 0.035*"solvent"
0.042*"layer" + 0.039*"shell" + 0.031*"flow" + 0.028*"electrod"
0.026*"modul" + 0.023*"optic" + 0.022*"system" + 0.018*"devic"
```

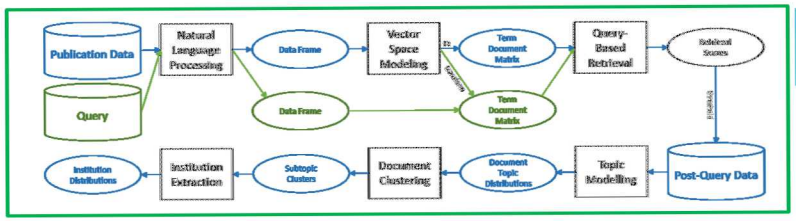
Methods: Document Clustering



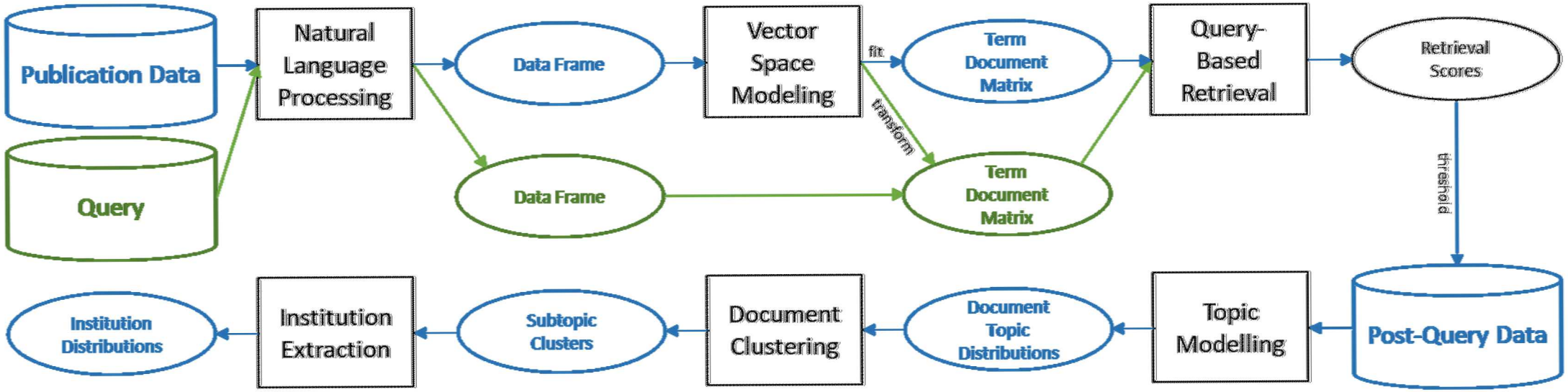
- Cluster documents using Louvain community detection algorithm
- Output: rank-ordered list of documents that best summarize a cluster
- Also provide rank-ordered list of summary topics and terms



Methods: Institution Extraction



- Extract the institutions associated with the documents in each cluster
- Provide histogram of the distribution of document counts per institution



Case Study: U.S. Department of Energy Publications

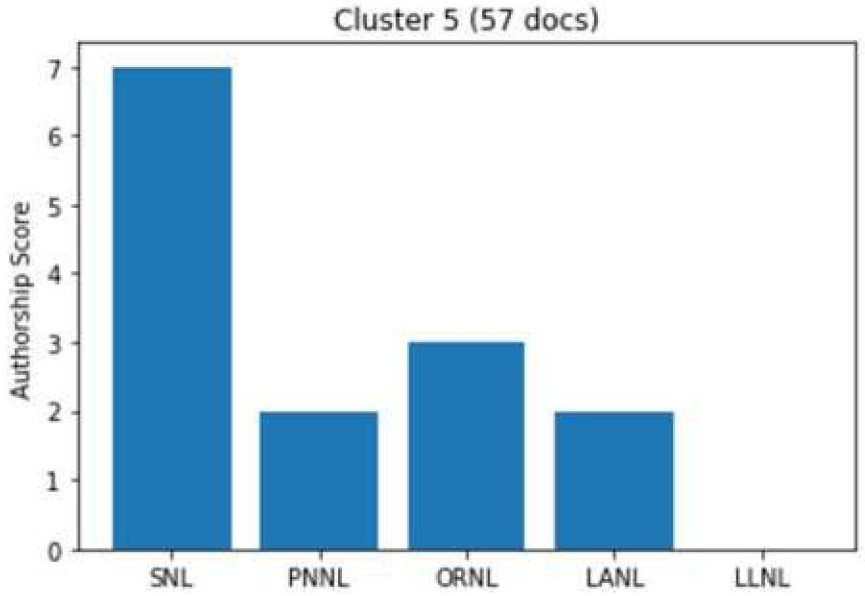
Case Study: U.S. Department of Energy Publications

- Generated an expert-derived search string related to nuclear fuel reprocessing

```
alloy, anion, arms, boxes, bromine, cadmium, calcination, cell, ceric, chloride,
chopping, concrete, cuts, decladding, density, electrolysis, exchange, extraction,
fluoride, fluoridized, fuel, fuels, glass, glove, handling, hardened, hexone,
high, inconel, irradiated, isobutyl, ketone, lead, lithium, manipulator, master,
metal, methyl, MIBK, molten, nitrate, nuclear, oxidation, plutonium, precipitation,
purification, radiation, redox, reduction, remote, reprocess, rods, salts,
separation, shield, slave, solvent, spent, trifluoride, uranium, volatility, window
```

- Data:
  - 2,000 Office of Science and Technical Information (OSTI) documents from 2017
  - Seven relevant "tracer" documents, identified by subject matter experts
- Main question to answer using our approach:
  - Do these documents cluster together?

Results

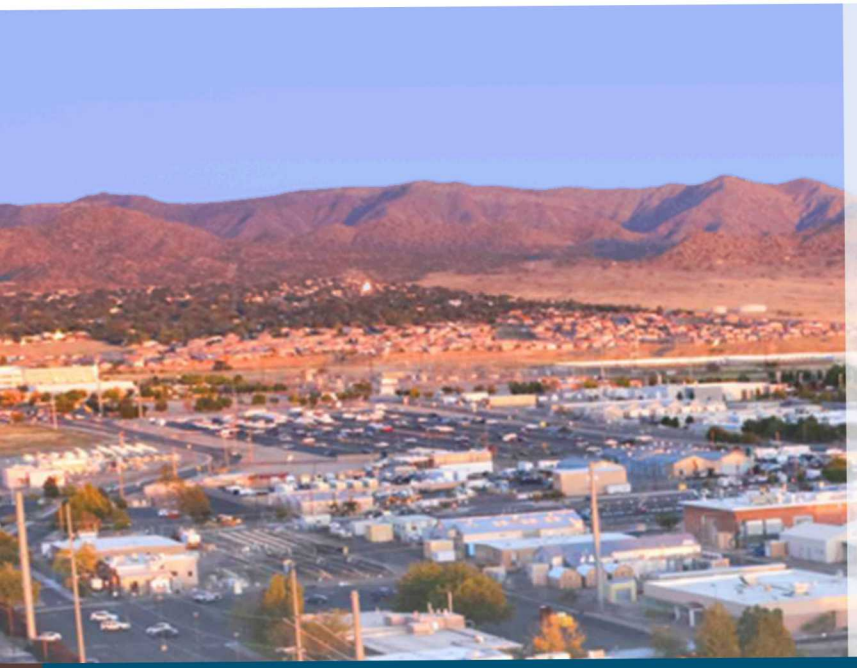


- Of the seven tracer documents, five of them appear in the same cluster
- Benefit of our approach:
  - Analysts can quickly scan the topics, terms, and titles to determine relevancy of each cluster

```
Cluster 5 topics and terms

0.058*"compon" + 0.047*"contain" + 0.041*"polym" + 0.035*"solvent"
0.042*"layer" + 0.039*"shell" + 0.031*"flow" + 0.028*"electrod"
0.026*"modul" + 0.023*"optic" + 0.022*"system" + 0.018*"devic"

Cluster 5 titles

Production of alpha-hydroxy carboxylic acids and esters from higher sugars using tandem catalyst systems
Electrochemical Nucleation and Growth of Uranium and Plutonium from Molten Salts
Variants of polypeptides having cellulolytic enhancing activity and polynucleotides encoding same
```

Future Work

- Uncertainty Quantification: how certain are we about the stability and reliability of our choices of algorithms?
- Different vector space models
- Alternative document retrieval methods



Questions?

Jon Bisila

jbisila@sandia.gov