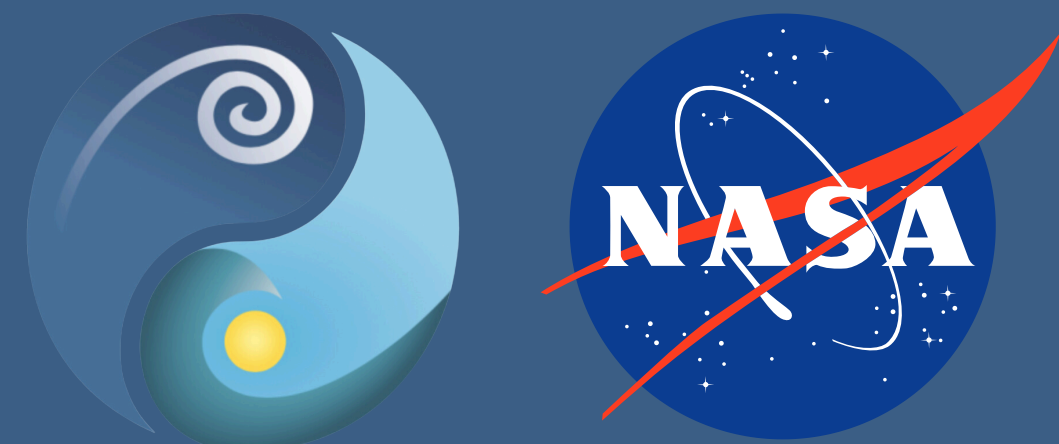




AGU 2024
WASHINGTON DC
DECEMBER 13, 2024

IMPROVING GES DISC DATA SEARCH AND DISCOVERY THROUGH AI METADATA AUGMENTATION

KENDALL GILBERT^{1,2}
ARMIN MEHRABIAN^{1,2}
IRINA GERASIMOV^{1,2}

¹ CODE 619, GODDARD SPACE FLIGHT CENTER,
GREENBELT MD, USA
² ADNET SYSTEMS INC., LANHAM MD, USA

NASA GODDARD EARTH SCIENCES DATA AND INFORMATION SERVICES CENTER

BACKGROUND

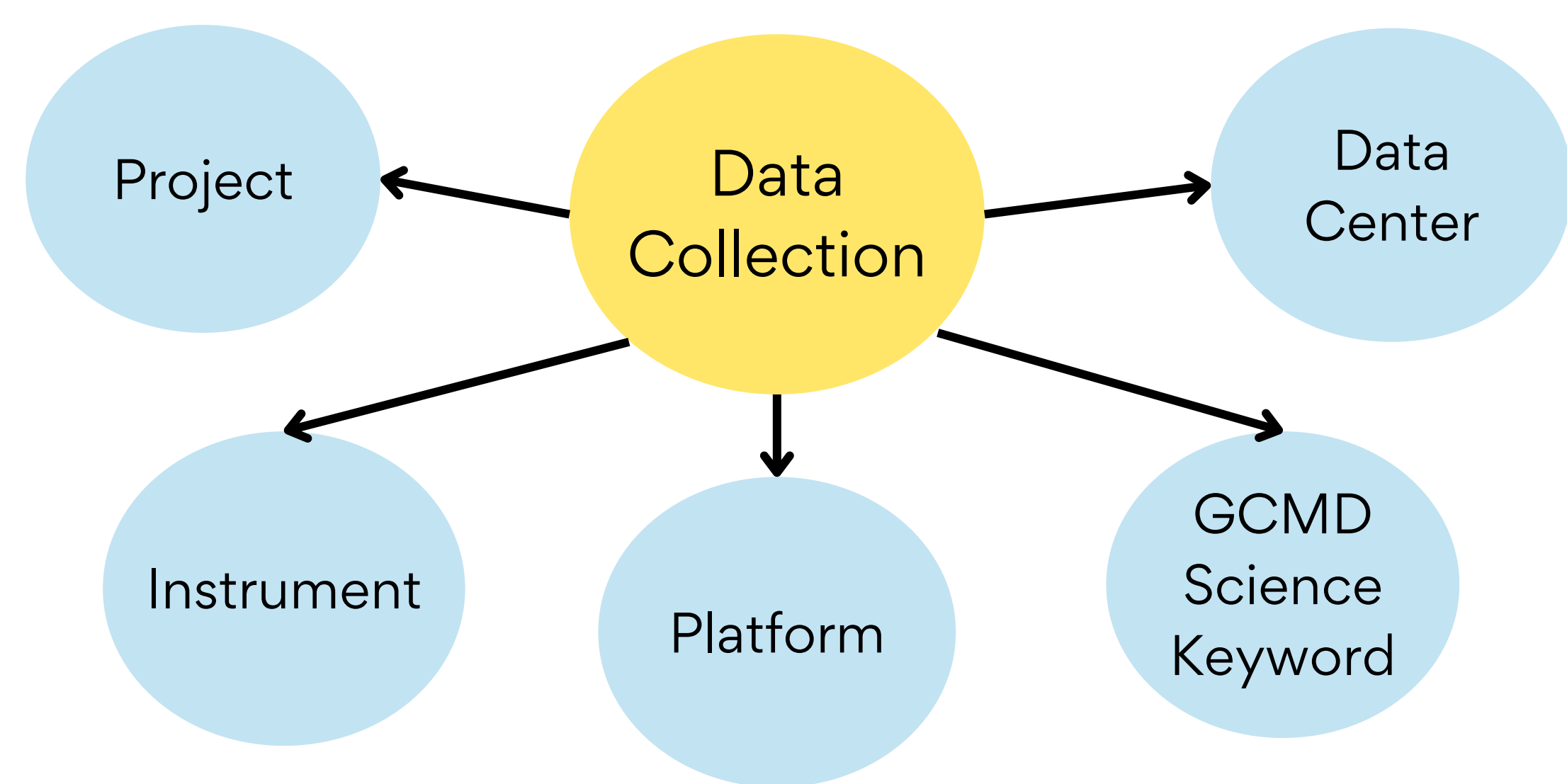
NASA's Goddard Earth Science (GES) Data and Information Services Center (DISC) is one of twelve earth science data centers in NASA's Science Mission Directorate (SMD), providing vital earth science data to a diverse user base. To enhance the discoverability of this data, GES DISC employs a keyword search system, which leverages scientific keywords from the Global Change Master Directory (GCMD) embedded in data collection metadata. However, the evolving nature of scientific applications of our data necessitates regular review and augmentation of these keywords. Currently, data curation scientists can review and augment data collection keywords with suggestions from a tool by NASA IMPACT (<https://impact.earthdata.nasa.gov>) that utilizes a large language model (LLM) trained on data collection descriptions. As part of an ongoing effort to understand the relationships between data collections at GES DISC, team members have developed a knowledge graph mapping GES DISC metadata. Here we describe an investigation into using a link prediction model, trained on a metadata knowledge graph, to augment science keyword suggestions.

SUGGESTED KEYWORD EXAMPLE

Data Collection	https://doi.org/10.5067/URTYDAGTM548
Current Keywords	Earth Science > Atmosphere > Atmospheric Water Vapor > Water Vapor Profiles
Suggested Keywords	Earth Science > Atmosphere > Atmospheric Water Vapor > Water Vapor Indicators > Water Vapor
Link to All GCMD Science Keywords	https://gcmd.earthdata.nasa.gov/KeywordViewer/

MAPPING METADATA IN A KNOWLEDGE GRAPH

All metadata, including science keywords associated with GES DISC data collections are stored in NASA's Common Metadata Repository (CMR) (<https://cmr.earthdata.nasa.gov/search/>). The knowledge graph was created by querying CMR for all GES DISC data collection metadata. The data retrieved from CMR was transformed into nodes and edges in the knowledge graph. This graphical representation of GES DISC metadata served as the training data input to the keyword prediction model.



KNOWLEDGE GRAPH SCHEMA

MODEL DEVELOPMENT AND RESULTS

The model for predicting absent science keywords in the data collection metadata was developed using Neo4j's Graph Data Science (GDS) library. Node features were generated using graph algorithms like fastRP, degree centrality, and PageRank, which capture structural and contextual properties. These features were combined using the Hadamard operation to create edge representations, which served as input for the model training. The model was trained and evaluated using **Area Under the Precision Recall Curve (AUCPR)** metrics to ensure robust prediction of relationships in the knowledge graph. We evaluated all the GDS training methods and selected the Random Forest algorithm which provided the highest AUCPR scores on average.



NODE PROPERTY STEPS

The tables below summarize the average AUCPR scores achieved during training and validation. The consistent scores around 94-95% for our validation, train, test, and outer train data collections highlight the model's strong performance for predicting the relationship between a data collection and keyword, even for portions of the knowledge graph not used in training. Additionally, the hyperparameters selected for the Random Forest classifier are provided, showcasing the configuration that provided consistent and high AUCPR scores.

AUCPR AVERAGE SCORES (%)

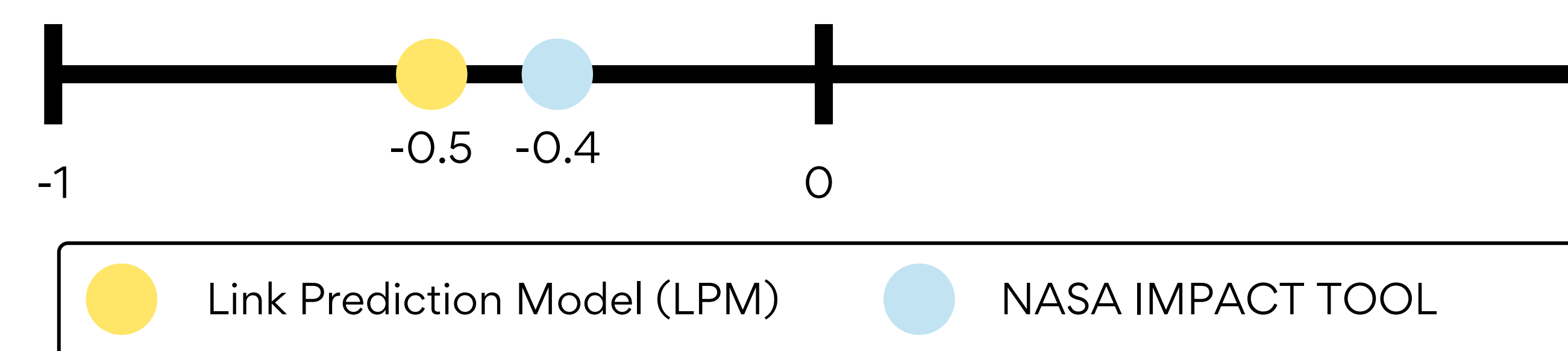
Validation	94 %
Train	95 %
Test	94 %
Outer Train	95 %

BEST MODEL PARAMETERS

Algorithm	Random Forest
Trees	100
Sample Ratio	0.07
Max Depth	Unlimited
Criterion	Gini Impurity

MODEL EVALUATION BY DATA CURATION SCIENTISTS

To understand how effective the predictions from our model are, we randomly sampled about 1.5 % of our data collections. For each sampled data collection we asked the GES DISC data curation scientist who reviews it to assess the suggested keywords from both the link prediction model and the current NASA IMPACT tool. The data collections sampled had around 5 - 15 science keywords associated with them. Our model generated an average of two suggested keywords for the sampled collections while the IMPACT tool generated an average of 5 suggested keywords. Data curators were asked to rank for each keyword whether they would agree with adding it to the data collection metadata (one point), disagree with adding it to the data collection metadata (negative one point), or feel neutral about adding it to the data collection metadata (zero points). Both tools had relatively similar scoring on average, with the NASA IMPACT tool having a slightly higher score, indicating slightly better average performance.



AVERAGE SCORES OF RECOMMENDATIONS

We also evaluated how similar each model's predictions were to the other model. Both models had a very low percentage similarity of their suggestions to the other model. This indicates that each model is providing relatively unique recommendations, despite scoring about the same in average accuracy of their keyword recommendations.

PERCENTAGE SIMILARITY OF RECOMMENDATIONS

Similarity of LPM to LLM Results	6.9%	Similarity of LLM to LPM Results	3.8%
----------------------------------	------	----------------------------------	------

FUTURE DEVELOPMENT

In the future we hope to combine the recommendations from both our tool and the NASA IMPACT tool to assist data curators with the review and augmentation of science keywords associated with our data collections. The more accurately we can describe our data collection's measurements and applications with science keywords, the easier it will be for users to find relevant GES DISC data.