AIDR 2020 Poster Session Abstracts

* denotes presenting author

#1: Standardizing Racial and Ethnic Categories to Assess Equity of Clinical Guideline Evidence

Jay Franklin*, Rensselaer Polytechnic Institute, NY
Miao Qi, Rensselaer Polytechnic Institute, NY
Shruthi Chari, Rensselaer Polytechnic Institute, NY
Morgan Foreman, IBM
Kristin Bennet, Rensselaer Polytechnic Institute, NY
Oshani Seneviratne, Rensselaer Polytechnic Institute, NY
Amar Das, IBM
Deborah McGuinness, Rensselaer Polytechnic Institute, NY

Clinical practice guidelines are widely used to support evidence-based healthcare decisions. However, these guideline recommendations may not be equitable since the evidence from cited clinical studies may be racially and ethnically underrepresented. We investigate the equitability of clinical guidelines by analyzing whether the combined cohort population from referenced studies, which we call the 'metacohort,' is representative of the population that has the medical condition. We build on methods we have developed for study cohort modeling, knowledge extraction, and equity computation to allow for measure of equity of race and ethnicity in a guideline metacohort against those of a benchmark population derived from the nationally representative National Health and Nutrition Examination Survey (NHANES).

We selected 13 US-based clinical trials cited in the cardiovascular comorbidity chapter of the American Diabetes Association (ADA) Standards of Medical Care guideline 2019 as our sample set. We then compute the racial and ethnic equity between the resulting metacohort and the NHANES population. To extract study cohort information reported in tables in research publications, we have previously developed a study cohort extraction pipeline. The pipeline begins by using the IBM Corpus Conversion Service to extract tabular data from publication PDFs. It then identifies hierarchies of table rows, annotates text segments with medical terms or other knowledge elements, and semantically relates knowledge elements to one another. The final structure the pipeline outputs is an RDF knowledge graph modeling the study cohort information and conforming to the Study Cohort Ontology, ideal for representing the oftentimes complex information within a study cohort table.

The extracted study cohort information cannot yet be combined into one metacohort and compared to a representative population, as different publications may report race and ethnicity categories that are incompatible with one another or with the representative NHANES population. Hence, all such information must first be aligned and standardized where possible, at which subpopulations can be confined and direct comparisons can be made. We have begun work on an extension module to the Human Health Exposure Analysis Resource (HHEAR) ontology. We are using this extension module to create a unifying standard between RDF knowledge graphs created by the pipeline and the benchmark population calculated from NHANES. We are implementing an algorithm to account for the differences between HHEAR and NHANES and create metacohorts that can be used for comparison.

Once standardized, the race and ethnicity categories of the metacohort can be analyzed by a previously implemented trial equity algorithm that can quantify the racial and ethnic equity of the metacohort against a target representative population. Hence, the equity of recommendations in the clinical practice guideline for specific underserved populations can be evaluated. Our existing equity algorithm compares the rate of patients under each racial and ethnic category in the metacohort to the hypothetical ideal rate of the same subgroup in an equitable RCT estimated from NHANES. Overall by leveraging the study cohort extraction, standardization, and equity algorithms, we can measure the equity gap in the guideline metacohorts and identify where more equitable clinical trial evidence is needed.

#2: Analyzing Time Series Clustering of COVID-19 Cases for US Counties

Qixuan Jin*, California Institute of Technology, CA

This poster presents a k-means clustering algorithm applied to the time series data of confirmed COVID-19 cases at the US county-level. The effect of data preprocessing methods such as outlier removal and normalization on the clustering algorithm was systematically studied. We further investigated the differing utilities of dynamic time warping and Euclidean distance as clustering distance metrics. Dynamic time warping enables the clustering of counties with similar outbreak trends that vary in temporal speed. The resulting clusters are useful as inputs to downstream prediction models. We demonstrate the better empirical performance of a Monte-Carlo dropout neural network trained in clusters for 4-week predictions against the baselines of training in individual counties and by state. The clusters computed with Euclidean distance are more suited for data visualization and analysis. Euclidean distance enables the clustering of counties with similar features on matching timelines. This poster displays the evolution of computed clusters from March 1, 2020 to August 23, 2020. With a custom curve-fitting algorithm, we automated the decomposition of individual time series into skewed Gaussian or Voigt distributions. Each distribution captures a discrete "onset wave" of increasing or decreasing COVID-19 cases in the county. We present the decompositions along with state-level policy data and the clusters for a detailed analysis of geographic regions with high case counts. The COVID-19 data is from the publicly available New York Times dataset (https://github.com/nytimes/covid-19-data). We hope that the communication of this work can aid other researchers in their efforts to address the COVID-19 pandemic.

#3: Graph Neural Networks for COVID-19 Drug Discovery

Mark Cheung* and Jose Moura, Carnegie Mellon University, Pittsburgh, PA

Deep learning techniques have led to major improvements in fields like natural language processing, computer vision, and other Euclidean data domains. Yet, in many important domains data are irregular, requiring graphs or manifolds to be explicitly modeled. One such domain is drug discovery/repurposing. Recently, research has found that using graph neural network (GNN) models, given enough data, can perform better than using human-engineered fingerprints or descriptors in predicting molecular properties of potential antibiotics.

We explore these state-of-the-art AI models on predicting desirable molecular properties for drugs that can inhibit SARS-CoV-2. We build upon these GNN models with ideas from recent breakthroughs in geometric deep learning, inspired by the topologies of the molecules. In this poster, we will present an overview of GNNs along with some preliminary results on the AID1706 dataset (which contains 290,767 molecules of which 446 are found experimentally to be inhibitors of the SARS coronavirus 3C-like Protease).

#4: Scaling the Provision of AI-Supported Micro-Credentials to Library Learners

Huaibin Zhang*, Emily Rimland, Victoria Raish, and Anne Behler, Penn State University, PA

Artificial intelligence (AI) has the power to disrupt but also to revolutionize our services and profession so that libraries can continue to contribute positive changes to the world. We are experimenting with AI as a way to scale the integration of popular information literacy micro-credentials to large numbers of students. With the assistance of Natural Language Processing and Deep Learning model, we developed an AI-based web app to predict students' performance based on their reflection of their prior experience related to badge-earning behaviors. The performance is categorized into 4 levels: 'Failed', 'Ok', 'Good', and 'Exceptional'. Before this web app was developed, this process was conducted by the instructors or learning assistants manually, which is time-consuming and inefficient. Limited by the great efforts invested in this process, only a small proportion of students can benefit from the evaluation. The aforementioned web app supported by a deep learning model could potentially be used to generalize this evaluation to a wide audience. This web app applied the artificial neural network (ANN), which has proved to perform well in the text classification tasks. The model is fitted on a dataset with 8,548 reflections and human coded grades, in which 7,123 are used as training data and 1,425 are used as testing data. After a grid search among parameters, an optimized model with a three-layer artificial neural network has been selected as the final model, each layer containing 64 neurons. The result indicated an accuracy rate of 79.9%. In addition to the deep learning supported prediction feature, the web app can also automatically generate feedback based on the predicted performance level in the textbox for the instructor or learning assistant to edit. The automatically generated feedback has significantly saved the time of the evaluators of the students' work.

By developing and applying this web app, we will be able to scale up the micro-credential program to evaluate more student work, gain data richness about our learners, innovate at the micro level, and augment and improve our teaching practices. This session will demonstrate an innovative application of AI to the new form of academic currency–the micro-credential–and discuss how this application could lead to broader uses in libraries and academia.

#5: AITom: Open-Source AI Platform for Cryo-Electron Tomography Data Analysis

Min Xu*, Carnegie Mellon University, Pittsburgh, PA

Cryo-electron tomography (cryo-ET) is an emerging technology for the 3D visualization of structural organizations and interactions of subcellular components at near-native state and sub-molecular resolution. Tomograms captured by cryo-ET contain heterogeneous structures representing the complex and dynamic subcellular environment. Since the structures are not purified or fluorescently labeled, the spatial organization and interaction between both the known and unknown structures can be studied in their native environment. The rapid advances of cryo-electron tomography (cryo-ET) have generated abundant 3D cellular imaging data. However, the systematic localization, identification, segmentation, and structural recovery of the subcellular components require efficient and accurate large-scale image analysis methods. We introduce AITom, an open-source artificial intelligence platform for cryo-ET researchers. AITom provides many public as well as in-house algorithms for performing cryo-ET data analysis through both the traditional template-based or template-free approach and the deep learning approach. Comprehensive tutorials for each analysis module are provided to guide the user through. We welcome researchers and developers to join this collaborative open-source software development project.

#6: Machine Learning for the Automated Detection and Classification of Seabirds, Waterfowl, and Other Marine Wildlife from Digital Aerial Imagery

Kyle Landolt*, USGS - Upper Midwest Environmental Sciences Center
Timothy White, Bureau of Ocean Energy Management
Mark Koneff, U.S. Fish and Wildlife Service
Jennifer Dieck, USGS - Upper Midwest Environmental Sciences Center
Travis Harrison, USGS - Upper Midwest Environmental Sciences Center
Luke Fara, USGS - Upper Midwest Environmental Sciences Center
Larry Robinson, USGS - Upper Midwest Environmental Sciences Center
Enrika Hlavacek, USGS - Upper Midwest Environmental Sciences Center
Brian Lubinski, U.S. Fish and Wildlife Service
Dave Fronczak, U.S. Fish and Wildlife Service
Lucas Spellman, Univ. of Wisconsin – La Crosse
Simon Wagner, Univ. of Wisconsin – La Crosse
Stella Yu, Vision Group at the International Computer Science Institute at Univ. of California – Berkeley
Tsung-Wei Ke, Vision Group at the International Computer Science Institute at Univ. of California – Berkeley

Avian and wildlife population surveys can aid in informing regulatory decisions, environmental assessments, and impact analyses of offshore energy development projects. Additionally, low-flying ocular aerial surveys have historically been used to estimate waterfowl populations, but place agency personnel at risk of injury and survey results are prone to bias and misclassification. The U.S. Geological Survey (USGS), in collaboration with the Bureau of Ocean Energy Management (BOEM) and the U.S. Fish and Wildlife Service Division of Migratory Bird Management (USFWS-DMBM), is advancing the development of deep learning algorithms and tools to automate the detection, enumeration, and classification of seabirds, waterfowl, and other marine wildlife. High resolution aerial imagery collected from the Atlantic Outer Continental Shelf and the Great Lakes will provide data for algorithm development. An annotation database, seeLIFE, is being developed that contains targets (i.e., birds and other marine wildlife) from the imagery with other relevant attributes (e.g., species, age, sex, and activity). OpenCV's Computer Vision Annotation Tool (CVAT) is providing the framework for an web-based interactive GUI, the Wildlife Annotation Tool, allowing wildlife experts to efficiently create annotations and support annotation database development. Our research also explores the impacts of image glare and background color on the performance of object detection models.

Using our interactive GUI, we labeled approximately 30,000 bird objects in 169 images. Of these, approximately 10,000 bird objects were annotated with attributes of species, age, sex, and activity by trained biologists. Using the 10,000 annotated bird objects, we developed an object detection model using a MaskRCNN framework. The model explicitly detects birds without any further classification. Current performance is benchmarked at a mAP (mean average precision) of 0.47 and an AR (average recall) of 0.57. Clustering images by background color also affects model performance, with images of lighter blue tones performing better compared to images of darker values. Images with an increased presence of glare performed slightly better in our dataset, as mAP and AR were 0.6 and 0.7, respectively, for images with 0 – 0.1% glare while mAP and AR were 0.66 and 0.73, respectively, for images with ¿ 1% glare. Further work will be done to implement a multi-object detection model to detect and predict objects of certain species, age, sex, and activity while taking other co-variates like ground sampling distance into account. Lastly, we find that using deep learning algorithms to detect birds reduces the time manually annotating bird objects by 95%, rapidly accelerating annotation development. We expect to use machine-generated annotations as additional training data for future model development as we confirm its accuracy.

#7: Supporting Interdisciplinary Research Reviews with Multi-Level Topic Maps

Sara Lafia*, University of Michigan, Ann Arbor
Werner Kuhn, University of California, Santa Barbara
Kelly Caylor, University of California, Santa Barbara

Research reviews lack systematic methods for generating high-level, narrative insights in interdisciplinary contexts; it remains challenging to quantify, let alone compare, research accomplishments across academic disciplines. Abstraction methods that enable the "distant reading" of corpora are increasingly important for knowledge discovery in both the sciences and humanities. We examine the utility of a spatial approach for abstracting and mapping topics emerging from an interdisciplinary earth research institute (ERI). As an organized academic unit, ERI supports researchers spanning 24 academic departments that encompass the full breadth of earth and environmental sciences. This study strives to capture and represent institutional knowledge at ERI in support of a multi-year review of its interdisciplinary research activities. We present a systematic, reproducible, and data-driven approach for eliciting and spatially representing cross-cutting topics from ERI's body of research. This approach first models cross-cutting topics from publication and project metadata at multiple levels of detail using non-negative matrix factorization (NMF). The topics are then used to co-locate related research products with t-distributed stochastic neighbor embedding (t-SNE), which yields topic maps at multiple levels of detail that reveal the latent thematic structure of the corpus. ERI's researchers evaluate the topic maps by "reading" ERI's body of research at multiple levels of detail (i.e. at a distance). We find that the researchers gain insights from identifying and interpreting the positions of their own work relative to the work of their colleagues over time; this demonstrates the potential for spatial abstraction to complement the metrics on which many institutional research reviews currently rely. Our approach produces a decision support tool that exposes areas of thematic expertise, relationships among researchers, topical distributions and clusters of work, and importantly, the evolution of these aspects over time.

#8: Inventing Curriculum using Pointer-Generator Network

Gajendra Deshpande*, KLS Gogte Institute of Technology, India

The curriculum in general and undergraduate curriculum, in particular, is one of the most important pillars of an education system. The undergraduate curriculum has two main objectives i.e. employability and higher education. The greatest challenge in designing an undergraduate curriculum is achieving a balance between employability skills and laying the foundation for higher education. Generally, the curriculum is the combination of core technical subjects, professional electives, humanities, and skill-oriented subjects. We used natural language processing and machine learning packages in Python to build a curriculum design system.

The steps to build a curriculum design system are described below:

1. The dataset was built from the job profiles from different job listing websites like stackoverflow.com, indeed.com, linkedin.com, and monster.com. Also from the syllabus of competitive exams and qualifying exams for higher education.

2. On the dataset, we applied natural language processing techniques to identify the subjects and subject content. For natural language processing, we used spaCy an industrial strength Natural Language Processing package in Python.

3. To generate syllabus content for a particular subject, a pointer-generator network was used. The pointer generator network is a text summarization technique that combines extractive and abstractive summarization techniques. The extractive summarization technique extracts keywords from the dataset, whereas the abstractive summarization technique generates new text from the existing text. The pointer-generator network was implemented using the scikit-learn machine learning package in Python.

4. The generated curriculum was then compared with the existing curriculum to get insights like, how much percent of the curriculum is industry oriented, how much percent of the curriculum is aimed at higher education, and job-oriented skills.

5. The above steps can be repeated with modified parameters to get better insights and curriculum. This also gives us an idea of how we can have an evolving curriculum that can help us bridge the gap between industry and academia.

We used ROGUE (Recall-Oriented Understudy Gisting Evaluation) metric to compare the generated curriculum against reference/proposed curriculum.