Research Statement

KMA Solaiman

Real-world use-cases in data-centric applications (including societal, healthcare, or education) with minimal computational resources often have an unprecedented influx of unstructured and noisy data from multiple sources and modalities. Extraction of meaningful information from such heterogenous and changing datasets requires achieving the complementary functionalities of *cross-modal matching, scalable data management system (mostly search engines and databases),* and *situationally-aware data recommendation.* My research goal is to understand how these functionalities can be achieved by designing interactive algorithms and robust systems that solve the problem of **situational knowledge on demand in open-world** from multiple forms of information, regardless of whether presented as text, images, videos, audio, or other modalities, while achieving data-driven and resource-aware *data management* and *data integration* capabilities.

My later works branched into developing scientific principles to *quantify* and *characterize novelty* (significant and unexpected events) *in open-world domains*, while creating scalable and efficient AI systems that *react to novelty* in those domains.

Developing intelligent AI systems including situational knowledge recommenders for open-world environments involves tackling multiple *system design challenges*:
(1) Resource-aware Data Management: Data management systems require a comprehensive understanding of the data properties, user requirements, and limitations imposed by open-world to achieve optimal performance and scalability for multimodal data recommendation. Emerging multimodal applications impose challenges for traditional system (hardware and software) design choices ranging all the way from input (heterogeneous sources, context and modalities) and output (delivery-on-time, quick throughput and diverse users), changing information needs (knowledge base creation and query), to computational resources (lack of annotations, domain-specific feature extractors or human resources). For instance, we showed that transfer learning for fine-grained semantic concept extraction from videos turned out to be ill-suited in large-scale systems [4].
(2) Data Integration: Data integration from various sources to answer queries over a single view of the data to users, is confronted with a multitude of heterogeneity issues. These problems arise from differences in data attributes names that hold similar data [9, 7], communication problems, and variations in data schema and types [6]. As data volume increases and the necessity to share existing data intensifies, data integration becomes more prevalent. My first approach for data integration, EARS, delivers integrated query results over time using a mediation approach and schema mapping [9], solving the problem of scalability and quick throughput. The second approach, FemmIR, learns a co-ordinated graph representation of the data samples comprised of their semantic features to deliver approximate matches [7]. The third approach, WesJeM, uses Contrastive Learning to embed data-objects and their semantic properties in a high-dimensional space using higher-level semantic features in a data sample as weak labels [6], allowing zero-shot similarity matching and data discovery of multimodal data in open-world environments.
(3) Dealing with Open-world Novelties: To construct intelligent AI systems, it is necessary to adapt to evolving scenarios. However, conventional AI systems face restrictions when it comes to managing unexpected events or "novelties" that were not previously seen or modeled. We need to characterize, detect and adapt to novelties at various stages of the AI life cycle, such as data integration, relevance learning, environment modeling, feature extraction, and inherent domain properties. Novelty characterization and difficulty estimation is required in a plethora of AI systems ranging from multimodal information retrieval [6] (distribution change and concept drift), dataset complexity [1], to visual (object detection in video and image [2]) and planning domains (games [5] and war).

My work has shown positive results on open societal problems that previously required large-scale human endeavor and computational resources, such as Missing Person Search [9], Dataset Complexity Estimation [1], Search and Rescue in Disasters [8], and Medical Triage. The impacts of my work in academia span across diverse areas, including Multimodal Information Retrieval, Data Discovery, Building Robust AI Agents, and Data Completion, thereby creating a significant impact.

My research on Novelty Adaptation and Situational Knowledge Delivery has the capability to **transform information retrieval for everyone by empowering AI systems** to promptly and precisely retrieve the information required by users, even in situations where the resources are minimal, data sources are in a state of constant change, and involve various modalities, interconnections, and predictive arguments.

RESEARCH PHILOSOPHY. I believe that the ultimate objective of AI is not just to enhance performance on isolated tasks but to complement human abilities in solving persistent issues in real world, minimize human labor through responsible action, harness the power of data for our benefit, and make decisions that have long-lasting positive effects. My work is driven by a desire to integrate these ideas into a context that prioritizes social good and encourages the consumption and sharing of healthier information. This involves leveraging multimodal techniques to better understand

complex situations, adapting to different situations, and providing trustworthy and easily understandable information for humans.

RESOURCE AWARE DATA MANAGEMENT FOR MULTIMODAL APPLICATIONS

In designing data management systems for modern applications, such as missing person search, disaster resource management, triage, and emotion recognition, the focus has shifted to account for data-at-rest and streaming input while ensuring scalability to handle increasing information needs and data ingestion. To address these challenges, during our collaboration with local police department and MIT for building Missing-person Query engine [9] and Human-in-the-Loop Video Querying system [4], we proposed a novel multimodal knowledge querying system called **SKOD** (Situational Knowledge on Demand) [3, 4]. SKOD leverages entity-centric higher-level semantic concepts (such as objects, object types, physical relations, e.g., a person wearing blue shirt, time and place of an incident), and the functionalities of distributed systems and RDBMS to query domain-specific information needs in practical multimodal applications. SKOD was developed in collaboration with Northrop Grumman, MIT, and CMU and demonstrated at Northrop Grumman TechFest in 2019. The project has been funded by Northrop Grumman for three consecutive years since 2019, renewing the funding every year.

Heterogenous Data Ingestion, Scalability, and Delivery-on-demand

We used Postgres as the backend architecture for both data storage and on-time delivery. Building on top of the RDBMS allows us to scale to practical data volumes, as well as using the querying interface with query-by-example and query-by-features dramatically lowers the human costs of multimodal and visual domain search [4]. For consuming data from heterogeneous sources (both at rest and streaming), I integrated Kafka producers and consumers on top of SKOD [3]. Any query to the system was formulated and considered as an *incident in real life*. For delivery-on-demand from incomplete modalities, we used Postgres Trigger functionality, which is activated whenever an insert occurs that matches a certain incident (any matching data). This feature allows us to deliver incomplete information need and complete it later when new matching data is encountered, while being capable of adapting to changing information requirements. Queries in SKOD can be both standing queries or one-shot queries. To deal with the changing requirements, we proposed to build a ***query-drive knowledge base*** for each user, where all queries can relate to a single incident. SKOD speeds up the data delivery by storing frequent incidents by caching hot queries, and recently used data.

Resource-constrained Feature Extraction

Task-specific querying systems face challenges during the data preparation stage due to low-quality data sets and a lack of labeled training samples. Although large-scale language models have made significant progress, there exists very few task-specific attribute extractors for text. Our team addressed these issues in SKOD by implementing a *priority polling system* that selects candidate data samples for feature extraction from videos and images, instead of immediately processing features for batch inputs. This feature, coupled with trigger functionality, enables us to provide information needs on-demand and complete them with time. Additionally, we developed a **cloth-color extractor** for videos using common-sense reasoning and color and shape analysis [4] on top of YOLO. To identify attributes in unstructured text, I propose a model called **HART** [7], which solves the problem in two stages: (i) **candidate sentence identification** by transforming the problem into a similarity-search problem using pre-trained language representation models (SBERT) and lexical knowledge bases, and (ii) **semantic attribute understanding** using syntactic characteristics and lexical meanings of the tokens in the candidate sentences. This approach can be generalized for any domain and lays the groundwork for intelligent document processing.

LABEL-EFFICIENT DATA INTEGRATION AT HIGHER SEMANTIC LEVEL

View-based Data Integration

Traditional data integration approaches suffer because of heterogeneity among data sources and incomplete modalities. Machine learning models for multimodal data fusion learn joint representations to exploit complementarity and redundancy of multiple modalities, but overlooks the information needs based on higher-level semantic concepts. With the use of Postgres trigger and by using a mediated schema for each queried incident, SKOD delivers integrated query results over time. Since the number of properties-of-interest are quite moderate, using similar approach to the *Global as View* data integration, I proposed to employ **schema mapping** between the *mediated schema* and *local schema* from different data sources. The proposed approach, **EARS** [9] adopts an entity-relationship-attribute schema for each new data source, and a wrapper is designed to translate the source schemas to the mediated schema. The queries are

translated into conjunctive queries between features among data sources and a SQL-Join is performed at run-time to integrate all the relevant sources. Using the versatility of Postgres, we achieve the scalability and speed that is required for time sensitive use-cases, with minimal amount of computational resources.

Approximate Matching using Graph Representation Learning

While the SQL-JOIN based relational DBMS approach allows a lot of flexibility, it does not utilize the historical knowledge of previous queries, and cannot perform approximate matching. Considering the sensitivity of some open-world application domains, it is desirable to search for approximate relevance between different modalities and sources. Motivated by representation-invariant properties of graph representation models combined with the existing works on approximate graph matching techniques, I propose **co-ordinated graph representation learning of the data samples comprised of their semantic features** [7], where it learns to approximate a novel Edit distance metric, *CED*, based on the multiplicative comparison of the *Hierarchical Attributed Relational Graph* representations.

Weakly Supervised Metric Learning for Cross-Modal Matching

For real-world systems, *data discovery from heterogenous modalities* and *explanation of the relevant properties among similar data objects* is of equal importance. Since in these applications, manual annotation is not feasible or they lack annotation resources, we need alternative supervision techniques for cross-modal matching. Motivated by the advancement in translation and captioning models (video/audio $\rightarrow$ text), I propose to embed data-objects and their semantic properties in a high dimensional embedding space via Contrastive Learning. After extracting the interaction among entity-centric higher-level semantic features (such as, topics, events, entities, triplets) from texts and other translated modalities, a *data information network* is built by connecting data samples to their features via their interactions. Finally, I construct a structure-infused representation for the data-objects from all modalities in **WesJeM** [6], by jointly embedding the data samples, the features, and the available similarity labels, in a single space. For learning, I defined a multi-task learning objective capturing the interaction information, by aligning the representation of the data samples, defined by their textual content, with the representation of features, based on their common relations. For open-world environment where data and information-need keep changing, along with the dynamic data sources, WesJeM opens up the path for **Zero-Shot similarity matching** and **Data Discovery** of multimodal data.

ADAPTION TO OPEN-WORLD NOVELTIES

AI systems are often limited by their inability to handle unexpected events that are not part of their training data or well-defined environments. These significant changes or events are referred to as **'novelties'** under DARPA SAIL-ON project, and their characterization and adaptation is critical for real-world applications. To build robust and intelligent AI systems, I developed novelty characterization and adaptation techniques at various stages, including data integration and relevance learning, environment modeling, feature extraction, training, and domain or data level.

Novelty Characterization, Detection, and Difficulty Estimation

I characterized the **novelties encountered in multimodal information retrieval** in [6] and proposed how WesJeM can be adapted for changing data patterns and incomplete or noisy modalities in data integration and relevance learning stage. Moreover, motivated by the information-theoretic approach for difficulty estimation of novelties, I proposed an empirical framework for novelty characterization and difficulty estimation in **planning domains** [5]. For a reinforcement-learning based Monopoly agent, graphically modeling the environment to augment the state and action space allow to integrate graph edit distance as a novelty difficulty metric.

Robust Feature Extraction with Dataset Augmentation

The efficiency of entity-centric machine learning models in response to novelties depends on the efforts during the model training, design and data collection stages. We proposed a **novelty generation framework** [2] at the data preparation stage of training a model to assure its robustness and reduce the bias. We augmented the original dataset in a domain-agnostic and budget efficient manner with generated novelties for visual modalities, and improved the **novel object detection** performance with the augmented dataset.

Intrinsic Domain Complexity Estimation for Distributed AI Systems

Understanding of the inherent characteristics of the domain is essential for novelty characterization and model adaptability. We proposed an **application-independent domain complexity** measure for the AI systems in perception domain [1] using **federated learning as the reference paradigm** to handle distributed dataset operations. Build-

ing upon intrinsic dataset properties such as dimensionality, heterogeneity and sparsity for singular environment, we created a complexity metric *for the distributed environment*, showing efficacy for classification task.

FUTURE RESEARCH AGENDA

Currently, we are experiencing a thrilling era for multimodal information processing and robust AI research since it is highly supported by the core programs in NSF's Division of Information and Intelligent Systems (IIS) and by "Harvesting the Data Revolution (HDR^2)" idea - second wave of one of the 10 big ideas by NSF for long-term research.

My **long-term goal** is to create intelligent systems that can reason, learn and cooperate with humans to improve the standard of living by utilizing the vast amounts of data available in the modern era. My focus is to devise new algorithms and methods that can make a significant impact on society, leverage existing scientific advancements, and address real-world challenges. To that end, I plan to continue my research on *multimodal data management in real world* by approaching from the following directions:

User Preference Modeling

To complete the life-cycle of *situational knowledge delivery*, we still have challenges in modeling user's information need in a robust and efficient manner in multiple directions [9]: (1) user requirement is not always obvious or explicitly stated, (2) user can be interested in multiple types of events and knowledge bases with varying probabilities, (3) learning algorithms need to *adapt to changing user preferences with time*. I aim to develop novel algorithms using techniques such as active learning and reinforcement learning that can accurately capture and predict users' preferences based on their behavior, interactions, and feedback. Understanding the features that drive user preferences, and leveraging this knowledge to improve personalized recommendations and user experience, has applications in education (student advising, classroom teaching), e-commerce, healthcare, etc. To achieve this research goal, collaborations with researchers in **human-computer interaction, psychology, and marketing** will be essential.

Explainability and Trustworthiness in Data Recommendation

As the amount of multimodal data generated and consumed by users is increasing, there is a growing need for users to understand the basis for recommendations [7] and the saliency and trustworthiness of the information being consumed. This is especially important in sensitive domains such as *healthcare, finance, and legal decision-making* to allow for tracking, cross-checking with social contexts and verification. To achieve this goal, collaboration across multiple areas is necessary, including **data science, natural language processing, computer vision, human-computer interaction, and ethics**. With this, we can ensure that these models are designed with the user in mind, taking into account their cognitive and perceptual abilities. This collaboration can also lead to the *development of ethical guidelines and principles for designing trustworthy systems*, ensuring that users' rights and privacy are protected.

Privacy preserving Data Dissemination and Federated Learning

To address the growing concern over data privacy, particularly in medical and identity contexts, research in privacy-preserving multimodal data dissemination and federated learning is crucial, as identified in SKOD framework [3]. Further research to integrate the use of local data processing and remote federation with multimodal machine learning techniques is needed to ensure this new requirement in information processing, while understanding and formalizing the resource requirements. Collaboration across various fields such as **information security, statistics, data management, law, ethics, and public policy** is vital to advance research in this area.

Information Completion and Data Democratization

As data becomes increasingly important in all domains, there is a need for new techniques that enable individuals and organizations to efficiently extract insights from data and complete missing information. To address this challenge, future research should focus on developing advanced machine learning models that are able to perform well even with incomplete data, as well as methods for effective data integration and knowledge transfer within organizations. Collaboration is needed between **machine learning experts, data management specialists, and domain experts in various fields** to achieve a comprehensive and effective solution for data democratization and information completion.

COLLABORATION AND FUNDING

My future research vision requires collaboration with expert researchers in many fields, including natural language processing, computer vision, machine learning, data mining, social science, human computer interaction, systems and databases. I gained extensive expertise in overseeing and directing major projects, encompassing teams of over

12 individuals and collaborating with various universities and institutions. I led multiple masters and undergraduate students, collaborated with multiple Ph.D. students and coordinated with 5 professors from different universities to participate in the REALM project. I am fortunate to have close collaborations with professors from multiple universities and research institutes, such as Massachusetts Institute of Technology (MIT), University of Michigan (UMichigan), University of Southern California (USC), Information Sciences Institute (ISI), Institute for Defense Analyses (IDA), University of Massachusetts (UMass), Middle East Technical University (METU), etc. I also have had the fortune to work closely with researchers from databases and applications, along with end-users and program managers to conduct interdisciplinary research. I intend to maintain my current collaborations while actively cultivating new partnerships to advance the establishment of robust principles that underpin research in multimodal knowledge and novelty in learning models.

During my Ph.D., my work has been mainly supported by the Northrop Grumman Corporation (NGC), DARPA, ARFL, and Sandia National Lab. Additionally, I have contributed significantly to the writing of grant proposals, including idea generation, method design, idea illustration and visual aid creation, such as DARPA ITM project, and DARPA Triage Challenge. As a future faculty, I will continue to seek funding opportunities in the future from early career fellowships and various funding agencies (e.g., DARPA, ARL, AFRL, IARPA, NSF, NIH, DOE, DOD) and industries (e.g., NGC, Microsoft, IBM, Ford, Meta, Google, Intel).

REFERENCES

[1] Shafkat Islam*, **KMA Solaiman***, Ruy De Oliveira, and Bharat Bhargava. Domain complexity estimation for distributed ai systems in open-world perception domain. *Submitted to Artificial Intelligence*, 2023. ***Co-first Author**.

[2] Alina Nesen, **KMA Solaiman**, and Bharat Bhargava. Dataset augmentation with generated novelties. In *2021 Third International Conference on Transdisciplinary AI (TransAI)*, pages 41–44. IEEE, 2021.

[3] Servio Palacios*, **KMA Solaiman***, Pelin Angin, Alina Nesen, Bharat Bhargava, Zachary Collins, Aaron Sipser, Michael Stonebraker, and James Macdonald. SKOD: A framework for situational knowledge on demand. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare with VLDB*, pages 154–166. Springer, 2019. ***Co-first Author**.

[4] Michael Stonebraker, Bharat Bhargava, Michael Cafarella, Zachary Collins, Jenna McClellan, Aaron Sipser, Tao Sun, Alina Nesen, **KMA Solaiman**, and Ganapathy Mani. Surveillance video querying with a human-in-the-loop. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics with SIGMOD*, 2020.

[5] **KMA Solaiman** and Bharat Bhargava. Measurement of novelty difficulty in monopoly. In *Proceedings of the AAAI Spring Symposium on Designing Artificial Intelligence for Open Worlds*, 2022.

[6] **KMA Solaiman** and Bharat Bhargava. Open-learning framework for multi-modal information retrieval with weakly supervised joint embedding. In *Proceedings of the AAAI Spring Symposium on Designing Artificial Intelligence for Open Worlds*, 2022.

[7] **KMA Solaiman** and Bharat Bhargava. Multi-modal information retrieval for systems with explicit information needs and object properties (FemmIR). *Submitted*, 2023.

[8] **KMA Solaiman**, Md Mustafizur Rahman, and Nashid Shahriar. AVRA bangladesh collection, analysis & visualization of road accident data in Bangladesh. In *International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 1–6. IEEE, 2013.

[9] **KMA Solaiman**, Tao Sun, Alina Nesen, Bharat Bhargava, and Michael Stonebraker. Applying machine learning and data fusion to the *Missing Person* problem. *IEEE Computer*, 55(6), 2021.

Teaching Statement

KMA Solaiman

1 TEACHING PHILOSOPHY

Accessible Learning My teaching philosophy revolves around active learning and student well-being. Student well-being in a classroom can depend on their sense of autonomy, achievement, and participation. For example, in one of the courses I taught, Network Programming, the students were given a choice for their project topics from reproducible papers, novel idea implementation, or ongoing projects. Also, for the final presentation deliverable, they were asked to choose from multiple modalities i.e., video presentation, demo, or blog write-up. These choices gave the students a sense of autonomy. Satisfying these needs could intrinsically motivate the students to participate and grow in the classroom actively.

Relationship with the Students At the beginning of my classes, I make a conscious effort to know the students' backgrounds and their preferred outcomes from that course. This helps me understand them better, formulate relevant lectures during the teaching, and allow me to adjust the course outline accordingly. *"If you cannot explain it simply, you do not understand it well enough."* - I truly believe that and want that as an outcome for my students. To that end, I ask for *frequent feedback* from my students to learn more about their understanding and thought processes and identify any alternate conceptions students may have that need to be addressed.

Growth-focused Course Design (i) *Course Materials:* I prefer to design courses in a way that gives the students the ability to measure their progress. The courses consist of both the foundational CS knowledge and the application of the concepts to problem-solving. Depending on the audience type, the course will have varying styles of materials. For students aspiring for industry positions, it is imperative that they can apply academic knowledge to solve real-life problems. It is essential for those pursuing research careers to come up with solutions to new problems from existing knowledge. I grew very fond of the concept of *learning by doing* from my courses and prefer to utilize this. Allowing students to do different types of written or programming assignments alongside the relevant lectures bolsters their understanding of theoretical knowledge. (ii) *Assessment:* During grading, I always design fine-grained rubrics with lower stakes and provide detailed feedback to the students so that they can understand their mistakes and learn from them. Along with testing the fundamental and applied knowledge, I frequently add problems during the coursework that make them think on a deeper level.

2 TEACHING EXPERIENCE

My teaching experience consists of hands-on teaching for six years, dating back to 2014, and taking training courses such as ***Effective teaching in CS*** or ***Foundations of College Teaching***. The graduate TA training after COVID-19 has helped me familiarize myself with online teaching while learning about essential tools such as Brightspace, Campuswire, Gradescope, etc.

My first experience teaching undergraduates was in Bangladesh, starting in August 2014. **I instructed in multiple labs of graphics, data structure, and programming language.** While later two involved explaining to students how the core concepts are applied in problem-solving, *graphics* allowed me to mentor students to build tangible outputs. My second experience in Bangladesh was on a much broader scale teaching ***Network Programming***, ***Database,*** **and** ***Software Engineering*** with a maximum class size of 143, divided into two sections. **My duties included delivering the lectures, designing labs and course materials, conducting labs with assistants, grading, and advising.** By handling large classes and multiple sections, I learned the process of guiding and evaluating everyone in a balanced manner. All these courses required building full-stack projects. It taught me how to help students throughout semesters while giving continuous feedback. The satisfaction I shared with my students seeing the final projects bolstered my choice of being a teacher. As for academic services, I contributed to curriculum development for several courses, participated in departmental activities, and contributed to accreditation.

After joining Purdue, I became a teaching assistant for **object-oriented programming**, a freshman-level course. By the third semester of teaching this course, I have taken the role of structurally developing this course for the long run. **During 2016-23, I was a teaching assistant for two graduate-level courses along with three undergraduate courses.** The biggest difference from my previous teaching experience was the multicultural and diversified international student body. At Purdue, my responsibilities as a GTA included designing, testing, and grading programming assignments, projects, and written homework, teaching labs, and PSO sessions, assisting in creating and grading exams, and advising students during office hours and in online forums. Through this process, I have realized how differently students at senior and junior levels learn and communicate with instructors. For lower-level undergraduate courses, I was responsible for overseeing undergraduate TAs. For upper and graduate-level classes, I had to handle

complex situations like discussing students' fundamental research questions or mentoring for research reproduction. My goal for the weekly PSO sessions is to help students utilize the concepts learned in the lectures for their assignments with an in-depth understanding. As soon as an assignment is released, I go through the logic behind it and how the core concepts build up to the final outcome. During the grading, I tried my best to leave detailed feedback on their implementation issues so they could learn from their mistakes. The student's mental health is essential to me. I encourage and practice empathy in my class, which includes being aware of their hesitancy to ask questions or being inquisitive about their learning process. Most recently, one of my students in network programming class felt comfortable enough to talk to me about his anxiety. I have tried my best to accommodate and comfort him with Purdue's ongoing support for mental health management.

I have guest lectured on *situational knowledge, knowledge graphs, and multimodal information retrieval* where I talked about *cross-correlation learning, metric learning, decoder-encoder, and attention networks*. The lectures involved *feature extraction from multiple modalities, graph embedding,* and *graph matching techniques*. I have always enjoyed public speaking, which has helped me to deliver presentations at a large scale numerous times during my Ph.D. I have presented our research works at Northrup Grumman Corporation Review Meetings and Techfest (with 100+ attendants), JPL Nasa, and Darpa Review Meetings.

Mentoring: During my time in Bangladesh, I mentored seniors for their final year thesis and projects and served as an external member of the evaluation committee. I also participated as a coach for *ACM-ICPC* at AUST. At Purdue, I have mentored 13+ masters and undergraduate students for independent research.

3 TEACHING PLANS

I am excited to teach both undergraduate and graduate-level students. I am comfortable teaching software engineering, databases, networks, compilers, information retrieval, data management systems (SQL and NoSQL systems), and machine learning courses to both undergraduate and graduate students. I know the differences between them from my teaching experience at both levels. For undergraduate students, I would be happy to teach core courses, including programming languages, data structure, theory of computation, and computer graphics. Specifically, for graduate students, I can offer advanced information retrieval, natural language processing, distributed database systems, data science, or data mining. Moreover, I have plans to offer seminar courses related to multimodal information retrieval and adaptable and explainable AI to students with a research interest. These courses will be based on recently published papers in top conferences in machine learning, information retrieval, OpenAI, and XAI conferences. The course projects will be designed in a way that can lead to publications and can mimic peer reviewing.

Besides that, I would also like to create a new course *"Applied Machine Learning for Open World Systems"* to extend the current curriculum based on my research experience. Tentative topics would include data cleaning, handling lack of annotations, scalability, weakly supervised learning, multimodal information retrieval and feature extraction, intrinsic and extrinsic complexity of system domains, and case study of real-world use cases. Since learning algorithms must be adaptive to real-world situations, it will also include detection, adaptation, and difficulty analysis of novelties in machine learning algorithms.

Diversity Statement

KMA Solaiman

My commitment to encouraging diversity, equity, and inclusion in our community emits from my realization of the barriers that exist in higher studies. During my journey from a small-town high school student to a first-generation international graduate student at a public research university, I was fortunate enough to be inspired and guided by many altruistic role models. This showed me the importance of having a supportive and inclusive environment in higher education for students to achieve their full potential.

Transitioning from being a student to a researcher or leaving home for a foreign country is a daunting task in and of itself. From my experience of coming from a developing country and being friends with students from Colombia, Ghana, Thailand, and South Korea, I learned about different socio-economic hurdles people from under-represented countries had to overcome for a chance at higher studies. **Lack of information and accessibility** often makes it harder for under-represented groups (URG) even to take the first step towards it. The culture of inclusion during my Ph.D. has helped me identify the micro-aggressions and inequalities I witnessed growing up. During my undergraduate in Computer Science (CS), only 10% of our class were girls, along with only 3% of ethnic minorities. This gender inequality has emerged **either from the lack of female leaders in CS in Bangladesh or from the misconception that CS is a difficult career to pursue. For the ethnic minority groups, the issue was the lack of preparation opportunities for the very competitive admission tests compared to the students in large cities.** Although the situation has improved, these numbers still necessitate conscious and assertive steps toward inclusiveness and diversity in higher education.

PAST ACTIVITIES. My advisor has been a great advocate for equity at Purdue University and in the Department of Computer Science. During my Ph.D., having him as a mentor has helped me learn from him and participate in promoting diversity and inclusion.

***Promote Opportunity for All*:** During 8^{th} grade, I realized kids from my neighboring tenement home could not afford basic primary education due to their economic conditions. Being close to them, I started teaching them primary subjects for a year before I moved. Many of them have completed their college education now and still is in touch with me. During my bachelor's, I tutored a large number of students for college admissions in Engineering in remote areas of the city allowing them to have more time for preparation rather than travel.

***Collaborations beyond Borders*:** Collaborating with people from different countries such as the US, Panama, Ukraine, China, India, and Turkey and across multiple disciplines has broadened my view in terms of working style, personality, culture, and outlook. Paired with a senior colleague from Honduras with an extensive background in Software Engineering, I adapted to engineering-heavy research. We published the first paper for the REALM project while working with people from Boston, Indiana, and Turkey. I took the initiative to lead the paper and had to juggle between different timelines while traveling to Bangladesh. This led to multiple collaborations with a female colleague from Ukraine who has completed her Ph.D. and is working for Target now. After working through two successful publications, my collaborator from MIT has completed his Master's degree and has returned to China as an Engineer.

***Access through Mentoring*:** As our research group harbors a culture of mentoring, I have had the opportunity to mentor **13 bright Masters and Bachelors** students from diverse backgrounds in **gender** (5 are female students), **race**, **geographical location** (including USA, Turkey, and India), and **universities** (MIT, Purdue, METU). Among the 6 master's students in the REALM project, **3 were female students** and one of them joined META in Spring'23. **The master's student from the Czech Republic** is working at WePay now as Software Engineer. He worked with me for developing a novel human attribute recognition model from texts. **One of the two undergraduate female students from Turkey has started her master's program in Germany.**

***Outreach*:** Besides my advisees, I have helped multiple students from URG communities and my alma mater to apply to graduate schools, including revising their application materials, suggesting research directions, preparing for interviews, etc. After my Ph.D. admission, I wrote blogs describing the application process and provided free access to the preparation materials.

***Teaching*:** Having taught large classes for a long time and on various levels, I have slowly identified the biases and micro-aggressions that happen in a classroom. *Recognizing diversity can come in many forms, such as recognizing different accents, genders, races, preferences, disabilities, mental issues, etc.* I have always tried to have an interactive classroom where everyone has a voice. Asking students about their understanding of the materials or their well-being always makes them feel included, and as a teacher, we can improve and re-adjust. After knowing about the students' backgrounds and their goals in my courses, I was able to provide more relatable examples in the classroom. In addition, I actively tried to identify my own cultural and educational biases so that it does not affect my class.

***Socially Impactful and Ethical Research*:** The primary goal of my research has always been to work on problems that have an impact on complex societal issues. For instance, *multimodal information retrieval* has been applied to

finding missing persons or helping people with PTSD. We also explored identifying human anxiety or stress symptoms based on gait and emotion detection. *These have a direct implication for accommodating people with disabilities*. Furthermore, *identifying intrinsic and extrinsic domain complexity* before any large ML model is applied to new datasets can help reduce the hidden bias of the AI models. As a Database+ML+NLP researcher, my research involves working with people from diverse backgrounds, including people from essential services. This kind of interdisciplinary setting makes diversity a natural requirement and helps progress the scientific community.

FUTURE PLANS. My goal as a faculty is to create a welcoming and safe learning space for everyone, irrespective of their backgrounds. Along with continuing my efforts from the past, I would take the following tangible steps to promote diversity, equity, and inclusion in the context of a new faculty member:

***Research*:** (1) Professors have a tendency of recruiting from their alma mater due to institutional or affinity bias. My goal is to recruit capable students from all over the world, especially from under-represented countries. (2) Practising and promoting a growth-based peer-to-peer relationship among the mentees. Interaction among group members would be of continual learning, rather than competition. (3) Providing support and fair access to all students for traveling to conferences so everyone can gain experiences. (4) Continue my research on societal problems that impose hindrances to an inclusive and supportive environment in education and society. I aim to understand the complex patterns in multi-modal data while avoiding hidden bias in the decision-making process. (5) Continue working for promoting diversity and inclusion in database and ML conferences such as SIGMOD, VLDB, AAAI, ECML, etc.

***Teaching*:** (1) Promote an environment that ensures equity and inclusion in my classes. Every student should feel a sense of belonging, and I would ensure that the classroom stays free from bias, stereotypes, or prejudices. (2) Each student perceives lessons differently based on their background, and allowing them to freely express that brings out their full potential. (3) Implement methods that promote student participation and a fair assessment of the class. I will also want to take frequent feedback from the students to ensure accountability on my part. (4) Build an interpersonal relationship with them through active participation in and outside the classroom. They should feel at ease to share their concerns and needs. (5) Train the teaching assistants to practice diversity and inclusion in my classes. (6) Use of *inclusive language* in the classroom and my teaching materials. (7) Keep myself updated with relevant research literature and class etiquette. Since most of my expertise revolves around data management and ML, I would make the best effort to teach students how to practice diversity, fairness, and inclusion in research and life.

***Outreach and Academic Services*:** (1) Being part of the 'UNICEF Guardian Circle' or 'Save the Orphans', I have seen suffering children from all over the world. I want to connect with them for the next phase of their development and provide them with tools to develop themselves. (2) Continue my research on societal problems that impose hindrances to an inclusive and supportive environment in education and society. (3) Actively serving in *affinity group workshops* in Conferences such as NeurIPS to support under-represented groups. (4) Promote the inclusive culture to academic community and students, especially to recruit students from under-represented groups and countries. (5) Actively participating in workshops and organizations that promote diversity in the institution. (6) Continue to foster relationships with local minority-serving institutions.