

³ Unified Medical Language System (UMLS)

consequence, it is difficult to rank documents/contents when the query is about an emerging topic with minimal co-occurrences (e.g. long tail entities [WWW 2019]). Moreover any statistical AI algorithm functions on latent dimensions making ranking of documents/contents hard to explain. Algorithms in *summ* have a hard time in modeling knowledge constraints causing the end result to significantly differ from useful and actionable summaries [JMIR 2021]. In *convAI*, an intrinsic task of any statistical AI algorithm (preferably deep learning) is to understand user behavior during interactive search and later improve accuracy during search sessions. Research *convAI* has been hampered by lack of datasets that involve process knowledge, an approach experts follow during any formal conversational setting. Furthermore, algorithms trained on some standard benchmark datasets, such as General Language Understanding and Evaluation (GLUE) become rigid and lack reusability across other domains. These challenges broadly highlight the need of knowledge-intensive datasets and KiML algorithms that provide multi-hop knowledge traversal [ISWC 2018], domain-specific knowledge infusion [CSCW 2019], executable in low-resource domains [AAAI 2020], and support symbolic knowledge integration [JMIR 2021]. By developing the KiML paradigm, the dissertation answers the question: **Can the incorporation of domain knowledge enhance performance and explainability of data-intensive learning models?**

KiML paradigm steward knowledge graphs, a machine readable structured representation of knowledge consisting of entities (entity and entity type) and relationships in various forms (e.g., labeled property graphs and RDFs). **Figure 1**, illustrates two knowledge graphs; (a) An *Empathi Ontology*⁴ constructed to support emergency

responders in identifying crucial events on time-evolving information from social media streams. (b) An *Educational Knowledge graph* constructed from epub's of Amazon books, and other course textbooks to assess a student's learning outcome and suggest ways to improve. These domain-specific knowledge graphs provide necessary information aid for machine learning/deep learning algorithms for domain adaptation and reasoning over the outcomes. There are various ways to incorporate external knowledge which my dissertation categorizes into (a) Shallow Infusion, (b) Semi-Deep Infusion, and (c) Deep Infusion. *Shallow knowledge infusion* contextualizes the training examples with expert knowledge to capture meaningful patterns. Some of the shallow infusion examples include contextual modeling [CSCW 2019], entity normalization [WWW 2019] and explainable clustering [AMIA 2021]. *Semi-deep knowledge infusion* guides the model's attention in the learning process. It utilizes expert knowledge concepts as weights or constraints to guide an explainable learning process. This strategy falls short in assisting deep learning models adjust the high-level abstractions learnt through multiple layers. *Deep Knowledge Infusion*, combines the stratified representation of knowledge at varying abstraction levels to be transferred in different layers of deep learning models [AAAIc 2020]. I present methodological architectures of these strategies in **Figure 2**. This research statement focuses primarily on the utility of KiML algorithms in following applications:

(A) Domain Adaptation and Inference (recsys): Dependence on expert labeling is not only impractical but leads to spurious feature selection and implausible explanations [WWW 2019]. To generate semantic labels for unlabeled social media data automatically without expert supervision, I developed a novel semantic encoding and decoding (SEDO) algorithm (semi-deep infusion) that would learn a weight matrix (W) between concepts in unlabeled data and concept classes in knowledge graph (KG). To test the efficiency and reliability of SEDO, I curated natural testbeds in collaboration with psychiatrists in **Wright State Psychiatry**, and **Weill Cornell Medicine**. In comparison with the traditional statistical model, SEDO reduced false alarms by 92%, while achieving 84% annotator agreement. This was the first KG-based zero-shot (problem scenarios with no label data) approach with explainable outcomes [IC 2019, IC 2020, IC 2021, CIKM 2018] and has been extended to detect suicide risk from a

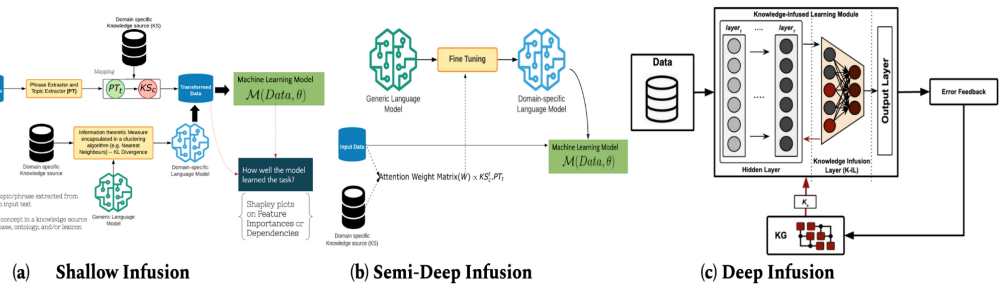


Figure 2: Different architectural variants of Knowledge infused Mining and Learning



Figure 3: Recommending support providers to support seekers for mental health assistance

⁴ <https://shekarpour.github.io/empathi.io/>

mix of supportive and at-risk users posting on Reddit [WWW 2019]. The testbed has >3400 downloads and has been employed by researchers in the intersection of explainable AI and healthcare. Furthermore, it has opened new research directions in facilitating formation of dynamic peer support groups on social media. A knowledge-infused siamese network was trained and adapted to understand the conversations between support seekers and support providers on Reddit for suitable matching. A natural language inference layer was appended at the end to classify the pair of users as; (a) both are support seekers (entailment), and (b) one is support seeker and other is support provider (contradiction). An illustration in **Figure 3** shows the matching [ICHI 2021]. The study received attention from various non-profits in the US, state department on mental health in south carolina, and industry-oriented meetings such as [Geekle.us](#). Prior research at the intersection of social media, explainable AI, and healthcare did not focus on incorporating external knowledge in model learning and anchor into costly human resources for annotation, verification, and validation. The seamless incorporation of clinical expertise and guidelines in computational mental healthcare research is an open research problem that I would be interested in exploring further. Moreover, my tenacity in healthcare and social computing research would complement the research directions of [Prof. Sameer Singh](#), [Prof. Carter Tribble Butts](#), and [seek potential interdisciplinary collaboration with faculties affiliated with Center for Health Care Management & Policy](#).

(B) Weighted constraints conditioned on time-evolving events (I2r, recsys): The deep learning approaches to pattern detection in a static world do not readily support inference over time, or prediction, essential to policy decisions. In a crisis event, as time evolves, new sub-events emerge [AAAIa 2020]. We want to learn a function that maps the world's state (Number of people in a Shopping Mall) to a policy (High traffic, a candidate for lockdown). The goal is to train a sequence of models that can estimate the rise in infection and help generate a policy: "Lockdown the place where this event is happening." I develop a semi-deep infusion-based framework that extends an epidemiological model and sequence it with a novel Knowledge-infused Policy Gradient (KiPG) module, which seamlessly integrates bayesian and conditional gradients [AAAIb 2021]. Real-world knowledge (e.g. social media chatter, news articles) is infused as weighted constraints conditioned upon the time-evolving events. Bayesian gradients reflect on events that are here to last, and conditional gradients reflect on immediate events with the potential to increase infection cases. The model could precisely estimate the rise in infection rate 15-25 days before it occurs—for instance, information on gathering events, pre-symptomatic patients, or asymptomatic patients. The model output's amenability to an analysis by the expert makes it attractive for real-world use e.g., the Government of Rajasthan [KDD 2020]. The developed KiPG forms a critical part of the Knowledge-infused Reinforcement Learning pipeline which has been extended to a recommendation setting on benchmark datasets [ECML 2021].

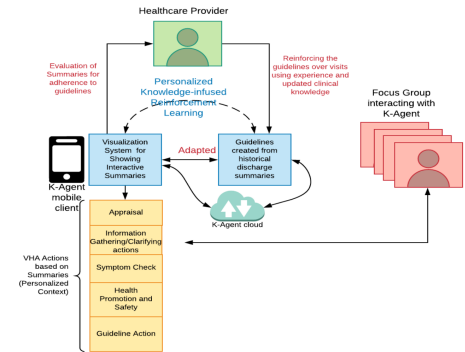


Figure 4: A proposed approach to build virtual assistants for mental health using Knowledge-infused Reinforcement Learning. The assistant will have the capability to generate questions on the fly based on interactive summaries for personalized health suggestions.

It has also been applied to mental health virtual assistant recommendations to assist clinicians in practice [ICSC 2019a] (see **Figure 4**). Moreover, I would be interested in discussing more pressing issues at the interface of virtual health assistant and NLP general healthcare with [Prof. Rina Dechter](#), [Prof. Padhraic Smyth](#), and [Prof. Jeff Krichmar](#).

(C) Matching and Ranking (I2r, recsys): I am excited to see tangible real-world impacts resulting from my research at Datamir Inc. and Jose De Mello Saude Hospital (through University of Chicago's DSSG Fellowship).

(i) Historical constraints on pattern matching: In the US and Europe, patient's appointments with General Practitioners (GPs) are typically decided based on the doctor's availability. This creates a scenario where patients fail to develop trust with a healthcare professional. So far, understanding the development of trust between patients and their family physicians largely relies on survey-based measures while seldom examining the actual consultation history. I developed a semi-deep K-IL system that models a patient's trust of GPs using the knowledge of consultation history and ICD-10 graphs which describe interaction and change in the severity of condition between interactions. The system can recommend >80% of patients with relevant GPs compared to existing heuristic-based systems in JMS that can do 37% [DSAA 2018]. The first ICD-10 knowledge-driven trust-based recommender system went into development at Jose De Mello Saude (JMS) Hospital Network in Portugal. This work was supported by the prestigious **Eric and Wendy Schmidt Data Science for Social Good Fellowship**. The work was featured in **VentureBeat** and described in **Dr. Eric Topol's** seminal paper in *Nature Medicine*. *(ii) Abstract knowledge to Rank sub-events and Risks for real-time alerts:* From the influx of millions of tweets during a significant event, humanitarian organizations can't extract emerging risk and high-impact sub-events that are worthy of alerts. Furthermore, the connection between sub-events is challenging to unfold without background information. I built a shallow infusion framework that uses a disaster ontology that I created from Federal Emergency Management Agency archives [ICSC 2019b] to cluster potentially connected tweets with mentions of impactful sub-events and risks, as described in the ontology. The system could explain the outcomes through concepts and relationships known to humanitarian organizations [AAAI 2020]. Later, it evolved into an integrable module within the alerting application framework and extended with multimodality and the above-described summarization capabilities. The

time in mental health. Further, my research has also attracted the attention of the public through workshops at **ACM SIGKDD, AAAI ICWSM, and Knowledge Graph** conferences. I have been interviewed by Ontotext, TheRegister, and my papers have been covered by global media outlets, including **Computing Research Organization (CRA), Healthline, The Conversation, and Venturebeat**.

Vision For the Future

My long-term research goal is to build the next generation of end-to-end neuro-symbolic artificial intelligence frameworks that are explainable, interpretable, and reasonable. Particularly, I want to exercise conscious efforts in developing deep knowledge infusion methods, making technical contributions in *convAI*, *recsys*, *l2r*, and *summ* that support applications in **healthcare, education, cyber social threats, digital security and internet policy, and bioinformatics**. Below, I discuss my future research directions in Core AI and interdisciplinary research that have funding potential.

(i) Core AI Research: So far, I have motivated the applicability of shallow and semi-deep knowledge infusion to provide explainable outcomes in scenarios with no labeled data and it is infeasible to create high-quality testbeds (A). Further, it is computationally hard for a statistical learning approach to make explainable inference in temporal problems with uncertainty (B). In my vision, what is more profound is the attainment of flexibility, consistency, and robustness in KiML algorithms to perform well in cross-domain applications. For instance, a KiML algorithm capable of discerning the stratified knowledge in musical chairs processes can help in understanding Enzyme Kinetics, a process in biochemistry. This is called “Learning by Analogy”, an instance of Deep KiML that can help any DL model in providing a prediction and explanation without the necessity of labels [TUT 2020] (see Figure 5; an illustration of analogy between coin pushing arcade game and oxidative phosphorylation). I see its immense potential in education technology, crisis management, and various medical disciplines. Moreover, this vision is of immediate relevance to NSF IIS and NSF CISE focusing on human-center explainable AI. As a prospective faculty in informatics, I seek to explore and devise innovative methods in analogy-based learning and education technology that interface KG and human-centered computing for a variety of application domains.

(ii) Interdisciplinary Research: Interdisciplinary research experiences have broadened my research perspectives and motivated many innovative ideas. For instance, In collaboration with Embibe, an *education* technology enterprise, I am inspecting neural and non-neural KiML algorithms that support mathematical reasoning while solving **math word problems (MWP)** straight from the English narratives. Here the role *proknow* is eminent. However, this would require collaboration with faculties in linguistics and education, which I would seek in the university I will be joining. I am also interested in studying **time-evolving user networks in spontaneous events**, such as pandemic, crisis, etc. In particular, I am curious about mathematical modeling of short-lived **communities of support seekers and support providers** that are formed on social media using KiML algorithms for dynamic **peer-support group formation**. I believe this research in computational social science would find relevance with **Prof. Carter Tribley Butts, Prof. Richard Futrell, and Prof. Alexander Ihler**. Through tutorials in computational data science conferences, I found overwhelming responses in these research directions [TUT 2021a, TUT 2021b, TUT 2020]. With faculties in bioinformatics and biostatistics, I would be interested in exploring the utility of **KG, Causal AI, and Statistics** to study large-scale high-dimensional data on **Early Colorectal Cancer (Bioinformatics)**. In particular, seek ways to generate causal explanations with knowledge by extracting causal attributions from blackbox AI models and corroborating it with information in scientific literature. On technical subjects Semantic Web, Biomedical Informatics, and Bioinformatics, I see synergy with **Prof. Pierre Baldi**.

(iii) Research Funding: I have had the good fortune to closely work on several proposals with my academic advisors. Some of the pending or soon to be submitted grants led by my advisor to which I have significantly contributed include:**(a)“Analogy-based learning in Biochemistry,”** Agency: NSF Medium, Collaborator: University of Tennessee; Proposed research examines compositional student-generated analogies for a novel AI-assisted interactive learning and assessment with a user logged feedback **(pending)**. With HCI as the subject area, I would seek potential collaboration with

Prof. Padhraic Smyth, Prof. Bill Tomlinson, and Prof. Pierre Baldi.

(b) “Virtual Health Assistant (VHA),” Agency: NSF SCH Medium, Collaborator: UofSC School of Medicine: Proposed research implements the principles of self-monitoring, self-appraisal, and self-management in VHA for personalized and

| Vision | Collaboration and Research Grant Funding | Outreach: Education and Training |
|--|---|--|
| <p>Methods:</p> <ul style="list-style-type: none"> Personalized Knowledge Graph Construction Deep Knowledge Infusion Neuro-Symbolic AI Safe, Explainable, Interpretable, and Tractable AI Systems Reinforcement Learning <p>Applications:</p> <ul style="list-style-type: none"> Healthcare Education Technology Finance Online Culture, Social Harms and Crisis Response | <p>Funding Agencies of Interest:</p> <ul style="list-style-type: none"> NIH (NIMH, NIDA)(e.g. R21, R03) NSF (e.g. IIS, SCH, CISE) DoD, DoJ, ONR Facebook, Google, Samsung Research, Bloomberg, Microsoft <p>Current Collaborations:</p> <ul style="list-style-type: none"> Weill Cornell Medicine Alan Turing Institute, UK University of Kentucky Northeastern Harvard CRCS Wright State and more. | <ul style="list-style-type: none"> Recruitment of undergraduates, graduates, and Ph.D. students (with Diversity and Inclusion) Coordination with Center for Teaching Excellence Workshops/Tutorials in ACM, IEEE, AAAI, AMIA, and NIH/NSF-sponsored meetings Invited Keynotes, Talks, and Teaching seminar-level courses |

Figure 6: Vision for implementing and executing research in CS and Interdisciplinary fields.

reflective support to patients with moderate mental health conditions. The proposal also lays emphasis on individual and societal implications of information systems and ethical usage of digital assistants **(pending)**. Likewise, **(c) “MAESTRO: Psychiatric Process-Guided, Safety Constrained, and Explainable Language Generation in VMHA to Screen and Triage Depression, Anxiety, and PTSD”**, Agency: SONY AI Research, Collaborator: Alan Turing Institute UK **(pending)** is a proposal for developing virtual mental health assistants. In the context of (b) and (c), I intend to explore and extend the problem design in these proposals with **Prof. Ramesh Jain, Prof. Lisa Pearl, and Prof. Michael Lee**. **(d) “Infant Autism Detection System: Development of an Instrumented, Intelligent Infant Interaction Laboratory for the Prediction of Autism Spectrum Disorder”**, Agency: UofSC Internal seed with planned R01, Collaborator: UofSC Center of Excellence on Autism and Neurodevelopmental Disorders **(Funded)**. I would be focusing on extending this project and seeking CS and cross-disciplinary collaborations with **Prof. Hal Stern, Prof. Ramesh Jain, and faculties in Center for Autism Research and Translation (CART)**.

These experiences enable me to think big, find novel and important fundamental or practical research problems, and convert my creative thinking into clear and convincing grant proposals. With the future research interests given above, I plan to write proposals to get funding support from NSF CRII, NSF CAREER, NSF Core IIS Programs, NIH K21, DoD (ONR, AFRL, AFOSR, etc.), DoE, and others to support my research and to support and train the next generation researchers. I have also been fortunate to have collaborated with a number of stellar researchers both in academia and industry, many of whom have been my mentors in research. The research agenda is very often shaped in wonderful ways through such collaborations, as well as through mentoring and advising graduate students (see **Figure 6**). To that end, I intend to keep nurturing and strengthening my ongoing collaborations and pursue collaboration with scholars within and outside my field.