Research Statement
Ching Nam Hang

My research interests lie at the intersection of big data analytics and artificial intelligence (AI), with an emphasis on their implementation in healthcare and education. Specific research projects include the theoretical exploration and practical applications of machine learning for pandemic response, the management of misinformation during crises, and the augmentation of self-learning in large-scale educational systems. My work is devoted to turning data-driven insights into practical, real-world applications, notably improving our response to global health crises and nurturing autonomous learning. My professional journey in big data analytics and AI consistently deepens my understanding and sharpens my problem-solving skills, providing a solid foundation for the development of innovative solutions with substantial and wide-ranging impacts.

Ph.D. Research

Machine Learning for Pandemic Response

In the face of the COVID-19 pandemic, an array of digital contact tracing (DCT) strategies has been proposed to curb virus transmission. Existing literature, however, largely focuses on DCT application design and deployment, with less emphasis on machine learning optimization. In [1], we conduct a comprehensive review of machine learning- enhanced DCT strategies, formulating relevant research questions and outlining potential solutions using existing machine learning methodologies. We introduce a new taxonomy categorizing DCT strategies into forward, backward, and proactive contact tracing. We examine the integration of privacy-preserving strategies, such as federated learning, with big data and machine learning to enhance the privacy and security of DCT systems. Additionally, we demonstrate how convolutional neural networks, graph neural networks (GNNs), and transformer-based learning optimize DCT strategies. As a solution to limited data availability in the early phase of pandemics, we show how generative adversarial networks (GANs) can be used to tackle such challenges in DCT.

Machine Learning for Infodemic Risk Management

The COVID-19 pandemic has led to an infodemic of unprecedented scale with the rapid spread of misleading or false information via online social networks. Network analysis plays an indispensable role in fact-checking efforts, modeling and learning infodemic risks through large-scale graph computations and statistical processes. Drawing insights from our work in [2], which showcases the efficacy of the proposed graph algorithm in handling efficient computations on large-scale graphs, we identify a unique intersection where these methodologies can be applied to decode the complexities of infodemic propagation. Notably, our contributions in [2] received the Honorable Mention Award at the MIT/Amazon/IEEE Graph Challenge 2018. Building upon the foundational principles established in [2], in [3], we present MEGA (*M*achine Learning-*E*nhanced *G*raph *A*nalytics), a novel framework enabling automated machine learning for mega-scale graph datasets. Within MEGA, we employ a blend of feature engineering and graph neural networks to enhance learning performance and enable parallel computation on extensive graphs. Infodemic risk analysis serves as a unique application of MEGA, encompassing spambot detection through triangle motif counting and influential spreader identification via distance center computation. Performance evaluations using the COVID-19 Infodemic Twitter dataset underscore MEGA's superior computational efficiency and classification accuracy. Further examining this issue, in [4], we introduce TrumorGPT, a maximum truth-seeking AI. TrumorGPT synergizes large language models (LLMs) like GPT and knowledge graphs to facilitate automated fact-checking, demonstrating efficacy on large-scale datasets.

Artificial Intelligence for Large-Scale Learning

The COVID-19 pandemic has significantly reshaped learning environments and modalities, elevating the importance of self-learning in STEM education. In [5], we offer a comprehensive review of transformer-based LLMs for AI-assisted programming, showcasing the potential of LLMs in facilitating programming learning. In [6], we introduce an AI-assisted programming tool that integrates LLMs with Xcode, enhancing developer productivity in the Apple software ecosystem. Through advanced natural language processing techniques and a chat interface, our software facilitates code generation, autocompletion, and interactive decision-making, as evidenced by case studies demonstrating its efficacy with LLMs like GPT. In [7], we propose an AI chatbot system designed to optimize flipped learning. This system employs deep learning techniques to blend peer instruction with just-in-time teaching. Additionally, we formulate optimization problems to ideally balance the allocation of pre-class and in-class learning time. In further pursuit of these aims, in [8-9], we develop AI-powered software tools to foster interactive self-learning in STEM education.

Future Research Directions

While big data analytics and AI are ubiquitously and extensively used in various forms in modern society, research on automated machine learning is widely acknowledged to be a challenging field. I plan to undertake the following long-term directions in my future research to meet these challenges:

- Large language models for richer data interpretation: LLMs such as GPT-4 have shown exceptional prowess in understanding and generating human-like text. However, capitalizing on their potential to comprehend the nuanced meaning within vast and diverse data domains remains a considerable challenge. I intend to investigate novel techniques such as, preference-based reinforcement learning, to unlock the full potential of large language models for richer data interpretation across various applications.
- Generative AI for synthetic data generation: Generative AI, especially GANs, has shown tremendous potential in creating synthetic data that mimic real-world distributions. This is particularly important in scenarios with limited data availability or privacy concerns. To propel the advancements in this research domain, it is important to develop more robust and versatile generative models that not only address the current challenges but also are adaptable to a broad spectrum of machine learning applications, ensuring efficiency and scalability in various scenarios.
- Graph-based machine learning: GNNs represent a powerful tool for processing data structured as graphs. This is particularly relevant in a world increasingly connected, where relations and networks dictate how information flows. The implications range from social networks to biological networks, recommendation systems, and the interpretation of molecule structures. It is crucial to push the boundaries of this field by exploring scalable and efficient algorithms for graph-based learning, as well as studying how to combine these approaches with large language models for enhanced understanding of complex, interconnected data.

I hope to contribute significantly to the rapidly evolving landscape of machine learning, specifically in the areas of large language models, generative AI, and graph-based machine learning. Beyond the immediate scope of these areas, I am also keen on diligently exploring the integration of machine learning with other disciplines, such as healthcare and education, to maximize the broad societal impact of AI advancements. While having specific milestones is crucial to guiding research efforts, I remain open-minded and constantly alert to emerging research opportunities. As the fields of big data and AI continue to progress and evolve, I am committed to ongoing learning, adeptly adapting to new challenges, and leading innovative research at the frontiers of technology.

References

[1] **C. N. Hang**, Y. -Z. Tsai, P. -D. Yu, J. Chen and C. W. Tan, Privacy-Enhancing Digital Contact Tracing with Machine Learning for Pandemic Response: A Comprehensive Review, ***Big Data and Cognitive Computing, special issue on Digital Health and Data Analytics in Public Health***, Vol. 7, No. 2, 2023.

[2] C. -Y. Kuo, **C. N. Hang**, P. -D. Yu and C. W. Tan, Parallel Counting of Triangles in Large Graphs: Pruning and Hierarchical Clustering Algorithms, ***IEEE Conference on High Performance Extreme Computing (HPEC), MIT/Amazon/IEEE Graph Challenge (Honorable Mention)***, 2018.

[3] **C. N. Hang**, P. -D. Yu, S. Chen, C. W. Tan and G. Chen, MEGA: Machine Learning-Enhanced Graph Analytics for Infodemic Risk Management, ***IEEE Journal of Biomedical and Health Informatics***, accepted and to appear, 2023.

[4] **C. N. Hang**, P. -D. Yu and C. W. Tan, TrumorGPT: Query Optimization and Semantic Reasoning over Networks, to be submitted to ***IEEE Transactions on Big Data***, 2023.

[5] M. F. Wong, S. Guo, **C. N. Hang**, S. W. Ho and C. W. Tan, Natural Language Generation and Understanding of Big Code for AI-Assisted Programming: A Review, ***Entropy, special issue on Statistical Machine Learning with High-Dimensional Data and Image Analysis***, Vol. 25, No. 6, 2023.

[6] C. W. Tan, S. Guo, M. F. Wong and **C. N. Hang**, Copilot for Xcode: Exploring AI-Assisted Programming by Prompting Cloud-based Large Language Models, ***IJCAI Symposium on Large Language Models (LLM@IJCAI'23)***, 2023.

[7] C. W. Tan, **C. N. Hang** and L. Ling, Optimal Flipped Learning: Blending Peer Instruction with Just-in-Time Teaching using AI Chatbots, submitted to ***IEEE Transactions on Learning Technologies***, 2023.

[8] J. Li, C. W. Tan, **C. N. Hang** and X. Qi, A Chatbot-Server Framework for Scalable Machine Learning Education through Crowdsourced Data, ***ACM Conference on Learning at Scale***, 2022.

[9] C. W. Tan, L. Ling, P. -D. Yu, **C. N. Hang** and M. F. Wong, Mathematics Gamification in Mobile App Software for Personalized Learning at Scale, ***IEEE Integrated STEM Education Conference***, 2020.