**Dr. Qianru Sun**
Assistant Professor of Computer Science
School of Computing and Information Systems

# Research Statement

This statement provides a detailed overview of my academic journey and accomplishments to date and delineates my research direction for the next five to ten years. Before delving into the specifics, I show an overview of the evolution of my research interests in the following figure.

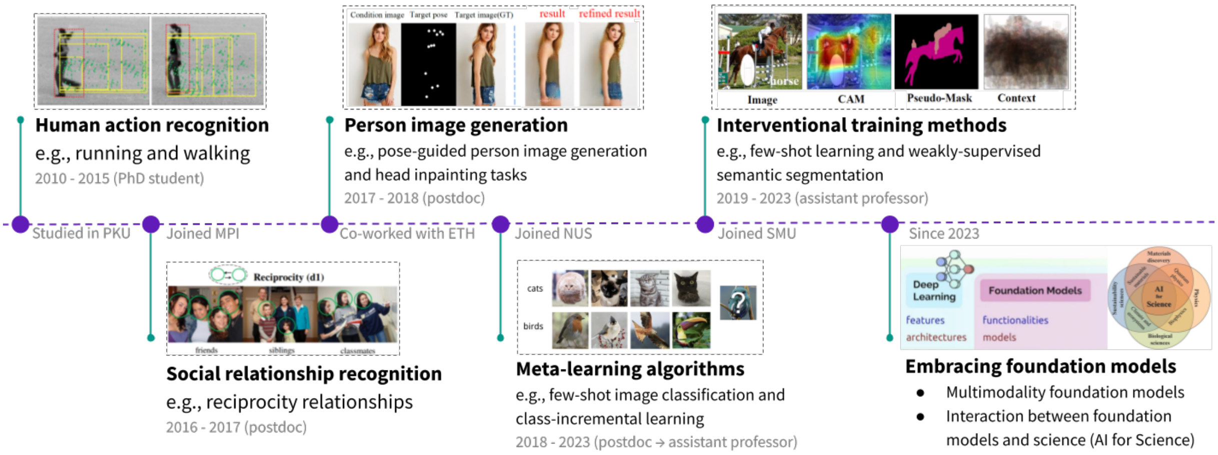


*Figure 1. During the PhD study at Peking University (PKU), I concentrated on non-deep-learning techniques for recognizing human actions in videos. My postdoc research at the Max Planck Institute for Informatics (MPII) marked a shift in my focus from human actions to high-level human activity, pose, and attribute recognition, especially in the context of classifying human social relationships and generating person images (collaborating with ETH). Upon joining the National University of Singapore (NUS), I dedicated more into the fundamental algorithms and training methods that address the challenges of deep learning datasets, especially the problems of data imbalance and shortage I encountered in my previous research. After joining SMU as an Assistant Professor, I continued this research by building a passionate team, where "I" becomes "We", to probe wider facets of visual recognition challenges. Projecting into the next five to ten years, the interests of our research team are multimodality foundation models that integrates attributes from various data modalities and in turn offers comprehensive and reasoning representations of any modality data for downstream recognition tasks. In terms of applications, we are very excited about the true future directions of AI such as AI for Science.*

**Backgrounds**. During my research, I have witnessed an astounding change through the evolution of AI, e.g., the transition from small-/mid-size to gigantic-size models, and the blooming of Transformer that reshapes the landscape of AI by unifying computer vision (CV) and natural language processing (NLP). As an AI researcher, I am proud to be part of this change, and I will position my research in this big picture and highlight the significance of my research. It's an acknowledged fact in the AI community that deep learning models are always hungry for training data. My research interest is efficient visual recognition with limited training data, also called learning with insufficient data, encompassing various application-driven visual recognition problems, e.g., few-shot learning, where each test case contains only a very few training samples, class-incremental learning, where old classes with limited examples are often forgotten when the model is adapting to the new classes with ample data, and weakly-supervised learning, which has only weak forms of supervision that are easy to obtain. The proposed algorithms in my research team encompass two key methodologies: transferring pre-trained model knowledge to learn limited data (Meta-Learning) and leveraging causality theory to achieve invariant learning with limited data (Interventional Training).

In this statement, I will introduce the motivation for why I delved into "visual recognition with limited data" based on my research experiences during 2016-2018, elaborate my research contributions on the two methodologies during 2018-2023, and finally look ahead for the future.

## 1. Data Limitation Problems in Prior Works (2016-2018)

A groundbreaking development in deep convolutional neural networks emerged with the proposal of Residual Networks (ResNet) [He et al., 2016] in the year 2015. This amazing network achieved victory in the prestigious ILSVRC 2015 competition, showcasing its prowess in image classification, localization, and detection tasks. This breakthrough sparked tremendous excitement within the Computer Vision (CV) community, leading to the widespread deployment of ResNet for various visual recognition applications, e.g., semantic segmentation [Long et al., 2015] and pose estimation [Ma et al., 2017]. At that time, I just joined the computer vision and deep learning group at Max-Planck Institute for Informatics (MPII) as a postdoctoral researcher. Our team was engrossed in a privacy implication project, and one of the topics was to explore the capabilities of state-of-the-art deep models (e.g., ResNet) in learning high-level social semantics and relationships from daily-life photos. Specifically, my work is to investigate whether deep models could discern social relationships such as classmates, lovers, friends, and more, within the context of everyday images. As I delved deeper into the realms of ResNet and its applications, my curiosity and determination to unravel the intricacies of high-level social semantics grew stronger.

### 1.1 Social relationship recognition

In this project, I (and my co-workers) constructed the first social relationship dataset and published it in [Sun et al., 2017]. For the dataset, we “borrowed” the image resources from a real-world person recognition dataset, People In Photo Albums (PIPA) [Zhang et al., 2015] where only person IDs were available. Our data annotation task was to label the classes of social relationships. It’s essential to note that this dataset was not originally collected for social relationships, so it has in a natural distribution of human social relationship classes (i.e., a long-tailed distribution). For instance, the dataset comprises numerous photos of “friends” and “lovers” while containing fewer images of “leader-subordinate” relationships. From such a natural distribution dataset, we found that the key challenge is not any specific relationship classes but mainly the ones with scarce training data, due to the natural distribution—it is more often for people to upload the photos of “friends” than “leader-subordinate” to social networks. It is worth mentioning that this social relationship recognition work [Sun et al., 2017] provides the social psychological definition of the holistic conceptualization of social life and contributes the first dataset and the first model to recognize social relations in daily-life photos [Sun et al., 2017]. For constructing the model, we proposed the novel idea of leveraging high-level individual as well as pairwise attributes such as ages, genders, clothing activities, and proximity as cues to enable the model to comprehend and distinguish various relationships effectively. From the ablation study, we found that human activities, represented through body poses and contexts, emerged as key attributes influencing the model’s performance of recognizing social relationships. Meanwhile, we found that the deployed pose estimator has the issue of poor performance for scarce-data pose classes such as “squatting” (compared to a sufficient-data one like “running”).

### 1.2 Person image generation

As manual annotation on real images is tedious and expensive, I questioned “how about using synthesized data?” in my mind. An intuitive idea is to leverage the pose information to generate person images and augment data for rare poses as well as rare social relationships. I and my co-workers started from the basic one, i.e., rare pose data generation, in our following works. In [Ma et al., 2017], we designed an innovative model using person pose data as a condition to generate natural images (called pose-guided person image generation). During the training of generation models, we again found that the result was unsatisfactory if not enough data on a particular pose was given to the models. It is worth mentioning that this work has gained 850+ Google Scholar citations and is of one the Top 100 Most Cited NeurIPS Paper over the Last Five Years (link). In our follow-up work [Ma et al., 2018], i.e., where we changed to augment features rather than images, we found that unfortunately the rare class issue also exists in the space of features, and thus only incremental results were obtained. Specifically, we found that the improvement of data augmentation using the generated features is marginal, e.g., 7.5% improvement over baseline model but only 1.4% over the state-of-the-art model who adopted other domain data [Ma et al., 2018]. The reasons are in two folds. First, the generative model suffers from the problem of category bias, like the recognition model, and thus is difficult to generate high-quality images for the classes with scarce training data. Second, the model trained on one dataset cannot generate images with “novel” variance outside this dataset. Meanwhile, we found that if the model cannot be provided with “fresh data” meaning data labeled by humans, its output quality will be significantly affected. This aligns with the recently discovered problem called Model Aging Dilemma (MAD) [Alemohammad et al., 2023]. MAD has been confirmed to make effects on all AI models and its impact has been verified on autoencoders, Gaussian mixture models, and large language models.

In the above works, I identified three visual recognition tasks that are significantly impacted by data limitation problems: social relationship recognition, pose estimation, and pose-guided person image generation. They encompass the areas of classification, regression, and generation, respectively. To address the challenges, my subsequent works are focused on the method of transferring data variations and acquired knowledge from large pre-trained models. The fundamental concept behind this is referred to as “meta-knowledge transferring”. The learning process involves two distinct forms of “meta” knowledge. The first entails understanding the model’s learning behavior, which is influenced by hyperparameters and can be traced by bi-level optimization, i.e., meta-learning. The second form pertains to comprehending the model's reasoning behavior. For this purpose, we construct machine learning models based on a key principle of causality theory: interventional training.

## 2. Meta-Learning (2018 - 2023)

Meta-learning aims to learn the training behavior of deep models on similar tasks. This ability is well-handled by humans, e.g., a person who knows how to ride a bike can learn to ride a motorcycle fast with a little demonstration, while it remains challenging for deep learning models that require large-scale training processes for good performance. In my research, I introduced several new meta-learning algorithms to solve the data-insufficient problems, e.g., few-shot learning and class-incremental learning tasks.

### 2.1 Few-shot learning tasks

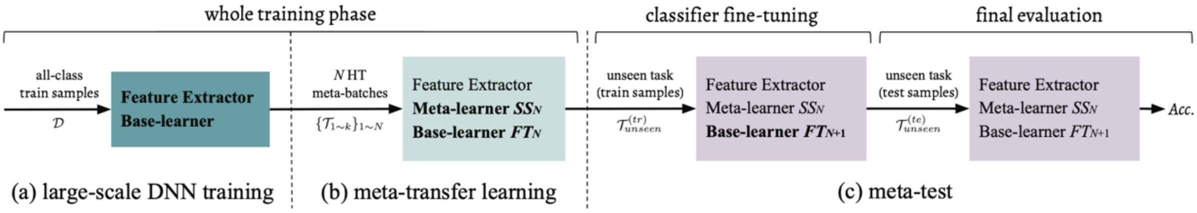


*Figure 2. The training and testing phases in Meta Transfer Learning [Sun et al., 2020]. The model training contains two steps: pre-training on the large-scale data (i.e., by merging all small-data task data) and meta-transfer learning on individual small-data tasks.*

Meta-learning for few-shot learning settings has the key idea of leveraging many similar few-shot tasks to learn how to adapt a base-learner to a new task for which only a few labeled samples are available. As deep neural networks (DNNs) tend to overfit using a few samples only, typical meta-learning models use shallow neural networks, thus limiting their effectiveness. To achieve top performance, previous works tried to use the DNNs pre-trained on large-scale datasets but mostly in straight-forward manners, e.g., (1) taking their weights as a warm start of meta-training, and (2) freezing their convolutional layers as the feature extractor of base-learners. In our CVPR 2019 work [Sun et al., 2019] and an improved version in [Sun et al., 2020], we propose a novel approach Meta Transfer Learning (MTL), as shown in Figure 2. MTL learns to transfer the large-scale pre-trained weights of DNNs for learning novel few-shot tasks. Specifically, "meta" refers to training multiple tasks, and "transfer" is achieved by learning the scaling and shifting functions of DNNs' pre-trained weights. It is worth noting that MTL is a highly efficient method of leveraging large-scale pre-trained models for tackling small data tasks. Its operations (scaling and shifting model weights) are rather generic and can be plugged into the models of many related learning tasks (such as the class-incremental learning tasks [Liu et al., 2020b, Liu et al., 2021a] introduced in the following sections). This work has achieved 1,000+ Google Scholar citations since published in 2019. It is also worth noting that its extended version is published in the top computer vision journal IEEE Transactions on PAMI in 2020 [Sun et al., 2020].

After this work, my research team members respectively delved into another two perspectives of meta-learning: 1) learning to customize and combine multiple deep neural networks [Liu et al., 2020a] (120+ Google Scholar citations); and 2) learning to self-train deep neural networks with semi-supervised data [Li et al., 2019] (270+ Google Scholar citations), in order to obtain robust and efficient fast adaptation results on novel few-shot learning tasks. It is worth noting that the second work deployed MTL [Sun et al., 2019] as one of the basic operations in learning its model backbone.

### 2.2 Class-incremental learning tasks

The class incremental learning task aims to learn new concepts by incrementally updating a model trained on previous concepts. However, there is an inherent trade-off to effectively learning new concepts without catastrophic forgetting of previous ones who have insufficient training samples in later phases. To alleviate this issue, it has been proposed to keep around few-shot examples of the previous concepts, but the effectiveness heavily depends on the representativeness of these examples. In our work [Liu et al., 2020b], we designed a novel meta-learning algorithm we call mnemonics training, where we parameterize few-shot exemplars and make them optimizable in an end-to-end manner. We train the framework through bi-level optimizations, i.e., model-level and exemplar-level, automatically, as shown in Figure 3. We conduct extensive experiments on three class-incremental learning benchmarks. Quite intriguingly, we found that the meta-learned few-shot exemplars tend to be located on the boundaries between classes, which is desirable for training discriminative models. This work has gained 250+ Google Scholar citations since its publication in CVPR 2020.

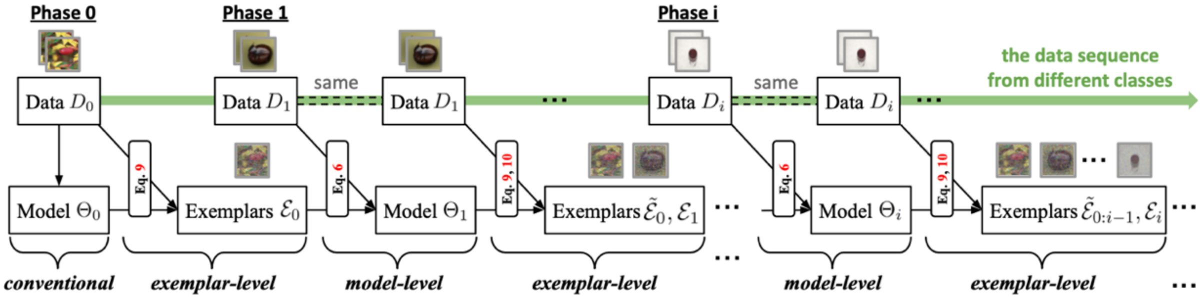


*Figure 3. Mnemonics Training [Liu et al., 2020b]. It alternatively optimizes the parameters of models and few-shot exemplars. The equations formulate the bi-level optimization problems.*

After this work, my team members developed a few more efficient meta-learning (or reinforcement learning in non-differentiable settings) algorithms on this task. The algorithms include: 1) learning to combine stable and plastic networks where stable network learns to keep the memory of past classes and plastic network learns to adapt to new coming classes [Liu et al., 2021a]; 2) learning to allocate the optimal amount of memory for each class in an online manner [Liu et al., 2021b]; and 3) learning to compress training samples and achieve memory-saving by reducing memory costs [Luo et al., 2023]. It is worth noting that the first and second works deployed MTL [Sun et al., 2019] as one of the basic operations for learning the model backbones.

## 3. Interventional Training (2019 - 2023)

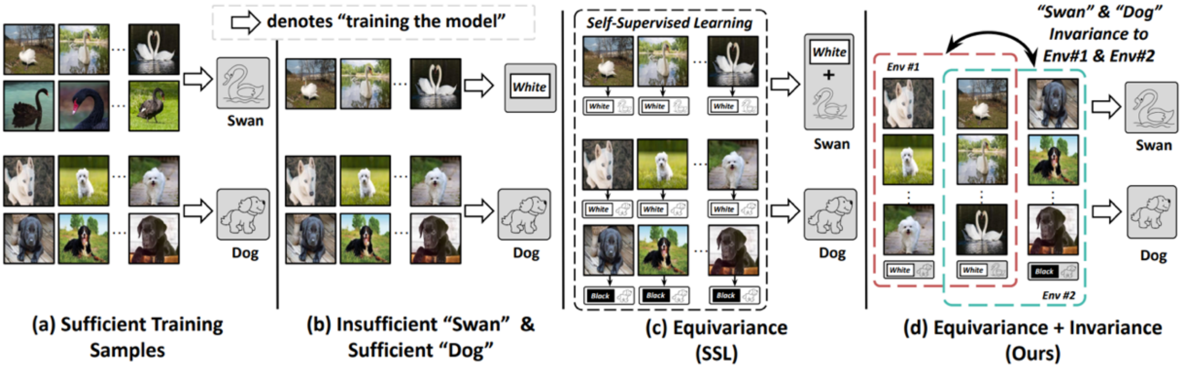
Intervention is a core research direction of causality theory [Pearl, 2009]. In human society, causality lets us stand on common ground where we humans can communicate with each other, learn from others, instruct, and motivate others. In the research of AI, equipping computing machines with causality helps to establish stable, invariant, causal relations among data, algorithms, and models without needing much training data. The training method based on intervention theory (we call Interventional Training) is a way of implementing causality in AI and making it computable to yield causal and efficient AI models. Interventional training includes three key steps: constructing a causal graph for a specific problem (e.g., a few-shot learning problem): identifying the confounder that hampers the learning of the causal model; and intervening in the learning process to mitigate the confounder while deriving the causal model.

### 3.1 Few-shot learning tasks

As I have stated in the section on meta-learning, one important way to few-shot learning is transferring pre-trained knowledge to the training of unseen tasks. In the work [Yue et al., 2020], we first pointed out that the bottleneck in such kind of method is the contextual bias in the pre-trained knowledge does evil in learning subsequent few-shot classifiers. We tackled the issue by proposing a novel interventional training paradigm: Interventional Few-Shot Learning (IFSL). Our theory assumes of the causalities among the pre-trained knowledge, few-shot samples, and class labels. We proposed a structural causal model of the causalities for few-shot learning and then developed three practical implementations based on a key technique of causality called backdoor adjustment. We diagnosed the classification accuracy comprehensively across query hardness and showed that IFSL improves all the baselines across all the hardness. It is worth highlighting that the contribution of IFSL is not only about improving the few-shot learning performance but also offering a causal explanation of why IFSL works well: IFSL is a causal approximation to many-shot learning. Note that few-shot learning uses a special setting to evaluate the models' adaptation ability. Its training is based on the availability of multiple few-shot tasks. My team also researched in more general settings where the model is learned on a single task with limited data.

### 3.2 Data-insufficient problems in image classification tasks

In the work [Wang et al., 2022], we touched on a general data-insufficient problem in classification: insufficient environments. For example, in Figure 4, if all the training "swan" samples are "white" (one of the environments of the class "swan"), the classification model may wrongly use the "white" environment to represent the intrinsic class "swan". Besides, we were interested in learning robust models without the need for any pre-trained model checkpoints, as we have found the evil they caused in few-shot learning settings.


denotes "training the model"
Self-Supervised Learning
"Swan" & "Dog" Invariance to Env#1 & Env#2
White
White +
Swan
Dog
Env #1
Env #2
Black

(a) Sufficient Training Samples

(b) Insufficient "Swan" & Sufficient "Dog"

(c) Equivariance (SSL)

(d) Equivariance + Invariance (Ours)

*Figure 4. Illustration of how the equivariance and invariance inductive biases help learning from insufficient data. Cartoon figures denote the class feature. Boxed words denote environmental features. Grey-boxed figures denote the learned model.*

In this work, we first show why insufficient data renders the model more easily biased to the limited training environments (that are usually different from testing), compared to the case of sufficient data. Then, we justify that equivariance inductive bias can retain the class feature while invariance inductive bias can remove the environmental feature, leaving only the class feature that generalizes to any testing environmental changes. To impose them on learning, for equivariance, we demonstrate that any off-the-shelf contrastive-based self-supervised feature learning method can be deployed; for invariance, we propose an interventional training method called class-wise invariant risk minimization (IRM). This method efficiently tackles the challenge of missing environmental annotation in conventional IRM. This work was accepted by ECCV 2022 with high ratings and its application in natural language understanding (a basic task in NLP) was published in EMNLP 2022 [Yu et al., 2022].

### 3.3 Label-insufficient problems in semantic segmentation tasks

Compared to image classification, semantic segmentation is another key computer vision task, and it aims to segment all objects from the image. The key challenge is the laborious and expensive annotation for pixel-level labels (called segmentation masks), especially when the scene in the image is complex, e.g., a crowded street scene in the application of autonomous driving. My research team has proposed several solutions [Zhang et al., 2020, Chen et al., 2022, Chen and Sun, 2023] by using weak labels such as image-level labels to automatically generate pixel-level labels, i.e., pseudo segmentation masks. Taking [Zhang et al., 2020] as a teaser example, we inspect the baseline method of generating pseudo masks (called Class Activation Maps---CAM) and attribute the cause of the ambiguous boundaries of pseudo masks to the confounding context, e.g., the correct image-level classification of "horse" and "person" maybe not only due to the recognition of each instance but also their co-occurrence context, making the mask generation model hard to distinguish between the boundaries. Inspired by this, we propose a structural causal model to analyze the causalities among images, contexts, and class labels. Based on it, we develop a new interventional training method: Context Adjustment (CONTA), to remove the confounding bias in image-level classification and thus provide better pseudo-masks as ground truth to train the final segmentation model. This work has gained 260+ Google Scholar citations since its publication in NeurIPS 2020 (Oral Presentation). The follow-up works in my team include ReCAM [Chen et al., 2022] which intervenes in the mask generation process by improving the objective function; and LPCAM [Chen and Sun, 2023] further optimizing the normalization step in the process to achieve the state-of-the-art performance. Both are published in one of the top venues of computer vision: CVPR.

## 4. Future Works

We have seen the NLP community advocating fully to the pre-training of Transformers since the year of 2017, i.e., the year when "Attention is all you need" paper was published on NIPS. Based on this, a variety of NLP tasks have been unified. We have seen the amazing performance presented by ChatGPT. In the computer vision community, we have also seen quite a few of pioneering foundation models such as Contrastive Language-Image Pretraining (CLIP) [Radford et al. 2021], Stable Diffusion [stability.ai] and Segment Anything (SAM) [Kirillov et al. 2023] emerged. I realized the fact that the above proposed data-limited learning methods are merely towards any key direction for the future, given the fact that the basic visual representation will be perfectly transferred from a large pre-trained foundation model (rather than training models from scratch with small data or weak labels). In other words, simple data fitting approach through large-scale Visual Transformers can produce good enough visual representation, and the representation itself is disentangled and can be basically fixed for hands-on deployment. So, I believe now it is the time to go ahead to perform high-level comprehension on 1) multi-modality data (e.g., a web page composed of images, videos, texts, graphs, etc.) for daily-life application, and 2) scientific data (e.g., a new and sustainable material of cloth) for interdisciplinary research. I am very interested in working towards these two directions.

**Towards a multimodality foundation model.** Emerging foundation models build upon deep learning and aim to unify models at the architecture level, such as Generative Pre-trained Transformers (GPT) for language generation, diffusion-based deep models for image generation, and Bidirectional Encoder Representations from Transformers (BERT) for text feature extraction. My plan is to break down the barriers between different modalities by creating a unified foundation model that accepts inputs from various modalities (e.g., image, text, video, and graph) and flexibly generates outputs of different modalities. The overall idea is that everything is represented as language tokens. It is inspired by the success of commercial language foundation models such as ChatGPT. Language is designed by humans, and it is thus naturally logical and structural. In contrast, other modalities like images and audio (including the voices from the nature world) are relatively random. To this end, my team members are actively discussing the novel idea of "Modality Unified Tokenization" (MUT), where we can learn an encoder to map data from different modalities into natural language tokens. These tokens can then be used by a diffusion model to decode and reconstruct the original data in the respective modalities. We do have to face quite a few of research questions: 1) “How to encode different modality data from various modalities into efficient feature representations?”, 2) “How to map different modality data from feature representations into unified text tokens?”, and 3) “How to align the semantics of the text tokens mapped from different modality data into one model?”. We have done a preliminary work (under review) for answering question 1 and will go ahead to explore possibilities for accomplishing all of them. Besides, since our models will be unifying different modalities, e.g., those on the internet, it is crucial for us to consider ethical and bias problems (e.g., societal and cultural biases) seriously. We believe that ensuring fairness and mitigating biases is an important aspect of any AI research.

**Towards a dynamic interaction between AI and the scientific process.** It is well-known that science is a falsifiable theory drawn from experimental justifications which are expensive and really depending on expertise in specific areas. My research team is wondering if we can use AI to automatically discover the scientific law connecting various scientific factors. We are discussing quite a few of thoughtful and comprehensive approaches to integrating AI with scientific discovery and understanding: 1) *Tentative Cause-Effect Hypotheses*: Just as humans use their commonsense knowledge to form initial hypotheses about cause-and-effect relationships, AI can use its learned patterns to create initial models connecting scientific factors. This is like the early stage of science discovery where researchers propose hypotheses based on expertise. 2) *Imagination and Hypothesis Testing*: AI's ability to imagine and simulate future scenarios is akin to the thought experiments scientists often conduct to test their hypotheses. This anticipatory process allows AI to generate predictions and hypotheses that can be verified or falsified when new data becomes available. 3) *Verification and Falsification Loop*: The heart of scientific discovery lies in verifying or falsifying hypotheses using experimental data. AI's capability to update its models based on new evidence is analogous to the scientific method's iterative process of refining theories based on empirical results. 4) *Integration with Physics-based Knowledge*: Connecting AI models with physics-based knowledge, such as mathematical equations and system symmetries, bridges between data-driven insights and established scientific principles. This integration can aid in making AI's discoveries more interpretable, explainable, and aligned with existing scientific frameworks. So, in summary, what we are interested in are three folds: 1) Data-Driven Discovery, 2) Incorporation of Expertise, and 3) Integration with Scientific Computing Models. Via these, we will work towards a dynamic interaction between AI and the scientific process, leveraging AI's strengths in pattern recognition, prediction, and hypothesis generation to complement humans’ “scientific curiosity”.

## 5. References

[Alemohammad et al., 2023] Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Ali, S., and Baraniuk, R. G. (2023). Self-consuming generative models go mad. arXiv, 2307.01850.

[Chen and Sun, 2023] Chen, Z. and Sun, Q. (2023). Extracting class activation maps from non-discriminative features as well. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Chen et al., 2022] Chen, Z., Wang, T., Wu, X., Hua, X.-S., Zhang, H., and Sun, Q. (2022). Class reactivation maps for weakly-supervised semantic segmentation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[He et al., 2016] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Li et al., 2019] Li, X., Sun, Q., Liu, Y., Zheng, S., Chua, T.-S., and Schiele, B. (2019). Learning to self-train for semi-supervised few-shot classification. In The Conference on Neural Information Processing Systems (NeurIPS).

[Liu et al., 2020a] Liu, Y., Schiele, B., and Sun, Q. (2020a). An ensemble of epoch-wise empirical bayes for few-shot learning. In European Conference on Computer Vision (ECCV).

[Liu et al., 2021a] Liu, Y., Schiele, B., and Sun, Q. (2021a). Adaptive aggregation networks for class-incremental learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Liu et al., 2021b] Liu, Y., Schiele, B., and Sun, Q. (2021b). RMM: Reinforced memory management for class-incremental learning. In The Conference on Neural Information Processing Systems (NeurIPS).

[Liu et al., 2020b] Liu, Y., Su, Y., Liu, A.-A., Schiele, B., and Sun, Q. (2020b). Mnemonics training: multi-class incremental learning without forgetting. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Long et al., 2015] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Luo et al., 2023] Luo, Z., Liu, Y., Schiele, B., and Sun, Q. (2023). Class-incremental exemplar compression for class-incremental learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Ma et al., 2017] Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., and Van Gool, L. (2017). Pose guided person image generation. In Advances in Neural Information Processing Systems 30 (NIPS).

[Ma et al., 2018] Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., and Fritz, M. (2018). Disentangled person image generation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Pearl, 2009] Pearl, J. (2009). Causality: Models, reasoning, and inference. Cambridge University Press, 2nd edition.

[Sun et al., 2020] Sun, Q., Liu, Y., Chen, Z., Chua, T.-S., and Schiele, B. (Aug 2020). Meta-transfer learning through hard tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44:1443–1456.

[Sun et al., 2019] Sun, Q., Liu, Y., Chua, T.-S., and Schiele, B. (2019). Meta-transfer learning for few- shot learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Sun et al., 2017] Sun, Q., Schiele, B., and Fritz, M. (2017). A domain based approach to social relation recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Wang et al., 2022] Wang, T., Sun, Q., Pranata, S., Jayashree, K., and Zhang, H. (2022). Equivariance and invariance inductive bias for learning from insufficient data. In European Conference on Computer Vision (ECCV).

[Yu et al., 2022] Yu, S., Jiang, J., Zhang, H., Niu, Y., Sun, Q., and Bing, L. (2022). Interventional training for out-of-distribution natural language understanding. In The Conference on Empirical Methods in Natural Language Processing (EMNLP).

[Yue et al., 2020] Yue, Z. Y., Zhang, H., Sun, Q., and Hua, X.-S. (2020). Interventional few-shot learning. In The Conference on Neural Information Processing Systems (NeurIPS).

[Zhang et al., 2020] Zhang, D., Zhang, H., Tang, J., Hua, X.-S., and Sun, Q. (2020). Causal intervention for weakly-supervised semantic segmentation. In The Conference on Neural Information Processing Systems (NeurIPS).

[Zhang et al., 2015] Zhang, N., Paluri, M., Taigman, Y., Fergus, R., and Bourdev, L. (2015). Beyond frontal faces: Improving person recognition using multiple cues. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Radford et al. 2021] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G. (2021). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[Kirillow et al. 2023] Kirillov, A., Mintun, E., Nikhila, R., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R. (2023). Segment anything. In The International Conference on Computer Vision (ICCV).