Max Planck - Amazon Science Hub 2024 Research Symposium

Monday, September 16th - Day 1 at MPI-IS, Max-Planck-Ring 4, 72076 Tübingen			
Time	**Event**	**Speaker**	**Location**
10:30 am	Registration		MPI-IS Tübingen, Lobby
11:00-11:40 am	MPI-IS Lab Tour	Pre-registration required (limited availability)	MPI-IS Tübingen
11:45-12:00 pm	Welcome Remarks	Michael J. Black, Managing Director, MPI for Intelligent Systems Michael Hirsch, Applied Science Manager, Amazon	MPI-IS, N0.002
12:00 -1:00 pm	*Lunch Break*		
1:00-1:30 pm Session Chair: Michael J. Black	Keynote Talk	***Haptic Intelligence*** Katherine J. Kuchenbecker, Director, Haptic Intelligence department, MPI for Intelligent Systems	MPI-IS, N0.002
1:30-2:30 pm 10+2 mins Session Chair: Michael J. Black	Project Talk	***Learning clothing pressure fields via physics-based simulations*** Gokhan Serhat, Assistant Professor at KU Leuven and Research Scientist at MPI for Intelligent Systems; Yashwardhan Deshmukh, Research Software Engineer, Haptic Intelligence department, MPI for Intelligent Systems	MPI-IS, N0.002
	Project Talk	***Physically Plausible Object Pose Refinement in Cluttered Scenes*** Jörg Stückler, Professor for Intelligent Perception in Technical Systems at the University of Augsburg	MPI-IS, N0.002
	Project Talk	***What do foundation models know about 3D humans?*** Gerard Pons-Moll, Professor, Department of Computer Science, University of Tübingen (remote)	MPI-IS, N0.002
	Project Talk	***Computer vision for exploring nonverbal human communication*** Carolin Schmitt, Research Engineer, Optics and Sensing Laboratory, MPI for Intelligent Systems Gökce Ergün, Research Engineer, Optics and Sensing Laboratory, MPI for Intelligent Systems	MPI-IS, N0.002
2:30-2:50 pm	*Coffee Break*		*Lobby*

2:50-3:20 pm 10+2 min Session Chair: Michael Hirsch	Project Talk	***Root Causal Analysis (RCA) at Amazon: some results and challenges*** Sergio Herna Garrido Mejia, Ph.D. Student, Empirical Inference Department, MPI-IS	MPI-IS, N0.002
	Project Talk	***Improving the AI recycling rate with thermodynamics-guided techniques*** Waleed Mohammed, Ph.D. Student, Department Microstructure Physics and Alloy Design, MPI for Sustainable Materials	MPI-IS, N0.002
3:20-3:50 pm Session Chair: Michael Hirsch	Talk 1	***Physics-Inspired Fit-Aware Virtual Try-On*** Ming Lin, Amazon Scholar ***Generative 3D and Novel View Synthesis*** Sunil Hadap, Principal Applied Scientist Amazon Fashion and Fitness	MPI-IS, N0.002
3:50-4:50 pm	Poster Session & Networking		MPI-IS, N0.002

Tuesday, September 17th - Day 2 at Amazon Research & Development Building, Tübingen

Time	Event	Speaker	Location
9:30 am	Welcome 15 min	Introduction: Betty Mohler Tesch, Principal Applied Scientist Amazon Fashion and Fitness Michael Hirsch, Applied Science Manager Amazon	Amazon
9:45-10:15 am	Keynote Talk	***Adversarial Training Should Be Cast as a Non-Zero-Sum Game*** Volkan Cevher, Amazon Academic Research Consultant, Amazon Artificial General Intelligence Foundations	Amazon
10:15-10:35 am	Keynote Talk	***Title tbc*** Chris Markus-Kratz, Director, AWS Automotive	Amazon
10:35-10:55 pm	Talk 2	***AI in Sports*** Ianir Ideses, Senior Manager Applied Science Amazon Prime Video	Amazon
10:55-11:15	*Coffee Break*		
11:15-11:35 am	Talk 3	***Making Time Series Classification Understandable using Multiple Instance Learning*** Jas Kandola Senior Manager Applied Science Amazon Prime Video	Amazon

11:35-11:55 am	Talk 4	***Packaging Optimization at Amazon Scale: Tackling Scientific Challenges with Machine Learning and ML Operational Excellence*** Matthias Polozek, Principal Applied Scientist Amazon Robotics	Amazon
11:55-12:15 pm	Talk 5	***REXEL: An end-to-end model for document-level relation extraction and entity linking*** Shubhi Tyagi, Senior Applied Scientist Amazon Artificial General Intelligence	Amazon
12:15-1:15 pm	*Lunch Break*		
1:15-1:35 pm	Talk 6	***Generative AI for Ambient Sound: Fusing Text, Melody, and Context for Personalized Experiences*** Hassen Dhrif, Senior Applied Scientist, Amazon Devices	Amazon
1:35-1:55 pm	Talk 7	***Distributed and Adversarial Training of Large Machine Learning Models*** Michael Mühlebach, Research Group Leader, Learning and Dynamical Systems Group, MPI-IS	Amazon
1:55-2:15 pm	Talk 8	***Theoretical Foundations of Deep Selective State-Space Models*** Antonio Orvieto, Principal Investigator at the ELLIS Institute Tübingen and independent group leader at MPI-IS	Amazon
2:15-2:35 pm	*Coffee Break*		
2:35-2:55 pm	Talk 9	***When do adversarial attacks against language models matter?*** Jonas Geiping, Principal Investigator at the ELLIS Institute Tübingen and independent group leader at MPI-IS	Amazon
2:55-3:15 pm	Talk 10	***Improving Decision Making with Machine Learning, Provably*** Manuel Gomez-Rodriguez, Faculty at the MPI for Software Systems	Amazon
3:15-3:35 pm	Talk 11	***Generative AI for Research: Global Examples how researchers leverage cloud services to accelerate their work*** Roberta Piscitelli, Academic Research Lead EMEA Andreas Wagner, Sr. Account Manager for Research in Germany	Amazon
3:35-3:50 pm	Talk 12	***Inference optimization for foundational models on AI accelerators*** Kailash Budhathoki, Sr Applied Scientist , AWS Deep Engine-Science	Amazon

Monday, September 16th **Day 1 at MPI-IS, Max-Planck-Ring 4, 72076 Tübingen**		
Time	**Speaker**	**Talk Details**
1:00-1:30 pm		***Haptic Intelligence*** Katherine J. Kuchenbecker, Director, Haptic Intelligence department, MPI for Intelligent Systems **Abstract**: Touch plays a crucial role in the sensorimotor systems of humans and animals. However, this sense is far less understood than vision or hearing, since what you feel greatly depends on how you move, and since engineered haptic sensors, actuators, and algorithms typically struggle to match biological capabilities. My team works to sharpen our understanding of haptic interaction while simultaneously inventing helpful human-computer, human-machine, and robotic systems that take advantage of the unique capabilities of the sense of touch. I will elucidate these ideas of haptic intelligence by showcasing several systems we have created and evaluated in recent years. **Bio**: Katherine J. Kuchenbecker is a Director at the Max Planck Institute for Intelligent Systems (MPI-IS) in Stuttgart, Germany, and an Honorary Professor at the University of Stuttgart. She earned her Ph.D. at Stanford University in 2006, did postdoctoral research at the Johns Hopkins University, and was an engineering professor in the GRASP Lab at the University of Pennsylvania from 2007 to 2016. Her research blends haptics, teleoperation, physical human-robot interaction, tactile sensing, and medical applications. She delivered a TEDYouth talk on haptics in 2012 and has been honored with a 2009 NSF CAREER Award, the 2012 IEEE RAS Academic Early Career Award, a 2014 Penn Lindback Award for Distinguished Teaching, elevation to IEEE Fellow in 2021, and 19 best paper, poster, demonstration, and reviewer awards. She co-chaired the IEEE Haptics Symposium in 2016 and 2018 and is Editor-in-Chief of the 2025 IEEE World Haptics Conference. Furthermore, she has led the International Max Planck Research School for Intelligent Systems (IMPRS-IS) since its founding in 2017 and is the Robotics Institute Germany (RIG) Spokesperson for MPI-IS.

1:30-2:30 pm Project Talks		***Learning clothing pressure fields via physics-based simulations*** Gokhan Serhat, Assistant Professor at KU Leuven and Research Scientist at MPI for Intelligent Systems **Abstract:** Accurate and rapid estimation of garment comfort is a desirable feature in online shopping applications. To achieve this capability, the applicability of physics-based tools is limited due to the intricate nature of body-cloth contact. Such complex interactions complicate the modeling process, hinder robustness, and often induce long simulation times. We address these significant challenges by using deep neural networks to learn body and garment deformation fields from clothing simulations. We present an elaborate analysis workflow that builds on the SMPL-X database for 3D human modeling, Blender for processing the models, and Houdini for simulating the body-garment contact. To reduce the problem, we focus on the upper body and select t-shirts as the cloth type. The adjustable parameters for the human body include gender, height, and weight; while size and material properties can be altered for the t-shirt. We leverage modern physics engines that allow us to approximate the stress fields that emerge as the fabric conforms to the body and reaches an equilibrium state. The resulting stretch stress distribution will be used to define a metric for clothing comfort estimation. We expect the developed framework to accurately predict garment pressure for various scenarios almost in real-time and be a valuable tool for clothing comfort assessment. **Bio:** Gokhan Serhat is an Assistant Professor of Mechanical Engineering at KU Leuven in Belgium and a part-time research scientist at the Max Planck Institute for Intelligent Systems in Germany since 2022. Prior to his current position, he worked as a research scientist (2021-2022) and postdoctoral researcher (2019-2021) at the Max Planck Institute for Intelligent Systems. He previously received a PhD degree in mechanical engineering from Koc University (2018), an MS degree in computational mechanics from the Technical University of Munich (2014), and a BS degree in mechanical engineering from Middle East Technical University (2011). His research interests include computational mechanics, applied machine learning, structural dynamics, numerical methods, and composite materials.

Physically Plausible Object Pose Refinement in Cluttered Scenes

Jörg Stückler, Professor for Intelligent Perception in Technical Systems at the University of Augsburg

Abstract: Estimating the 6-DoF pose of objects from images is a fundamental task in computer vision and a prerequisite for downstream tasks like augmented reality or robotic grasping applications. This task becomes particularly challenging in cluttered scenes, when many objects are present in the image in close proximity and occlude one another. However, the close proximity between objects also provides additional cues about the objects, as objects in physically plausible scenes do not intersect one another and thus occluding objects constrain the ones they occlude. We present a novel approach for utilizing this information in 6-DoF object pose refinement of known objects. Our formulation extends RAFT-based pose refinement to reduce penetrations between objects to a large degree and leads to more plausible object poses with less penetrations. We evaluate our approach quantitatively and qualitatively on two benchmark datasets, and demonstrate improvements over baselines.

Bio: Joerg Stueckler holds a professorship for Intelligent Perception in Technical Systems at the University of Augsburg. Prior to this position, he led the Cyber Valley Max Planck research group on Embodied Vision at the Max Planck Institute for Intelligent Systems in Tuebingen from April 2018 until July 2024. He conducted his postdoc research in the Computer Vision groups at RWTH Aachen University (2015-2017) and TU Munich (2014-2015) and has been visiting professor at TU Munich in the winter term 2017/2018. In 2014, he obtained his PhD from University of Bonn with summa cum laude. His dissertation thesis has been awarded with the Georges Giralt PhD Award in 2015 by euRobotics aisbl.

what do foundation models know about 3D humans?

Gerard Pons-Moll, Professor, Department of Computer Science, University of Tübingen (remote)

Abstract: Reconstructing 3D humans from monocular or sparse input views has been a main research area for several years. Only in recent years we have witnessed methods capable of reconstructing 3D humans, including clothing from images. While conceptually powerful, such models have limited generalisation due to the small-scale training. Foundation models have revolutionised many computer vision sub-areas. But how much do these models know about 3D humans in clothing? Are they 3D consistent? Can we use them to our advantage to deal with monocular images? In the second half of the talk, I will try to answer these questions and present a new model which nicely combines foundation models with good old 3D computer vision.

Bio: Gerard Pons-Moll is a Professor at the University of Tübingen endowed by the Carl Zeiss Foundation, at the department of Computer Science. He is also core faculty at the Tübingen AI Center, senior researcher at the Max Planck for Informatics (MPII) in Saarbrücken, Germany, and faculty at the IMPRS-IS (International Max Planck Research School - Intelligent Systems in Tübingen). He is also program director of the new ELLIS program for Vision and Graphics. His research lies at the intersection of computer vision, computer graphics and machine learning -- with special focus on analyzing people in videos, and creating virtual human models by "looking" at real ones. His research has produced some of the most advanced statistical human body models of pose, shape, soft-tissue and clothing (which are currently used for a number of applications in industry and research), as well as algorithms to track and reconstruct 3D people models from images, video, depth, and IMUs.

His work has received several awards including the prestigious Emmy Noether Grant (2018), a Google Faculty Research Award (2019), a Facebook Reality Labs Faculty Award (2018,2024), the German Pattern Recognition Award (2019), which is given annually by the German Pattern Recognition Society to one outstanding researcher in the fields of Computer Vision and Machine Learning. His work got Best Papers Awards BMVC'13, Eurographics'17, 3DV'18, 3DV'22 and CVPR'20, ECCV'22 and has been published at the top venues and journals including CVPR, ICCV, Siggraph, Eurographics, 3DV, IJCV and PAMI. He serves regularly as Area chair for the main conferences in vision and learning.

Computer vision for exploring nonverbal human communication

Carolin Schmitt, Research Engineer, Optics and Sensing Laboratory, MPI for Intelligent Systems

Abstract: In my talk I am going to discuss how we can use recent computer vision methods to support research tackling the analysis of human-to-human non-verbal communication. Towards this goal, I present our recently released NICE Toolbox: An easy-to-use framework for exploring nonverbal human communication. It aims to enable the investigation of observable signs that reflect the mental state and behaviors of the individual. Further, I am going to present our ongoing efforts of developing comprehensive evaluation strategies. These translate into concrete recommendations for data capture setups and behavior analysis methods that likely produce more accurate and reliable predictions.

Bio: I am a computer vision research engineer working in the Optics and Sensing Laboratory (OSLab) at the Max-Planck Institute for Intelligent Systems. Here, I am developing a toolbox for vision-based human-to-human communication analysis from single or multi-view camera data. Prior to this work, I obtained a doctoral degree in the intersection of computer vision and computer graphics. Specifically, I worked on photo-realistic 3d reconstructions of geometry and materials of indoor scenes from data acquired by hand-held multi-sensor devices.

2:30-2:50 pm	*Break*	
2:50-3:20 pm		***Root Causal Analysis (RCA) at Amazon: some results and challenges*** Sergio Hernan Garrido Mejia, Ph.D. Student, Empirical Inference Department, MPI-IS **Abstract:** Root Cause Analysis (RCA) is a common task in complex large systems where failures arise as a result as disturbance in a node on the system. Although the name suggests it, this task has not always been tackled from a statistical causal perspective. In this talk I will talk about where this problem arises in Amazon systems and some solutions, we have proposed using statistical causal methods. We will conclude the talk by presenting some of the current challenges of RCA in Amazon systems and possible fruitful directions of research. **Bio:** Sergio is an industrial ELLIS PhD student at Amazon and the Max Planck Institute for Intelligent Systems, supervised by Dominik Janzing and Bernhard Schölkopf. His work focuses on the intersection of causality and machine learning. In particular, he studies what are the causal implications of data merging. Recently, Sergio broadened his interests to include causal abstraction and representation, as well as applications of causality to environmental and social problems. His research is supported by an Amazon Science Hub grant.

Improving the Al recycling rate with thermodynamics-guided techniques

Waleed Mohammed, Ph.D. Student, Department Microstructure Physics and Alloy Design, MPI for Sustainable Materials

Abstract: The primary production of aluminum (Al) metal from its ore contributes to approximately 3% of all global greenhouse gas emissions due to the energy-intensive electrolysis process. The rising focus on sustainable practices has prompted a shift towards recycling scrap materials. Al recycling can save up to 95% of the energy used per ton of Al. However, a significant portion of the scrapped material is post-consumer scrap with high levels of elemental contamination. Consequently, the remelting of scrap during the recycling process leads to the accumulation of impurities in the recycled alloys. This can negatively impact their properties and limit their potential applications. Moreover, the removal of these impurities is often energy-intensive and time-consuming. In our research, we employ thermodynamics-guided techniques to investigate the possibilities of producing impurity-tolerant recycled Al alloys with high fractions of scrap and less compromised properties.

Bio: Waleed Mohammed is a researcher at the Max Planck Institute for Sustainable Materials, specializing in aluminum alloys. His primary focus is on the design and processing of scrap-based aluminum alloys, aiming to advance sustainable practices in metallurgy. Waleed completed his Master's degree in Metallurgical Engineering with distinction from RWTH Aachen University in 2023. In recognition of his exceptional academic achievements, he was honored with the Springorum-Denkmünze award.

3:20-3:35 pm		***Physics-Inspired Fit-Aware Virtual Try-On*** Ming Lin, Amazon Scholar **Abstract:** Digital try-on is still not practical and easy-to-use enough to replace physical try-on, mostly due to the gap in modeling and in determining the fit between the digital twin and the actual body. The estimation of the hidden parameters of the garments plays an important role in closing the gap. In this talk, we will focus on faithful estimation of garment materials via learning and optimization, and briefly discuss user-friendly recovery of garments, and fast, realistic visual rendering of animated try-on results. Although previous methods have made some progresses on these under constrained problems, learning-based approaches have shown tremendous potential in making notable impact. **Bio:** Ming C. Lin received her B.S., M.S., Ph.D. in EECS from University of California, Berkeley. She is a Distinguished University Professor, Dr. Barry Mersky and Capital One E-Nnovate Endowed Professor, and former Elizabeth Stevinson Iribe Chair of Computer Science at University of Maryland at College Park, as well as John & Louise Parker Distinguished Professor Emerita at University of North Carolina at Chapel Hill. She is also an Amazon Scholar. She has received several honors and awards, including NSF Young Faculty Career Award, UNC Hettleman Award for Scholarly Achievements, Beverly W. Long Distinguished Term Professor, IEEE VGTC VR Technical Achievement Award, Washington Academy of Sciences Distinguished Career Award and several best paper awards. She is a Fellow of National Academy of Inventors, ACM, IEEE, Eurographics, ACM SIGGRAPH Academy, and IEEE VR Academy.

3:20-3:35 pm		***Generative 3D and Novel View Synthesis*** Sunil Hadap **Abstract:** Fashion and Fitness (F2) is committed to delivering top-tier immersive shopping experiences. We've successfully launched industry-leading 3D experiences such as AR Virtual Try-On (AR) and View in 3D (V3D), resulting in high customer satisfaction (CSAT) and conversion rates. However, scaling these immersive experiences across Amazon's vast catalog presents unique technical challenges. This talk introduces F2's comprehensive Generative 3D and Novel View Synthesis strategy to acquire VTO ready 3D assets at scale and required fidelity. **Bio:** Sunil Hadap is a Principal Applied Scientist in Amazon's Fashion & Fitness organization and heads the science for 3D Virtual Try-On and AR. Before that he was a principal scientist at Adobe Research, and R&D staff at Dreamworks Animation before that. He is credited for his simulation work on several block-buster DreamWorks movies including Shrek2, Madagascar, Kung Fu Panda. He is passionate about everything and at the intersection of CG, CV, DL, Simulation and Generative AI.
3:50-4:50 pm	Poster Session & Networking	

Tuesday, September 17th Day 2 at Amazon Research & Development Building, Friedrich-Miescher-Str. 4, 72076 Tübingen		
Time		**Speaker**
9:30 am	Welcome	Introduction: Betty Mohler Tesch, Principal Applied Scientist Amazon Fashion and Fitness Michael Hirsch, Applied Science Manager Amazon (TBC)
9:45-10:15 am Keynote Talk		***Adversarial Training Should Be Cast as a Non-Zero-Sum Game*** Volkan Cevher, Amazon Academic Research Consultant, Amazon Artificial General Intelligence Foundations **Abstract:** One prominent approach toward resolving the adversarial vulnerability of deep neural networks is the two-player zero-sum paradigm of adversarial training, in which predictors are trained against adversarially-chosen perturbations of data. Despite the promise of this approach, algorithms based on this paradigm have not engendered sufficient levels of robustness, and suffer from pathological behavior like robust overfitting. To understand this shortcoming, we first show that the commonly used surrogate-based relaxation used in adversarial training algorithms voids all guarantees on the robustness of trained classifiers. The identification of this pitfall informs a novel non-zero-sum bilevel formulation of adversarial training, wherein each player optimizes a different objective function. Our formulation naturally yields a simple algorithmic framework that matches and, in some cases, outperforms state-of-the-art attacks, attains comparable levels of robustness to standard adversarial training algorithms, and does not suffer from robust overfitting. **Bio:** Volkan Cevher received the B.Sc. (valedictorian) in electrical engineering from Bilkent University in Ankara, Turkey, in 1999 and the Ph.D. in electrical and computer engineering from the Georgia Institute of Technology in Atlanta, GA in 2005. He was a Research Scientist with the University of Maryland, College Park, from 2006-2007 and also with Rice University in Houston, TX, from 2008-2009. He was also a Faculty Fellow in the Electrical and Computer Engineering Department at Rice University from 2010-2020. Currently, he is an Associate Professor at the Swiss Federal Institute of Technology Lausanne and an Amazon Scholar. His research interests include machine learning, optimization theory and methods, and automated control. Dr. Cevher is an IEEE Fellow ('24), an ELLIS fellow, and was the recipient of the ICML AdvML Best Paper Award in 2023, Google Faculty Research award in 2018, the IEEE Signal Processing Society Best Paper Award in 2016, a Best Paper Award at CAMSAP in 2015, a Best Paper Award at SPARS in 2009, and an ERC CG in 2016 as well as an ERC StG in 2011.

10:15-10:35 am Keynote Talk		***Title tbc*** Chris Markus-Kratz, Director, AWS Automotive

10:35-10:55 pm Talks		***AI in Sports*** Ianir Ideses, Senior Manager Applied Science Amazon Prime Video **Abstract:** In this talk, we will present our team's work on developing cutting-edge AI-powered experiences for sports broadcast. We'll showcase a range of novel analytical features and enhanced graphics that can help fans gain deeper insights and get more immersion from the game. The lecture will provide a glimpse at our computer vision and machine learning (CV/ML) powered technologies that are used to enhance our viewers experience and provide our production team with unique story telling tools. **Bio:** Ianir currently heads the Science teams for the Advanced Visual Experiences group at Amazon Prime Video Live Events. In his previous role at Amazon, Ianir led the development of CVML algorithms in the field of Health and Wellness at Amazon Lab126. Prior to joining Amazon, Ianir was Chief Scientist at Gett Systems, where he headed the development of marketplace algorithms in transportation. Ianir holds a PhD from Tel-Aviv University in the field of 3D Visualization.

10:55-11:15	*Break*	
11:15-11:35 am		***Making Time Series Classification Understandable using Multiple Instance Learning*** Jas Kandola Senior Manager Applied Science Amazon Prime Video **Abstract:** Time Series Classification is widely used in many domains such as e-commerce, medicine, and industrial monitoring. A lot of work has explored applying machine learning to these problems, but how can we understand and trust decisions made by such automated systems? We will explore how deep learning systems can be made 'inherently interpretable' - the process of transforming black box methods into approaches that can be understood without any post-hoc analysis. In this talk I will introduce a new Time Series Classification framework, that shows very strong performance in providing explainability and transparency through the use of Multiple Instance Learning. **Bio:** Jas Kandola is a Senior Applied Science Manager leading the Studios AI Lab within Prime Video Amazon Studios. Prior to Amazon he worked as a senior quantitative analyst for leading hedge funds and investment banks as well as in the healthcare domains. He holds a Ph.D. in machine learning from the University of Southampton, with postdoctoral work at MIT and Microsoft Research.

11:35-11:55 am		***Packaging Optimization at Amazon Scale: Tackling Scientific Challenges with Machine Learning and ML Operational Excellence*** Matthias Polozek, Principal Applied Scientist Amazon Robotics **Abstract:** Packaging plays a critical role in the customer experience, logistics, and Amazon's broader sustainability efforts but it also presents complex problems that require innovative solutions. The 'right' choice of packaging protects products from damage in addition to improving the sustainability of the delivery, e.g., by decreasing 'shipped air' and weight, and reducing packaging waste, in particular, plastics. In the first part of the talk, I will discuss the challenges of employing Machine Learning (ML) to make granular decisions at Amazon scale while striving for explainability and transparency. ML systems like this provide a multitude of business-critical functions to Amazon Stores, including content generation and placement, pricing, product discovery, to name a few. Their dependency on external data and the dynamic environment they are deployed in often lead to hidden performance degradation and failures. While mechanisms from traditional software OE are necessary, they do not suffice to assert correctness after deployment. Thus, we have developed solutions to enhance the operational excellence of data and ML that I will discuss in the second part of the talk, including recent wins and current challenges. **Bio:** Matthias Poloczek is a Principal Scientist in Stores Monetization at Amazon where we he works on large-scale automated decision-making using methods from ML, optimization, and causal inference. Recently, he has focused on content recommendation, advertising, and packaging eligibility. Moreover, he is passionate about AutoML and ML Operational Excellence. In 2022, he led a broad initiative to improve ML Excellence across Amazon. Previously, Matthias was a Senior Manager at Uber AI where he founded Uber's Bayesian optimization team and led a cross-org effort to build a company-wide service to tune ML models at scale.

11:55-12:15 pm		***REXEL: An end-to-end model for document-level relation extraction and entity linking*** Shubhi Tyagi, Senior Applied Scientist Amazon Artificial General Intelligence **Abstract:** Extracting structured information in the form of facts from unstructured text is one of the fundamental backbones for knowledge-based Q&A systems. LLM variants like RAG systems often rely on external structured knowledge sources like knowledge graphs for robust factual Q&A. In this work, we introduce REXEL a highly efficient and accurate model for extracting structured facts from web scale unstructured text for enrichening knowledge graphs. REXEL is 11 times faster than competitive existing approaches while being state of the art at the task for information extraction. The combination of speed and accuracy makes REXEL an accurate cost-efficient system for extracting structured information at web-scale. **Bio:** Shubhi is a Sr. Applied Scientist from Amazon working in their Artificial General Intelligence (AGI) group. She has 10 years of industry experience working as an AI applied researcher

12:15-1:15 pm	*Lunch Break*	
1:15-1:35 pm		***Generative AI for Ambient Sound: Fusing Text, Melody, and Context for Personalized Experiences*** Hassen Dhrif, Senior Applied Scientist, Amazon Devices **Abstract:** In this talk, we will showcase the latest advancements in generative AI for ambient sound creation. By leveraging deep learning models that fuse text, melody, and user context, we can now generate rich, contextual soundscapes that are seamlessly tailored to individual preferences. We will discuss our innovations in hybrid search, blending lexical and semantic approaches, to enable intuitive sound discovery and recommendation. Furthermore, we will demonstrate techniques to infer detailed sound characteristics based on user utterances, history, and situational context, paving the way for truly personalized and immersive audio experiences. Through these cutting-edge, multi-modal technologies, we aim to empower users to effortlessly craft their ideal sonic environments, transforming the way they interact with and experience ambient sound. **Bio:** Dr. Hassen Dhrif, Senior Applied Scientist, Amazon Hassen is a seasoned AI leader and Generative AI expert with over 20 years of experience in software engineering, data analytics, and machine learning. As a Senior AI Scientist at Amazon, Hassen leads a multidisciplinary team focused on advancing the state-of-the-art in multi-modal generative AI. Hassen has deep expertise in developing and deploying high-performing large language models, integrating multimodal data sources, and optimizing data pipelines for domain-specific applications. Previously, Hassen built cutting-edge AI applications for Alexa Skills, business trend analysis, and user experience improvement, using multimodal data (text, image, audio). Dr. Dhrif is passionate about leveraging the power of AI to solve real-world problems and create innovative solutions that enhance efficiency and cost savings across various industries. His pioneering work has been published in top-tier conferences and journals, and he has led a multimodal AI project for cardiology at Weill Cornell Medicine, using computer vision and AR. Hassen is a sought-after speaker, advisor and adjunct faculty member, sharing his cutting-edge expertise and research insights with academic and industry audiences.

1:35-1:55 pm		***Distributed and Adversarial Training of Large Machine Learning Models*** Michael Mühlebach, Research Group Leader of the Learning and Dynamical Systems Group at MPI-IS **Abstract:** Optimization has played an essential role in machine learning by providing a conceptual and practical framework on which algorithms, systems, and datasets were brought together at unprecedented scales. However, recent developments towards even larger models with billions of parameters pose new challenges for optimization and new opportunities for cross-fertilization between theory and practice. My talk will highlight this cross-fertilization on two examples related to distributed and adversarial training. I will highlight how foundation models such as RoBERTa can be vulnerable to adversarial label poisoning and introduce a robust training algorithm that solves a bilevel optimization problem to address this vulnerability. The second part of the talk introduces an event-based communication protocol for distributed training that has strong convergence guarantees (acceleration and robustness to non-i.i.d. data), can be applied to distributed optimization problems over graphs, and achieves a better communication vs. accuracy trade-off than federated learning algorithms (e.g., federated averaging or FedProx). **Bio:** Michael Muehlebach studied mechanical engineering at ETH Zurich and specialized in robotics, systems, and control during his Master's degree. He received the B.Sc. and the M.Sc. in 2010 and 2013, respectively, before joining the Institute for Dynamic Systems and Control for his Ph.D. He graduated under the supervision of Prof. R. D'Andrea in 2018 and joined the group of Prof. Michael I. Jordan at the University of California, Berkeley as a postdoctoral researcher. In 2021 he started as an independent group leader at the Max Planck Institute for Intelligent Systems in Tuebingen, where he leads the group "learning and dynamical systems". He is interested in a variety of subjects, including machine learning, dynamical systems, and optimization. During his Ph.D. he developed approximations to the constrained linear quadratic regulator problem, a central problem in control theory, and applied these to model predictive control. He also designed control and estimation algorithms for balancing robots and flying machines. His more recent work straddles the boundary between machine learning and optimization, and includes the analysis of momentum-based and constrained optimization algorithms from a dynamical systems point of view. He received the Outstanding D-MAVT Bachelor Award for his Bachelor's degree and the Willi-Studer prize for the best Master's degree. His Ph.D. thesis was awarded with the ETH Medal and the HILTI prize for innovative research. He was also awarded a Branco Weiss Fellowship, an Emmy Noether Fellowship, and an Amazon Fellowship, which fund his research group.

1:55-2:15 pm		***Theoretical Foundations of Deep Selective State-Space Models*** Antonio Orvieto, Principal Investigator at the ELLIS Institute Tübingen and independent group leader at MPI-IS **Abstract:** Structured state-space models (SSMs) are gaining popularity as effective foundational architectures for sequential data, demonstrating outstanding performance across a diverse set of domains alongside desirable scalability properties. Recent developments show that if the linear recurrence powering SSMs allows for a selectivity mechanism leveraging multiplicative interactions between inputs and hidden states (e.g., Mamba (2) and Hawk/Griffin), then the resulting architecture can surpass attention-powered foundation models trained on text in both accuracy and efficiency, at scales of billion parameters. In this paper, we give theoretical grounding to the selectivity mechanism, often linked to in-context learning, using tools from Rough Path Theory. We provide a framework for the theoretical analysis of generalized selective SSMs, fully characterizing their expressive power and identifying the gating mechanism as the crucial architectural choice. Our analysis provides a closed-form description of the expressive powers of modern SSMs, such as Mamba, quantifying theoretically the drastic improvement in performance from the previous generation of models, such as S4. Our theory not only motivates the success of modern selective state-space models, but also provides a solid framework to understand the expressive power of future SSM variants. In particular, it suggests cross-channel interactions could play a vital role in future improvements. **Bio:** Antonio Orvieto is an independent group leader at the Max Planck Institute for Intelligent Systems and a principal investigator at the ELLIS Institute Tübingen, Germany. He holds a Ph.D. from ETH Zürich and spent time at Google Deepmind, Meta, MILA, INRIA Paris, and HILTI. His main area of expertise is optimization for Deep Learning. He has published in NeurIPS, ICML, ICLR, AISTATS, and CVPR and organized the ``Optimization for Data Science and Machine Learning'' session at the International Conference on Continuous Optimization (ICCOPT) in 2022, as well as the ICML 2024 Workshop on Next Generation of Sequence Modeling Architectures. In his research, he strives to improve the efficiency of deep learning technologies by pioneering new architectures and training techniques grounded in theoretical knowledge. Central to his studies is exploring innovative strategies for decoding patterns in sequential data, with implications in natural language processing, biology, neuroscience, and music generation. His LRU architecture is the basis for some of Google's Gemma language model variants.

2:15-2:35 pm	*Break*	
2:35-2:55 pm		***When do adversarial attacks against language models matter?*** Jonas Geiping, Principal Investigator at the ELLIS Institute Tübingen and independent research group leader at MPI-IS **Abstract:** Adversarial attacks can be optimized to attack large language model applications, such as conversational chatbots. This is often used to "jailbreak" these models, that is to circumvent the post-training modifications made to the model to increase the safety of its answers. However, this is practically not a threat with current-generation language models. Yet, as soon as these models are used for any task that goes beyond chatting and simulating text, practical security problems arise. **Bio:** Jonas Geiping leads a joint research group at the Max Planck Institute for Intelligent Systems and the ELLIS Institute Tübingen. Before this, he studied Mathematics at the University of Münster, received a PhD degree from the University of Siegen, and spent time at the University of Maryland as a postdoc. His group is interested in questions of safety and efficiency in modern machine learning. This topic covers questions of safety such as robustness against adversarial attacks, but also efficiency, for example regarding the training and serving of large language models.

2:55-3:15 pm		***Improving Decision Making with Machine Learning, Provbly*** Manuel Gomez-Rodriguez, Faculty at the MPI for Software Systems **Abstract:** Decision support systems for classification tasks are predominantly designed to predict the value of the ground truth labels. However, these systems also need to make human experts understand when and how to use these predictions to update their own predictions. Unfortunately, this has been proven challenging. In this talk, I will introduce an alternative type of decision support systems that circumvent this challenge by design. Rather than providing a single label prediction, these systems provide a set of label prediction values, namely a prediction set, and ask experts to predict a label value from the prediction set. Moreover, I will discuss how to use conformal prediction, online learning and counterfactual inference to efficiently construct prediction sets that optimize experts' performance, provably. Further, I will present the results of a large-scale human subject study, which show that, for decision support systems based on prediction sets, limiting experts' level of agency leads to greater performance than allowing experts to always exercise their own agency. **Bio:** Manuel Gomez Rodriguez is a tenured faculty at the Max Planck Institute for Software Systems (MPI-SWS). His research interests lie in the development of human-centric machine learning models and algorithms. He has co-authored over 50 publications in top-tier conferences (NeurIPS, ICML, UAI, AISTATS, AAAI, KDD, WWW) and journals (PNAS, Nature Communications, JMLR, TMLR). Manuel is an ELLIS Fellow and has received several recognitions for his research, including an ERC Starting Grant, an outstanding Paper Award at NeurIPS and a Best Research Paper Honorable Mention at KDD and WWW. Manuel holds a M.S. and a Ph.D. in Electrical Engineering from Stanford University (2009 and 2013) and a B.S. in Electrical Engineering from University Carlos III in Madrid, Spain (2006).

3:15-3:35 pm		***Generative AI for Research: Global Examples how researchers leverage cloud services to accelerate their work*** Roberta Piscitelli, Academic Research Lead EMEA **Abstract:** Generative AI and High-Performance Computing are changing the way research is conducted. We will showcase examples from the global research community how cloud services are being used to accelerate research. We will also provide information how researchers can easily access cloud resources in Germany. **Bio:** Roberta Piscitelli is Academic Research Lead EMEA at Amazon. Roberta worked 18+ years in technology and research, Ph.D. in Computer Science from University of Amsterdam.  Andreas Wagner, Sr. Account Manager for Research in Germany **Bio**: At Amazon he is the Senior Account Manager for Research in Germany. He worked 7+ years at Amazon and Amazon Web Services with a focus on public customers, including Research and Nonprofit.

3:35-3:50 pm		***Inference optimization for foundational models on AI accelerators*** Kailash Budhathoki, Sr Applied Scientist, AWS Deep Engine-Science **Abstract:** Larger foundation models tend to do better across a wide range of standard benchmarks. But their inference is expensive and slow. In this talk, I will briefly cover various state-of-the-art optimization strategies for cost-effective and fast inference **Bio:** Kailash currently is leading a science team on optimizing foundation models for inference at AWS AI DS3 (Deep Science for Systems and Services) org. He joined Amazon Research Lablet Tübingen (part of AWS AI) in 2020, where he developed algorithms / tools to help businesses explain complex cause-effect relationships underlying their business problems, and led cross-org effort within Amazon to launch them in production. Briefly he also led the cross-org effort within Amazon to deliver bias mitigation solutions for Amazon's in-house vision foundation models, towards the re:Invent 2023 release. He has a PhD in Computer Science from the Max Planck Institute for Informatics.