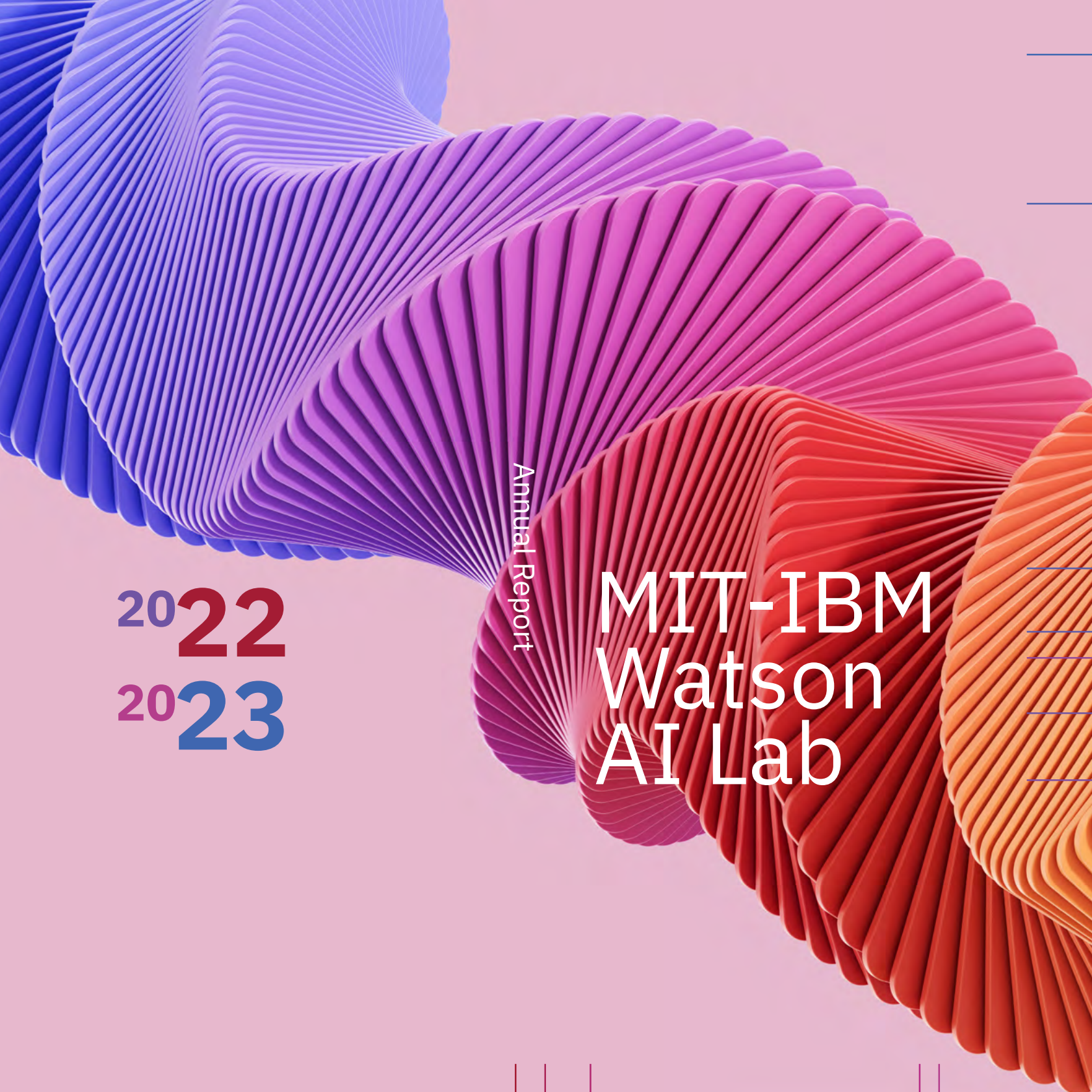


MIT-IBM Watson AI Lab

Annual Report

2022
2023

Table of Contents

Message from the Directors

In the last year, artificial intelligence entered the zeitgeist, transforming technology, culture, and economies overnight. While projections of its valuation vary, some estimate that generative AI alone will add trillions to the global economy annually, let alone with the addition of other forms of AI, machine learning, and deep learning.

Not only was the MIT-IBM Watson AI Lab prepared for this accelerated transition, but we had already established a full-stack environment for students, researchers, and member companies to thrive and adapt to their changing needs. The Lab's research portfolio regularly ranges from finance, efficient computing, and hardware, to health care, robotics, multimodal processing, and natural language processing. This year, the Lab focused more of our resources on foundation models and generative AI to much success.

Our commitment to fostering the next generation of AI researchers remains unwavering. This year, 12 MIT undergraduate, Masters of Engineering, and PhD students from the MIT Electrical Engineering & Computer Science participated, through the 6A Program, in our research projects on novel, application-driven solutions; numerous other MIT graduate students and interns contributed ideas and their creativity in the Lab. The fresh perspectives and innovative ideas from these young researchers are invaluable to our Lab's mission. Furthermore, the mentorship and engagement opportunities we offer to students have been instrumental in ensuring that the future of AI is bright.

In the coming years, generative AI and foundation models, and their impact on many other fields, will continue to be at the forefront of our research endeavors. With the active participation of our member companies, we've embarked on projects that are not only academically intriguing but also have the potential to revolutionize industries.

Finally, we would like to extend our thanks to each member company for your continued support and investment in our mission and shared vision. Together, we are not just witnessing the evolution of AI; we are actively shaping it for the better.

David Cox and Aude Oliva

IBM Research MIT

Lab by the Numbers

Over
1,000
peer-reviewed papers

Over
800
news and media mentions

Over
430
project proposals submitted from MIT and IBM

$240M
10-year investment to found a joint lab

169
researchers across MIT and IBM

Over
100
active projects

54
patent disclosures

Over
80
member co-funded projects

6
current member companies

Selected Projects

The MIT-IBM Watson AI Lab's research portfolio draws on academic rigor and industry ingenuity to bring results and solutions in the Lab to bear in transformative applications across various time scales. Established to redefine AI's frontiers, the Lab dives deep into fundamental AI challenges that can revolutionize industries and improve the human experience. While our research portfolio is vast and flexible, it consistently emphasizes core and ambitious research, as well as the convergence of AI with other disciplines, including quantum computing, hardware, neuro-symbolic reasoning, and algorithmic fairness. In the last year, the Lab pivoted to feature generative AI and foundation models more prominently in our strategy, underscoring their enterprise potential and their significant role in our longer outlook. Further, our research philosophy and approach allow our member companies to collaborate with us to develop novel technologies, prototypes, and metrics for near- and long-term impacts on business and/or society.

Here, we showcase the myriad ways the Lab thinks about key research and technology questions and the novel methods we're developing to tackle them.

Sherrie Wang, Chris Hill (MIT)
Yada Zhu, Campbell Watson,
Hilde Kuehne, Kommy Weldemariam (IBM)

AI to improve local weather forecasting

Foundation model concepts that underpin popular generative AI tools have utility beyond natural language processing. Researchers from the Lab are looking at how foundation model approaches can improve short-term, targeted weather forecasts for specific locations, when trained on multimodal data taken from satellite imagery, numerical forecast models, and observations. They are investigating how a foundation model from IBM can be fine-tuned to leverage and adapt reinforcement learning ideas that have been deployed in natural language processing. The system will be evaluated against forecasts for selected sites relevant to improving renewable energy operations.

> "Foundation models can be applied to scientific computing challenges, like targeted weather forecasting. Our MIT-IBM Watson AI Lab team is researching how foundation models can be fine-tuned to make targeted short-term weather forecasts for specific locations. The overall approaches and concepts developed could have broad applicability to the environmental forecasting and planning needs of sectors including agriculture, transportation, energy, and more."

Multimodal foundation models: pushing the frontiers of trustworthy, compositional, and generative AI

Aude Oliva (MIT)
Rogerio Feris,
Leonid Karlinsky (IBM)

Pre-training large language and vision models on massive-scale datasets has led to major advances in AI, albeit accompanied with shortcomings related to privacy, usage rights, data protection, bias, and ethical issues. In addition, existing models struggle with compositional reasoning and understanding concepts beyond object nouns, such as attributes and relationships. As part of the Moments in Time project, Lab researchers have addressed these challenges by proposing novel methods for pre-training video models solely based on synthetic data, closing the gap between simulation and real data performance. They have also leveraged synthetic data to enhance the compositional and concept understanding capabilities of multimodal foundation models, achieving state-of-the-art results in standard datasets. Finally, they developed a system for automatically producing commentary for sports videos using generative AI. Their work turns multimodal data into fluent commentary using a watsonx large language model trained with trillions of tokens at IBM. This system was showcased at the Wimbledon 2023 tennis tournament.

> “We are pushing the frontiers of multimodal foundation models, making them more trustworthy, enhancing their compositional abilities, and exploring innovative applications, such as automatically producing dynamic and engaging sports commentary.”

Bringing deep learning to “internet of things” devices

Song Han (MIT)
Chuang Gan, John Cohn (IBM)

The branch of AI that curates your social media feed and serves up search results could soon check your vitals or set your thermostat. To enable this, Lab researchers are working to bring deep neural networks to the tiny computer chips in wearable medical devices, household appliances, and the billions of other gadgets that make up the “internet of things” (IoT). One system from this team, called MCUNet, designs compact neural networks that allows AI applications to run smoothly on IoT devices despite their limited memory and processing power. The technology could facilitate the expansion of the IoT universe, while saving energy and improving data security.

> “Today’s AI is too big. We need model compression and acceleration techniques that bridge the gap between the supply and demand for AI computing, optimizing performance, reducing resource usage, and democratizing access.”

Yoon Kim (MIT)

Yang Zhang, Kaizhi Qian (IBM)

Efficient and aligned language models

Language evolves quickly, so current large language models need to keep up. This means researchers need to able to adjust them over time, so that the models can reflect new language preferences and uses, in a way that's helpful and not harmful. Additionally, the methods need to be computationally efficient, since training can be expensive. Addressing this need, Lab researchers are applying several methods that help models to learn more like humans do, including multimodal self-supervised algorithms, regularization to improve learning new information, off-policy and direct preference optimization to pick up information from broader, external experiences or explicit feedback, and improved architectures.

> Our project aims to develop methods for **practical, efficient, and safe deployment** of systems based on language models, in order to broaden access and maximize their beneficial impact.

Learning to conduct a cyberattack like a human

In order to protect connected networks from attackers, organizations might employ white hat hackers to test a system's security and identify vulnerabilities — attempting to gain system access, gather intelligence, steal data, and disrupt service, while avoiding defensive measures. In this work, Lab researchers are building a machine-learning agent that can perform opportunistic, adversarial threat behaviors and "think" like a human attacker. Employing both crystallized and fluid intelligence, the agent seeks to exploit the network to understand what it is seeing in an unknown situation; learn from its current environment and remember this for future actions; draw on prior knowledge about cybersecurity; and plan and contingency plan future intentions, goals, and behaviors, while receiving feedback from the simulated network and evolving defensive actions that are being taken during the attack.

Una-May O'Reilly (MIT)

Masataro Asai (IBM)

> Fluid intelligence powers cyber adversaries as they reason, learn, and adapt to unforeseen network configurations. **Neuro-symbolic algorithms** support agent-based simulations of cyber plans and threat scenarios.

Representation learning as a tool for causal discovery

Caroline Uhler, Devavrat Shah (MIT)
Kristjan Greenewald, Akash Srivastava,
Karthikeyan Shanmugam (IBM)

As humans, we are adept at examining a raw object or phenomenon, and intrinsically understanding some of its characteristics and features, further, how that object might behave in or influence another scenario or environment. Massive data collection holds the promise of a better understanding of complex phenomena and ultimately, of better decisions. An exciting opportunity in this regard stems from the growing availability of interventional data (in medicine, advertisement, education, etc.). However, these datasets are still miniscule compared to the action spaces of interest in these applications (e.g., interventions can take on continuous values like the dose of a drug or can be combinatorial as in combinatorial drug therapies). Here, representation learning has been shown to excel at predicting how an object may behave, but it often fails at inferring causal effects.

Therefore, Lab researchers are developing a framework for using representation learning as a tool for causal discovery to deal with such exploding action spaces. The framework will allow these neural networks to discover and map cause and effect, and apply that to new scenarios, to help predict and optimize unseen intervention outcomes, such as medical treatments, policies, and advertising.

> “Large-scale representation learning techniques are the main workhorses behind many of the recent advancements in deep learning, including some of the foundation models. However, most of these techniques learn correlations, not causal relationships. This may limit their ability to make reliable predictions when conditions change, like predicting the effects of new medical treatments. The method we propose enables deep learning models to uncover causal structures hidden in the data.”

A better way to compare and contrast: AI taking shortcuts

Dina Katabi, Piotr Indyk (MIT)
Rogerio Feris (IBM)

> “We are developing novel methods for self-supervised learning across domains, as well as targeted, supervised contrastive learning, aimed at improving the performance of long-tail recognition tasks.”

There are many ways to group items in a set by similarities and differences. The task becomes much more difficult when many items are rare and we lack sufficient data about them. Like humans, AI may take shortcuts or disregard some information to boost categorization. Here, Lab researchers aim to improve this contrastive learning on images and other modalities by developing a new framework that assesses and addresses these issues. Further, they seek to enhance the AI’s generalizability in order to apply it to areas, such as data modalities that are not interpretable by humans like radio signals or sensors, or ones that require expert knowledge like medical data, satellite images, or limited datasets.

Toward building smaller, efficient AI

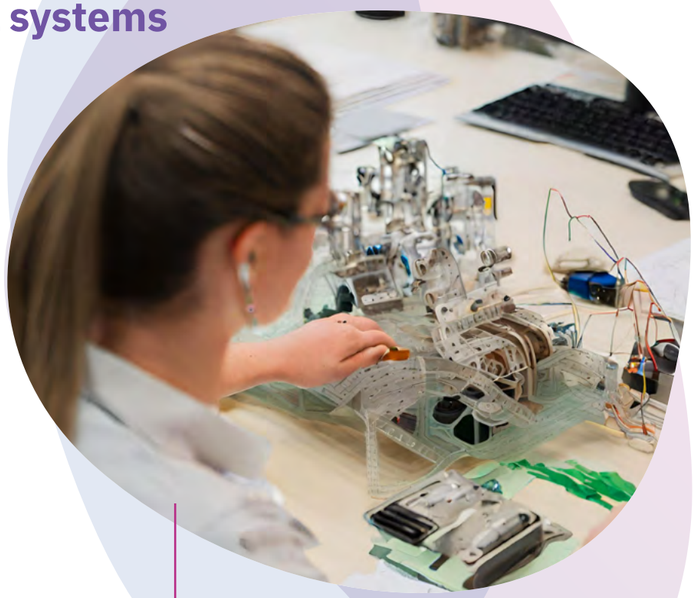
Jonathan Ragan-Kelley (MIT)
Rameswar Panda (IBM)

In order to perform a particular task and “work smarter, not harder,” researchers need to use the right tool; the same goes for AI models. This work searches for programs that can solve a task and trains them by optimizing their parameters, in order to perform the task better. In this project, Lab researchers are achieving this through multi-objective optimization of models that have trade-offs between task performance/accuracy and computational cost, particularly by designing specific compilers and automatic GPU scheduling algorithms. In this way, the programs will generate a set of efficient solutions for the models and pair that with the appropriate hardware to make deep learning more cost-effective and memory efficient to run faster and on smaller devices.

> “Our goal is to bridge the gap between the powerful but expensive deep-learning models and efficient classical algorithms by automatically combining machine learning components with domain-specific modules. The results of this could be potentially applicable to areas of robotics and autonomous navigation.”

Faez Ahmed (MIT)
Akash Srivastava, Dan Gutfreund (IBM)

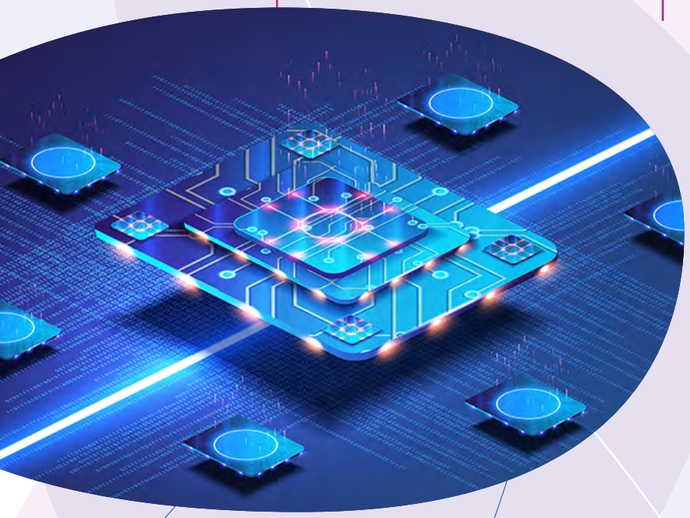
Designing an improved innovation process for mechanical systems



Designing is a trial-and-error process that requires creativity; understanding of real-world physical constraints; expert knowledge; and data that’s highly dimensional, structured, and heterogeneous. AI can help optimize and automate this process. To do this, Lab researchers are combining deep generative models that are good at mimicking training data, like creating deep fakes, design theory, and symbolic learning that can learn relationships between objects and concepts to build better engineering designs. The team is using their framework for exploring inverse design algorithms for constrained engineering applications, like human-mobility and mechanical power-transmission systems.

> “The current generation of generative AI and foundation models have democratized content creation. Anyone with access to the internet can now create blogs, write copies, design websites, and program small applications without requiring a lot of specialized training. However, the application of such models is very limited in the field of engineering and sciences. One of the main reasons for this is the lack of precision in today’s probabilistic generative AI that is required in these fields. Our project is tackling this problem of a lack of precision in today’s generative AI. Our objective is to create models like ChatGPT for engineering and sciences with the goal of democratizing these fields.”

> Embarking on new challenges, from chip design to edge machine learning, our innovative team leads with secure in-memory computing, enabling extended battery life, faster speeds, enhanced privacy, and forward-looking technology migration.



Anantha Chandrakasan (MIT)
John Cohn, Xin Zhang (IBM)

Machine learning to optimize tiny AI chip design

Analog/Mixed-signal integrated circuits used in AI applications currently exist in the 10nm realm, but new applications and technologies require shrinking the hardware down to 2nm scales, which comes with numerous challenges and the need for an efficient and predictable process. Lab researchers are endeavoring to do this — they're using AI to design a way to prototype circuits and to predict the performance and scalability. The team is exploring machine-learning models (reinforcement learning and active learning) for migrating circuits to smaller scales. The resultant circuit designs can be fabricated and tested within neural network frameworks to validate the technology migration model's prediction.

A memory-first perspective for deep learning

In order to understand relationships and interactions between data points, like in a social network, it's helpful to be able to illustrate them with a graph establishing connections. However, these exceedingly large datasets currently need to be stored in a computer cluster or an external high-capacity machine, making computation inefficient. Lab researchers are addressing this issue to streamline graph-based deep learning models — graph neural networks (GNNs) — by developing new GPU-based hardware and software that optimizes parallel processing and bandwidth usage, while minimizing redundancies and latency. It can be dropped into current AI infrastructures, improving GNN applications to areas like drug discovery, financial security, and social media recommendations.

Arvind (MIT)
Jie Chen (IBM)

> From power grids to transaction networks, graph-based methods can be used to learn their relational structures and detect anomalies. Our objective is to train and deploy models with massive graphs on affordable hardware without compromising accuracy.

People

Aude Oliva

MIT Director, MIT-IBM Watson AI Lab

Director of Strategic Industry Engagement, MIT Stephen A. Schwarzman College of Computing

Anantha Chandrakasan

MIT Chair, MIT-IBM Watson AI Lab

Dean, MIT School of Engineering

Vannevar Bush Professor of Electrical Engineering and Computer Science

Dario Gil

IBM Chair, MIT-IBM Watson AI Lab

Senior Vice President and Director of Research, IBM

Daniel Huttenlocher

MIT Co-Chair, MIT-IBM Watson AI Lab

Dean, MIT Stephen A. Schwarzman College of Computing

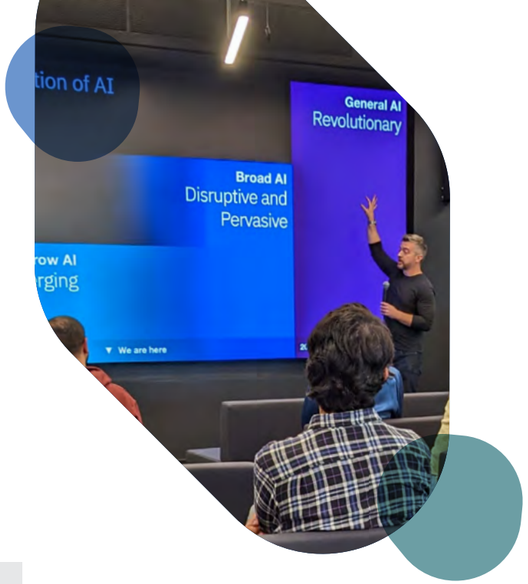
Henry Ellis Warren (1894) Professor of Electrical Engineering and Computer Science

David Cox

IBM Director, MIT-IBM Watson AI Lab

Vice President for AI Models, IBM

The MIT-IBM Watson AI Lab community has been a vibrant one with strong engagement, particularly during events. Throughout the year, researchers, students, and member companies interfaced across seminars, client and research meetings, mentorship opportunities, and networking events. Additionally, the Lab helped to sponsor several activities for students, postdoctoral researchers, and Lab researchers in service to the Lab's mission to improve the field of AI and those who are a part of it. The following provides a high-level view of the Lab's ongoing work.



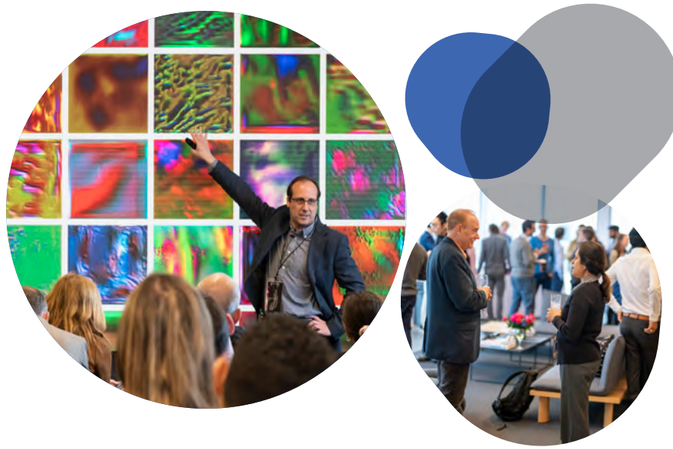
Principal Investigator Networking

As AI evolves, the Lab continues to innovate and adapt to the needs of our senior researchers and members. Accordingly, the Lab invited the MIT and IBM community to network, discuss this year's proposal process, and the marked shift to emphasize collaborative projects in foundation models, generative AI, and applied machine learning.

Fall Student Outreach

Few things excite us more than seeing up-and-coming AI researchers thrive while tackling tough and important problems in the Lab. To help ensure the continuation of this mentorship pipeline, the Lab regularly reaches out and opens its doors to undergraduates, graduate students, and postdocs.

Last October, nearly 50 young researchers joined us to learn from Lab co-directors Aude Oliva and David Cox about research projects, how to become involved with the Lab, and insights into our goals and focuses. The floor then opened up for poster demonstrations and networking with the Lab's investigators.



Ask the experts

In just weeks, AI applications, projects, and research in machine- and deep-learning within and external to the Lab jumped by leaps and bounds. To ensure that our industry members understand how we're thinking about solutions and utilizing the latest available work, the Lab invited industry members to discuss and hear from four Lab experts at the forefront of their respective fields, in areas across large language models and natural language processing, health care, neural scene representation, and computer vision. Here, corporate members were encouraged to leverage the opportunity to expand their knowledge, and suggest topics for future seminars.

MIT-IBM Watson AI Lab Industry Showcase

Gathering in the Lab this past spring, industry members and researchers explored the transformative journey and potential of AI in the modern era and embraced the advantages that an academia-industry partnership, like the Lab's, can offer. Throughout the day, members met to discuss current projects and leadership expounded on the opportunities now available that AI has penetrated diverse software realms, from predictive analytics to image generation. At the height of the day, the Lab showcased its latest thrusts into generative and foundation models. Further discussions revolved around the necessity for trustworthy and explainable AI, given the pace of its integration into various industries. Highlighted were creative and novel ways in which the Lab is building data-efficient AI systems to address challenges faced by our members.

Students, who pass through the doors of the MIT-IBM Watson AI Lab, seek to create and better technology and society in the same measure. With direction from MIT and IBM mentors and with industry-scale computing resources, undergraduates up through PhD candidates develop skills to scope projects, as well as research, experiment, and analyze AI technologies and applications. Self-driven, they tackle novel and real challenges together with machine-learning innovators — some for a summer internship, others for a year or more.

Graduate students in the MIT 6A Program pursue their research passions to advance computing in the Lab, with the added experience of collaborating on industry-focused projects. Teams composed of our researchers and students work together to develop solutions for near- and longer-term problems. Their innovations build the foundation for their thesis work and real-world impacts.

When their time at the Lab comes to a close, we can see the mark our students have made, from novel AI models and improved efficiencies, to prototypes and new use cases. Future leaders and 6A MEng students Irene Terpstra SB ’22 and Rujul Gandhi SB ’22, and 6A PhD candidates Andi Peng SM ’23 and Aniruddha Nrusimha, are still on their journey in the Lab, but their influences can already be seen, from original planning to execution.

> “Graph transformers have the potential to become foundation models but are currently limited by their quadratic computational complexity. As a 6A intern, I am working on benchmarking graph transformers toward scalability for large graphs.”
> **– Katherine Lim, 6A Program participant**

Gandhi and Peng both appreciate the power that natural language can bring to machine learning, particularly when it comes to the realm of robotics. Fascinated with linguistics, Gandhi is parsing language into a domain that robots can understand. Leveraging the structure of language, she’s converting English instructions into a logical representation, from which an action plan can be algorithmically derived for robots and understood by the user. Working with the Lab and advisors Yang Zhang of IBM and MIT Assistant Professor Chuchu Fan, Gandhi is designing the representation to be general, so that it works for a variety of instructions, but those especially targeted for complex and collaborative tasks that robots might take on in industries like aviation, automotive, and warehouse and home settings. Gandhi and her team are achieving this using a pre-trained encoder-decoder model for text-to-text formatting. They’re fine-tuning and augmenting it with natural language instructions from WikiHow for how to perform different tasks. The methodology allows for the identification of objects, tasks, and subtasks, as well as logical operators, so a robot could understand how to, for example, pick up an apple and put it on the table until a light turns red. An advantage, Gandhi says, to using a dataset like WikiHow is that it has information about our physical world and dependencies baked into the natural language instructions.

In the Lab, Gandhi also focuses on developing speech models for languages that lack a written form or low resource languages. Her research group infers words and word boundaries from streams of unlabeled speech and derives a pseudo-language to train a language model. This requires examining context and deciphering which sounds occur frequently together.

Passionate about safe, ethical, and equitable AI, Peng leverages language as well as simulated environments to build machine-learning agents and systems that can learn from abstract human knowledge and human-robot interactions to assist people. Her advisory team, with Chuang Gan of IBM and MIT’s H.N. Slater Professor in Aeronautics and Astronautics Julie Shah, considers people who might have constraints on what they are able to do, particularly around the home — some might not be able to reach a high shelf, others might have a mobility or dexterity limitation. Robots can supplement these actions, Peng says, providing support and autonomy to the user. To do this, Peng infers different disabilities


Irene Terpstra


Andi Peng


Aniruddha Nrusimha


Rujul Gandhi

> “I enjoyed working with the folks at IBM in collaboration with the Probabilistic Computing group at MIT. I was able to learn from excellent engineers, work on interesting problems, and understand what it is like to work on a well orchestrated team.”
> **– Ian Limarta, MEng student**

people might have. She notes that large language models capture semantic priors well, when it comes to human abilities and scoping the problem. Her group then uses that information to construct a helper agent — within the Lab's simulated world, called ThreeDWorld — that can understand motivation and deduce how best to assist the individual in need, also represented by an agent. Here, the helper agent learns bidirectional communication, sequential decision-making, and how to perform actions in a "human-like way," that humans can understand. The techniques include diagnosing the issue, collecting feedback, and adapting to needs. During this work, Peng has experienced how current solutions fall short and enjoys exploring opportunities for innovation and developing resolutions.

Machine learning and natural language processing also play a significant role in Irene Terpstra's research. As an undergraduate at MIT, Terpstra designed ocean sensors and autonomous vehicles while studying robotics and machine learning, leading her down the path to designing with AI. Working toward her MEng, she creates ways to customize integrated circuits. Terpstra's Lab team, including advisors MIT School of Engineering Dean Anantha Chandrakasan and IBM's Xin Zhang, is generating a systemic workflow to use AI models to better design computer chips. To build better chips, she employs a pre-trained large language model that's geared toward the circuit descriptor language NGspice, which Terpstra familiarized herself with. The language model can then be provided with text prompts and queries for how to modify the physical computer chip and, leveraging prior information about circuit design, the model can return suggested guidance for updating the construction, in terms of parameter adjustments. These are then passed into a reinforcement learning model, which contains a simulator that outputs the new design and parameters of the design. The process can be iterated on until an ideal configuration is achieved — it's these practical and tangible results that Terpstra enjoys seeing.

"Through MIT-IBM collaborations students have the opportunity to perform large-scale experiments and also see their research translate to impactful, real-world applications."
– Yoon Kim, MIT assistant professor

"Without a doubt, collaborating with students is the best part of my job. This past year, I have had the opportunity to work with a lot of very talented student researchers across 10+ topics via the MIT-IBM Watson AI Lab and other grants. These collaborations have not only led to quite a few impactful papers, but also techniques that are being incorporated into IBM production offerings."
– Akash Srivastava, Lab research manager

"In the MIT-IBM Watson AI Lab, I embarked on two pivotal projects: the development of cutting-edge toolkits for graph machine learning and a distinct venture leveraging GPT's linguistic strengths for effective graph learning. Both have been instrumental in reshaping my AI research trajectory and building powerful real-world applications."
– Eva Yi, 6A undergraduate student

"Using machine learning in practice often involves making choices: what model to use, how to measure success, etc. In our project, students have developed methods to help interrogate the sensitivity of major conclusions drawn from a machine-learning analysis to real-life choices."
– Tamara Broderick, MIT associate professor

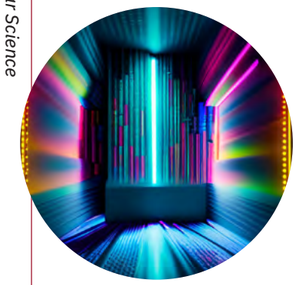
The Lab cares deeply about optimizing processes and AI methods, and Aniruddha Nrusimha's research and focus are no exception. He has been working to make machine-learning models run faster and more efficiently, now with a concentration on large language models. Nrusimha wants to take radical steps at the system-level for large gains. He says that through training, large language models learn data outliers to improve their performance, but this makes them less stable and compressible. Nrusimha and his advisors, MIT Assistant Professor Yoon Kim and Rameswar Panda of IBM, want to incentivize models to be more compressible using activation quantization. Language models, using transformers, are primarily a matrix of large numbers of multiplications: activations and their weights. He is training and experimenting on 1 billion parameter models with the goal of using 8 or 4 bit weights for integers rather than the current models which use 16 bits; state-of-the-art activations are currently 8 bits. Nrusimha is trying to get each weight down to 4 bits by examining the intrinsic properties of language models. If his team is successful, these models could run on more devices and cutting-edge deep learning innovations could be accessible to researchers and the public, enabling numerous applications.

In the Media

The MIT-IBM Watson AI Lab and its researchers have been cited and featured across numerous media outlets, including *Forbes, TechCrunch, Bloomberg Law, The Washington Post, WIRED,* and *The Wall Street Journal.*

Popular Science

This AI can harness sound to reveal the structure of unseen spaces



Researchers from the MIT-IBM Watson AI Lab are modeling spatial acoustics, with a focus on reverberations. For this, they developed a neural network model, called a neural acoustic field (NAF), which simulates how sound changes as a listener moves through a space and considers the sound's source, listener's position, and room geometry. It can even deduce room structures enabling potential applications for virtual reality to enhance sound perception and augment sensors and devices underwater or in low light.

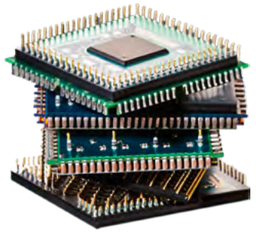
Yonhap News

The MIT-IBM Watson AI Lab and co-director David Cox hosted the Korean media, providing a look under the hood into the how the Lab operates and insights into the evolution of technologies, from narrow AI through broad AI to general AI. Human-like fluid AI and foundation models will help to get us there, but they need to be fair and ethical, which is necessary for their successful deployment.

ChatGPT blast, artificial intelligence in the palm of your hand… 'human-like AI' in 30 years

The Register

MIT boffins cram ML training into microcontroller memory



As the need for running machine learning on increasingly smaller devices grows, so does the demand for computational efficiency. Research out of the MIT-IBM Watson AI Lab has produced a method that empowers a computer vision model to run on tiny microcontrollers, which are internet-of-things (IoT) devices. The technique allows for on-device AI training using less than a quarter megabyte of memory, significantly less than other techniques. This approach, which emphasizes efficient algorithms, enables faster, privacy-preserving training.

Quanta Magazine

Research from the field of machine learning informs how neuroscientists understand the human brain and its ability to store and recall memories on spatial data. Work from Dmitry Krotov from the MIT-IBM Watson AI Lab and his colleagues noticed how their new Hopfield networks can hold and retrieve memories effectively due to an attention mechanism, like transformers can. This led to the finding that transformers mirror patterns of behavior observed in the grid cells of the hippocampus, important for location mapping.

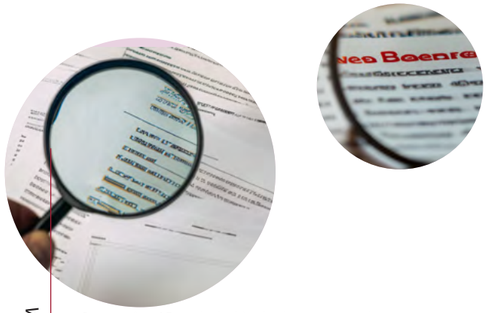
How transformers seem to mimic parts of the brain

The Next Web

Synthetic data is the safe, low-cost alternative to real data that we need

Synthetically generated data provides a solution to the challenges of real-world data collection. Developed by MIT-IBM Watson AI Lab researchers, the Task2Sim model alleviates some of these issues by creating synthetic images for multi-task training, making data collection efficient and cost-effective. While synthetic data offers advantages like privacy and reduced bias, its combination with real data has shown promising outcomes, suggesting a future where synthetic data plays a pivotal role in AI training.

Vanity Fair France



Generative AI exploded in the last year, and a notable area of influence was writing. New applications emerged for creating text for students' homework, marketing, etc. Accordingly, a need arose to be able to determine what content was original and manmade, and what came from an algorithm. One tool developed by researchers from the MIT-IBM Watson AI Lab, called the giant language model test (GLTR), examines text and, based on the randomness of the word choice, determines how it was created.

How to detect texts created by artificial intelligence

Politico

5 questions for MIT's Neil Thompson

MIT-IBM Watson AI Lab researcher Neil Thompson investigates how new technologies influence and are implemented in industry, as well as challenges faced in their application. Thompson shared how the capacity of computation has evolved, ways the U.S. could maintain its computational prowess, and how large language models have been used since their widespread debut last year.

Forbes

The uses for foundation models and generative AI extend beyond text creation to many other practical applications and breakthroughs, producing new scientific tools and improving processes to benefit industries across the board. Research from IBM and MIT-IBM Watson AI Lab researcher Payel Das demonstrates how they've been able to leverage synthetic data and transformers and develop methods and generative systems that are trustworthy and fair.

IBM demonstrates groundbreaking artificial intelligence research using foundational models and generative AI

Case Studies

The power in analyzing data is not about revealing what happened in the past but enabling the anticipation and, even, determination of what happens in the future. For businesses, uncovering insights to improve current services and products is essential; now, with advances in AI, corporations are unlocking novel use cases and averting risks better.

As a key player in this space, MIT-IBM Watson AI Lab understands that CEOs and decision-makers globally are looking to implement AI to gain an edge and appreciates how guidance from experts like those in the Lab can enable strategic applications.

Optimizing dynamic resource allocation in health care

For companies, supply chain forecasts are a necessity, but often inaccurate, resulting in unnecessary costs and downstream impacts. Products may be over- or under-stocked or people and resources suboptimally distributed. This may be due to limited time-series data or undiscovered influences within the supply chain. Additionally, many of today's prediction methods do not assess multiple variables in tandem, limiting what leaders can see as the "complete picture". Machine learning can assist in many ways, like identifying customer request patterns, optimizing dynamic assets and prioritizing their application, and integrating corporate objectives into decision-making. To improve an industry member's pipeline forecasting and hedge against risk, the Lab developed a custom, fine-grained, transformer-based prediction approach that combines machine learning and inventory optimization in a single step to evaluate multiple drivers of their supply chains for various products and services. The approach emphasized interpretability and resilience to uncertainty or noisy data.

Key Metrics:

- 5-10 percentage points more accurate near-term forecasts for pilot families
- $20M+ predicted annual savings due to corresponding safety stock reductions for a fully scaled solution

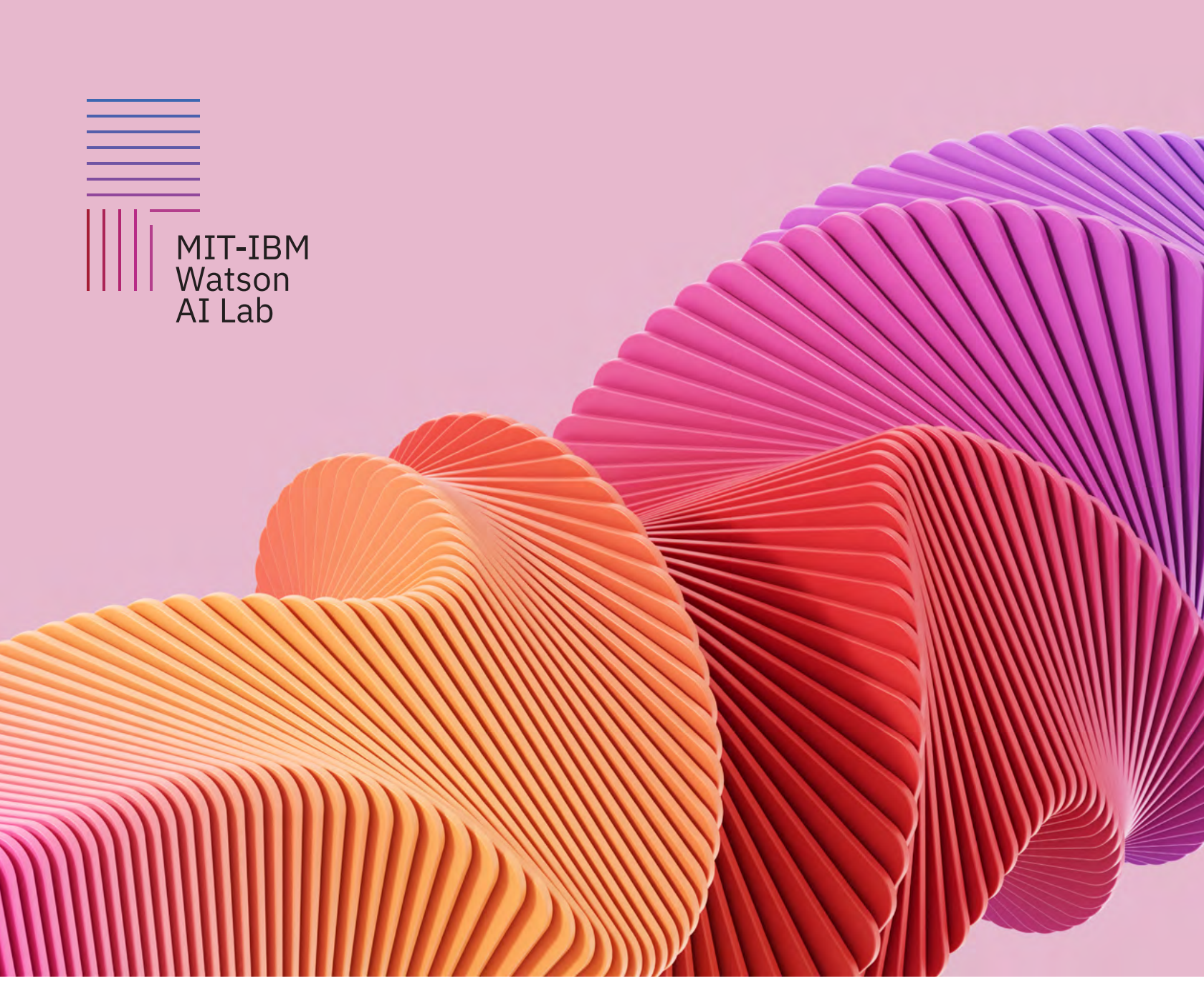
Accelerating design optimization in graph spaces

Graphs can provide a meaningful and cost-effective way to represent complex, physical systems and their features, like computers and circuits, networks, or polymers; however, when exploring them to create new ones, the design space can become intractable. The issue can become compounded by a scarcity of available labeled data in a particular domain, making property prediction difficult. A member company sought to develop a workflow to leverage the capacities of graphs to design new materials with prescribed properties. To address this, the Lab developed a pipeline that uses reinforcement learning to generate a set of production rules for graphs, like grammar for language, to learn material structures that are similar, and therefore likely have similar properties. When combined with tailored graph neural networks, the Lab's approach could efficiently search the material design space, optimize trade-offs and desired objectives, and generate a set of materials that are physically similar and intrinsically close to the training examples.

Key Metric:

- 50%+ reduction in the training set required for one dataset

Boston Scientific — Advancing science for life™ · EVONIK — Leading Beyond Chemistry · NEXPLORE — EXPANDING HORIZONS · WELLS FARGO · woodside



MIT Schwarzman College of Computing

MIT School of Engineering

IBM Research