Computer Science and Artificial Intelligence Laboratory

The Computer Science and Artificial Intelligence Laboratory (CSAIL) pioneers research in computing that improves the way people work, play, and learn. CSAIL serves the MIT community, the country, and society at large by creating a positive future enhanced by computer science through contributions of ideas, artifacts, and people.

CSAIL's current research addresses some of the grand challenges of the 21st century, including developing personalized learning, securing cyberspace, advancing health informatics, reverse-engineering the brain, enhancing virtual reality, developing tools for scientific discovery, improving urban infrastructure, and ensuring the health of our environment. Computing is central to solving these challenges and CSAIL contributes to making computing more capable by addressing fundamental algorithmic and systems questions at the core of computing, and broadening the scope of computing to address important social challenges that confront us.

Key CSAIL initiatives currently underway include tackling the challenges of data governance, developing new machine learning (ML) applications, securing computers and the cloud against cyberattacks, rethinking the field of artificial intelligence (AI), and developing the next generation of robots. Advanced software-based medical instrumentation and medical informatics systems to aid clinical decision making is being investigated. Advancements in biological research are also under way, including developments in the field of computational biology and the application of ML to the interpretation of complete genomes and understanding gene regulation.

CSAIL leadership spearheaded several other efforts that benefited the broader MIT community:

Achievements

The following are CSAIL initiatives and corresponding impacts for FY2020.

College of Computing Transition

In September 2019, CSAIL focus groups worked on transitioning its services and organization to MIT Schwarzman College of Computing, and now report to the deputy dean of research.

Communities of Research

Launched on January 1, 2020, nine CSAIL Communities of Research (CoR)—a new system of self-governance and identity—support the CSAIL community in social and intellectual impact through $30,000 a year in funding, administrative support for CoR heads, and through developing values for the community.

Symposium

CSAIL hosted the first-ever International Symposia: TEDxMIT featuring an all-female lineup. CSAIL also held the Third AI and the Future of Work Symposium. These were campus collaboration events to raise visibility and present research.

Space

The CSAIL space committee implemented improvements to the facilities to increase the quality of space for the laboratory's faculty, staff, and students. Space continues to be a challenge in light of personnel growth, research group demands, and the cost impact of building, renovating, and maintaining CSAIL's space in the Stata Center.

COVID-19 Response

COVID-19 has accentuated the fact that robots have an important advantage over humans: they cannot catch or transmit disease. As part of the COVID-19 response, CSAIL participated in the Research Ramp Up Lightening Committee to develop programs to support the community and to address their needs. Highlighted research and activities include:

- Private Automated Contact Tracing (Ronald Rivest, Daniel Weitzner, and Julie Shah)
- AI cures and drug discovery (Regina Barzilay)
- COVID-19 vaccine design (Dave Gifford)
- E-vent (Daniela Rus and Alex Slocum)
- UV robot disinfects Greater Boston Food Bank (Saman Amarasinghe and Daniela Rus)
- Modeling campus mobility (Nicholas Roy)
- Organizing approximately 30,000 PPE donations to MIT and hospitals (Daniela Rus and Victor Zue)

New Lab-wide Research Initiatives

CSAIL launched and funded initiatives to address basic research challenges of resilience and unsupervised learning in AI, such as the Air Force AI Accelerator, Defence Science and Technology Agency, MLA@CSAIL, and more.

CSAIL Growth

In the past 10 years we have had a 46% growth in headcount, and an 88% growth in research volume.

With a total combined research volume (primary, secondary, and World Wide Web Consortium) of $84,532,403 for FY2020, CSAIL is the Institute's largest interdisciplinary research laboratory. CSAIL continues to have the highest research volume amongst MIT Interdisciplinary Departmental Laboratories. The table comparison below reflects research volume data 2015 through 2020 between Vice President for Research laboratories. In FY2020 CSAIL moved to the MIT Schwarzman College of Computing (SCC). FY2020 Institute volume data is not yet published. In 2019 CSAIL research volume represented approximately 11% of the MIT total campus volume.

CSAIL had a healthy research volume growth of approximately 33.5% for the last five years. CSAIL manages approximately 580 active research awards and over 115 principal investigators (PIs), with appointments across 11 MIT departments. Through AY2019 we had 590 graduate students with research assistant appointments in CSAIL, and 194 Undergraduate Research Opportunity Program (UROP) students.

CSAIL research is sponsored by a large number of diverse sources, from US government contracts to the private sector. United States government sponsors include the following:

- Air Force Research Laboratory and the Air Force Office of Scientific Research
- Army Research Office
- Defense Advanced Research Project Agency
- Department of Defense Research and Engineering
- Food and Drug Administration
- Department of Education
- Department of Energy
- Intelligence Advanced Research Projects Activity, National Institutes of Health
- National Institute of Justice
- National Science Foundation
- US Navy (including the Office of Naval Research, Naval Air Systems Command)
- Space and Naval Warfare Systems Center

US and international nonfederal sponsors include the following:

- Accenture
- Advanced Technology Laboratories
- Aptima
- BAE Systems
- BBN Technologies
- Boeing
- BMW of North America
- Delta Electronics Foundation
- DSTA
- Ford Motor Company
- Foxconn Technology Group
- Honda R&D
- IBM
- Intel Corporation
- iFlytek
- JD.com
- Jaguar Land Rover Limited
- Lockheed Martin
- Microelectronics Advanced Research Corporation
- Mitsubishi Electric Corporation
- National ICT Australia Limited

- Nissan Motor Company, Ltd
- Nippon Electric Company
- Nippon Telegraph and Telephone Corporation
- Northrop Grumman Corporation
- Omron
- Ping An Technology
- Pfizer
- Qatar Computing Research Institute
- Quanta Computer
- Rakuten
- Raytheon
- Samsung Electronics
- Siemens
- Steelcase
- Suzhou Industrial Park
- Systems & Technology Research
- Takeda
- Toyota Research Institute
- Wistron

Other organizations sponsoring research include the following:

- Aarhus University
- Battelle Memorial Institute
- DSO National Laboratories
- Epoch Foundation
- The Hong Kong University of Science and Technology
- Industrial Technology Research Institute
- Nanyang Technical University
- Singapore-MIT Alliance

Research Projects

Within CSAIL we have many single- and multi-investigator projects, as well as a number of virtual centers and large-scale projects. CSAIL developed 10 lab-wide collaborations with a total of 67 projects, out of which 58 are housed in CSAIL and nine are housed in other DLC across MIT. The total number of PIs participating in all of these projects is 109, out of which 72 are CSAIL PIs and 37 are PIs from DLC across MIT. The large-scale projects and collaborations include the following:

Air Force AI Accelerator

The MIT-Air Force AI Accelerator is a five-year program that brings together the expertise and resources of MIT and the United States Air Force to perform fundamental research aimed at enabling rapid prototyping, scaling, and application of AI algorithms and systems. Established in 2019 and with an annual funding of $15 million, the accelerator launched 10 interdisciplinary projects in January 2020, involving researchers

from MIT, MIT Lincoln Laboratory, and the air force. The three-year projects, which encompass a total of 15 research workstreams, are advancing research in a broad range of areas, including the use of AI to advance weather modelling and visualization, the optimization of training schedules, and the enhancement of guardian autonomy for augmenting and amplifying human decision making.

The project workstreams and their leads are::

- Guardian Autonomy for Safe Decision Making
    - PI: Professor Daniela Rus (Electrical Engineering and Computer Science [EECS], CSAIL)
    - Lincoln Laboratory Leads: Ross Allen (G01) and Ho Chit Siu (G76)
- Fast AI
    - PI: Professor Charles E. Leiserson (EECS, CSAIL)
    - Lincoln Laboratory Lead: Vijay Gadepally (Lincoln Laboratory Supercomputing Center)
- ML-Enhanced Data Collection, Integration, and Outlier Detection
    - PI: Professor Tim Kraska (EECS, CSAIL)
    - Co-PIs: Professor Sam Madden (EECS, CSAIL), Professor Michael Stonebraker (EECS, CSAIL), Professor Manya Ghobadi (EECS, CSAIL)
    - Lincoln Laboratory Lead: Benjamin Price (G57)
- Transferring Multirobot Learning through Virtual and Augmented Reality for Rapid Disaster Response
    - PI: Professor Sertac Karaman (Aeronautics and Astronautics; Laboratory for Information and Decision Systems [LIDS]; Institute for Data, Systems, and Society [IDSS])
    - Co-PI: Professor Luca Carlone (Aeronautics and Astronautics, LIDS)
    - Lincoln Laboratory Lead: Mark Mazumder (G104)
- Conversational Interaction for Unstructured Information Access
    - PI: James Glass (CSAIL)
    - Co-PI: Boris Katz (CSAIL)
    - Lincoln Laboratory Lead: Charlie Dagli (G52)
- AI for Personalized Foreign Language Education
    - PI: Professor Shigeru Miyagawa (Linguistics and Philosophy, Open Learning)
    - Co-PI: Professor Emma Teng (Global Studies and Languages)
    - Lincoln Laboratory Lead: Douglas Jones (G52)

- Multimodal Vision for Synthetic Aperture Radar
  - PI: Professor Phillip Isola (EECS, CSAIL)
  - Co-PIs: Professor Taylor Perron (EAPS) and Professor William Freeman (EECS, CSAIL)
  - Lincoln Laboratory Lead: Miriam Cha (G01)
- AI-Assisted Optimization of Training Schedules
  - PI: Professor Hamsa Balakrishnan (Aeronautics and Astronautics, Operations Research Center [ORC], IDSS)
  - Co-PI: Professor Caroline Uhler (EECS, IDSS)
  - Lincoln Laboratory Lead: Michael Snyder (G104)
- The Earth Intelligence Engine
  - PI: Professor Dava Newman (Aeronautics and Astronautics)
  - Co-PIs: Chris Hill (Earth, Atomospheric and Planetary Sciences [EAPS]) and Stephanie Dutkiewicz (EAPS)
  - Lincoln Laboratory Lead: Mark Veillette (G43)
- Continual and Few-Shot Learning: Transferring Knowledge to New Low Resource Domains and Tasks
  - PI: Professor Pulkit Agrawal (EECS, CSAIL)
  - Co-PIs: Professor Regina Barzilay (EECS, CSAIL) and Professor Marin Soljačić (Physics, Research Laboratory of Electronics [RLE])
  - Lincoln Laboratory Lead: Olga Simek (G52)
- Explainable Machine Learning for Decision Support
  - PI: Professor Aleksander Madry (EECS, CSAIL)
  - Co-PIs: Professor Arvind Satyanarayan (EECS, CSAIL) and Professor Antonio Torralba (EECS, CSAIL)
  - Lincoln Laboratory Leads: Theodoros Tsiligkaridis (G01), Steven Gomez (G57), and Kevin Nam (G45)
- RAIDEN (Robust AI Development ENvironment)
  - PI: Professor Asu Ozdaglar (EECS, LIDS)
  - Co-PIs: Professor Aleksander Madry (EECS, CSAIL) and Professor Pablo Parrilo (EECS, LIDS)
  - Lincoln Laboratory Lead: Olivia Brown (G01)

- Objective Performance Prediction and Optimization Using Physiological and Cognitive Metrics
    - PI: Professor Thomas Heldt (EECS, Institute for Medical Engineering and Science [IMES], RLE)
    - Co-PIs: Professor Vivienne Sze (EECS, CSAIL, RLE), Professor Tamara Broderick (EECS, CSAIL)
    - Lincoln Laboratory Lead: Gregory Ciccarelli (G48)
- Robust Neural Differential Models for Navigation and Beyond
    - PI: Professor Alan Edelman (Mathematics, CSAIL)
    - Co-PI: Chris Rackauckas (Mathematics, CSAIL)
    - Lincoln Laboratory Leads: Michael O'Keeffe (G89) and Jonathan Taylor (G52)
- AI-Enhanced Spectral Awareness and Interference Rejection
    - PI: Professor Greg Wornell (EECS, RLE)
    - Co-PI: Professor Yury Polyanskiy (EECS, LIDS)
    - Lincoln Laboratory Leads: Binoy Kurien (G62), Jarilyn Hernandez Jimenez (G51)

Toyota-CSAIL Joint Research Center

Toyota established a collaborative research center with CSAIL in 2015 to further the development of autonomous vehicle technologies, with the goal of reducing traffic casualties and potentially even developing a vehicle incapable of getting into an accident.

Today, a car crash occurs every five seconds in the United States. Globally, road traffic injuries are the eighth leading cause of death, with about 1.24 million lives lost every year. In addition to this terrible human cost, these crashes take an enormous economic toll. The National Highway Traffic Safety Administration has calculated the economic cost in the United States at about $277 billion per year. Putting a dent in these numbers is an enormous challenge—and it is one that is motivating the research of the Toyota-CSAIL Joint Research Center, which was kicked off in September 2015. The center is in collaboration with the Toyota Research Institute, led by Gill Pratt.

Imagine if your car could tell you were having a bad day and turned on your favorite album to improve your mood. Imagine if your car could talk to your refrigerator, figure out that you are out of milk, and suggest where to stop on your way home. Imagine if your car knew that you forgot to call your parents yesterday and issued a gentle reminder on the way home, and that making the call was easy because you could turn the driving over to the car on a boring stretch of highway. These are just a few of the possibilities when we bring together cars and computer science. They are motivating the research at the Toyota-CSAIL Joint Research Center.

The objective of the Toyota-CSAIL Joint Research Center is to advance AI and robotics research, develop a safe and intelligent car, and improve mobility and transportation by advancing the science of autonomy and machine intelligence. The CSAIL researchers are working on new tools for collecting and analyzing navigation data with the objective to learn from humans; perception and decision-making systems for safe navigation; systems that can handle difficult driving situations: congestion, high-speed driving, and inclement weather; predictive models that can anticipate the behavior of humans and vehicles; more intelligent user interfaces; and, human support robots, whereby robots can work safely and seamlessly with and around humans in performing tasks, establish effective communication with humans and recognize human intent.

More specifically the projects and principal investigators active in the Toyota-CSAIL Joint Research Center during the AY2020 are:

- Machines that Can Introspect (PIs: Nick Roy and Boris Katz)
- Understanding Human Gaze (PIs: Antonio Torralba and Wojciech Matusik)
- Exploring the World of High Definition Touch (PIs: Ted Adelson and John Leonard)
- Formal Verification Meets Big Data Intelligence in the Trillion Miles Challenge (PI: Armando Solar-Lezama)
- The Car Can Explain! (PIs: Gerald Sussman and Lalana Kagal)
- Crossing the Vision-Language Boundary (PIs: Jim Glass and Antonio Torralba)
- Analysis by Synthesis Revisited: Visual Scene Understanding by Integrating Probabilistic Programs and Deep Learning (PI: Josh Tenenbaum)
- Using Deep Learning to Speed Up Deep Learning (PIs: Saman Amarasinghe and Fredo Durand)
- Decision Making for Parallel Autonomy in Clutter: Addressing Intent, Interactions, Rules of the Road, and Safety (PIs: Daniela Rus and Sertac Karaman)
- A Parallel Autonomy System: Data-Driven and Model-Based Parallel Autonomy with Robustness and Safety Guarantees (PIs: Sertac Karaman and Daniela Rus)
- Driver Perception and the Car-to-Driver Handoff (PI: Ruth Rosenholtz)
- Differentiable Computer Graphics for Training and Verification of Machine Perception (PI: Fredo Durand)
- The Robotic Manipulation Data Engine (PI: Alberto Rodriguez)
- Dense, Freeform Tactile Feedback for Manipulation and Control (PI: Wojciech Matusik)
- Sensible Deep Learning for 3D Data (PI: Justin Solomon)
- All Terrain Mobility and Navigation (PI: Sangbae Kim)
- Inner Vision: Camera-Based Proprioception for Soft Robots (PIs: Edward Adelson and Daniela Rus)

- A Safety Interlock for Self-Driving Cars (PIs: Daniel Jackson and Armando Solar-Lezama)
- Automation for Everyone (PI: Brian Williams)
- A CAD Tool for Designing Superintelligent Human-Computer Groups (PI: Thomas Malone)

Wistron-CSAIL Research Collaboration

Good health—both mental and physical—is one of the most pressing social and economic issues of the day. A healthier population makes for a happier society and a more productive economy. Today, people are surrounded by an explosion of sophisticated and increasingly affordable information devices, from laptop computers, e-book readers, and smart glasses to mobile phones, smart watches and health trackers. We monitor stock prices, weather forecasts, and traffic patterns through websites and apps, share our thoughts and experiences through emails, Facebook, and Twitter, and increasingly learn within online communities. These technologies open so many new opportunities for improving how we live, work, and play. But how do they empower us and at what costs? Recent studies show that we consume 11 to 14 hours of technology each day. And this often involves multitasking, which in turn retrains our brains, reduces concentration, and increases stress (e.g., studies show that the brains of heavy technology users show similar patterns to those of people who suffer from substance abuse). Finding ways to reduce stress and technology's negative impact on a workforce is critical to our well-being in the future.

The multiyear research program between Wistron and CSAIL focuses on rethinking how we compute and communicate in the digital age to ensure that health and well-being are at the core of our lives, and that our use of technology accelerates this objective. Some of the questions we pose and the answers we seek include: How to design the next generation of computers and communication systems to minimize our body's exposure to electromagnetic radiation? How should we rethink computer and communication architectures for sustainability? How to develop systems that deliver appropriate lighting? How to develop systems that reengineer email? How to develop algorithms that can help with information overload? How to use computing and communication in support of individual and community well-being? And, how to build computer and communication systems that are friendlier to our environment?

Our vision is to develop new computing and communication hardware and software platforms and supporting algorithms for modeling, controlling, and making decisions that will bring wellness to our use of technology. One thrust of this program focuses broadly on the computer and communication platforms. The second thrust focuses on using these novel platforms to promote healthier living. More specifically, the three projects that are currently active in the Wistron-CSAIL Research Collaboration are as follows:

- Individualized Prediction and Interpretation of Risk: Predicting Trajectories of Chronic Disease and Recovery (PIs: Polina Golland and Peter Szolovits)
- Smart Homes that Monitor Breathing, Heart Rate, and Life Quality (PI: Dina Katabi)
- Using Machine Learning to Build Better Clinical Support Tools (PI: John Guttag)

CSAIL-Microsoft Trustworthy and Robust AI Collaboration

The Trustworthy and Robust AI Collaboration (TRAC) between MIT CSAIL and Microsoft Research is working toward fostering advances on robustness and trustworthy AI, which spans safety and reliability, intelligibility, and accountability. The collaboration seeks to address concerns about the trustworthiness of AI systems, including rising concerns with the safety, fairness, and transparency of technologies.

The collaboration leverages the mutual interests of Microsoft and MIT CSAIL on achieving AI systems in both the autonomous, semiautonomous, and collaborative realms, but centered on the vision of extending and augmenting the abilities and intellect of people. The research engagement includes funded projects between Microsoft researchers and MIT Professors and students. The currently active projects are as follows:

- Safe Online Reinforcement Learning in Networked Systems (PI: Mohammad Alizadeh)
- Machine Learning with Theoretical Guarantees (PI: Tamara Broderick)
- Exploration of Robust Machine Learning for High-Stakes Predictions (PI: John Guttag)
- Explainability and Interpretability at Scale (PI: Stefanie Jegelka)
- Uncertainty and Robustness at Scale (PI: Aleksander Madry)
- Robustness Meets Algorithms (PI: Ankur Moitra)
- State-Based Approaches for Verifying and Testing Neural Networks (PI: Martin Rinard)
- Compression for Interpretability at Scale (PI: Daniela Rus)
- Distributed, Private and Efficient Machine Learning (PI: Vinod Vaikuntanathan)
- Off-Policy Evaluation for Risk-Aware Autonomous Systems (PI: Cathy Wu)

In addition, this collaboration gave rise to three workshops on November 14–15, 2019, and February 8, 2019 (both organized at MIT), and June 22–23, 2020 (held virtually).

Qatar Computing Research Institute

In 2012, CSAIL started a research collaboration with Qatar Computing Research Institute to collaborate on a wide range of research topics in computer science. After eight years, the program has come to a successful close. The program included the following projects in the 2020 program year:

Accurate Map Making Using Mobile Sensor Data

Led by Professor Hari Balakrishnan, Professor Samuel Madden, and Assistant Professor Mohammad Alizadeh, this project aimed to develop accurate map-making techniques using crowd-sourced methods to overcome challenges related to creating and maintaining street maps, especially in a rapidly developing environment, such as Doha, Qatar, leveraging data primarily from mobile phones and investigating current limitations due to sensor noise, outages, and data sparsity.

Arabic Speech and Language Processing

Led by Senior Research Scientist Jim Glass, this project aimed to develop key speech and language processing technology enabling users to search for verified facts and claims, in both written and video repositories of English and Arabic, using questions posed in natural and spoken language. The research addressed four essential cross-cutting topic areas to achieve this objective. First, methods that enable rich annotation of Arabic multimedia content. Second, language processing methods to analyze open-ended, user-generated content—such as dialogs—and perform veracity assessment and inference. Third, speech and language methods for processing low-resource Arabic dialects. And finally, interpretation and debugging techniques to improve machine translation between English and Arabic.

Database Management

Led by Adjunct Professor Mike Stonebraker and Professor Samuel Madden, this project dealt with database management. Specifically, the project team investigated Data Civilizer, a system to support data scientists.

HealthyForce—Promoting a Better Lifestyle with Reinforcement Learning in Type 2 Diabetes and Obesity

Led by Professor Peter Szolovits, this project focused on the development of a lifestyle recommendation system eventually intended to reduce the risk of obesity and type 2 diabetes. The project team explored the use of reinforcement learning with a new healthy lifestyle and behavioral change representation initially focused to recommend activity patterns that maximize the user's quality of sleep. These recommendations will be used to create both a new model for behavioral change (which will be incorporated into a health coaching system to provide just-in-time recommendations to increase the user's quality of sleep), as well as a new analytics system to support coaching by health-care professionals.

Humanitarian Image Analysis

Led by Professor Antonio Torrabla, the goal of this project was to conduct applied and core computer science research and to build innovative technologies that can be used by decision makers, NGOs, affected communities, and scholars to improve the effectiveness of humanitarian strategies, such as preparedness, mitigation, and response during humanitarian crises and emergencies. The core of this project focused on developing a multimodal data processing system for understanding disaster scenes and situations from social media.

Optimal Transport Models for Health Care Analytics

Led by Associate Professor Justin Solomon, the main focus of this project was to discover causal relationships in (multivariate) sequence of states (e.g., in health data) and to uncover the complex dependency structures from high-dimensional time-series encoded as sequence of states. In particular, the project team addressed the following challenges using optimal transport methodology: machine learning techniques to extract sequences of states from time-series data; causality analysis from sequence of states data; explanatory

models for sequence of states data; supervised learning methods to predict categorical or continuous output from sequence of states input data; unsupervised learning methods for sequence of states data; and factor analysis for sequence of states data.

Quanta-CSAIL Research Collaboration

Quanta and CSAIL started a new five-year, $12.5 million collaboration. In this phase of the collaboration the parties will be focusing on long-term research to create future platforms focusing on the delivery of improved health care, using computer science in general, and AI in particular. We are currently pursuing the following three projects:

Using Machine Learning to Curb Infectious Disease

Led by Professor John Guttag, this project is based on using ML to further our understanding of what makes people more or less susceptible to different kinds of infections; how different kinds of infections are spread; and how to design and test interventions that reduce the spread of infectious diseases. Our initial experiments have been in the area of health care-associated infections. These are infections that are spread within medical settings. We are studying them for the following reasons:

- They tend to have high mortality and morbidity, in part because the pathogens are often resistant to treatment
- Because we can acquire detailed data about hospitalized patients, we can study what makes people vulnerable to infectious diseases and how they are spread
- Because a hospital is a closed and relatively small environment, we can test interventions in a controlled way

Since inception of this project, we have made progress on both methodological and experimental fronts.

Methodological progress

We have worked on three critical issues: separating correlation from causality; incorporating prior knowledge in the ML process; and, understanding the role of latent spreaders. In the health care domain, typical ML models capture correlations between attributes of people (such as their medical history) and outcomes (bad outcomes, such as becoming infected with a disease, and good outcomes, such as being cured of a disease). For example, our current model for predicting who is at risk for contracting a *Clostridium difficile* (C. diff) infection (CDI) demonstrates an association of both chlorhexidine and metronidazole with CDI. Though the model makes no distinction between the two, any infectious disease specialist knows that chlorhexidine is highly unlikely to be causative, and therefore should not be avoided when trying to reduce risk. We recently developed a method designed to estimate aspects of the impacts, both positive and negative, of treatment.

Medical practitioners and researchers already know quite a lot about diseases and treatments. Yet, existing ML models are driven entirely by data. This means that they cannot be constructively applied to problems on which there is limited data. It also means that the models produce results without providing justifications related to existing knowledge. We recently developed a method that starts to address both of these

issues. The best way to limit the spread of an infectious disease is to avoid transferring pathogens from those who are carrying the disease to those who are vulnerable to contracting the disease. A major challenge in doing this is identifying latent carriers (as we have seen with spread of COVID-19). We are currently in the process of enhancing our previously published method for doing this.

Experimental progress

We continue to be deeply engaged with our colleagues at Massachusetts General Hospital (MGH). We are currently collaborating closely with the two most senior physicians working on infection control within the hospital, and also with technical specialists helping us to understand the hospital's data systems. Much of our most recent work has been translating models we built using a previous version of the hospital's electronic medical record system to the current version. We have succeeded in pulling data from the new system and understanding what the various fields mean. We now have a running model that works on the current system. It has been tested and evaluated on a retrospective data plan.

We also worked with our clinical colleagues in infection control and in oncology to design an experiment that will shed light on latent spreaders. The experiment involves testing patients on admission to an oncology unit for colonization with the C. diff bacterium. MIT is transferring funds supplied by Quanta to pay for part of this testing. The rest of the expense will be covered by MGH's oncology department. We expect testing to start this coming spring. The clinical trial was supposed to have started April 1, but was delayed due to COVID-19. Our trial coordinator was reassigned to manage the MGH trial of remdesivir.

Learning to Assess Breast Cancer Risk to Enable Early Detection and Prevention

Led by Professor Regina Barzilay, the group's goal is the development and clinical deployment of ML-based models for risk assessment and early detection of breast cancer. We visited Quanta and Chang Gung Memorial Hospital (CGMH) in early February to complete our breast cancer risk project. There, working with Drs. Yung-Liang Wan and Gigin Lin from Chang Gung Memorial Hospital, and Johnnie Huang, Robe Yang, and the Quanta team, we've demonstrated that models trained on MGH data to predict future breast cancer risk performed equally well on Taiwanese women. We are excited to pursue collaborations both clinically implementing these models in Taiwan and also scaling these technologies to new organ systems. A group member also met with Huang from National Taiwan University Hospital (NTUH), thanks to help from Quanta, and we are eager to work together on predicting lung cancer risk. On the MIT side, we've obtained 45,000 lung CT scans from the National Lung Screening Trial dataset and have begun developing new algorithms to predict lung cancer risk. As these algorithms become more mature, we look forward to working with Quanta to specialize them to Taiwanese populations in collaboration with CGMH and NTUH. The subcontract and collaboration with MGH are still ongoing.

Revolutionizing the Care of Patients with Cardiovascular Disease

Led by Professor Collin Stultz, the intersection of modern cardiovascular medicine and information technology presents an opportunity to establish a new paradigm for the treatment of patients with cardiovascular disease. Risk stratification and early warning systems are most powerful when coupled with intelligent, patient-specific therapeutic recommendations. We will use sophisticated machine learning methods to build models that suggest personalized therapeutic interventions that minimize adverse outcomes in patients with cardiovascular disease. Moreover, we strive to build models that not only have predictive power, but also provide clinically meaningful insights that help explain how each model arrives at a particular result. This represents a collaborative effort between MIT and MGH and seeks to transform the way in which clinical care is provided to patients with cardiovascular disease.

First, we will explore how computation can enable researchers to significantly improve the diagnoses and treatments of some of the most pervasive diseases, such as cancer. Second, we will apply AI techniques to improve the institution-specific delivery of patient care. Third, we will investigate how the interface between a patient and a primary care physician can be streamlined, so that the data collection and resulting analytics can be more reliable. While the focus of this collaboration will largely be defined by the troika of patient, hospital, and doctors, other aspects such as privacy, security, nutrition, and self-care may also be investigated as time and resources permit.

Industrial Outreach

CSAIL Alliances

The CSAIL Alliance Program

The CSAIL Alliance Program (CAP) is a gateway into the lab for industry, organizations, and governmental institutions seeking a closer connection to the work, researchers, and students of CSAIL. The program provides a proactive and comprehensive approach to developing strong connections with all CSAIL has to offer. Leading organizations come to CSAIL to learn about our research, to recruit talented graduate students, and to explore collaborations with our researchers. Through this program, we are able to better provide our members with access to our latest thinking and our deep pool of exceptional human and informational resources. Overall, CAP supports the mission of CSAIL by connecting our researchers, students, and technological advances to industry and organizations across the globe.

Levels of membership

The CAP program provides a proactive and comprehensive approach to connect members to the whole lab—all 60 research groups spanning robotics, natural language processing, networks, databases, cryptography, web science, and more. CAP has three levels: Student Engagement, which is focused on connecting with students and postdocs for career opportunities; Affiliate, which provides lab visits, access to the annual meeting, recruiting assistance, research briefings, and professional education discounts; and, Partner, which includes all of the benefits of the Affiliate level as well as more expanded options with added access to research initiative meetings, custom faculty-led seminars, and expanded recruiting options.

Member companies

Currently there are over 85 member companies, including global brands such as Apple, BASF, Google, Samsung, JP Morgan, NASDAQ, and Microsoft. Members are headquartered in North America, South America, Europe, and Asia and represent a wide variety of industry verticals.

Online Courses and Professional Development

CSAIL Alliances also produces and manages online professional development courses in partnership with MIT's Professional Education, MIT's Office of Digital Learning (ODL), edX, Get Smarter, and Harvard Extension School. The following is a list of the programs to date, a brief description, number of offerings to date, and total enrollments to date. Total enrollment is now approximately 40,000 online learners.

Enrollments in CSAIL-Produced Online Courses

Course title	Description	Offered	Enrolled
Artificial Intelligence: Implications for Business Strategy	Offered in partnership with the Sloan School of Management, this course focuses on the organizational and managerial implications of AI technologies.	27	14,979
Machine Learning: Implementation in Business	This course is offered in partnership with the Sloan School of Management and aims to demystify machine learning for the business professional—offering a firm, foundational understanding of the advantages, limitations, and scope of machine learning from a management perspective.	8	1,022
Tackling the Challenges of Big Data (not offered FY2019)	Survey state-of-the-art topics in big data, looking at data collection (smartphones, sensors, the web), data storage and processing (scalable relational databases, Hadoop, Spark, etc.,), extracting structured data from unstructured data, systems issues (exploiting multicore, security), analytics (machine learning, data compression, efficient algorithms), visualization, and a range of applications.	10	11,431
Tackling the Challenges of Big Data—Taiwan (not offered FY2018)	The original course translated into traditional Chinese.	1	1,296
Tackling the Challenges of Big Data—Ilumno	The original course translated into Spanish and Portuguese, offered through Ilumno in collaboration with universities in South America.	3	1,032
Internet of Things: Roadmap to a Connected World—Ilumno	The original course translated into Spanish and Portuguese, offered through Ilumno in collaboration with universities in South America	3	608
Cybersecurity: Technology, Application, and Policy (not offered FY2019)	This six-week online course provides a holistic look at cybersecurity technologies, techniques, and systems.	8	3,199
HCI: Human Computer Interaction for User Experience Design	The six-week course was produced in partnership with Get Smarter and includes eight CSAIL researchers who review a host of cutting-edge HCI concepts, including voice-activated and spatially-aware computers, as well as speech and vision tools.	13	962
Introduction to the Challenges and Opportunities of Big Data, the Internet of Things, and Cybersecurity	A semester-long course combining our big data, internet of things, and cyber courses, offered for credit through Harvard Extension School.	6	912
Startup Success: How to Start a Technology Company in Six (Not So Easy) Steps (not offered FY19)	This course discusses the lessons learned by Michael Stonebraker and Andy Palmer during their start-up endeavors over a 30-year period. The lessons are distilled into six steps that any entrepreneur can follow to get a company going. Topics include the generation and assessment of ideas, the challenges of building a prototype, the recruitment of a talented team, the closing of the first financing round, and pursuing growth with the right business leadership.	2	359
Internet of Things: Roadmap to a Connected World	The course introduces both the broad range of internet of things technologies and the most recent developments in the space.	8	3,955
Total		**89**	**39,755**

SystemsThatLearn@CSAIL

The next decade will usher in a new frontier of sophisticated systems that perform complex, humanlike tasks, with complex inferences and predictions. Using data gathered from diverse sensors and mobile devices, computing power spread across embedded devices and data centers, as well as ubiquitous network connectivity, we will need new tools to realize the potential of learning systems. We are already seeing practical applications of these systems in areas such as autonomous vehicles and personalized health care that have the potential to transform industries and societies.

The goal of SystemsThatLearn@CSAIL is to accelerate the development of systems and applications that learn. We intend to accomplish this goal through combining our expertise in systems and ML to create new applications for understanding complex relationships unearthed by analyzing the avalanche of data available today. Presently, however, software systems that incorporate machine learning are hard to build, deploy, and maintain. They require a large and highly skilled workforce. Unlike traditional enterprise systems, once built, they often require thousands of hours of ongoing (and sometimes daily), maintenance to ensure that their predictions and behavior continue to be accurate and useful. Integrating ML systems into traditional enterprise architecture, testing and deployment processes are too complex—partly due to organizational silos that exist between systems engineers and data scientists. In application, many problems in large-scale software systems involve optimizations that benefit from predictions, such as scheduling, compilation, query planning, routing, data cleaning, and congestion control. Today, it is hard to apply machine learning tools to design this type of system software.

Our approach to designing, training, and deploying humanlike tools will focus on the following four areas of investigation:

Heterogeneous architectures

The data and features that drive learning in these systems and applications increasingly come from diverse distributed infrastructure, including phones, sensors, or other bandwidth and power impoverished endpoints. Therefore, even acquiring data for learning may require adaptive allocation of computation over heterogeneous infrastructure. Furthermore, the rise in heterogeneous hardware, such as graphics processing units and many-core processors, which excel at certain aspects of the learning pipeline, suggests a diversity of computational resources will be brought to bear.

Predictable composition

Successfully designing and training machine learning methods for the desired task once data is available (i.e., programming at the level of learning components, and reasoning about the behavior of the composition of such components), calls for skill and expertise that is not yet well-supported or automated.

Distributed execution

In terms of the underlying infrastructure, complex machine learning methods also demand considerable parallel resources to train effectively. Once trained, models may be deployed either on massive parallel infrastructures (e.g., data centers) or may have to be reduced and distributed back to the heterogeneous components to be utilized where needed (e.g., mobile devices), requiring new distributed algorithms and execution frameworks.

Seamless integration of training and deployment

Many machine learning solutions today are trained and deployed in well-separated phases of training and testing (deployment), but this will change. Learning will increasingly become an ongoing, integrated process. The tighter integration of learning and computer systems offer exciting possibilities in terms of new capabilities, but requires us to overcome challenging hurdles pertaining to programming abstractions, maintenance and monitoring, analysis, and performance guarantees. This includes, among other things, safeguards and ways of containing learning functions. In addition to building better systems for machine learning, we believe our focus on deployability of models will help us advance machine learning itself by developing new models designed to further the above democratization goals, while still providing excellent prediction accuracy. We expect many new tools and practices to be developed.

SystemsThatLearn@CSAIL is a large, multi-PI research program to accelerate the development of this next generation of systems. The primary focus is on developing a common infrastructure, specifically in the form of software that includes the following new theoretical advances:

- Tools to help data scientists and engineers understand their models, train them in a scalable fashion, monitor their results, and retrain models efficiently
- Useful models focusing on efficiently deploying models in distributed and data center settings, reusing and redeploying models, as well as creating development environments good for training and deployment
- Developing heterogeneously deployable models (i.e., models that can be decomposed across heterogeneous devices), or lower fidelity models that can run on sensors or smartphones and also on more powerful servers as well as developing models that are more interpretable
- Tools for statistical monitoring and performance prediction where machine learning is used to understand the performance of complex systems
- Tools and methods to implement and run systems that learn over an untrusted infrastructure

SystemsThatLearn@CSAIL is led by Professors Sam Madden and Tommi Jaakkola and includes 37 CSAIL researchers. It is structured as an industry consortium. On March 29, 2017, we launched this initiative with five founding members: BT, Microsoft, NOKIA Bell Labs, Salesforce, Schlumberger, and ScotiaBank. Additional members added in FY2018 include: BASF, Element AI, EY, and JP Morgan Chase. In FY2019, Facebook was added as a member. In FY2019, there were grants for 10 projects. In FY2020, SystemsThatLearn@CSAIL provided grants for 12 projects.

FinTech@CSAIL

Financial technology (FinTech) is disrupting many aspects of financial services, banking, and insurance, among other industries. Not only will infrastructure and operations be disrupted, but new technologies, business models, services, and even industries will be launched. FinTech holds promise not only for verified transaction systems, such as

blockchain, but also for technologies involving AI, security, data analytics/value extraction, machine learning, trust verification, risk management, and privacy advances as well.

The financial sector has many unique attributes and at the core of a company's success in this sector is trust, security, value, and efficiency. Current technology roadmaps aren't perceived as providing sufficient guidance for FinTech, and new players and technologies are constantly emerging. The shifting demands of customers are evident and pose both risk and opportunity. To stay competitive and stay ahead, companies need access to innovation, thought leadership, new technologies and high caliber talent.

FinTech@CSAIL will bring together industry, thought leaders, innovators, academics, disruptive technology development, and start-up companies that are reinventing global financial services. Through FinTech@CSAIL, we will work closely with industry partners in leveraging innovation from cutting edge research to develop the next generation of impactful technologies that will open up new business models, broaden access, gain new data insights, and improve security.

The breadth of research at CSAIL uniquely positions the lab to address a wide variety of challenges in the space. FinTech@CSAIL will include 15 researchers of MIT CSAIL, who have pioneered the fields of secure computation, ML, AI, data analytics, and risk management. The goal is to advance the state of the art in collaboration with select industry partners to address the hardest problems facing the finance industry today.

Through the rigorous research of our faculty coupled with our tradition of collaborating with industry, FinTech@CSAIL will address relevant business problems with long-term vision. Additionally, FinTech@CSAIL will draw from across the lab as well as other focused research initiatives in Cybersecurity@CSAIL and SystemsThatLearn@CSAIL, all of which come together for a very powerful collaboration to address the challenging issues that are emerging. We anticipate many opportunities for direct, active collaboration and knowledge sharing though events, projects, and directed research. By leveraging the research ecosystem at CSAIL, we will work to address the following:

- New approaches to efficient shared public ledger systems and digital currency
- Technologies to provide secure, multiparty computation as well as secure and private data extraction
- Advanced data analytics to apply to risk management and prediction
- Scalable, trusted systems
- Security of machine learning and AI systems
- ML and AI in automatic advising, compliance, and augmented assistance
- Applying natural language processing in the context of financial contracts
- Speech recognition applications
- Security of legacy systems

- Efficient processing and automation of tasks
- Anonymization of data and privacy
- Movement of datasets and sharing datasets securely
- New lending and payment technologies
- Addressing the changing role of banks as new technologies form the landscape of the future of the financial industry

FinTech@CSAIL was launched on July 18, 2018. The initiative is led by Professors Andrew Lo, Silvio Micali, Gary Gensler, and Randall Davis. Founding members include: NASDAQ, Ant Financial, Citigroup, Ryan Software, The London Stock Exchange Group, and Ripple Labs. State Street Bank, the Canadian Imperial Bank of Commerce, the Bank of International Settlements, Fidelity, ConsenSys, Itau, Bank Negara Malaysia, and Bank of New York Mellon have also joined. In FY2019, FinTech@CSAIL provided grants for 10 proposals. In FY2020, FinTech@CSAIL provided grants for six proposals.

Policy Dialogue and Public Engagement

The initiative will provide a strategically managed forum for dialogue amongst MIT researchers, policymakers, industry consortium members, and civil society partners.

- Education: The initiative will contribute to expanding academic offerings on privacy topics, and will create a professional education series delivered online and in person that is designed for privacy professionals in industry, civil society, and government.
- Leadership: The Future of Data, Trust and Privacy will be co-led by Daniel Weitzner and Professor Srini Devadas. These leaders in their respective fields will bring together a deep technological and policy understanding to the critical issues in data governance and management.

Internet Policy Research Initiative

Communication and information networks have become fundamental to our increasingly digital economy and society. Despite this importance, the technologists and policymakers who play key roles in supporting the transition to these networks approach issues from different perspectives and often do not speak the same technical languages. This disconnect can lead to uninformed policy making and misdirected research efforts. As such, there is a pressing need to bridge the gap between technical and policy communities.

The mission of the Internet Policy Research Initiative (IPRI), an Institute-wide initiative, is to work with policymakers and technologists to remedy this issue and increase the trustworthiness and effectiveness of interconnected digital systems, such as the internet. We accomplish this via a three-pronged approach: targeted engineering and public policy research; educational programs for students and policymakers; and outreach programs to build policy communities that facilitate communication.

IPRI's Research Efforts Cover Six Categories

Cybersecurity

IPRI's cybersecurity research focuses on the technical and policy aspects of cybersecurity issues as they relate to the communication networks and software systems affecting the global society and economy. This multidisciplinary research area encompasses encryption policy, accountability, cryptography, data sharing, securing core economic and social infrastructure, measuring cyber risk, and more. Major projects include the following:

- "Keys under Doormats," which has been cited favorably in leading policy documents, such as the House Encryption Working Group report
- The PACT (Private Automated Contact Tracing) project, a joint collaboration between CSAIL, IPRI, Massachusetts General Hospital Center for Global Health, MIT Lincoln Laboratory, and more; PACT aims to enhance contact tracing in pandemic response by designing exposure detection functions in personal digital communication devices that have maximal public health utility, all while preserving privacy
- An interdisciplinary project called SCRAM (Secure Cyber Risk Aggregation and Measurement), which designs and builds cryptographic tools and platforms that allow us to measure cyber risk more accurately, without putting participant data at risk of discovery by other participants or the platform host

AI Policy

Artificial intelligence and machine learning technologies are becoming increasingly prevalent in not only advertising and research, but also in traditionally regulated spaces, such as health care, finance, transportation, and employment. Current research areas include studying the role of AI in financial decision making, increasing access to new training datasets with policy, working with stakeholders on AI principles, and shaping global internet policy making via policymaker engagement and informing the public debate. As part of this work, IPRI co-hosted the first AI Policy Congress with the MIT Quest for Intelligence on January 15, 2019. At this event, OECD members, world-leading ML experts, global policymakers, and industry experts discussed how we should govern AI systems and how to enable AI systems to meet society's needs domestically and internationally.

Privacy

Work also focuses on privacy policy and its critical role in trustworthiness. The Privacy group has published work on such topics as privacy and security in home assistants, exposure minimization, and the data-sharing practices of smartphone apps. Current projects include the development of databases that are privacy aware.

Networks

The Advanced Network Architecture (ANA) group works to understand and shape the future of the internet. They achieve this goal with the understanding that the future of the internet is defined by the economic, social, regulatory, legal, and political

concerns involving the internet. As such, the ANA group is organized around five themes: internet architecture, internet security, internet economics, internet policy, and network management. In 2018, the ANA group had its first major release of data about interconnection congestion on the internet.

Decentralized Web

The Decentralized Information Group focuses on data and systems governance (primarily on the web) and explores both policy and technical issues. Current projects include a decentralized privacy-preserving platform for clinical research, evaluating the trustworthiness of autonomous systems, studying the relationship between privacy and machine learning, developing explanations for complex machines and models, securely aggregating distributed data, and developing smart contracts for data sharing.

App Inventor

The core goal of the MIT App Inventor team is to empower young people to develop useful apps that serve as novel digital solutions to the problems they face in their lives, communities, and world.

World Wide Web Consortium

The World Wide Web Consortium (W3C) was founded at MIT in 1994 by the inventor of the web, Tim Berners-Lee. W3C is responsible for developing and maintaining the standards that make the web work and for ensuring the long-term growth of the web. Nearly 450 member organizations, including most of the world's leading technology companies, are working to enhance the capabilities of web documents and create an open web platform for application development, available across a wide range of devices, enabling everyone on the planet to collaborate and share data and information. In recent years, a great many factors (people, devices, bandwidth, policy decisions, etc.,) have extended the reach of the web in society. Video, social networking tools, user-generated content, location-based services, and web access from mobile devices are transforming many industries, including mobile, television, publishing, automotive, entertainment, games, and advertising. This transformation has led to greater demands on W3C and other organizations to build robust technology that meets society's needs, in areas such as privacy, security, accessibility, and multilingual content. As the world has become more virtual and web based with the current pandemic, there have been new demands on the web platform.

Core Technology Focus

W3C standards define an Open Web Platform for application development that has the unprecedented potential to enable developers to build rich interactive experiences, powered by vast data stores that are available on any device. Although the boundaries of the platform continue to evolve, industry leaders speak in unison about how HTML5 is the cornerstone for this platform. The full strength of the platform relies on many more technologies that the W3C and its partners are creating, including cascading style sheets. The W3C's real-time communications spec is becoming more central to society as physical meetings are replaced by virtual, real-time meetings over the web.

In recent years, publicly noted security and privacy breaches have resulted in unprecedented attention to fixing web security and privacy. W3C addresses that both with specific solutions (such as the Web Authentication spec that authenticates users without passwords), and by conducting reviews of every W3C standard for security and privacy. There is new focus on how to support advertisers without the current approaches of having so much personal information available to advertisers. The growth of e-commerce has focused new attention on standardizing payment and e-commerce approaches. With immersive technologies, there is a strong focus on web solutions for virtual reality and augmented reality.

Industry Impact and Broadening the Set of Participants

In recent years, web technology is not only used by consumers and companies for information sharing, but increasingly the web is the delivery mechanism for companies to deliver their services. Examples of that include telecommunications (where web access is a key service), entertainment (which is increasingly delivered over the web), publishing, advertising, personal communications, and financial services. This has caused a diversification in the membership of W3C, and also has enriched the technical agenda to address new technical issues that arise. For example, web browser companies, credit card companies, and other FinTech stakeholders are using W3C to streamline web payments through a browser.

Research Highlights

In addition to the large-scale collaborative projects and center research, numerous individual and multi-investigator projects are under way. A sampling of the work is highlighted below:

Fast and User-Friendly Evaluation of Machine Learning Methods

Lead by Associate Professor Tamara Broderick, ML methods are increasingly deployed in our day-to-day lives, with the potential to impact individuals' health, safety, employment, and finances. So it is crucial that we understand whether these methods actually work as desired. Moreover, as ML methods become increasingly complex, black box, and generally inscrutable, it becomes prohibitively difficult to provide a priori theoretical guarantees on their quality. An alternative is to test their performance empirically. Note that ML methods typically require a dataset (so-called training data); ML discovers patterns in the training data that can then be used in the future. Unfortunately, testing a method on the same data that it is trained on is known to overfit—and thus may provide dramatically incorrect estimates of how well the method will perform on genuinely new data, where it is meant to be applied. Leaving out some separate data for testing gives more accurate estimates. Leaving out more data provides less noisy performance estimates but leaves less data for training. A way to ameliorate this trade-off is to leave a smaller amount out for testing, but repeat the process for multiple folds of data—and then average the performance across folds. This final idea, called cross-validation (CV), is the gold standard for evaluation in the ML literature and is taken as the de facto evaluation technique essentially across the practice of ML. But CV requires rerunning an ML algorithm multiple times—sometimes as many times as there are data points. When a single run of an ML algorithm is already hitting the limits of available computational resources, rerunning the algorithm is often too expensive to be practical.

To solve this problem, we have proposed to instead use an approximation. Namely, we consider each left out dataset as a reweighting of the original dataset and use a polynomial approximation to this perturbation to estimate the actual effect of leaving the data out. Our method takes just a few lines of code for the user and requires just a single run of the original machine learning algorithm (rather than many such runs), so it achieves orders of magnitude speed ups in practice. We have supported our method with theoretical guarantees on quality and have shown that it provides an excellent approximation to exact CV in practical problems. We have verified the accuracy of our approximation in a number of particularly challenging and complex domains in machine learning, including high-dimensional problems and structured problems.

Lottery Ticket Hypothesis

Led by Assistant Professor Michael Carbin, the Programming Systems Group conducts research across ML and programming languages. A key highlight of the group's work includes the development of the Lottery Ticket Hypothesis, a conjecture and empirical study that demonstrates that within the large, state-of-the-art neural networks in many domains, there exist counterparts that are an order of magnitude smaller, yet train just as well. These results directly contend with the contemporary understanding that neural networks need to be incredibly large to train from scratch. The results have sparked community-wide efforts to study the phenomenon, to develop techniques to efficiently identify these small neural networks, and to put in place new theoretical frameworks.

The overarching charter of the Programming Systems Group is to develop computer programming systems, such as programming languages and software analysis techniques, that enable developers to deliver performant, resilient, and correct programs in the presence of uncertainty. This uncertainty may arise from the inherent uncertainty of modeling a real-valued and probabilistic physical world with discrete, finite precision, deterministic approximations inside a computer; from our inability to fully predict aspects of the computational platform such as execution times or error rates; or from the composition of traditionally developed code with relatively opaque neural network components.

The Lottery Ticket Hypothesis work fits in the latter space, offering new directions to consider neural network interpretability and reasoning techniques for neural networks that necessarily require fewer parameters than are presumed to be required by state-of-the-art neural network applications. However, more broadly, as computing continues to become more integrated in modeling, perceiving, and processing the uncertain world, traditional computer systems will increasingly need to grapple with the concept of probability, approximation, and error resilience. The group's research directly addresses the presenting need for sound and efficient computing abstractions of these concepts.

Understanding Graph Neural Networks

Led by Associate Professor Stefanie Jegelka, ML tasks on graphs and networks are ubiquitous. For instance, one may want to predict interests or properties of people in a social network. Recommender systems can be modeled as link prediction in a network of users and items. Many applications arise in the sciences, too, for example, predicting properties of molecules for drug or materials design, or predicting interactions between

drugs, or drugs and proteins, in a network of drug and protein interactions. Or, we may want to learn about particle interactions in a physical system.

Specialized graph neural networks (GNNs) have shown big empirical promise for these applications. But their theoretical foundations are much less well understood (e.g., what kinds of predictions such GNNs can learn to make, or how many examples they need to see for learning, and what this depends on).

Our group has made progress toward answering these questions, and, based on this new understanding, developed more capable graph neural networks that have been adopted for multiple applications. To study the representational power of GNNs, we asked: What kinds of graphs can common GNNs distinguish? If the GNN cannot distinguish two graphs, it must make the same prediction for both. This discriminative power depends on the way the model aggregates information in a local neighborhood. The theoretical limit can be reached via specific aggregation operations, which result in better practical performance, too. Still, with only local aggregations, we showed that many common models cannot recognize simple graph motifs and graph properties.

For functions that GNNs can learn, we showed how the number of training data examples needed for learning depends on the GNN's size and the input graph structure, with perhaps surprising relations to neural networks for sequence data; and, the structure of the input task, especially for learning reasoning tasks and graph algorithms. Our results are among the first learning analyses for GNNs, they are in correspondence with empirical results, and offer guidance for designing neural networks for specific tasks.

Supertech Research Group

Led by Professor Charles Leiserson, the Supertech Research Group has produced an initial beta release of the open-source OpenCilk parallel-programming platform for developing high-performance parallel code, which is freely available on GitHub. The release includes beta versions of the OpenCilk compiler, the OpenCilk runtime system, the Cilksan race detector, and the Cilkscale scalability analyzer. The release supports parallel computing using the three Cilk keywords (cilk_spawn; cilk_sync; and cilk_for) in C and C++ programs. OpenCilk supports advanced programming features, including C++ exceptions and parallel reductions using the reducer hyperobject library from Intel Cilk Plus. The beta release targets Unix/Linux x86-64 systems, and it has been tested on a variety of operating systems, including Ubuntu 18.04, FreeBSD 12.1, Fedora 30, and Mac OSX 10.15. Leiserson's team is continuing to develop OpenCilk for future releases, including a beta release suitable for teaching classes this fall, and a version 1 release around the end of the calendar year.

Supertech has been advancing the technology for automatic differentiation with several projects. They have developed the PARAD system for reverse-mode autodifferentiation of parallel programs, which contains the first provably efficient algorithm for the problem. In addition, they have built a compiler-based differentiation tool, called Enzyme, which allows programs in any language that can be compiled to the LLVM intermediate representation to be automatically differentiated.

Supertech has made numerous other contributions to the science and engineering of fast code, including an algorithm for computing fast and accurate prefix sums for floating-point values, a comprehensive analysis of cup-filling games (a foundation of many parallel load-balancing problems), achieving near-zero variability in executing compute-intensive code in cloud environments, compressed representation of sparse tensors suitable for supporting parallel execution, and a simple compiler framework for parallel exception handling.

MIT Center for Deployable Machine Learning

Led by Professor Aleksander Madry, the fact we came to trust machine learning with just about every aspect of our lives does not mean that our machine learning tool kit is already worthy of that trust. Indeed, ensuring our self-driving cars reliably detect pedestrians and obstacles, making the Facebook/Twitter/YouTube content recommendation be resilient to manipulations, or gaining the ability to root out some of the undesirable biases in our data are just some of the challenges we have yet to tackle.

Professor Madry aims to change this state of affairs. His research has put forth a robust, optimization-based perspective on the question of resilience of machine learning models to so-called adversarial perturbations (i.e., slight changes of inputs that result in their complete misclassification). This perspective not only provided a conceptually unified view of this domain but also gave the first truly effective mechanism for training models that attain adversarial perturbation resistance. In particular, his latest research in this context enabled us to understand why models trained in a standard manner are so brittle and to identify the fundamental differences between the "standard" and "resilient" learning, paving a way toward building a machine learning tool kit that is reliable, interpretable, and easily adaptable to different classification tasks.

As a part of this endeavor, Professor Madry recently founded the MIT Center for Deployable Machine Learning. This center brings together an interdisciplinary team of faculty, researcher scientists, and students to facilitate a multipronged effort to revisit all the tenets of the existing machine learning framework and build the next generation of machine learning that can be deployed in the real world in a safe, secure, and responsible manner.

Intersection of Programming Systems, ML, and AI

Led by Professor Armando Solar-Lezama, the research in the Computer-Aided Programming group seeks to explore the intersection between programming systems, ML, and AI. Our original motivation for exploring this space was to leverage techniques from ML and AI to help with our long-term goal of bringing more automation into the programming process, and several of our projects over the past year have explored this direction. For example, in a paper published in the 2019 NeurIPS proceedings, we demonstrated a new approach toward neural guided program synthesis, where a neural network learns to construct a program incrementally, by observing the behavior of the partially constructed program as it goes along.

One intriguing aspect of this research direction is the way in which programming can serve as a model for other cognitively demanding tasks that require significant

education and training, such as writing, mathematics, and science. For example, one characteristic of such tasks is that they require a combination of explicit declarative knowledge with implicit procedural knowledge. Declarative knowledge is the kind of knowledge one learns from books and in the classroom; for example, knowing about dynamic programming and how it works. Implicit procedural knowledge, in contrast, is harder to characterize and is often built through years of experience. For example, procedural knowledge is what allows a skilled programmer to hear a problem description and immediately realize that this problem could be solved with dynamic programming. As part of the DreamCoder project—in collaboration with Professor Josh Tenenbaum and graduate student Kevin Ellis—we have been exploring how a machine could acquire and represent these different kinds of knowledge in the process of solving progressively more challenging programming problems. But the interplay between programming systems, ML, and AI goes beyond automated programming; one of the most exciting directions in the past year has been the application of ideas from program synthesis to help build models that are more interpretable and predictable than what can be generated with traditional neural networks. For example, in a 2020 ICLR conference paper, we showed how a combination of numerical optimization and program synthesis could be used to discover control policies for repetitive behavior that can generalize much better than policies learned through deep reinforcement learning.

In April of this year, we launched a National Science Foundation-funded expeditions project that aims to apply these ideas to accelerate scientific discovery. The goal of our project is to develop new learning techniques that can help automate this process of generating scientific theories from data. In particular, we are working to develop methods for learning neurosymbolic models that combine neural elements capable of identifying complex patterns in data with symbolic constructs that are able to represent higher level concepts. Our approach is based on the observation that programming languages provide a uniquely expressive formalism to describe complex models. Our aim is therefore to develop learning techniques that can produce models that look more like the models that scientists already write by hand in code. Our proposed techniques will more easily incorporate prior knowledge about the phenomena we want to model and produce interpretable models that can be analyzed to devise new experiments or to infer causal relations.

Laboratory Sponsored Activities

CSAIL Outreach

CSAIL regularly encourages the online community to submit questions about computer science and academia to its researchers in a series of Reddit "Ask Me Anything" (AMA) sessions. CSAIL's AMAs have spurred approximately 8,000 comments and questions, as well as more than 300,000 page views.

Media Outreach

CSAIL has a combined online following of 250,000 users across Twitter, Instagram, Facebook, LinkedIn, and YouTube; with growth in media coverage and viewership at a 900% increase in media hits since 2013 (567 vs. 53). The amount of CSAIL video that YouTube users have watched is now more than two million minutes, which is equal

to more than four years. Our YouTube viewership comparison with MIT News is as follows: CSAIL—10 videos, with an average of 37,000 views; MIT News—70 videos, with an average of 13,000 views.

CSAIL Hosted Lecture Series

Dertouzos Distinguished Lectures have been a tradition since 1976, featuring some of the most influential thinkers in computer science. Two speakers presented lectures during the AY2020 Dertouzos Distinguished Lecture Series:

- Yoshua Bengio, University of Montreal; Mila: "Learning High Level Representation for Agents," September 18, 2019
- David Patterson, University of California at Berkeley; Google: "Domain Specific Architectures for Deep Neural Networks: Three Generations of Tensor Processing Units (TPUs)" October 16, 2019

Hot Topics in Computing is a speaker series initiated by CSAIL in 2017, which convenes experts in computing to discuss emergent potential, perception, and problems associated with the proliferation of computation and machines. We held a much higher number of events (12 events in AY2020), in order to address coronavirus and COVID-19 related topics; see below:

- Dava Newman, Aeronautics and Astronautics, MIT: "Saving Spaceship Earth," October 2, 2019
- Professor Scott Aaronson, University of Texas at Austin: "Quantum Computational Supremacy and Its Applications," February 6, 2020
- Professor Michael Lin, Stanford University: "Coronaviruses and COVID-19: Basic Biology Behind the Epidemic," March 20, 2020
- Professors Ankur Moitra and Elchanan Mossel, Mathematics, MIT: "An Invitation to Computational Epidemiology," March 24, 2020
- Yaneer Bar-Yam, Media Lab, MIT: "Pandemic Updates and Recommended Actions," March 26, 2020
- Visiting Professor Esteban Moro, and Professor Alex Pentland, Media Lab, MIT: "Effectiveness of Social Distancing Strategies for Protecting a Community from Pandemic," March 31, 2020
- Amar Gupta, visiting scientist, MIT: "Telehealth and Coronavirus," April 7, 2020
- Professor Julie Shah, CSAIL, MIT: "Using Computing to Address Social Impacts of COVID-19," April 14, 2020
- Professor Retsef Levi, Sloan School of Management, MIT: "Targeted Risk Analysis to Fight COVID-19 Outbreak," May 5, 2020
- Brad Smith, president, Microsoft Corporation: "Conversation with Microsoft President Brad Smith," May 14, 2020

- Professor Ron Rivest, Associate Professor Julie Shah, and Daniel Weitzner, CSAIL, MIT; Professor Ramesh Raskar, Media Lab, MIT: Panel discussion on contact tracing activities at MIT, May 19, 2020
- Institute Professor Daron Acemoglu, Economics, MIT: "Multi-Risk SIR Model with Optimally Targeted Lockdown," June 09, 2020

CSAIL also co-hosted TEDxMIT/Operation Earth, held on December 6, 2019, featuring a series of 16 talks from researchers and faculty across the Institute.

Organizational Changes

Professor Daniela Rus has continued in her role as director of CSAIL and has also been named Deputy Dean of Research in the Schwarzman College of Computing. CSAIL director's duties include developing and implementing strategies designed to keep CSAIL growing and evolving, fund raising, determining laboratory policies, and examining promotion cases.

CSAIL's leadership team includes an associate director and a chief operating officer (COO), and the executive cabinet. These leaders are appointed by the laboratory's director and assist her with her duties. Professors Daniel Jackson and Charles Leiserson were the FY2020 associate directors. Professor Leiserson also served as chief operating officer, providing leadership and strategy for how we conduct our operations and initiatives, enabling the director to allocate more time to strategic planning. Victor Zue held the role of director of international relations, managing the engagements and oversight of various important CSAIL international contracts and negotiations.

The CSAIL executive cabinet met weekly to review and advise the director on policy, processes, activities within the laboratory, preparation for transitions related to the Schwarzman College of Computing. Members of the FY2020 executive cabinet included: Ted Adelson, Saman Amarasinghe, Randall Davis, Jim Glass, Daniel Jackson, Charles Leiserson, Sam Madden, Wojciech Matusik, Ronitt Rubinfeld, Daniela Rus, Bruce Tidor, and Victor Zue.

The CSAIL Enterprise Services team manages lab operations. There are seven units—Administrative Assistants, CSAIL Alliance Program, Communications, Finance, HR, Special Projects, and the Infrastructure Group—reporting to the CSAIL COO on all operational matters. Carmen Finn is the assistant director for Administration overseeing the Finance, HR operations, and Administrative Assistants team and serves on the Space Committee; Lori Glover is managing director of the CSAIL Alliance Program. Jack Costanza is the assistant director for infrastructure, overseeing information technology infrastructure and user support, building operations, and communications, and serves on the space committee.

Bruce Tidor oversees the space committee and manages the allocation of space within CSAIL. The space committee also implements improvements to the facilities that will increase the quality of the environment for the laboratory's faculty, staff, and students. The space committee also includes Assistant Director for Infrastructure Group Jack Costanza and Assistant Director for Administration Carmen Finn.

New faculty starting during AY2021 include:

- Assistant Professor Jonathan Ragan-Kelley, January 1, 2020
- Assistant Professor Henry Corrigan-Gibbs, July 1, 2020
- Assistant Professor Anand V. Natarajan, fall 2020
- Assistant Professor Sam Hopkins, January 1, 2021
- Assistant Professor Yoon Kim, winter 2021
- Assistant Professor Ashia Wilson (housed in LIDS), January 1, 2020
- Assistant Professor Dylan Hadfield-Menell, July 1, 2021

Faculty taking leave during AY2020 included:

- Mohammad Alizadeh, parental leave spring 2020
- Hari Balakrishnan, personal leave fall 2019 to fall 2020
- Tim Berners-Lee, sabbatical fall 2019
- Adam Chlipala, sabbatical fall 2019
- Shafi Goldwasser, professional leave fall 2019; personal leave spring 2020 to fall 2020
- Polina Golland, sabbatical spring 2020 to fall 2020
- Tommi Jaakkola, sabbatical fall 2019
- Tim Kraska, research leave spring 2020
- Nancy Lynch, sabbatical fall 2019 to fall 2020
- Sam Madden, personal leave spring 2020
- Silvio Micali, professional leave fall 2019
- Stefanie Mueller, research leave fall 2019
- Nir Shavit, personal leave fall 2019 to fall 2020
- Vinod Vaikuntanathan, sabbatical fall 2019 to fall 2020

Faculty changes and promotions during AY2020 included:

- Associate Professor Justin Solomon, July 1, 2020
- Full Professor Aleksander Madry, July 1, 2020
- Full Professor Armando Solar-Lezama, July 1, 2020
- Full Professor Richard Ryan Williams, July 1, 2020

Other PI promotion and hire

- Aude Oliva, promoted to senior research scientist, July 1, 2020
- Michael Cafarella, new hire, principal research scientist, June 15, 2020

Awards and Honors

Our faculty and staff have achieved many awards, including the following:

- Edward Adelson: 2020 Ken Nakayama Medal, Vision Sciences Society
- Saman Amarasinghe: 2019 fellowship, ACM
- Hari Balakrishnan: 2020 fellow, Institute of Electrical and Electronics Engineers
- Bonnie Berger: 2019 senior scientist, International Society for Computational Biology; and 2020 member, National Academy of Sciences
- Michael Carbin: 2019 Distinguished Paper, International Conference on Functional Programming; 2020 Sloan Research Fellow, Alfred P. Sloan Foundation 2020 Frank E. Perkins Award for Excellence in Graduate Advising, EECS/MIT; and 2020 research award, Facebook
- Joel Emer: 2020 member, National Academy of Engineering
- William Freeman: 2020 fellow, Association for the Advancement of Artificial Intelligence
- Stefanie Mueller: 2020 Sloan Research Fellow, Alfred P. Sloan Foundation; and 2020 Faculty Fellow, Microsoft Research
- Ronitt Rubinfeld: 2020 member, American Academy of Arts and Sciences
- Daniela Rus: 2019 Mass TLC Distinguished Leadership Award; 2019 Constantin Brâncoveanu International Award, Alexandrion Foundation; and 2020 IJCAI John MacCarthy Award
- Peter Shor: 2020 member, National Academy of Engineering
- Armando Solar-Lezama: 2020 Expeditions in Computing grant, National Science Foundation
- Joshua Tenenbaum: 2019 MacArthur Fellowship; and 2020 member, American Academy of Arts and Sciences
- Daniel Weitzner: 2019 fellow, National Academy of Public Administration

Key Statistics for AY2020

Category	Count	Women (%)
Faculty	108	18%
Postdoctoral associate/fellow	90	19%
Principal research scientist	10	50%
Research staff	27	26%
Senior research scientist	5	40%
Administration, technical, and support staff	80	63%
Graduate students	591	26%
Undergraduate Research Opportunities Program students	194	39%
Visitors	82	22%
Total	**1,187**	**29%**

Daniela Rus
Director, Computer Science and Artificial Intelligence Laboratory