WAM 2022 Yearbook

The Mathematics of Machine Learning

(hybrid)

Institute for Advanced Study, Princeton

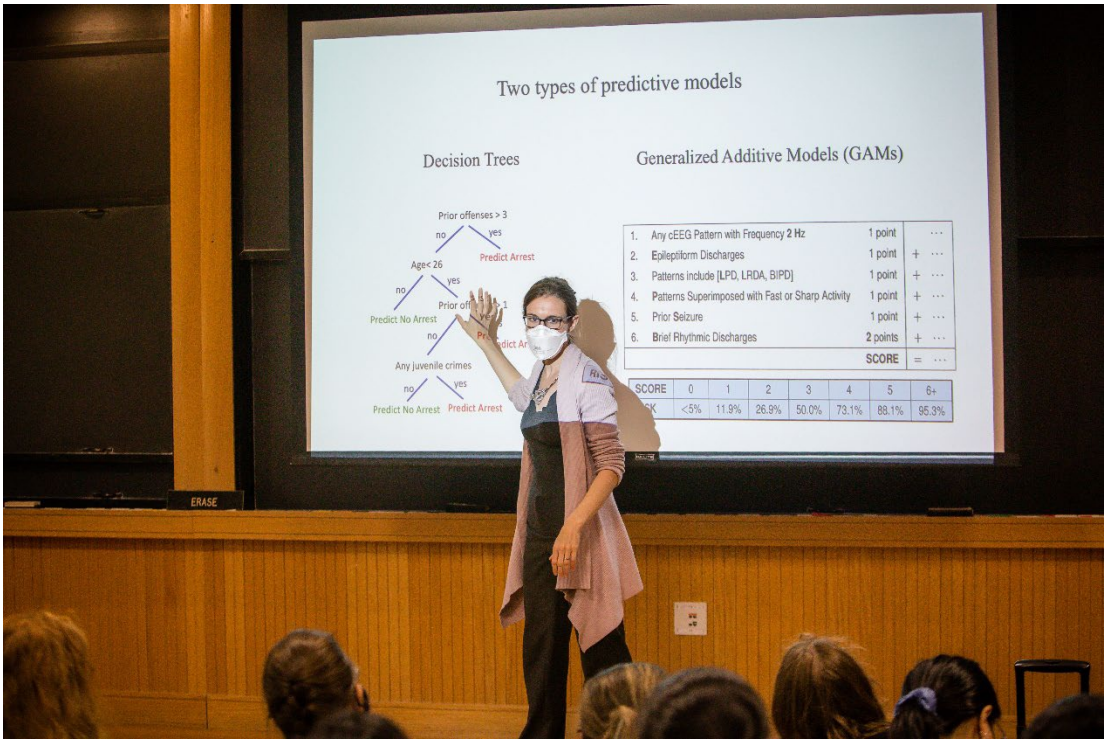
May 21-27, 2022

This is a compilation of activities and resources contributed by participants during the 2022 Women and Mathematics Program. We hope this can serve as a mathematical and professional reference guide for women mathematicians around the country.

Mathematical Talks

Terng Lectures:

Cynthia Rudin, Duke University, “Introduction to Interpretable Machine Learning”

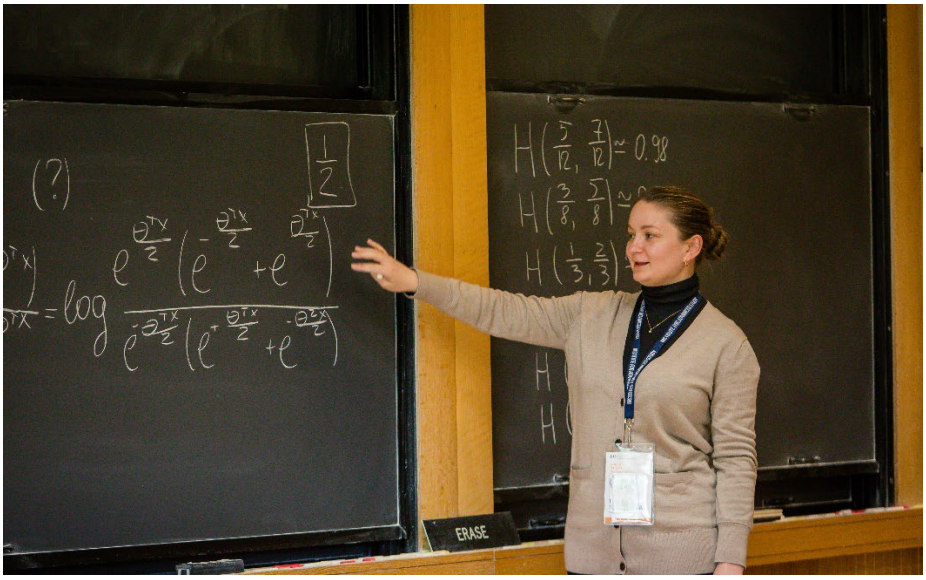


Abstract: Machine learning is now used throughout society and is the driving force behind the accuracy of online recommendation systems, credit-scoring mechanisms, healthcare systems and beyond. Machine learning models have the reputation of being "black boxes" meaning that their computations are so complicated that no human would be capable of understanding them. However, machine-learning models do not actually need to be black boxes. It is possible, with some mathematical sophistication, to derive algorithms for machine learning that produce models that are interpretable by humans.

C Rudin presented a self-contained short course that introduces the basic concepts of supervised machine learning, including loss functions, overfitting, and regularization, and how interpretability is a form of regularization. The focus was on two important topics in interpretable machine learning: (1) sparse decision trees and (2) interpretable neural networks. For both topics, a brief history of the field was presented, going all the way to

the state-of-the-art in current methodology, and introduce specific technical and mathematical challenges.

Terng problem sessions were run by Lesia Semenova, Duke University.



Lectures were recorded and can be found here: https://www.ias.edu/video-tags/2022-program-women-and-mathematics-mathematics-machine-learning

Uhlenbeck Lectures:

Maria Florina Balcan, Carnegie Mellon University, “Foundations for Learning in the Age of Big Data”

Abstract: With the variety of applications of machine learning across science, engineering, and computing in the age of Big Data, re-examining the underlying foundations of machine learning has become imperative. In this lecture, N Balcan described new models and algorithms for important emerging paradigms, specifically, interactive learning and distributed learning.

Most classic machine learning methods depend on the assumption that humans can annotate or label all the data available for training. However, many modern machine-learning applications have massive amounts of unannotated or unlabeled data. As a consequence, there has been a lot of interest recently in machine learning and its application to design active learning algorithms that intelligently select which data to request to be labeled, with the goal of dramatically reducing the human labeling effort. Recent advances on designing active learning algorithms with provable guarantees that are computationally efficient, noise tolerant, and enjoy strong label efficiency guarantees were discussed.

Also discussed was the problem of learning from distributed data and analyze fundamental algorithmic and communication complexity questions involved. Broadly, we considered a framework where massive amounts of data are distributed among several locations. Our goal was to learn a low-error predictor with respect to the overall distribution of data using as little communication and as few rounds of interaction as possible. Lastly, N Balcan discussed general upper and lower bounds on the amount of communication needed to learn a given class, as well as broadly applicable techniques for achieving communication-efficient learning.

Uhlenbeck Problem Sessions were run by Ellen Vitercik, University of California, Berkeley.

Lectures were recorded and found here: (scroll for dates and lectures)
https://www.ias.edu/video-tags/2022-program-women-and-mathematics-mathematics-machine-learning

References for both courses:

1. The classic book Machine Learning, Tom Mitchell, McGraw Hill, 1997.

http://www.cs.cmu.edu/~tom/mlbook.html

2. An introduction to computational learning theory, Kearns and Vazirani

https://mitpress.mit.edu/books/introduction-computational-learning-theory

3. Statistical learning theory, by Vapnik

https://www.amazon.com/Statistical-Learning-Theory-Vladimir-Vapnik/dp/0471030031

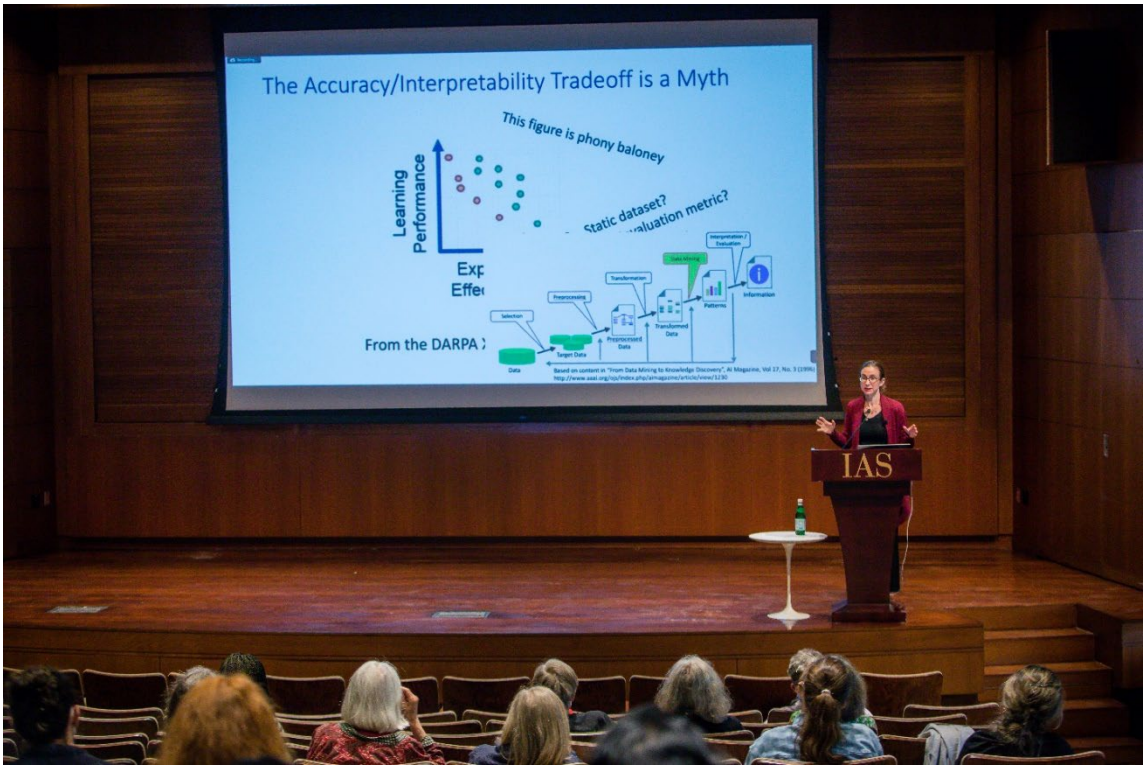
4. https://users.cs.duke.edu/~cynthia/teaching.html

Colloquium

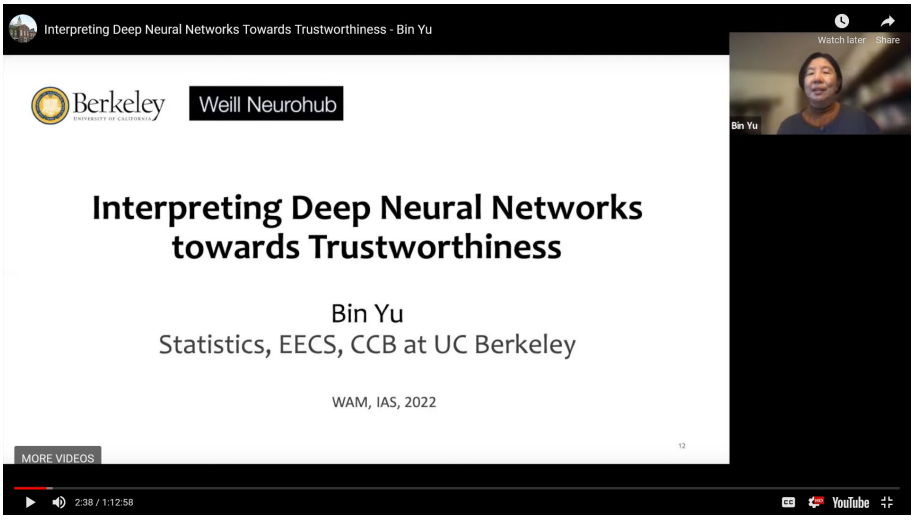
Kate Saenko, Boston University, "Learning Generalizable Visual Representations". Recording of the talk can be found here: https://www.ias.edu/video/learning-generalizable-visual-representations

Special Talks

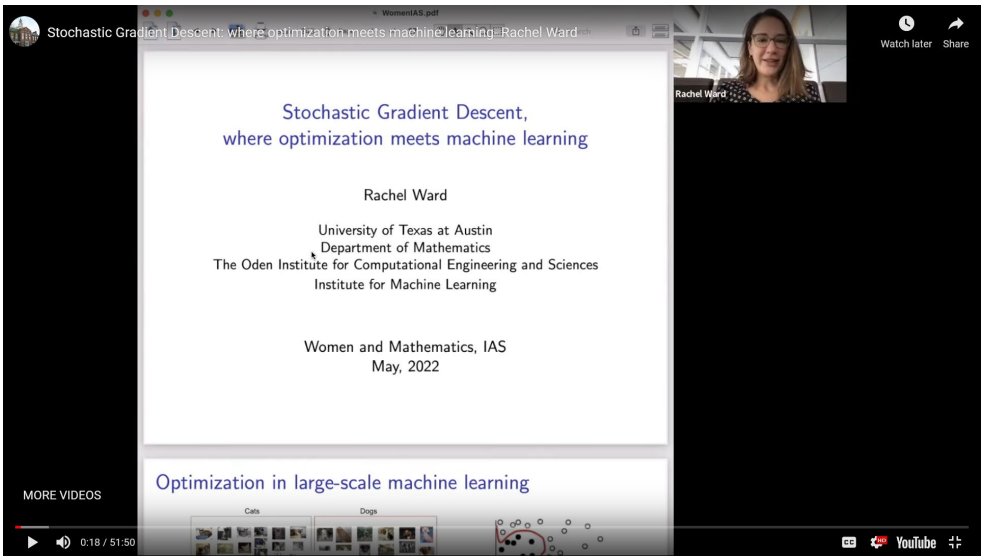
Public Lecture by Cynthia Rudin, Duke University, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead” https://www.ias.edu/video/stop-explaining-black-box-machine-learning-models-high-stakes-decisions-and-use-interpretable



Bin Yu, University of California, Berkeley, “Interpreting Deep Neural Networks towards Trustworthiness”. Recording of the talk here: https://www.ias.edu/video/interpreting-deep-neural-networks-towards-trustworthiness



Rachel Ward, University of Texas at Austin, "Stochastic Gradient Descent: where optimization meets machine learning" https://www.ias.edu/video/stochastic-gradient-descent-where-optimization-meets-machine-learning



E. **Young Researcher Seminar**

Rachel Freedman, University of California, Berkeley
Title: Objective Misspecification and Value Alignment
Abstract: As machine learning (ML) systems become increasingly capable and general, it becomes increasingly important to ensure that their behavior aligns with human values. However, specifying reward functions for ML agents that operate in environments without a natural reward signal can be challenging, and incorrectly specified rewards can incentivise degenerate or dangerous behavior. A promising alternative to manually specifying reward functions is to enable ML systems to infer them from human interaction and feedback, but this is sensitive to incorrectly specified models of human decision-making. In this talk, I'll discuss the *objective misspecification problem* and my work to improve reward inference in the face of human model misspecification.

Jessica Gronsbell, University of Toronto
Title: Statistical learning with high volume, high noise health data
Abstract: In this talk, I will give a brief overview of my background and my path to statistics and machine learning. I will also discuss my current research on the development of statistical learning methods for large observational health data sets such as electronic health

records. I will emphasize the importance of knowing your data and embracing its complex structure to make analysis more trustworthy and efficient.

Jamie Haddock, Harvey Mudd College
Title: Tensor Models, Methods, and Medicine
Abstract: There is currently an unprecedented demand for efficient, quantitative, and interpretable methods to study large-scale (often multi-modal) data. One key area of interest is that of tensor decomposition, which seeks to automatically learn latent trends or topics of complex data sets, providing practitioners a view of what is "going on" inside their data. This talk will survey tools for topic modeling on matrix and tensor data. These tools are of interest across the many fields and industries producing, capturing, and analyzing big data, but are of particular interest in applications where expert supervision is available and often essential (e.g., medicine). We will describe an application of these methods to medical data; an ongoing application to cardiovascular imaging data.

Longxiu Huang, University of California, Los Angeles
Title: Matrix Completion with Cross-Concentrated Sampling
Abstract: In matrix completion, uniform sampling has been widely studied and CUR sampling can be applied to approximate a low-rank matrix via row and column samples. Unfortunately, both sampling models are lack of flexibility for various circumstances in real-world applications. Recently, we propose a novel and easy-to-implement sampling strategy, coined Cross-Concentrated Sampling (CCS). By bridging uniform sampling and CUR sampling, CCS provides extra flexibility that can potentially save sampling costs in applications. Moreover, we propose a highly efficient non-convex algorithm, termed Iterative CUR Completion (ICURC),

for the proposed CCS model. In this talk, I will show the efficiency of our methods on both synthetic and real-world datasets.

Harlin Lee, University of California, Los Angeles
Title: Identifying future connections in machine learning topics
Abstract: The field of artificial intelligence (AI) and machine learning (ML) is exponentially growing with no sign of waning popularity. As the discipline gets larger, it is impossible for an individual researcher to be familiar with the entire body of literature. This forces them to specialize in a sub-field. Such insulation can hinder the birth of ideas that arise from new connections, eventually slowing down scientific progress. As such, discovering fruitful interdisciplinary connections by analyzing scientific publications is an important problem in the science of science, or metascience. We present dynamic-embedding-based methods for link prediction in a ML semantic network, where the nodes are concepts in machine learning, and the time-stamped edges indicate co-occurrence in scientific papers.

Alane Lima, Federal University of Parana
Title: Statistical Learning Theory and Sampling Algorithms for Complex Networks
Abstract: When dealing with problems in large scale graphs, using an exact algorithm may be inefficient in practice. A key idea in our work is to use a sampling-based approximation algorithm that takes into consideration the combinatorial structure of the problem. Such algorithms may achieve tighter sample size bounds in comparison with strategies that use classical bounds of Probability theory. The study of the sample complexity analysis of a particular problem, which uses tools that are in the core of statistical learning theory, such as VC-dimension, pseudo-dimension, and Rademacher complexity, relates the minimum size that is required for a particular problem and instance to attain desired parameters of quality and confidence of the solution. When applied in the design of approximation algorithms for well-known problems in graphs (e.g. shortest paths and a variety of centrality measures), this approach leads to efficient solutions for problems of practical interest.

Anna Ma, University of California, Irvine
Title: Approaches for working with Large-Scale Data
Abstract: Data is available at larger scales than ever before seen, making it an important task to be able to implement and understand the tools and techniques that utilize this data. For example, if data is too large to fit into a computer's memory, how can one use it to train a machine learning algorithm? What if you have the ability to pick and choose which data you want to utilize? How can you pick the best data for your problem? Also, what if your data goes beyond the standard data-as-a-matrix setting and you have multidimensional data? What can be done then? In this brief talk, we will speak broadly to these questions and discuss motivations, approaches, and open directions for working with large-scale data.

Mirjeta Pasha, Arizona State University
Title: On Deterministic and Statistical Methods for Large-scale Dynamic Inverse Problems
Abstract: Inverse problems are ubiquitous in many fields of science such as engineering, biology, medical imaging, atmospheric science, and geophysics. Three emerging challenges on obtaining relevant solutions to large-scale and data-intensive inverse problems are ill-posedness of the problem, large dimensionality of the parameters, and the complexity of the model constraints. In this talk we discuss efficient methods for computing solutions to dynamic inverse problems, where both the quantities of interest and the forward operator may change at different time instances. We consider large-scale ill-posed problems that are made more challenging by their dynamic nature and, possibly, by the limited amount of available data per measurement step. In the first part of the talk, to remedy these difficulties, we apply efficient regularization methods that enforce simultaneous regularization in space and time (such as edge enhancement at each time instant and proximity at consecutive time instants) and achieve this with low computational cost and enhanced accuracy [2]. In the remainder of the talk, we focus on designing spatial-temporal Bayesian models for estimating the parameters of linear and nonlinear dynamical inverse problems [1]. Numerical examples from a wide range of applications, such as tomographic reconstruction, image deblurring, and chaotic dynamical systems are used to illustrate the effectiveness of the described approaches.

1. S. Lan, S. Li, and M. Pasha. Bayesian spatiotemporal modeling for inverse problems. in preperation, 2022.

2. M. Pasha, A. K. Saibaba, S. Gazzola, M. I. Espanol, and E. de Sturler. Efficient edge-preserving methods for dynamic inverse problems. arXiv preprint arXiv:2107.05727, 2021.

Courtney Paquette, McGill University
Title: Optimization Algorithms in the Large
Abstract: In this talk, I will present a framework, inspired by random matrix theory, for analyzing the dynamics of optimization algorithms (e.g., 1^{st}-order methods, stochastic gradient descent (SGD), and momentum) when both the number of samples and dimensions are large. Using this new framework, we show that the dynamics of optimization algorithms on a least squares problem with random data become deterministic in the large sample and dimensional limit. In particular, the limiting dynamics for stochastic algorithms are governed by a Volterra integral equation. This model predicts that SGD undergoes a phase transition at an explicitly given critical stepsize that ultimately affects its convergence rate, which we also verify experimentally. Finally, when input data is isotropic, we provide explicit expressions for the dynamics and average-case convergence rates. These rates show significant improvement over the worst-case complexities.

Sui Tang, University of California, Santa Barbara
Title: Data-driven discovery of interaction laws in multi-agent systems
Abstract: Multi-agent systems are ubiquitous in science, from the modeling of particles in Physics to prey-predator in Biology, to opinion dynamics in economics and social sciences, where the interaction law between agents yields a rich variety of collective dynamics. We consider the inference problem for a system of interacting particles or agents: given only observed trajectories of the agents in the system, can we learn what the laws of interactions are? We would like to do this without assuming any particular form for the interaction laws, i.e. they might be “any” function of pairwise distances.

Melanie Weber, University of Oxford
Title: Geometric Methods for Machine Learning and Optimization
Abstract: A key challenge in machine learning and optimization is the identification of geometric structure in high-dimensional data. Such structural understanding is of great value for the design of efficient algorithms and for developing fundamental guarantees for their performance. This talk will discuss research on utilizing Riemannian geometry in machine learning and data science, motivated by the observation that many applications involve non-Euclidean data, such as graphs, strings, or matrices. First, we consider the task of learning a classifier in hyperbolic space. Such spaces have received a surge of interest for representing large-scale, hierarchical data, since they achieve better representation accuracy with fewer dimensions. Secondly, we consider the problem of optimizing a function on a Riemannian manifold. Specifically, we will consider classes of optimization problems where exploiting

Riemannian geometry can deliver algorithms that are computationally superior to standard (Euclidean) approaches.

Kara Ycoubou Djima, Amherst College
Title: Extracting Autism's Biomarkers in Placenta using Multiscale Methods
Abstract: The placenta is the essential organ of maternal-fetal interactions, where nutrient, oxygen, and waste exchange occur. In recent studies, differences in the morphology of the placental chorionic surface vascular network (PCSVN) have been associated with developmental disorders such as autism. This suggests that the PCSVN could potentially serve as a biomarker for the early diagnosis and treatment of autism. Studying PCSVN features in large cohorts requires a reliable and automated mechanism to extract the vascular networks. In this talk, we present a method for PCSVN extraction. Our algorithm builds upon a directional multiscale mathematical framework based on a combination of shearlets and Laplacian eigenmaps and can isolate vessels with high success in high-contrast images such as those produced in CT scans.

Outreach

Math Carnival - https://princetonlibrary.libnet.info/event/6491644

This year's outreach for WAM featured a Math Carnival organized by WAM Committee member Margaret Readdy assisted by WAM participants Princess Allotey, Sarah Brown, Nischita Kaza, Amber Lee, Anna Ma, Mansi Sood and Ria Stevens. The Math Carnival, held May 22, 2022 at the Princeton Public Library, attracted over seventy-five participants.

The seven activities included two from the area of Machine Learning: Quick Draw! demonstrates how a neural network recognizes doodles and Magic Sketchpad allows users to see if machine learning can correctly finish your drawing. The carnival also included challenging and unconventional math mazes from the game designer Robert Abbott who as featured in a recent issue of the Mathematical Intelligencer.


Playing Ninja

WAM Ambassadors

More information on how to apply for the WAM Ambassadorships please go here: https://www.ias.edu/math/wam/apply/ambassador-program

For the reports from the Ambassadorships for 2021-2022 please go here to watch the video summaries of the events: https://www.ias.edu/math/wam/program-years/2022-program-women-and-mathematics/2022-ambassador-reports


SARAH
BROWN


DUSA
MCDUFF


NISCHITA
KAZA