Stats News 2024-2025

Department of Statistics

HARVARD
Faculty of Arts and Sciences

Table of Contents

Letter from the Chair

Dr. Samuel Kou,

Chair of the Department

Greetings Statistics Community,

I open this letter by acknowledging that this is a very challenging time for our department and for our students, postdocs, faculty, and staff. Please know that we are committed, as a department, to supporting every member of our community and helping each of you achieve your academic and professional goals. The resilience of our community and the strength of our educational and research mission will guide us through these turbulent times.

Reflecting on the accomplishments of our students, postdocs, faculty, and staff over the past year, I feel honored to be part of this community and to have served as chair for the past three years. Our scholars are developing statistical methods that make a significant impact across fields. You can read more about their work in this newsletter, including stories about PhD student Souhardya Sengupta discovering an alternative to the t-test, Biyonka Liang designing a model for improved resource allocation in public health settings, and Professor Susan Murphy's development of a healthcare app built to increase patient adherence to medication regimens. In addition, the Awards section highlights the many recent accolades received by department members, including the following departmental awards: the Concentrator Prize (Kenneth Gu), the Concurrent Master's Prize (Danielle Paulson), and the Dempster Prize (Yufan Li and Tianle Liu).

Thanks to the innovative instruction and close mentorship of our faculty, our students are honing their analytical, critical-thinking, and creative problem-solving skills—preparing them for the future. Through student perspectives in this newsletter, you'll see common themes emerge: the impact of peers and faculty mentors on academic and personal growth and their growing appreciation for the power of statistics in the world. When asked to capture their Harvard Statistics experience, students described it as "meaningful," "empowering," and "inspiring." I couldn't agree more. Through our scholars' own words, we get a snapshot of our vibrant student-faculty connections, empowering pedagogy, and inspiring scholarship.

Beyond these stories and interviews, I would like to share some important updates about our undergraduate, master's, and PhD programs and other departmental initiatives.

Undergraduate News

A visiting committee of external faculty aptly described our undergraduate program as "arguably the best statistics program in the nation," noting that "its success has become an inspirational example for peer institutions." Under the leadership of Joe Blitzstein (Professor of the Practice and Director of Undergraduate Studies) and Kevin Rader (Senior Preceptor and Associate Director of Undergraduate Studies), the concentration continues to grow:

- Our numbers increased to 300 concentrators this year (compared to 187 in 2021).
- We had our largest cohort of thesis writers with 43 students, nine of whom were showcased in lightning thesis talks on May 2.
- Our course enrollments, the majority of which are undergraduate students, totaled 3,267.

We continue to refine and improve our curriculum offerings, including the introduction of a new Machine Learning (ML) track (in addition to our three other tracks). This addition reflects both student interest and the importance of statistical applications in ML and AI.

PhD News

Our PhD program continues to attract and support some of the most talented students from across the U.S. and the world. Led by Joint Directors of Graduate Study, Professor Natesh Pillai and Professor Susan Murphy (Director of Graduate Admissions), the program prepares students to be leading researchers and statisticians in academia and industry through rigorous core coursework and close faculty mentorship.

A highlight of the program is our weekly Stat 300 Research in Statistics seminar, where students present their work. Here are some milestones:

- We welcomed 4 incoming PhD students this year: Nicholas Barnfield, Aniket Jain, Zimeng Li, and Théo Voldoire (profiled in this newsletter).
- We graduated 6 PhD students in 2024: Louis Cammarata, Dieyi Chen, Dae Woong Ham, Buyu Lin, Yue Liu, and Jiaze Qiu (see their profiles in this newsletter), and are graduating 13 in 2025: James Bailie, Alexandre Bayle, Qizhao Chen, Zeyang Jia, Kuanhao Jiang, Yicong Jiang, Yufan Li, Biyonka Liang, Tianle Liu, Xiang Meng, Lisa Ruan, Yanke Song, and Yi Zhang.
- We hosted 30 PhD student research talks in 2025.

AM News

Under the leadership of Mark Glickman, Senior Lecturer and Director of Master's Study, the AM program has grown to include 57 undergraduates and four PhD students from other programs pursuing the degree concurrently. This is a substantial increase and is a testament to Dr. Glickman's stewardship, as well as to the value of an advanced degree in statistics across industries—whether finance, tech, life sciences, or government.

Community News

I would also like to recognize the work of key groups and committees that support inclusion

and community engagement, including our Equity, Diversity, Inclusion, and Belonging (EDIB) committee; GUSH (Group for Undergraduates in Statistics at Harvard); HSGC (Harvard Statistics Graduate Council); Women in Statistics; and the Data Adventure Planning Committee.

This year, the EDIB committee's goals have been to create better support systems for all our students and to organize more community-oriented events. Highlights include:

- The introduction of the Undergraduate Colloquium, which showcases different careers and applications of statistics;
- Stat Night, which helps undergraduates form study groups;
- Postdoctoral networking dinners; and
- A Peer Concentration Advisor program, which offers peer-to-peer mentorship.

Additionally, for the third straight year, we hosted Data Adventure Day, an all-day event with statistics and data science activities for about 120 high school students (and roughly 50 teachers and volunteers) from Boston and Cambridge public schools.

The student-led groups in our department—GUSH, HSGC, and Women in Statistics—continue to play an integral role in fostering community and providing professional development opportunities. HSGC sponsored social events throughout the year for our graduate students and organized a popular annual PhD retreat that combines faculty research talks with industry networking. GUSH offered a variety of events, including a graduate school panel (co-hosted with Women in Statistics) and a Women in Data Science panel. Women in Statistics hosted lunch and learn sessions with professionals from both academia and industry.

As part of our ongoing efforts to further strengthen the statistics community, I am pleased to announce that we will be moving to the Maxwell-Dworkin building this winter. This strategic move will consolidate our department from six floors to two, facilitating collaboration and enhancing interpersonal connections.

While this letter and newsletter attempt to capture the depth of accomplishments and talents within our community, they can only offer a glimpse of the whole. As you read its contents, I hope you find them as inspiring and rewarding as I do. In my final month as Chair, I want to express my deep admiration and gratitude to my colleagues, as well as my pride in the achievements of our students, postdocs, and alumni. I look forward to witnessing the new advancements and initiatives that will shape the next chapter of the Statistics Department under new leadership.

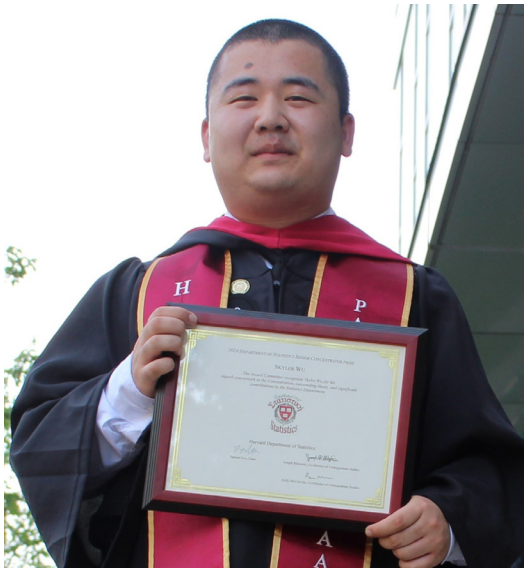
Sincerely,

Samuel Kou, Professor

Chair of the Department

Profile of Skyler Wu

May 2024 Senior Concentrator Prize Winner



In 2024, alum Skyler Wu received the 2024 Department of Statistics Senior Concentrator Prize at our Commencement Celebration on May 23, 2024. Wu graduated with an AB in Statistics and Mathematics (Joint) and an SM in Applied Mathematics. He was selected for the prize by the department based on the high quality of his coursework and thesis. After completing his degrees at Harvard, he started the next chapter in his statistical journey by joining the Statistics PhD program at Stanford University.

To learn more about Skyler Wu's challenges and triumphs in statistics as well as his thoughts on what he will miss most about the department, we had the following conversation with him, edited and excerpted below. Congratulations, Skyler!

Was there an experience that made math or statistics personally engaging?

Wu: Honestly, I entered college with a feeling of imposter syndrome (and COVID didn't help!) because everyone seemed to be on such a different level. When I heard about students placing into Math 55 [honors abstract algebra] and representing their countries in math competitions during high school, I wondered whether I belonged.

The experience that changed my perspective and gave me more confidence was taking Stat 110 Probability with Professor Joe Blitzstein. Although I originally signed up for Stat 110 because I figured I should fill a prerequisite (Stat 110 is a prerequisite for many STEM concentrations) while being cooped up inside, to my surprise, I ended up loving the class! Interacting with Professor Blitzstein and amazing teaching fellows (TFs) like Rachel Li (AB '23), Ginnie Ma (AB '23), and Yash Nair (AB '22) made me feel like I belonged and had a seat at the table.

During freshman spring, my interest in statistics and math really cemented when I took CS 181 Machine Learning. There was a lot of buzz in the air about machine learning, so I was curious and wanted to explore it. I quickly realized that what people call "machine learning" or "artificial intelligence" is really applied statistics and applied math. This core realization compelled me to pursue a path of studying statistics.

What initial memories do you have of getting to know the Stats Department?

Wu: As I mentioned, one of the most formative experiences was Stat 110, even though it was over Zoom. During COVID, we weren't allowed to interact in person much, but through Stat

110, I was able to meet a lot of phenomenal TFs virtually—there were basically office hours around the clock—who made me feel supported and welcomed in the Statistics Department. Of course, Professor Blitzstein, as the instructor of Stat 110, really reinforced this positive experience. I even told Professor Blitzstein that, before coming to Harvard, I held an image of faculty far removed, sitting in lofty ivory chairs. But Professor Blitzstein didn't match my image at all; instead, he directly interfaced with undergraduates in a manner that was very approachable, friendly, and supportive. He helped me transition into college and the stats department.

What motivated you to pursue a Statistics Concentration?

Wu: At first, I thought I would pursue a career in science policy and diplomacy. I wanted to combine technical skills with a background in policy to address big global challenges. Eventually, I realized that nearly every discipline requires the tools to work with data, understand uncertainty, and predict future patterns; they require the common backbone of statistics. I first heard the saying "The best thing about being a statistician is that you get to play in everyone's backyard" [attributed to the statistician John Tukey] from some of my professors, and it really hit home for me. Regardless of your field, statistics is a common denominator, which made it a darn good option to pursue as a concentration!

I'm also very interested in machine learning, but I like approaching it from a statistical angle. When examining a machine learning model from a statistics perspective, I appreciate that you need to be explicit about the assumptions that you make about the model and its data. Another reason why the concentration was a good fit was because I feel naturally drawn to questions such as, "To what extent do we believe this prediction?" and "What is the level of uncertainty about our results?"

In addition, I like the beauty and unity that statistics offers. For example, in Stat 230 [Multivariate Statistical Analysis], Professor Sam Kou (also my thesis advisor) introduced some modern statistical techniques during the last third of the semester, including principal component analysis, canonical correlation, and critical angles. Professor Kou emphasized that all these statistical tools derived from the same exact underlying theorem, which reveals the kind of beautiful unity and organization in statistical thinking that I so admire.

Did you encounter challenges (personal or academic) during your studies? How did you overcome these challenges?

Wu: One of my challenges was taking CS 181 Machine Learning during my freshman spring. Initially, I felt like I bit off more than I could chew; I barely knew python and thought maybe I should quit. However, Yash Nair, who was a teaching assistant for the course, encouraged me to stick with it and helped me rethink how I approached studying and problem-solving—he really helped me to become more resilient. I'm so glad I listened to his advice because I ended up enjoying the course. I even ended up TA'ing for the course the following year, and by junior year, I was one of two co-head TFs!

Another challenge came during sophomore fall, when I took Stat 210 [Probability I], my first graduate-level course, with Professor Blitzstein. Professor Blitzstein must have known that I needed some encouragement because he shared an article about the "Ben Franklin method," which uses the example of Ben Franklin improving his writing as a kid to argue for the importance of deliberate practice. The article reinforced the idea that learning doesn't just take place if you lock yourself in a room; rather, it's about identifying the parts you don't get and then structuring practice in that area. Long story short, this message stuck with me and the strategy worked out; I gained confidence and finished the semester strong.

How did you start collaborating with Prof. Sam Kou and how did you select your thesis topic?

Wu: Professor Kou, my thesis advisor, has been one of the most influential mentors in my time here. I connected with Professor Kou through Professor Mauricio Santillana at the Harvard School of Public Health. While working on a disease forecasting project, I expressed an interest in pivoting towards more methodological work, and Professor Santillana suggested Professor Kou.

When I took an applied math course [APMTH 216 Inverse Problems in Science and Engineering] with Professor Michael Brenner, I started to ask: "Could we use Bayesian approaches to analyze systems as complex and unpredictable as the Lorenz system?" My proposal to Professor Kou was to look at how his lab group's Bayesian method (called MAGI—Manifold-constrained Gaussian Process Inference) would work under the Lorenz system, which is used to model complex phenomena like climate.

To give a brief introduction to the Lorenz system, it is a dynamical system that has a bunch of variables that evolve over time and interact with each other. Some examples of variables within a Lorenz system are: 1) a disease-infected population vs. a recovered population and 2) a predator population vs. prey population. For these types of variables, it's very difficult to predict how they will evolve in the future; even if you are off a tiny bit numerically, it can cause your prediction to go awry.

For my senior thesis, we built upon the MAGI method by creating a version called Pilot Magi, which addressed some of the numerical instabilities in MAGI. By building Pilot Magi, we met our goal of building a tool that anyone working with dynamical systems could use. For example, in a disease forecasting setting, researchers might want to ask for a prediction of the rate of infection over time. Also, our method could help by taking noisy and sparse data and reconstructing what the ground truth would have looked like without this noise.

What do you value the most about your experience in the department and at Harvard?

Wu: "Inspired" is the word that best captures my experience at Harvard. There are two big takeaways for me about Harvard and the Stats Department; my experience has provided me

with a sense of purpose and meaningful connections to people. From interacting with various role models at Harvard and in the department—professors, grad students, and upper-class undergraduates—I have graduated with a real sense of purpose. I would like to become a professor of statistics and machine learning to advance cutting-edge research and to democratize these tools and knowledge so that they are available to the public as much as possible.

Equally important to me are the people that I met every day at Harvard who inspired me. Many of my classmates were graduate students whom I considered older academic siblings, and I looked up to faculty and my undergraduate friends. These classmates, mentors, and friends were wildly different in their passions and interests, but they all wanted to use their talents to make a positive difference in the world. The good part about being in this digital age is that, while I have left Harvard, these lifelong mentors, friends, and role models will stay with me.

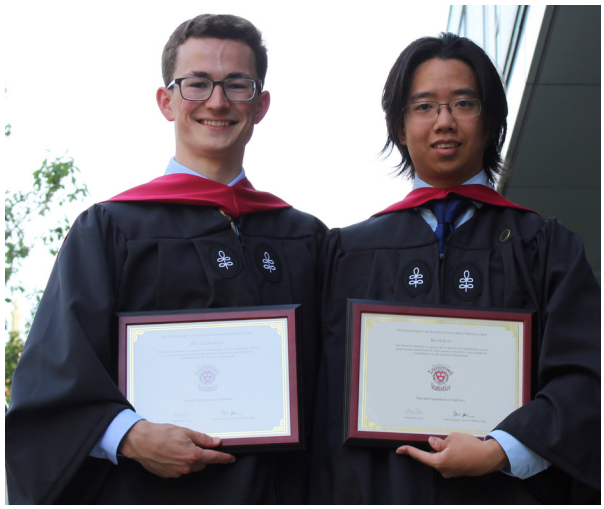
What are you excited to pursue this year?

Wu: Over the summer, I visited family in China, which was wonderful. On a fun note, I finally had access to a kitchen at home and enjoyed cooking for my family.

In the fall, I'm starting my PhD in Statistics at Stanford University. A short-term goal is to finish up research work with Professor Kou, with the aim of completing a journal submission. My long-term goal is to become a professor, like many of the incredible mentors I've had at Harvard, especially Professors Blitzstein and Kou. I'm excited to continue growing as a researcher and teacher and further pursue my aspirations of making statistics and machine learning accessible and impactful.

Profile of Kevin Luo

2024 Concurrent Masters Prize Winner



Congratulations to Kevin Luo, the May 2024 recipient (along with William Nickols) of the Department of Statistics Concurrent Master's Prize! The prize is awarded annually to up to two graduating students having completed the Concurrent Master's program in Statistics who have the best overall performance (as indicated by coursework results), have demonstrated achievements in Statistics outside of coursework, and have contributed significantly to the department.

To learn more about Kevin's inspiration, thesis experience, and sense of community within the department, we spoke with him. Highlights from our conversation are edited and excerpted below.

What sparked your initial interest in math or statistics?

Luo: My first exposure to statistics was when I completed a science project in eighth grade. The lakes in my neighborhood in Germantown, Tennessee had turned increasingly green over the years due to fertilizer runoff and, as a result, most of the fish disappeared (when I was little, I could reach my hand in the water and grab a fish!). My mom, who works in bioinformatics, suggested doing toxicity testing on the water and, in the process, introduced me to hypothesis testing. While I didn't fully understand hypothesis testing at the time, I started to frame statistics as a useful tool that lets you control for error, making it feel less rigid than the math I knew. This early impression influenced my decision to study statistics when I started attending university.

What are your initial memories of the Stats Department?

Luo: My first experience with the stats department was taking Stat 110 Probability during my freshman fall in 2020, a year that was entirely online due to COVID. Based on my schedule, I ended up in a 10:30 p.m. section for the course! It turned out to be an amazing section, run by Yash Nair ('22 AB alum) and Junu Lee ('22 AB alum), who later graduated and went on to top graduate programs. They were a big influence on my decision to pursue statistics; their section also helped me to have fun and meet other students during an isolating time.

What challenges did you face transitioning to Harvard and the Master's program?

Luo: When I arrived at Harvard, I experienced a lot of imposter syndrome because I looked at others' accomplishments in high school and did not feel as prepared. Not wanting to seem "dumb" initially made me hesitate to study with other students. COVID also didn't help because all of our classes were on Zoom and there were no gatherings in the dining hall, making it

hard to form friendships. Gradually, my fears faded away when I realized that in college, it no longer matters what you did in high school because you are free to choose a new path and go at your own pace. As a sophomore on campus in person, I made a bunch of friends, which also helped me to pursue my interests without worrying about others' judgment.

The transition to taking grad level courses was pretty smooth because there is a strong culture at Harvard of undergrads taking graduate courses. Going from undergraduate to graduate courses was also comfortable because Professor Joe Blitzstein teaches both Stat 110 and the first graduate level class Stat 210 Probability I. While the content of the two courses is different and Stat 210 is more challenging, they follow the same style and structure.

What motivated you to pursue the concurrent master's program?

Luo: The best advice that I received was from Kim Nguyen ('21 AM alum), who said, "Take the classes you're interested in, and then in senior year, figure out which degree you can graduate with by taking the fewest additional classes." I followed her advice and ended up not only being a statistics concentrator but also receiving a master's degree, which is a pretty good deal!

A specific course to highlight would be Stat 217 [Topics in High-Dimensional Statistics: Methods from Statistical Physics] – it changed my life in my junior year! Taught by Professor Subhabrata Sen, the course introduced me to statistical physics. Two contrasting ideas, which are really two sides of the same coin, were very cool to me. The first was universality, which roughly means that some large system behaviors are not sensitive to small details. One example of universality is the central limit theorem. The second was the concept of phase transitions and order parameters, which say that, on the other hand, there are certain parameters that entirely determine the large system behavior, and this behavior can change sharply when these parameters are slightly perturbed.

How did you start collaborating with Professor Pragya Sur on your thesis and how did you select your thesis topic?

Luo: With a broad interest in working on a project related to machine learning and theoretical statistics, I began skimming research papers of faculty and talking to them. I chose to work with Professor Pragya Sur because we discussed a project that I thought would help me grow and learn the most.

The original goal of my thesis was totally different from the final product. I started out working with Professor Pragya Sur and her student Yufan Li ('25 PhD alum) on random feature models. I was trying to prove something about the asymptotic mean squared error, but I failed because I didn't understand the techniques well enough. I remember having a tough time with the project over winter break, but it was a good first lesson in the research process and how to move forward after failure. I realized that you can't just expect a solution to unravel in front of you; instead, you need to work on small pieces of the problem, even if it feels far from

helping you solve the whole problem. It was a humbling experience.

We pivoted to a different, simpler problem: high-dimensional ridge regression under dependent data settings. Ridge regression is related to linear regression but is used to combat overfitting by adding a penalty to the cost function for complex solutions. Most existing high-dimensional ridge regression analysis assumes datapoints are independent and identically distributed (IID), but in real-world applications like medicine or finance, this is often not the case.

For example, medical data for a family, such as weight and height, are likely to be correlated, not independent variables. Using models that allow for data dependence, we explored what happens when data points are dependent and whether conventional tools still work. We found that many behaviors remained similar, but common cross-validation methods broke down in this setting with this data, so we proposed a new approach.

What did you value most about your experience in the department and at Harvard? If you had to choose one word to describe it, what would it be?

Luo: I most valued the opportunity to have one-on-one conversations and work closely with faculty. I never imagined this kind of relationship with faculty when I was in high school. My advisor, Pragya Sur, helped me to understand the landscape of statistical research, the problems researchers care about, and my research interests. When Pragya advised me on my thesis, she never seemed surprised about anything—she always understood the problems extremely well and had great intuition.

Related to my experience with Pragya and other faculty, I would use the word "defining" to sum up my Harvard experience. When I entered college, I had no clear idea of what I wanted to pursue (I even wrote my college essay about this!). In my first few years at Harvard, I dabbled a lot in different subjects, and this provided me with a strong understanding of my interests.

What are you excited to pursue? Describe some of your career, academic, and personal aspirations and plans.

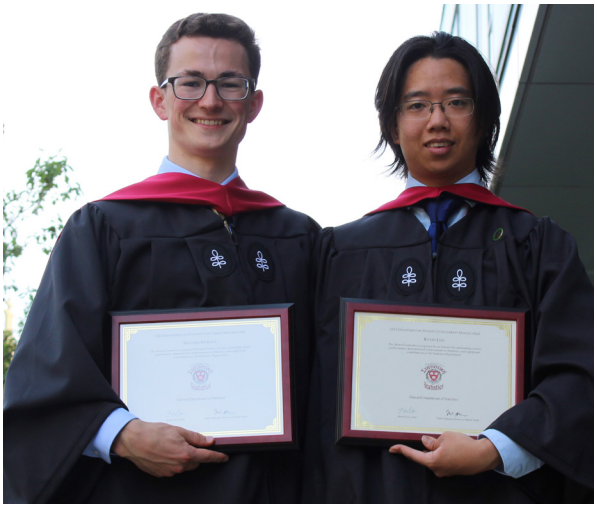
Luo: This summer, I'm excited to revisit the research project that initially failed, the one on random feature models. During the semester, there were some nights when I could not fall asleep because I was thinking about the problem and what I would do differently! I now realize what to do, so I would like to start working on the problem again. I'm also looking forward to spending time with some of my high school friends, whom I've stayed very close with especially because of my COVID freshman year.

In the fall, I'm excited to start as an algorithm developer at Hudson River Trading. My work will likely center on building systems for stock trading strategies.


Hello

Profile of William Nickols

2024 Concurrent Masters Prize Winner



Congratulations to William Nickols, the May 2024 recipient (along with Kevin Luo) of the Department of Statistics Concurrent Master's Prize! The prize is awarded annually to up to two graduating students from the Concurrent Master's program in Statistics who have the best overall performance (as indicated by coursework results), have demonstrated achievements in Statistics outside of coursework, and have contributed significantly to the department.

To learn more about Will's inspiration, thesis experience, and sense of community within the department, we spoke with him. Highlights from our conversation are edited and excerpted below.

What sparked your initial interest in math or statistics?

Nickols: I remember, in my high school junior year AP statistics class, learning about the German Tank Problem. Based on a true WWII story, the set-up was: suppose you've captured a random sample of your enemy's tanks, each of which have a random ID number from 1 to the total number of tanks. You want to estimate the total number of tanks your enemy has. How do you do it? My group thought about it for a while and then decided we should double the average of the observed tank IDs. Then, the class had a competition, drawing numbered slips of paper mimicking tank IDs, and our estimate ended up being just two away from the true number! Two years later in Stat 111 (Introduction to Statistical Inference at Harvard), I learned our approach had a name: the Method of Moments. I found that I enjoyed taking an intuition and turning it into a mathematical framework to address real questions—that's what really sparked my interest.

What are your initial memories of the Stats Department?

Nickols: Starting college during the pandemic, I remember finding community right away in Stat 110 and 111 through the faculty (Joe Blitzstein and Neil Shephard) and other students. In Stat 111, we had a virtual Pset group to get to know each other and work on homework problems together. In the spring, we scheduled a Zoom "classroom to table" conversation with Joe, where we talked about Joe's educational philosophy, chess, and biographies we had read. We really appreciated that he made time for us, even with all his obligations and hundreds of students to teach! On the other end, right before graduation, Joe hosted a dinner at Henrietta's Table with former teaching fellows (TFs) from his classes. We talked for about three hours about everything from hidden messages in the Stat 110 Probability book, to Joe's Youtube cult following, to MLEs (which half the table thought it meant maximum likelihood estimators, and the other half thought

it meant machine learning engineers—maybe a reflection on the state of statistics). These moments demonstrated how invested he was in us.

During my sophomore and junior years, there were many opportunities to connect with others in the department in person. When I took Stat 210 (Probability I) in my junior year, I formed a study group with 5-6 close friends, and we regularly got together to work on Psets. The department faculty and admin, along with GUSH (Group for Undergraduates in Statistics at Harvard), really invested time in creating an in-person community after COVID. When I was a teaching fellow for Professors Kevin Rader, James Xenakis, and Neil Shephard too, they often reached out to get coffee or a meal. There was a lot of faculty and peer support that created a real sense of community for me.

What motivated you to pursue the concurrent master's program?

Nickols: The concurrent master's program was a natural fit due to my interest in higher-level coursework, particularly in causal inference, an area of research that looks at the effect of treatment on an outcome. A standout course was Stat 286 (Causal Inference), which looked at how to determine the effects of particular interventions using modern statistical methods. For example, we learned methods to deal with panel data—data collected repeatedly over time from individuals. On an exam, we looked at whether per-pupil educational spending changed when teachers' collective bargaining laws changed, based on data across U.S. states over time.

A similar problem came up in a project for Stat 288 (Deep Statistics: AI and Earth Observations for Sustainable Development), where, inspired by previous work done by student Victoria Li, I looked at abortion rates after the Dobbs decision in June 2022. To answer the question of whether the Dobbs decision affected abortion rates overall in the US, I used similar statistical methods as in Stat 286 with panel data across states from before and after the court decision. This class exposed me to the cutting edge of applied research, especially causal inference applications.

Did you encounter challenges (personal or academic) during your studies? How did you overcome these challenges?

Nickols: A big jump for me was coming into college and taking Stat 110. I still remember the first question of the first homework assignment. The question asked how many choices there are for courses for a degree if a student can elect to take any 7 courses out of 20 with the condition that at least one course must be a stats course and 5 out of the total 20 courses are statistics courses. I thought about the problem for a while and came up with the answer 135,660. But, the next question on the homework was, "Explain intuitively why the answer is not 135,660"! At that point, I figured it was going to be a rough semester. However, over the next few months, I started recognizing patterns and understanding better how to learn in the course—what to focus on from the textbook and lectures and how to apply that to problems.

Being a TF for Stat 139 (Introduction to Linear Models) and Stat 111 during my junior year

provided an excellent chance to solidify what I had learned in prior semesters. Each week, I created and solved practice problems in preparation for teaching my section, providing me with a strong understanding of what goes on behind the scenes of teaching—how (and why) you formulate a problem for students and solve it. Teaching Stat 111 was initially more daunting since it focuses on statistical theory, a topic I had initially struggled with as a student, but as a TF, I could see how I had grown as a statistician. After only two years, the topics that had once kept me up late at night were now things I could explain like the back of my hand.

You completed your senior thesis with Professor Curtis Huttenhower at the Harvard T.H. Chan School of Public Health, working on statistical methods for the microbiome. How did you start this project?

Nickols: I started working with Prof. Huttenhower during the summer after my freshman year as part of the Program for Research in Science and Engineering (PRISE). I wanted to apply the skills I had been developing in my statistics and computer science courses to biological questions—a particular interest of mine. Microbiome research involves microbial communities made up of bacteria, fungi, and archaea that exist essentially everywhere including on human surfaces like the skin, mouth, and gut. In fact, the average person has 2-6 pounds of microbes in and on their body! (When I would present my senior thesis to non-microbiome audiences, I would start by saying "I'm Will... or I should say, I'm mostly Will, because I'm also about 3% microbes." This usually got groans, but Joe liked the joke so much he said he would've studied the microbiome just to use that line). These microbes have important implications for human health, and diseases like inflammatory bowel disease (IBD), Crohn's, and ulcerative colitis, which are linked to the gut microbiome.

To study the gut microbiome, researchers extract DNA from stool samples, map it to microbial genome databases to determine the abundance of different species (e.g., 10% species X, 15% species Y), and look at the differences in the gut microbiome between patients with and without IBD. These data have many zeros (species that are either not present or not detected in a sample), and historically researchers have just added a small "pseudo-count" to handle such zeros before running linear regressions.

My thesis improved the traditional approach in a few ways. First, we separated a microbe's abundance (how much of it there is, if it's present) from its prevalence (how likely it is to be present at all). This was an important step because we found that, for over 70% of the associations between the microbiome and disease outcomes in our data, the association was with the microbe's presence vs. absence rather than its abundance if present. Second, my thesis tackled the issue of relative abundances. Because sequencing data tell us a species' proportion out of the whole community, but not the actual number of cells of that species, if one microbe becomes more abundant, the other species' relative proportions decrease, even if their absolute counts stay the same. To

address this problem, we introduced statistical models to identify associations between health outcomes and absolute counts, either by incorporating additional experimental information or by estimating the total cell count. The tools we developed are already being used to address questions related to ulcerative colitis, bacterial vaginosis, pre-term birth, colorectal cancer, and pet nutrition.

What do you value the most about your experience in the department and at Harvard?

Nickols: I would describe my experience with statistics at Harvard as empowering. By taking courses like Stat 110 and 111, I developed an understanding of the fundamentals that gave me the skills to build new methods and tools. With more sophisticated tools, I felt empowered to answer complex questions and to evaluate the reliability of my answers. For example, after freshman year, I did research involving multiple databases with overlapping identifiers. I wanted to know: if I randomly sample from each database, how many overlaps should I expect purely by chance? Using the foundational tools that I learned in Stat 110 and Stat 111, I was able to work out the theory, write software, and answer my question. When you understand statistical fundamentals like distributions and random variables, you can dig deep into the methods you are using and why they work.

What are you excited to pursue this fall? Describe some of your career, academic, and personal aspirations and plans.

Nickols: I am excited to begin my PhD in Biostatistics at Harvard next year, where I will rotate in a variety of biostatistics and computational biology groups. I am also looking forward to continuing work on genetic methods that can improve our understanding of how infectious diseases are acquired, cleared, and transmitted. For example, if you have longitudinal genetic data from people who acquired malaria, you can determine previously unknowable things like how long a person has been infected with a particular parasite and whether that parasite is resistant to treatment. I'm also interested in bigger picture questions like how the development of artificial intelligence will speed up our ability to analyze data and produce useful and correct results.

After my PhD program, I'm open to different options, including positions in academia; government (e.g. the NIH or CDC); nonprofits; or hospitals—this will be my problem to solve over the next 4-5 years! Regardless, I intend to continue working on challenges related to real data, including developing and applying methods to solve meaningful problems.

Profile of Souhardya Sengupta

2024 PhD Dempster Prize Winner

In May 2024, PhD student Souhardya Sengupta received the 2024 Dempster Award for his exceptional paper titled "Leveraging Sparsity in the Gaussian Linear Model for Improved Inference," co-authored with Associate Professor Lucas Janson. The Dempster Prize is named in honor of Emeritus Professor Arthur P. Dempster and is given annually to a graduate student in recognition of outstanding research. In the following excerpted and edited interview, Sengupta talks with us about his academic journey thus far—from receiving early inspiration from his father to learning how to navigate the rigorous environment at the Indian Statistical Institute to relishing the community and intellectual freedom in his Statistics PhD program at Harvard.

Was there an experience or influence early in your life that sparked your interest in math or stats?

Sengupta: One influence on my interest in math was my father when I was young. When I was in elementary and middle school studying arithmetic and basic algebra, he used to help me with my math homework. I admired his wits as he would quickly solve problems I had struggled with for days—something I thought was really cool! I believe being able to do the same was one of my earliest inspirations, slightly tilting my preference toward math over other subjects. My interest then continued throughout high school and led me to pursue it in college.

What motivated your choice of undergraduate institution and your decision to pursue a PhD program?

Sengupta: In short, ISI [Indian Statistical Institute] was the most prestigious place I got into! Initially, I wanted to study math; I thought that I would pursue ISI's Bachelor of Mathematics program in another city. However, a few of my teachers convinced me to apply for the Bachelor of Statistics program instead at ISI Kolkata, which was also only about 45 minutes away from my home. So my decision to pursue statistics was not particularly well thought out and by some happy coincidence I ended up liking this subject.

While I was undecided about graduate school during my first two years at ISI, my decision to apply solidified during the COVID-19 lockdown. I had just taken a course on nonparametric statistics that I really liked. When I was in lockdown, I did some independent reading on some of the topics I liked, and I think that entire period built my appreciation for research that later on developed through my remaining years at ISI. I appreciated the clever and highly creative ways researchers have approached problem in the literature. An example of something I really liked was the literature on multivariate concepts

of ranks and ordering, where people have tried to extend the idea of ranking scalar data points to ranking multivariate data or tuples—that is, a collection of numbers rather than single values. Later on, I worked on some research projects, which convinced me that I wanted to continue doing this, and hence, pursuing PhD was an obvious choice.

What challenges have you experienced during your studies and what did you learn from these difficulties?

Sengupta: In my undergraduate and graduate experience thus far, I have been lucky that I haven't had a major setback that completely threw me off track, but I have had my shares of ups and downs and some tough personal times. Also, like most of us, I struggled with the disruptions, isolation, and the mental toll of the COVID lockdown.

In my academics, especially when I first joined ISI, I found much of my coursework to be challenging and quite a few times felt overwhelmed with it. Over time, I learned two major lessons though that I have found echoed in various aspects of my life and not just academics. First, I learned that most of the things that seem impossible today might not seem that difficult eventually. Second, I learned not to overemphasize certain negative events. Early on, I treated any failure like a disaster, but over time, I have learned to treat setbacks as an inevitable part of the process, perhaps, just like successes.

What memories do you have of getting to know the Stats Department?

Sengupta: Everything was new to me when I first arrived in the department, so my early experiences stand out. The wide range of social events, especially during my first year, made me feel welcome in the department. Often on Friday nights, we would gather in the lounge to play board games. Later in the year, a group of us started playing badminton at the Harvard MAC gym. Another great memory is our annual departmental PhD retreat, which features an outside speaker, faculty lightning talks, and a social. During my first year, the social included ice skating, which I tried for the first time; it didn't go well, but it was fun and I'm glad I tried it!

How did your collaboration with Professor Lucas Janson begin? What were some of your key findings in your paper "Leveraging Sparsity in the Gaussian Linear Model for Improved Inference"?

Sengupta: When I started my PhD program at Harvard, I was familiar with Professor Janson's research area and felt that my research interests aligned, so I just walked into his office and asked if I could work with him (and he agreed!).

Our paper revolves around the linear regression model, one of the most widely used statistical models to answer basic scientific questions. For example, suppose you want to study how age affects someone's risk of diabetes. You might want to study other factors too, such as blood biomarkers. The goal would be to quantify how each of these factors, also known as covariates, relates to a response variable (in this case, diabetes). Linear regression is a simple model that helps you study these relationships. In this

scenario, if you want to see whether a factor like age has an effect on diabetes that cannot be explained by the other covariates, you typically use the standard t-test.

Our contribution is an exact alternative to the t-test called the "ℓ-test." You can use the ℓ-test in any scenario in which you would apply the t-test, and often it has higher power (that is the probability that the test would correctly detect a true relationship) than the t-test. Our most surprising finding was that our ℓ-test achieves power close to that of the one-sided t-test, which apriori knows whether the relation between the covariate and the response is increasing or decreasing.

To illustrate, if we had prior knowledge that any significant relationship between age and diabetes could only be positive (i.e., increased age leads to increased risk), we would use a one-sided t-test to assess whether an increase in age corresponds with an increase in the likelihood of developing diabetes, unlike the usual t-test that tests whether an increase in age corresponds to any change in the likelihood of developing diabetes. It turns out that without any access to this prior knowledge, our ℓ-test still approximates the performance of a one-sided t-test, often very closely in the settings when there are a large number of covariates, with most of them having no effect on the response (this is called sparsity, and such a setting occurs frequently in many scientific studies, for example in genetics).

What aspects of your PhD program have you valued the most? Can you think of a word to encapsulate this experience?

Sengupta: If I had to choose one word to encapsulate my PhD experience, it would be: excitement. There's a lot of frustration, too, but it's all built around the excitement that something might finally work out. To be more specific, it's the multifaceted nature of my program as well as the novelty and intellectual freedom that have made this an exciting experience for me. I enjoy taking on different roles: sometimes I'm a student attending lectures and taking exams, other times I'm a teaching assistant running sections and grading homework, and then I'm a researcher, working through ideas and trying to solve open-ended problems.

I also value the novelty and intellectual freedom that comes with PhD research. You never know where an idea will lead, and each day holds a bit of unpredictability. Most attempts at a solution are usually futile but when something clicks and finally works, it's incredibly rewarding. That joy of discovery is unmatched!

What are you excited to pursue this fall (both personally and academically)?

Sengupta: In my academics, I have ongoing research problems that I'd like to make progress on during the fall—I'm curious to see what comes out of it! I also have a personal goal: I want to get back into playing the guitar. I started learning it in 2022, just before I moved to the U.S. Since starting grad school, I've just been too busy to pick it up again, but now that I have bought a guitar, I'm going to really try to practice!

Statistics Community Outreach

Data Adventure Day & Project Teach

Entering Hall D in the Harvard Science Center, I met a volley of questions, "Do you study inequality in society? Do you like programming?" Unfortunately, the answer was no, but I could point the students to another Harvard volunteer. Students from the John D. O'Bryant School of Math & Science, Brighton High School, and the StatStart Program had just arrived for the 3rd annual Data Adventure Day (known as "Florence Nightingale Day" in its first year), sponsored by Harvard Statistics and Biostatistics Departments and the Harvard Data Science Initiative.

Data Adventure Day (DAD), an all-day event on November 8th with statistics and data science related activities for ~120 high school students (as well as ~50 teachers and volunteers), is part of our Department's commitment to pursuing equity, diversity, and inclusion goals by connecting our faculty and students with youth from Boston and Cambridge. In addition, some of our faculty and students have participated in Project Teach, a Harvard program that partners with Boston public middle schools to provide exposure to college through on campus activities. By participating in Project Teach and hosting Data Adventure Day, our department aims to encourage a future generation of students to pursue statistics, data science, and related fields and to foster greater diversity in these fields. Now let's take a closer look at the impact of these initiatives!

Hands-on Introduction to Statistics at Data Adventure Day

While the trivia activity in Hall D was an icebreaker, it also gave students the opportunity to learn about a diverse group of statisticians. Students identified and answered questions about various headshots positioned around the room – from David Blackwell, the first African American full professor at UC Berkeley, to Florence Nightingale, who convinced Queen Victoria to improve sanitation in hospitals, to Joy Buolamwini, who researches how algorithms encode racial and gender bias. Students' hands-on learning continued when they visited Harvard's Museum of Natural History to investigate fun data facts (e.g., a specimen from the deepest part of the ocean) and to search for interesting stories about how scientists are collecting data.

Connecting to College Life at Data Adventure Day

After the visit to the museum, high school students attended a college panel moderated by Data Adventure Day Planning Committee members Oscar Mercado (Senior Statistics Concentrator) and Rebecca Hurwitz

(Biostatistics Master's alum), which highlighted the college experience of three panelists with different personal backgrounds, avenues into statistics, and career goals. They heard from Ace Mejía-Sánchez, a Senior Statistics and Government Concentrator who developed an interest in pursuing data science applications in the nonprofit sector, as well as from Raihana Rahman, a Junior Mathematics and Statistics Concentrator who discovered her passion for studying the theory behind stats and math, and Aseel Rawashdeh, a Junior Statistics and Computer Science Concentrator.

Reacting to the panel, student Kevin Dang shared, "DAD was fun, especially having the opportunity to talk to Harvard students and learn about their experience in college. Learning about the nuanced applications of stats opened my mind to new potential job prospects for myself." The panelists similarly found their exchange with the high school students to be rewarding. Summing up the value of participating in Data Adventure Day, Rahman reflected, "It's important for me to find ways to volunteer with local Boston communities and work with students, who (like me) are underrepresented in statistics; DAD, specifically the panel, gave me the opportunity to do just that!"

Exploring Real-World Applications at Data Adventure Day

The day culminated in two activities that highlighted the real-world applications of statistics and data science. Some students participated in an activity designed by Paul Schwein at LabXchange, a nonprofit organization created at Harvard University that develops digital STEM lessons and materials for classrooms. Through an activity focused on data related to music trends, students learned how data visualizations shape the story that the data tells. For example, students looked at the rankings of music genres between 2016 and 2024 and discussed how the creation (or deletion) of genres might change or leave gaps in the interpretation of genre popularity over time. Students then reviewed a graph displaying the range of rapper vocabulary, which led to questions such as, "Is there a relationship between the amount of words a rapper typically uses in their songs and their era or between the amount of words and their net worth?"

The second activity consisted of interactive presentations with Harvard researchers, faculty, PhD students, and alums on applications of statistics in their daily work. Presenters from our department included faculty Lucas Janson, Associate Professor, Kevin Rader, Senior Preceptor, and Joe Blitzstein, Professor of the Practice, as well as PhD student Yuzhou Lin, and AB alum Ethan Kahn. The topics illustrated the exciting breadth of statistics and data science applications, including research on the global burden of disease, gun violence causes in the US, the death toll in Puerto Rico from hurricane Maria, drug treatment effects in clinical trials, and sports analytics.

Describing a highlight of the presentations, Blitzstein shared, "Rafa Irizarry's presentation on Hurricane Maria was very engaging and showcased high impact work, and there was a great Q&A with students afterwards." While presenters explained how they apply statistical

methods to their field, they also shared their personal journeys, backgrounds, and growing pains, such as learning that it's okay to move on from job or career choice that is a bad fit. After attending the presentations, student Begine Derilien reflected on the activity: "I really liked the panel of different statisticians because it made me realize that, yes, stats is useful in real life. It's not another math class where you're wondering how will this help me in the future? The panel helped me understand the importance of stats in our daily life and future careers."

Diving into Statistics with Middle School Students at Project Teach

Plunk, plunk, plunk. Several hacky-sacks flew through the air and into a bucket in swift succession; a middle school student, a participant in Project Teach, grinned with satisfaction. Pausing the activity, the instructor, Lucas Janson, explained to students that this exercise was not just fun and games but was also an entry-point to thinking about how to collect and ask questions about data. Janson elaborated, "If you wanted to identify your favorite food or Tik Tok star, then you'd have to try out different options and write down your findings." He shared other examples, including sports team owners making decisions about players based on performance stats, Chat GPT generating recommendations based on internet data, and medical researchers determining the effectiveness of a treatment based on survival rates.

"So, is Lucas good at hacky-sack-bucket?" asked Janson as he tossed several hacky-sacks and landed one in the bucket on his third try. After conferring with each other, students agreed to test out Janson's mastery of the game by having all students record their number of tosses before reaching the bucket and then comparing their average results with Janson's. The results? On average, the students performed better than Janson, demonstrating that Janson was, unfortunately, not so good at hacky-sack-bucket! Through this hands-on activity for Project Teach, students gained valuable experience with formulating a question, designing an experiment, collecting the data, and using statistics to analyze the data.

Evaluating Data Visualizations at Project Teach

In another Project Teach classroom, Julie Vu, Preceptor in Statistics, displayed a graph with multi-colored dog figures located in four quadrants: "hot dogs" (popular with the general population as well as highly rated by dog experts), "overlooked treasures" (not popular but highly rated by dog experts), "the rightly ignored" (neither popular nor highly rated by dog experts), and "inexplicably overrated" (popular but not highly rated by dog experts). Turning to her middle school audience, Vu asked students what they noticed about the organization of the graph, the x/y axis, the labels, and the datapoints. "What story is the graph trying to tell us about these dogs?" asked Vu, "and do you think that it does a good job?" In response, students argued for both the benefits and detractors of the graph; some students thought that it was creative and engaging while others thought that it was cluttered with information and confusing.

After debating the effectiveness of this graph, students turned their attention to graphs posted throughout the room, which they marked with a sticker (a blue star for the most memorable graph, a yellow star for most informative, and a red frowny face for most confusing). In the group discussion that followed, students explained their reasoning behind their assessments, honing their skill at identifying and evaluating the choices made in these data visualizations.

From learning to evaluate data visualizations, to learning to develop questions about data, to connecting statistics and data science to real-world applications and careers, students participating in Project Teach and Data Adventure Day had the opportunity to see first-hand the power of statistical thinking.

*** Top Photos by Dayanara Torres. Bottom Photo by Stephanie Mitchell/Harvard University*

Our hope is that these students and the students that we host in future years harness their experiences at Harvard, as well as opportunities Teach!

Refence List

Daniels, Matt and Pera-McGhee, Michelle. "You should look at this chart about music genres." The Pudding, 2023, https://pudding.cool/2023/10/genre/

Daniels, Matt. "The Largest Vocabulary In Hip Hop." The Pudding, January 21, 2019, https://pudding.cool/projects/vocabulary/index.html

McCandless, David. "Best in Show: The Ultimate Data Dog." Information is beautiful, 2024, https://informationisbeautiful.net/visualizations/best-in-show-whats-the-top-data-dog/

Prof. Murphy & Dr. Xu

Like having a personal healthcare coach in your pocket: New apps for cancer patients, cannabis users, others make use of algorithms that continually customize support

Cancer patients who undergo stem cell transplantation face a long recovery, requiring medications with debilitating side effects and support around the clock. It's a difficult experience, with studies showing that more than 70 percent of patients don't adhere to drug regimens.

Statistician Susan Murphy spends her days trying to help people suffering from such challenging maladies. The Mallinckrodt Professor of Statistics and Computer Science and associate faculty at the Kempner Institute and her team address healthcare needs not through medicine, but by mobile apps.

Murphy's lab specializes in creating sophisticated computational instructions known as reinforcement learning algorithms, which form the technical backbone of next-generation programs to help people stick to a medication protocol, for instance, or regular tooth brushing, or reducing cannabis use.

And if this sounds like one of those ubiquitous apps that tracks steps or counts calories, think again.

"If you've ever downloaded a health app, those tend to be pretty dumb," Murphy said. "For example, you'll get a physical activity app, you'll sprain your ankle, and it'll continue to tell you to go for a walk."

Using advancements in artificial intelligence and sensing technologies to move beyond one-size-fits-all interventions, the lab's apps are capable of real-time personalization, meting out psychological rewards, and in some cases, leveraging social networks to help users stick to goals.

This approach is called "just-in-time adaptive intervention" because it aims to provide support at just the right time by registering changing needs and contexts.

Currently the Murphy lab is working with software engineers, cancer clinicians, and behavioral scientists to develop an app for stem-cell transplant patients and their primary caregivers, usually parents.

Health management, especially for the sickest, typically requires involvement of others. For instance, up to 73 percent of family-care partners have primary responsibility for managing cancer-related medications.

The researchers are in the early stages of developing the algorithm, to be deployed in a first-round clinical trial this year by collaborators at the University of Michigan and Northwestern University. The trial, called ADAPTS HCT, will focus on adolescent and young adult patients who've had stem-cell transplants in the 14 weeks post-surgery.

The algorithm will inform sequential decisions, including when and whether to send motivational prompts to the patient, and whether to send

messages and reminders to both patient and caregiver. The application includes a word-guessing game that fosters social support and collaboration between patient and caregiver.

"We hypothesize that in improving the relationship between patients and their caregivers, patients can function and manage their medications better," said Harvard postdoctoral fellow Ziping Xu, who is leading the ADAPTS HCT algorithm development.

The app will employ reinforcement machine learning, in which the software will "learn" from previous interactions. For example, rather than simply sending preset reminders about medications, the algorithm will tailor timing and content according to when they have been most useful to patients. That way there is less chance the notifications will be deemed irrelevant or ill-timed and eventually habitually ignored.

The Murphy lab is deploying its algorithmic expertise across other domains. With their University of Michigan collaborators, they've recently pilot-tested a program called MiWaves aimed at young adults who are abusing cannabis.

Like the ADAPTS HCT app, MiWaves continually learns and adapts from interactions with each patient to improve its decision rules, with the goal of helping them reduce their daily intake.

The lab is also several years into a project called Oralytics, which recently wrapped up a 10-week randomized trial to help refine the delivery of push notifications to help patients adhere to a tooth-brushing protocol: two sessions of two-minute duration daily, covering all four mouth quadrants.

The first Oralytics clinical trial included some 70 participants who all received the mobile app with a wireless-enabled toothbrush that sent data to the team's collaborators at Proctor and Gamble.

Graduate student Anna Li Trella, who led the Oralytics project through the first trial, said the recently collected data will help the team develop methods to better handle messy problems like missing data and software errors.

"There are many constraints to running an algorithm in real life," Trella said. "Now that we've conducted the first trial, we can make improvements to help the algorithm collect better data and learn better."

Murphy thinks of her lab as creating practical pocket coaches who can help people get where they want to go.

"Very, very few people can afford a human coach. And in fact, some people may not want such intensive human interaction," Murphy said. "That's where the idea for these digital supports comes in."

Credits

Manning, Anne J. "Like having a personal healthcare coach in your pocket: New apps for cancer patients, cannabis users, others make use of algorithms that continually customize support." https://news.harvard.edu/gazette/story/2025/04/like-having-a-personal-healthcare-coach-in-your-pocket/

Photo by Grace DuVal.

Dr. Biyonka Liang ('25 Graduate)

Develops Model for Improving Resource Allocation Decisions in Public Health

How can healthcare providers effectively decide which patients to provide extra support when resources, such as budget and staff, are constrained? ARMMAN, an NGO in India that runs multiple large maternal and child health programs (interventions implemented by ARMMAN have reached over 60 million women now, including 3 million via mMitra and the rest through Kilkari) and their collaborators in Computer Science and Statistics at Harvard and at Google Research in India, have been driven to answer this question in their research to improve health outcomes for pregnant women. While all women enrolled in mMitra receive weekly automated calls with maternal healthcare information, ARMMAN wanted a mechanism to help allocate live calls to the women who would benefit the most from them.

To address this problem, Biyonka Liang, Lily Xu, Milind Tambe, and Lucas Janson of Harvard University and Aparna Taneja of Google Research worked with ARMMAN to develop a model for allocating these live phone calls in a simulation study. The model, BCoR (Bayesian Learning for Contextual RMABs), introduces a novel approach by incorporating contextual information across beneficiaries to determine who would benefit the most from receiving a live call from a healthcare worker. To elucidate the significance of this new method and its potential applications, Biyonka Liang, a 2025 Statistics PhD graduate, met with us to discuss her work with her co-authors on their paper "Context in Public Health for Underserved Communities: A Bayesian Approach to Online Restless Bandits." This work was recently published as part of the proceedings of AAAI, a major AI conference.

Motivation of Public Health Collaboration

Interested in research motivated by applied problems, particularly healthcare and genetics-related applications, Liang jumped on the opportunity to collaborate with colleagues in CS, Dr. Lily Xu (a PhD alum) and Professor Milind Tambe, who had an ongoing collaboration with Dr. Aparna Taneja and ARMMAN. More specifically, Liang, with Associate Professor Lucas Janson, was interested in bringing a statistical modeling approach to this resource allocation problem, both for ARMMAN and for other settings. Reflecting on her motivation, Liang said, "The goals of the project were to determine how to allocate resources effectively and to learn the type of person who is most responsive to an intervention. To answer these questions, you have to learn how people will behave from observed data, which is exactly what our statistical modeling contributed to this project."

Previous approaches to this problem often relied on more theoretical frameworks without considering the practical constraints of a real-world setting. These methods often lacked the ability to account for specific time parameters (such as the 40-week duration of pregnancy) and contextual factors (like education and income levels) that could influence a woman's likelihood of listening to and benefiting from the healthcare information provided in the calls. In addition, while there was an earlier real-world implementation for ARMMAN, the offline nature of the model prevented it from learning from data in real time.

Reflecting on the importance of the collaboration with ARMMAN, Liang said, "It was very important to us throughout the project that we had ARMMAN's input and contextual information so that the design of our model would directly address their concerns." For example, for the simulation study, the nonprofit organization provided anonymized covariate data collected in 2022 from 24,011 enrolled women, including education level, income level, and phone ownership. This real-world data was crucial for incorporating context into the model and improving its ability to predict which women would benefit most from a live phone call.

Unpacking the BCoR Method

At a high level, BCoR (Bayesian Learning for Contextual RMABs) in the simulation study with ARMMAN works in the following way. During the first week, all women enrolled receive an automated call with information about how to sustain maternal and infant health. During the second week, the adherence to the treatment (whether women listened to the call) is observed and the algorithm uses this information and certain covariates to decide which women should receive a live call the next time. For example, if person A listened to the call, then she might be more likely to listen to calls in the future and may not need a live call. However, if person B did not listen to the phone call and is in a lower income and education bracket, then the algorithm uses this context to decide that the healthcare worker should call person B. In the third week, the algorithm tracks whether women A and B listened to the automated call and updates the probability of their future adherence.

To describe the method at a more granular level, BCoR employs restless multi-armed bandits (RMABs). A traditional multi-armed bandit problem uses reinforcement learning algorithms to optimize rewards over time. Liang illustrated this concept: "Imagine that you are in a casino and there are multiple slot machines that will generate a different reward depending on your selection. Each slot machine represents an 'arm' and when you pull the arm, there is a probability that you will or won't get money. Over time, from pulling different arms multiple times, you learn the reward distributions of the slot machines."

Restless multi-armed bandits, however, are more complex and often more applicable to real-life scenarios because the state of the arms, even for arms that aren't pulled, and the reward distributions are in flux. In the simulation, the women beneficiaries are represented by the arms, pulling an arm means giving a woman a live call, and the reward is whether the woman listens to the automated call the next time. Liang explained the application of RMABs in their simulation: "Because Person A is already listening to automated calls, she will likely continue to do so, with or without a live call. On the other hand, if I give a live call to person B, who has not listened to automated calls, there is a chance that she will continue to not listen. The reward distributions for person A and B (probability of listening to the next automated call) depends on their current state, which means that the reward distribution isn't fixed and can change over time."

Differentiating BCoR from Other Methods

Previous theoretical work in RMAB models has assumed that you have the ability to observe people forever, without a time limit. However, in the ARMMAN example, women are only enrolled for about 80 weeks (throughout pregnancy and until one year after childbirth), which means that "you're often in a situation where your model has to determine how to allocate resources across the entire pool of beneficiaries, when you haven't even observed how the vast majority of them would respond to this type of intervention," explained Liang. She further emphasized the advantage of BCoR: "The benefit of using BCoR is that you get much stronger empirical results when applying this method to finite time settings and small datasets from real life."

While other methods have attempted to address both resource allocation optimization and the identification of individuals most responsive to interventions, Liang highlighted a key difference: "Their approach is to treat every single arm as if it is independent from other arms, which assumes that there's nothing learned from one arm that could be applied to another arm." In contrast, BCoR leverages contextual data, such as demographic information, phone ownership, and adherence history, across all beneficiaries to inform its recommendations for live calls.

This allows the model to even make informed recommendations for women new to the study. For example, if a woman in her third week of the study does not own a phone, and the model has substantial data indicating that women in a similar situation tend to decrease their adherence, BCoR would likely recommend prioritizing a live call for her. Summing up the value of BCoR, Liang stated, "Our work is different from existing work because our model learns continuously from data how people are going to behave with a level of greater granularity, while in the past, other methods assumed that everyone behaves the same across time."

Importance of Ethics in the Project

Ethical considerations are paramount in public health research involving patient data. Liang and her co-authors explicitly addressed these concerns in their paper. They noted that the 24,011 women participants provided informed consent after receiving detailed information about the simulation study, including the anonymization process for their demographic data and ARMMAN's data privacy policy. Liang and co-authors only had read-only access to the anonymized data for the purpose of this research project. Reflecting on the ethical considerations, Liang said, "While our model wasn't yet deployed in real life, it was important to consider what the ethical implications would be for real-life deployment (since that is the ultimate goal)." She further clarified, "The purpose of our model was to determine who gets the stronger touch, based on who is at risk of not receiving the information; we never withheld or provided additional health information to women."

Future Exploration of Topics

Liang expressed enthusiasm about the potential for this project to lead to new and exciting research directions. A key goal is to see BCoR implemented in real-world settings. Expanding on this idea, Liang said, "I'd love to see BCoR deployed in real life and work through the challenges, which would include challenges outside of statistics, such as logistics and computing resources." There are also examples of other resource allocation problems, such as where to deploy park rangers to prevent anti-poaching, that Liang could potentially employ BCoR with collaborators and stakeholders.

The work of Biyonka Liang and her collaborators demonstrates the powerful synergy between machine learning, statistical methods, and public health expertise in developing innovative solutions for critical resource allocation challenges. Their research holds the promise of improving health outcomes for underserved communities worldwide.

Works Cited

Liang, B., Xu, L., Taneja, A., Tambe, M., & Janson, L. (2025). "Context in Public Health for Underserved Communities: A Bayesian Approach to Online Restless Bandits." Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI). https://doi.org/10.48550/arXiv.2402.04933.

Welcoming G1 PhD Students

Nicholas Barnfield

Previous Institution:
McGill University (undergrad)

Research Interests:
high-dimensional probability, theoretical machine learning, continuous optimization, information theory

Hobbies:
hockey, skiing, reading (classics and historical nonfiction)

Aniket Jain

Previous Institution:
Indian Statistical Institute Kolkata

Research Interests:
high dimensional statistics, diffusion models

Hobbies:
classical music, reading

Zimeng Li

Previous Institution:
University of Science and Technology of China (undergrad)

Research Interests:
high-dimensional statistics and text data analysis

Hobbies:
Yoga-like sports, Kpop

Théo Voldoire

Previous Institution:
PSL University (Dauphine, ENS), Sciences Po Paris (master's)

Research Interests:
high-dimensional statistics, probability and sampling, applications to social sciences

Hobbies:
listening to music, viewing modern art, reading non-fiction books

Awards, Appointments & Honors

NEW POSTDOCTORAL FELLOWS (starting fall 2024)

Yongkai Chen has joined Dr. Samuel Kou's research group. Dr. Chen's research interests are in nonparametric and Bayesian methods, machine learning, and bioinformatics.

Bhanu Teja Gullapalli has joined Dr. Susan Murphy's research group. Dr. Gullapalli's research interests are in mobile health sensing and substance use and addiction.

Zhaoyang Shi has joined Dr. Tracy Ke's research group. Dr. Shi's research interests are in network analysis, geometric and topological statistics, and nonparametric and kernel methods.

Ke Sun has joined Dr. Susan Murphy's research group. Dr. Sun's research interests are in statistical reinforcement learning and machine learning.

FACULTY AWARDS & APPOINTMENTS

Morgane Austern received a 2025 National Science Foundation's (NSF) Faculty Early Career Development Program (CAREER) Award for her project on "Distributional Approximation for Sharp Finite Sample Bounds with Applications to Dependent Data and Complex Estimators."

Iavor Bojinov, Affiliate Faculty in Statistics, won the 2024 Apgar Award for Innovation in Teaching at HBS for his outstanding work in developing and delivering the Data Science for Managers required course. Starting on July 1, 2025, Dr. Bojinov is being promoted to Associate Professor.

Mark Glickman received the 2025 Founders Award from the American Statistical Association for demonstrating exceptional dedication to advancing the mission of the association through his years of service.

Kosuke Imai was selected to be a 2024 Guggenheim Fellow based on his prior career achievement and exceptional promise and for his project proposal entitled "Improving Statistical Methodology for Evaluating and Reducing Racial Disparities."

Lucas Janson received a 2025 COPSS (Committee of Presidents of Statistical Societies) Emerging Leader Award for his contributions to statistical methodology, particularly methods that leverage machine learning for statistical inference on data that is high-dimensional or adaptively collected, and for commitment to teaching, mentorship, and outreach. He also received the 2024 American Statistical Association (ASA) Noether Early Career Scholar Award due to his exceptional work on nonparametric methodology for statistical inference on data that is high-dimensional or adaptively collected.

Tracy Ke received a 2024 COPSS (Committee of Presidents of Statistical Societies) Emerging Leader Award for her contributions to statistical text analysis, statistical methods for complex network data, sparse inference and rare/weak signals, and for her service to the scientific community.

Xihong Lin was elected to be a 2024 AAAS (American Association for the Advancement of Science) Fellow in the Section on Statistics for her achievements in her discipline and excellence in communicating and interpreting science to the public.

Jun Liu was elected to the National Academy of Sciences in 2025 in recognition of his distinguished and continuing achievements in original research.

Susan Murphy received the 2024 Mosteller Statistician of the Year Award from the Boston Chapter of the

American Statistical Association in recognition of her exceptional contributions to the field of statistics and outstanding service to the statistical community.

Mark Sellke received the 2025 Rollo Davidson Prize for his contributions to applications of probability, especially in the development and understanding of algorithms for high-dimensional optimization. He was also selected to be a 2025 Sloan Research Fellow in the area of mathematics due to his creativity, innovation, and research accomplishments.

Subhabrata Sen received a Harvard 2024 Roslyn Abramson Award for excellence and sensitivity in teaching undergraduates.

Pragya Sur received a 2025 National Science Foundation's (NSF) Faculty Early Career Development Program (CAREER) Award for her project on "High-dimensional Learning and Inference from Heterogeneous Data Sources."

Julie Vu received a Harvard 2024 Excellence in Teaching Prize, which selected her for her commitment to fostering an equitable and inclusive learning environment, orientation towards pedagogical innovation, and use of evidence-based teaching practices.

José Zubizarreta, Affiliate Faculty in Statistics, received the American Statistical Association's 2025 Health Policy Statistics Section (HPSS) Mid-Career Award, which recognizes his outstanding research and demonstrated promise of continued excellence at the frontier of statistical practice that advances the aims of HPSS.

POSTDOCTORAL AWARDS

Asim Gazi in Dr. Susan Murphy's lab, received a NIH National Institute of Biomedical Imaging and Bio-engineering (NIBIB) K99 Pathway to Independence Award for his project titled "Uncertainty-Informed Decision Making for Just-in-Time Adaptive Interventions (JITAIs)."

PHD AWARDS

The Arthur P. Dempster Fund, named in honor of Faculty Emeritus Arthur Dempster, recognizes promising graduate students within the Department of Statistics, in particular those who have made significant contributions to theoretical or foundational research in statistics:

- **Souhardya Sengupta** ('24 recipient for his paper titled "Leveraging Sparsity in the Gaussian Linear Model for Improved Inference," with Dr. Lucas Janson)
- **Yufan Li** ('25 recipient for his paper on "Spectrum-aware Debiasing: A Modern Inference Framework with Applications to Principal Components Regression," with Dr. Pragya Sur)
- **Tianle Liu** ('25 recipient for his paper on "A Heavily Right Strategy for Integrating Dependent Studies in Any Dimension," with Dr. Xiao-Li Meng and Dr. Natesh Pillai)

Louis Cammarata, a 2024 PhD alum, received the 2025 Lawrence D. Brown PhD Student Award from the Institute for Mathematical Statistics for his work on dynamic network analysis with his faculty advisor, Dr. Tracy Ke.

Nathan Cheng received a 2024 National Science Foundation's (NSF) Graduate Research Fellowship Program (GRFP) award in the "Mathematical Sciences – Statistics" category for research focused on selective and causal inference.

Alan Chung received a 2024 National Science Foundation's (NSF) Graduate Research Fellowship Program (GRFP) award in the "Mathematical Sciences – Statistics" category for research focused on probability and machine learning theory.

Matthew Esmaili Mallory received a 2024 National Science Foundation's (NSF) Graduate Research Fellowship Program (GRFP) award in the "Mathematical Sciences – Statistics" category for research focused on high-dimensional universality with dependence.

Yi Zhang received the Student Research Award at the 2024 New England Statistics Symposium for her paper "Individualized Policy Evaluation and Learning under Clustered Network Interference," co-authored with Professor Kosuke Imai.

MASTER'S AWARDS

The department awards a Statistics Concurrent Master's Prize every year for the best overall performance and having demonstrated achievements in Statistics outside of coursework and contributed significantly to the department.

- **William Nickols** ('24 recipient and alum)
- **Kevin Luo** ('24 recipient and alum)
- **Danielle Paulson** ('25 recipient)

CONCENTRATOR AWARDS

The department awards a Statistics Senior Concentrator Prize every year for the best overall performance and for making significant contributions to the department.

- **Skyler Wu** ('24 recipient and alum)
- **Kenneth (Kenny) Gu** ('25 recipient)

Audrey Chang was selected as a finalist for Harvard College's Three Minute Thesis competition. Her thesis was on the topic of classification in statistics.

Mridula (Mally) Shan received First Place Prize (the Ruth and Silen M.D. Award for Public Health, Epidemiology, or Biostatistics, Bioinformatics, Physics, Chemistry, or Engineering, and Clinical or Social Science) at the 2024 New England Science Symposium, where she presented a poster on her health security dashboard project.

Skyler Wu received a 2024 National Science Foundation's (NSF) Graduate Research Fellowship Program (GRFP) award in the "Mathematical Sciences – Statistics" category for research focused on Bayesian identification, reconstruction, and forecasting of dynamical systems.

Phi Beta Kappa Recipients (2024-2025)

Raul Bodrogean, Katrina Brown, Jack Bruce, Clara Chen, Jarell Cheong Tze Wen, Jameson Cohen, Sreetej Digumarthi, Ethan Jasny, Evgeni Kayryakov, Brian Lee, Danielle Paulson, Ryan Xia, Doris Yang, Michael Zhao

Hoopes Prizes

The following students received the award for their outstanding scholarly work and theses:

2024 Recipients

- **Lauren Chen** for her submission entitled “The Graphex Model for Multilayer Networks,” supervised and nominated by Dr. Subhabrata Sen.
- **Beverly Fu** for her submission entitled “A Study of the Mutational Signatures of Structural Variation in Human Cancer,” supervised and nominated by Dr. Peter Park.
- **Andrew Garber** for his submission entitled “Average Ranks in Manipulable School Matching,” supervised and nominated by Dr. Yannai A. Gonczarowski and Dr. Morgane Austern.
- **Kevin Luo** for his submission entitled “High-Dimensional Linear Interpolation for Structured Data,” supervised and nominated by Dr. Pragya Sur.
- **Edis Memis** for his submission entitled “Beyond Empirical Averages: Berry-Esseen Bounds in Transport Distances,” supervised and nominated by Dr. Morgane Austern.
- **Ivan Specht** for his submission entitled “Reconstructing Viral Epidemics: A Random Tree Approach,” supervised and nominated by Dr. Michael Mitzenmacher and Dr. Pardis Sabeti.
- **Erik Zou** for his submission entitled “Paint and Pixel: Reimagining Impressionist Color in Monet’s Rouen Cathedral Series through Statistics,” supervised and nominated by Dr. Jeffrey Hamburger and Dr. Alexander Young. Zou also received the 2024 Harvard Taliesin Prize.

2025 Recipients

- **Elliot Chin** for his submission entitled "Population Encoding Dynamics within the Mouse Auditory Cortex,” supervised by Dr. Anne Takesian and Dr. Subhabrata Sen.
- **Kenneth (Kenny) Gu** for his submission entitled “Bootstrap Methods for High-Dimensional Linear and Logistic Regression,” supervised and nominated by Dr. Pragya Sur.
- **Brice Laurent** for his submission entitled “Epistasis between Somatic Mutations in the Normal Blood,” supervised and nominated by Dr. Kamila Naxerova.
- **Victoria Li** for her submission entitled “Abortion after Dobbs: A Causal Inference Approach to Changes in Abortion and Fertility,” nominated by Mr. William Nickols and supervised by Dr. Stephen Sachs and Dr. Kevin Rader.
- **Jamie Liu** for his submission entitled “Accelerating Inference: Mitotic Stein Variational Gradient Descent for Bayesian Analysis of Dynamical Systems,” supervised and nominated by Dr. Samuel Kou.
- **Jacob Miller** for his submission entitled “Minding the Attention Gap: Statistical Approaches to Inattentive Survey Respondents,” supervised and nominated by Dr. Isaiah Andrews.
- **Moses Stewart** for his submission entitled “Constructing an Instrument as a Function of Covariates," supervised and nominated by Dr. Rahul Singh.
- **Michael Zhao** for his submission entitled “Words Speak as Loudly as Actions: Deep Learning Methods for Stance-Based Ideal Points from Congressional Speeches,” nominated by Mr. António Câmara and supervised by Dr. Ariel Procaccia and Mr. António Câmara.

Sample of 2024 PhD Graduates:
Where They Are Now...

Louis Cammarata

Dissertation Title: Connecting the Dots: Network Testing, Community Estimation, and Genomic Applications

Dissertation Committee: Dr. Tracy Ke, Dr. Caroline Uhler (MIT), Dr. Morgane Austern

Where I am Now: Consultant at Bain & Co's Boston office - I look forward to continued learning and getting to know my colleagues!

Favorite Stats Memory:
So many! If I had to pick one, it would probably be my time on the social committee, and especially co-organizing the department retreat with some of my classmates :).

Dae Woong (David) Ham

Dissertation Title: Design-Based Causal Inference: Applications to Social Sciences and Industry

Dissertation Committee: Dr. Kosuke Imai, Dr. Lucas Janson, Dr. Iavor Bojinov, Dr. Luke Miratrix

Where I am Now: Assistant Professor at the University of Michigan

Favorite Stats Memory: My dissertation presentation was one of my favorites. It was a bonus getting the opportunity to acknowledge the faculty and seeing their expressions! I've thoroughly enjoyed working with all the faculty from Harvard.

Jiaze Qiu

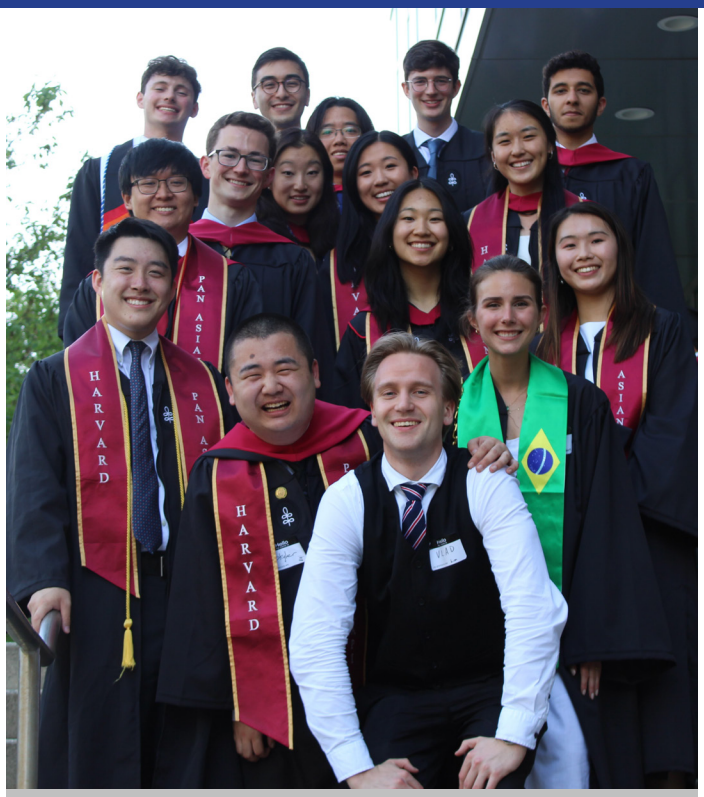
Dissertation Title: Variational Inference in High-dimensional Bayesian Regression Models

Dissertation Committee: Dr. Subhabrata Sen, Dr. Jun Liu, Dr. Yue Lu, Dr. Sumit Mukherjee

Where I am Now: Five Rings LLC, Quantitative Researcher

Favorite Stats Memory: Poker games on the 9th floor of the Science Center.

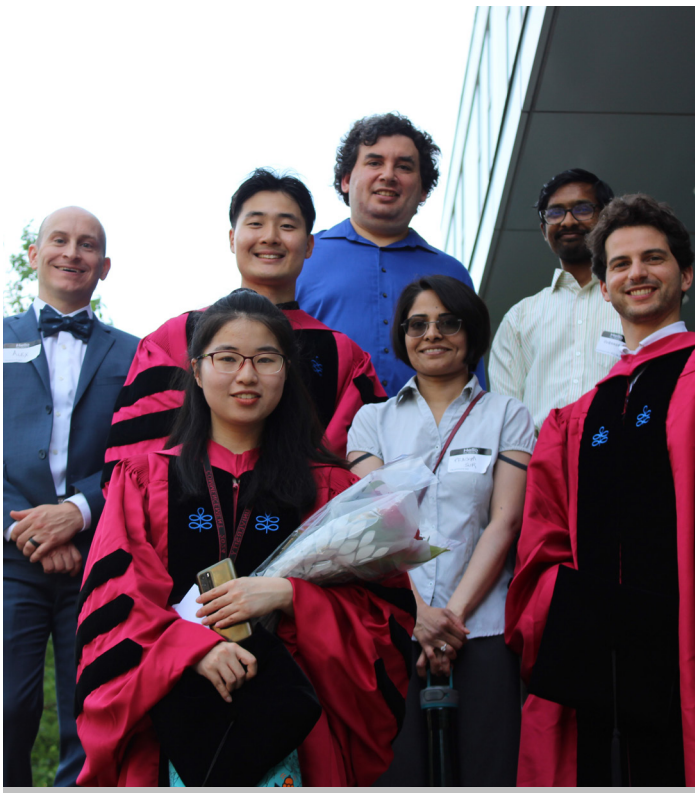
Commencement Celebration - May 23, 2024



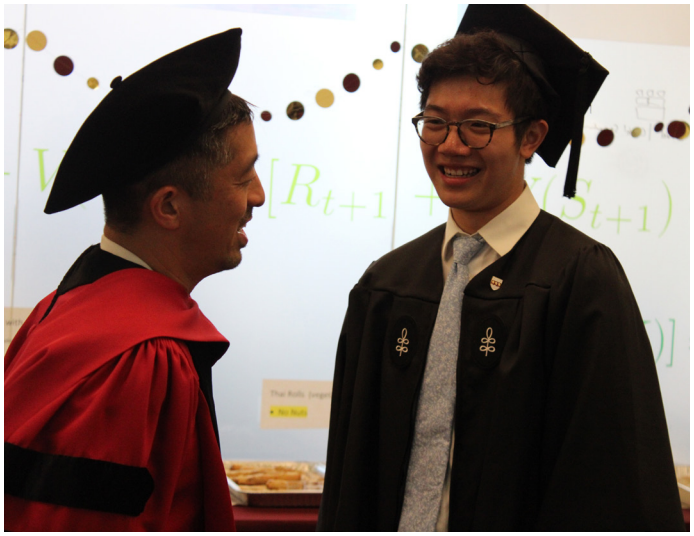
Bachelor of Arts Graduates

Master of Arts Graduates

PhD Graduates



PhD Graduates with Faculty



Commencement Celebration – May 29, 2025

Bachelor of Arts Graduates

Master of Arts Graduates

PhD Graduates

PhD Graduates with Faculty


EXIT

ADDRESS:	1 OXFORD STREET, SUITE 400 CAMBRIDGE, MA 02138
PHONE:	617-495-5496
EMAIL:	statistics@fas.harvard.edu
WEBSITE:	https://statistics.fas.harvard.edu
SOCIAL MEDIA:	

Support our Department's next generation of leaders:

Give