Harvard Data Science Review • Issue 7.2, Spring 2025

Statistics and AI: A Fireside Conversation

Xihong Lin[1] Tianxi Cai[1] David Donoho[2] Haoda Fu[3] Tracy Ke[1] Jiashun Jin[4] Xiao-Li Meng[1] Annie Qu[5] Chengchun Shi[6] Peter Song[7] Qiang Sun[8,9] Wenyi Wang[10] Hulin Wu[11] Bin Yu[12] Heping Zhang[13] Tian Zheng[14] Harrison Zhou[13] Jin Zhou[15] Hongtu Zhu[16] Ji Zhu[7]

[1]Harvard University, Cambridge, Massachusetts, United States of America,
[2]Stanford University, Stanford, California, United States of America,
[3]Amgen, Thousand Oaks, California, United States of America,
[4]Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America,
[5]University of California at Irvine, Irvine, California, United States of America,
[6]London School of Economics and Political Science, London, England, United Kingdom,
[7]University of Michigan, Ann Arbor, Michigan, United States of America,
[8]University of Toronto, Toronto, Ontario, Canada,
[9]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates,
[10]University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America,
[11]University of Texas Health Science Center at Houston, Houston, Texas, United States of America,
[12]University of California at Berkeley, Berkeley, California, United States of America,
[13]Yale University, New Haven, Connecticut, United States of America,
[14]Columbia University, New York, New York, United States of America,
[15]University of California at Los Angeles, Los Angeles, California, United States of America,
[16]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

The MIT Press

DOI: https://doi.org/10.1162/99608f92.c066fe9c

ABSTRACT

A 3-hour webinar titled “Statistics and AI – A Fireside Conversation” was held on Sunday, March 17, 2024, attracting an online audience of approximately 1,000. The event featured three sessions aimed at engaging the statistical community on key topics in the AI era: addressing statistical challenges and opportunities (Panel I), evolving the publication process (Panel II), and advancing next-generation statistical pipelines and resources (Panel III). Panel I examined issues such as dwindling talent, shifting funding landscapes, and AI's rapid rise, highlighting the need for statistical rigor, interdisciplinary collaboration, and innovative approaches to shape the future of AI. Panel II emphasized the importance of streamlining the publication process, fostering impactful research, and prioritizing workflows and data quality. Panel III focused on modernizing statistical education by integrating AI and deep learning, promoting interdisciplinary collaboration, and maintaining foundational principles such as uncertainty and reproducibility. These discussions collectively outlined a strategic roadmap for ensuring the relevance and advancement of statistics in the age of AI.

These discussions were organized by (in alphabetical order) Xihong Lin (Harvard University), Tracy Ke (Harvard University), Tian Zheng (Columbia University), Jing Zhou (University of California at Los Angeles), and Hongtu Zhu (University of North Carolina at Chapel Hill).

In the dynamic landscape of statistical science, the fireside chat organized by the Stats Up AI Alliance (https://statsupai.org/) and the International Chinese Statistical Association (ICSA) emerged as a seminal event, bringing together leading experts to explore the evolving role of statistics in the era of artificial intelligence.

Keywords: artificial intelligence, statistical research, publication culture, statistics education, reproducibility, team science

---

Tian Zheng: I would like to begin by expressing my gratitude toward my co-organizers and the esteemed panelists, for sharing your insights on this important topic. Having all the panelists and participants spending their Sunday morning together demonstrates how our community recognizes the significance of AI to the future of our field. The session is being recorded and live-streamed on YouTube (https://www.youtube.com/live/iTupgaXHiaY?si=tR-7S6foVD6glaXD; www.youtube.com/@StatsUpAI). It is meant to generate discussions and dialogues in the statistical community, reflecting our collective excitement and concerns about integrating statistics and AI. I would also like to point out the 2019 National Science Foundation–sponsored report *Statistics at Crossroads: Who is for the Challenge?*, which laid the groundwork for today’s discussion. Let’s continue our efforts to empower statisticians in this rapidly changing domain. Today’s panels will address the challenges and opportunities that arise as statistics intersect with AI, fostering a conversation that could shape the future of the field.

Xihong Lin: Statistics has a critical role to play in understanding and shaping both scientific research and societal decision-making in the age of artificial intelligence. We're living through a time where AI has had a transformative impact on how we use data and learn from it, and this transformation is filled with both excitement and concerns within our field. Today's discussions will reflect this, and we have structured the event around three key sessions focused on the challenges and opportunities in statistical research, publication, and education in the AI era. I want to use this occasion to stress the importance of being open-minded, thinking outside the box, and embracing risk as we move forward. To me, the integration of statistics and the rapid advancement of AI is twofold: first, to harness AI's transformative power to strengthen statistical science, and second, to make AI more trustworthy by embedding it with the rigorous principles of statistics. Ultimately, our goal is to push the boundaries of real-world scientific discovery and its applications.

Panel I: The Challenges and Opportunities for the Statistics Community in the Era of AI

Moderator: Xihong Lin (Harvard University)

Panel:
- **Hongtu Zhu (University of North Carolina at Chapel Hill)**
- **Jiashun Jin (Carnegie Mellon University)**
- **Tianxi Cai (Harvard University)**
- **Haoda Fu (Eli Lilly)**
- **Tracy Ke (Harvard University)**
- **Harrison Zhou (Yale University)**

Discussant: David Donoho (Stanford University)

Hongtu Zhu: I see three significant challenges: 1) Diminishing pool of talents of students in statistics and biostatistics departments, 2) the dwindling opportunity for new and existing funding opportunities, and 3) the emerging new domains with few statisticians involved. Technological foundations have been driving the AI wave, especially the transformative impact of academic innovations like ImageNet and convolutional neural networks, which were later embraced by the private sector. The private sector has invested heavily in developing products that have gained significant attention across social, government, and public sectors. Statistical theory plays a critical role as a collection of mathematical tools and formulations that validate statistical methods under reasonable assumptions. Theoretical statistics, such as empirical processes, have been vital in advancing statistical methods in key applications. However, statistical theory has struggled to keep pace with the complexities brought on by the AI era—complexities in both data and methodologies, as Leo Breiman discussed back in 2001 (Breiman, 2001), particularly regarding stochastic data models versus algorithmic models. It's therefore important to recognize that statistical models are simplified representations of reality, and their usefulness depends on how accurately they approximate the real world. We need deeper theoretical

insights into emerging and challenging topics like random forest, deep learning, and deep reinforcement learning—areas that are crucial for understanding and advancing these complex methods. It is also vital to have a system for evaluating methods, which now includes data sharing, code sharing, and competitive challenges, as David Donoho discussed in 2024 (Donoho, 2024). This shift is changing how we assess and validate statistical methods in contemporary research, and it's something we must continue to explore and adapt to.

Jiashun Jin: I'd like to share some observations and suggestions regarding AI and theoretical statistics. While I acknowledge the benefits AI brings to areas like editing, visualization, and research (such as in protein folding), I also have serious concerns about its broader impact. One major concern is AI's potential threat to job markets in fields like statistics, operations research, and applied mathematics. Additionally, AI research is often costly and produces few significant breakthroughs compared to its costs. Another issue I see is AI's tendency to overshadow other fields, sometimes implying their irrelevance and risking the erasure of important existing literature. A critical problem is that AI frequently provides incorrect answers. From my experience working with Google researchers on the nonnegative inverse eigenvalue problem, I have seen firsthand how AI can summarize complex materials but fail to grasp the underlying mathematics, often producing outputs that are subtly flawed. This highlights AI's limitations when dealing with intricate concepts, and it's a reminder to be cautious when relying on AI in complex research. At the same time, I want to recognize the value of theoretical statisticians, especially in industries like finance, where their skills and the 'white box' algorithms they develop are highly appreciated. However, there are concerns, many of which have been raised in the WeChat group, about the slow progress in theoretical statistics research. The heavy investment in unrealistic settings and the reliance on journal publications for evaluation can often penalize unconventional research. To tackle these challenges, I propose we foster an environment that encourages unconventional research, reforms curricula and publication methods, and perhaps even establishes a Hilbert's 23 problems for data science. This would allow us to focus resources on solving key problems and position our field to have a greater social impact.

***Tianxi Cai*:** There are vast opportunities for statisticians. With AI's growing influence, there are also pressing actions we need to take. We're seeing a significant increase in the amount of data—across different modalities, sizes, and types—which opens up exciting opportunities to develop new methods for analyzing nonstandard data like images, text, and voice. But with these opportunities come challenges: we're facing more heterogeneity, greater complexity, and the need to develop robust and privacy-conscious approaches. The potential we now have to make a real impact is higher, both in research capacity and in terms of social influence. This means we need to focus on end-to-end product development. There's still a gap between statistical theory and AI, and we feel the pressure to redefine what our discipline's major contributions and strengths are, especially as machine learning and computer science continue to advance. Statisticians today have more tools than ever to access, analyze, and interpret different data types, but we still face a significant

challenge when it comes to applying these methods to real-world problems. I agree with the idea that statisticians should aim to become 'full stack data scientists,' capable of working across the entire data science spectrum—from data collection all the way to creating impactful end products. It's crucial that we close this 'last mile' gap, where our statistical innovations too often remain stuck in academic publications instead of making their way into practical, real-world applications.

***Haoda Fu*:** I would like to discuss, as an example, the development and significance of XGBoost, one of the leading algorithms for structured data analysis, and what it means for theoretical statistics. Its origin can be traced back to Schapire's (1990) concept of weak classifiers coming together to form a robust committee through majority voting. This idea laid the foundation for AdaBoost, developed in 1997 by Freund and Schapire (1997), which was initially hailed as the best off-the-shelf classifier. However, AdaBoost's performance was inconsistent across different scenarios, which caused some confusion. A pivotal paper in 2000 connected AdaBoost to minimizing exponential loss, providing a theoretical explanation and sparking the development of gradient boosting and other advancements (Friedman et al., 2000). Over the next 15 years, gradient boosting became the standard, but it faced challenges with handling large data sets and lacked parallel computing capabilities for modern GPU architectures. This gap was addressed in 2016 by Tianqi Chen, a PhD student at the University of Washington, who developed XGBoost (Chen, 2016). He designed it to leverage GPU support and incorporated insights from computer architecture, addressing those earlier limitations. As illustrated by this example, I believe it's essential for us to adopt an architecture-algorithm codesign approach. This involves mastering areas like low-level programming, data structures and algorithms, optimization, and design patterns. Also important is the proficiency in tools like Git and involvement in open-source projects to push the field forward. To further advance the field, I suggest a few key actions: embracing open-source educational materials for more efficient teaching, revamping conference formats with journal-style reviews to speed up the publication process, and hosting focused workshops to create immersive learning environments for scientists. Adopting a Hilbert's 23 problems approach for data science could also help us identify and tackle key theoretical and applied challenges. Lastly, I'd like to reflect on Leo Breiman's (2001) "Two Cultures" paper, which highlights the value of algorithm-based data analysis methods. Rich Sutton's (2019) *Bitter Lesson* reinforces this, emphasizing that general methods leveraging computation are often the most effective. As statisticians, we should lead the way in developing general models, while mastering modern high-performance computing environments. Ultimately, there is a strong need for a multidisciplinary approach to statistics and machine learning, with a focus on bridging theory and practice to increase the field's impact through a product-oriented mindset.

***Tracy Ke*:** I would like to highlight the intersection and distinct roles of statistics and AI by sharing an experiment on topic modeling using large language models (LLMs). My team and I conducted an experiment where we fine-tuned a pretrained LLM, which required manually labeling documents. We then tried assigning

topic weights using GPT-4. The process turned out to be quite expensive, with an estimated cost of over $50,000 to process a specific data set, especially when compared to the minimal cost and time required by statistical methods. Interestingly, we found that even ChatGPT eventually reverted to using traditional statistical methods for topic modeling. This experience highlighted that, despite AI's impressive capabilities in advanced tasks, it doesn't necessarily excel in more classical tasks where economic efficiency is key—tasks that require unsupervised, low-cost, and reliable methods. In many of these settings, statistical methods still hold the advantage. This raised the question of how we can combine AI and statistics, especially in decision-making contexts such as college admissions, medical treatments, or piloting flights, where AI's rich outputs need to be transformed into actionable decisions through statistical modeling and inference. I proposed that we should think of AI as a new type of data source—one that requires dimension reduction, signal detection, and new statistical models tailored to AI-generated data. Furthermore, statistics can offer a critical role that helps inspire new methods and identify limitations, in contrast to AI, where the role of theory is less clear and more often driven by engineering efforts. There's a need for new AI-related papers and criteria to assess their intellectual merit. Lastly, AI is very resource-intensive and partnerships between universities and tech companies have become ever more important. These partnerships could provide access to essential resources like computing power, nonpublic training data, and engineering support, which would greatly enhance our ability to make significant contributions in AI-related research.

***Harrison Zhou*:** When it comes to statistical theory, AI foundations, and their applications in data science, I thought about Wald's (1939) contributions to statistical estimation and hypothesis testing. This work laid a clear foundation for theoretical statistics, while the foundation of AI remains less well-defined. While the past decade has witnessed the significant growth in statistics departments, the rapid development of AI, especially large language models (LLMs), has caught more public attention. I believe LLMs have great potential for the statistics community, particularly through fine-tuning pretrained models and handling data collection. However, there are challenges in defining the theoretical foundation of AI. Statisticians might not fully understand the inner workings of LLMs or AI, but I argued that, much like Wald in 1939, we can still contribute meaningfully even without a complete understanding of the underlying mechanisms. In fact, gaining insight into the mysterious emergent abilities of LLMs could eventually transform how we approach their architecture, pretraining, and fine-tuning. There is a long-standing tradition of collaboration in statistics departments. Now is the time to capitalize on new developments in AI, and foster deeper collaborations in areas like data collection and model fine-tuning; statisticians will probably need to develop stronger coding skills to be effective in these areas. As a concrete step, statistics departments should begin hiring specialists in LLMs or AI, viewing AI as a natural extension of data science. This could open up more collaborative opportunities across a range of disciplines, including the humanities, law, and social sciences. We should also invite AI researchers into the editorial boards of statistics journals and encourage junior faculty to publish in

top AI conferences. It's crucial for statisticians to take part in the AI evaluation process. Integrating AI into statistics and adopting a more collaborative approach will allow us to leverage the strengths of both fields.

***David Donoho*:** These panel discussions, I believe, are significantly shaping the future of our discipline. One of the key issues facing statistics today is the gap between the concerns of faculty and graduate students, and I appreciated how the panel focused on topics important to the younger generation. When it comes to AI's foundation, we should note the difference in how AI and statistics approach problems. AI often learns from examples, sometimes without any clear structure, which contrasts with the more structured methods in statistics, such as Wald's foundational work in theoretical statistics. Can statisticians still contribute to the field, without fully understanding AI? The expansion of statistics departments over the last decade is remarkable and our community should be lauded for our tradition of collaboration. While some may feel that AI threatens to overshadow statistics, I suggest that we take a more welcoming stance toward AI—maybe even embrace it more openly. Let us 'wear AI t-shirts' to show our willingness to participate and collaborate. Looking at data from Google Books Ngram Viewer, I noticed how the popularity of terms like 'statistical analysis,' 'data science,' and 'AI' has fluctuated, with AI now being widely appropriated by other fields. But rather than viewing AI and statistics as separate entities, maybe we see AI as something that often integrates successful techniques from other fields, including our own. Let's not overlook the financial resources pouring into AI research. Statisticians can tap into these through collaboration, especially with experts in computer science or medical fields who are working on AI applications. This could give us access to large language models for various uses, presenting an incredible opportunity for statisticians to merge their expertise with AI advancements. Let me be clear: there's enormous potential for synergy between AI and statistics, and we statisticians should engage proactively with the developments in AI to harness that potential.

Panel II: Publication Process of Statistics in the Era of AI

Moderator: Heping Zhang (Yale University)

Panel:
- **Annie Qu (University of California at Irvine)**
- **Ji Zhu (University of Michigan)**
- **Chengchun Shi (London School of Economics and Political Science)**
- **Xiao-Li Meng (Harvard University)**

Discussant: David Donoho (Stanford University)

Annie Qu: I agree that the challenge of slow journal publication poses a serious issue in our rapidly advancing era of AI. There's often a gap of years between conducting research and publishing it, which can make the findings feel outdated by the time they reach readers. For research to be accepted in top-tier journals, it needs to present innovative ideas and sound theory, backed by strong empirical data and clear, accessible writing.

There have been some promising improvements over time, but more is needed. On the other hand, in fields like medical sciences that publish quickly, the rapid turnaround may sometimes compromise review quality. One way to address this is to advocate for the early rejection of papers with limited potential and incentives and rewards for productive and responsive associate editors and referees who provide faster, and higher quality reviews. High-quality papers that tackle important problems with a significant impact in other fields is as essential as papers that address major problems within our field. We should encourage papers that address original real-world data, rather than unconvincing or trivial examples. Our research and our journals should both pursue continuous innovation and creativity. Here, I would like to announce a special issue on Statistics in the Age of AI of the *Journal of the American Statistical Association*. I hope it will foster integration of statistics into AI to drive scientific discovery forward.

Ji Zhu: I agree that innovation is crucial not only from the authors' perspective but also for journal editors, associate editors, and reviewers. Landmark machine learning papers by statisticians, [on] random forests, MARS [Multivariate Adaptive Regression Splines], and the statistical view of boosting, made significant contributions and helped establish statisticians as key players in machine learning, despite lacking full theoretical explanations or asymptotic theorems. These groundbreaking works were once published in top statistical journals. I am wondering whether such innovation would be as welcome today. Editorial boards of top statistical journals could collectively consider how we could encourage more forward-thinking research in our evaluations. One constant challenge these days is finding good reviewers. Incentives for timely reviews are helpful, but they may not be enough. Maybe, we could consider the idea of open reviews after acceptance, which could foster more discussion and increase visibility for published work. Regarding *The Annals of Applied Statistics*, our mission is to publish application-driven research that advances the field. We are looking to include more data-centric and discussion papers. *The Annals of Applied Statistics* is also interested in publishing the statistical methodology behind important scientific discoveries. This will increase the impact of statistical journals and reach a broader audience. In other words, while scientific discoveries should be published in domain-specific journals, the sophisticated statistical modeling behind these findings could be featured in statistical journals. It is time for the statistical community to move beyond discussion and take concrete steps to enrich the field of statistics.

Chengchun Shi: I would like to offer a comparison of the publication processes between statistics and machine learning. Looking at the data from a quick Internet search, we could see the number of papers presented at ML (machine learning) conferences has grown substantially faster than those in leading statistical journals. ML conferences have a much faster review process, unlike the extended, multi-round reviews common in statistics, where response letters sometimes reach 30 pages. This lengthy process can be inefficient. Here are a few ideas of mine to improve the publication process in statistics: 1) shorten review cycle: we could consider publishing shorter papers and setting page limits on response letters, encouraging authors to address

only major comments to keep the process more focused; 2) reduce the number of revision rounds: sometimes referees provide nonessential feedback that doesn't significantly enhance the paper; we should encourage reviewers to focus on essential comments to save time and streamline the review cycle; 3) increase the number of cited references: references shouldn't be counted toward page limits; this will directly encourage interaction within the field and broaden our impact; 4) broaden journals' scope: we should be open to publishing emerging AI topics. This will require openness among associate editors and reviewers. This will attract submissions from other domains and increase the impact of statistical journals. Our journals should adapt to an evolving scientific landscape in the age of AI. To begin, let's improve the efficiency of our publication process.

Xiao-Li Meng: I am going to share what I call 'The Three Ps for Statistical Journals,' drawing from my experience with the *Harvard Data Science Review*. This experience has given me some unique perspectives on the intersection of statistics, data science, and AI. Statistical journals should be more proactive in shaping statistical research within the data science ecosystem, rather than simply publishing whatever work is submitted. *Harvard Data Science Review* has taken this proactive approach, with the hope to define and influence the field of data science. *Promoting* statistics, or communication with the broader data science community, is very important. Discussion articles, while uncommon in AI and computer science, can be powerful tools for sparking deeper reflection and engagement in our field. Additionally, I suggested that journals focus more on the *Processes* in data science, not just the final outputs like algorithms or theorems. This shift can foster meaningful discussions about the data science process itself, which requires intricate statistical insight. For example, I've invited contributions from various fields to discuss the data life cycle and its broader implications, encouraging statisticians to explore the initial stages of data conceptualization and the final stages of data reuse, both of which are crucial for scientific reproducibility. Another example is the industry's focus on data quality, where data scientists often spend a majority of their time on data cleaning. We should put a greater emphasis on *Preserving* the integrity of data. Journals should include a required section on data quality in every article involving data analysis, highlighting the importance of high-quality data from the outset. This is an area where statisticians can make a profound impact on the future of statistics, data science, and AI.

David Donoho: We should reflect on the evolution of statistical publication and research, especially in light of the rapid advancements in computer science and artificial intelligence. I would like to offer an analogy between the traditional statistical paper—rooted in a century-old model of deriving theorems from probabilistic generative models—and an F-35 fighter jet: highly sophisticated and expensive, yet potentially outpaced by newer, more cost-effective technologies like drones. Similarly, traditional models may struggle to compete with the flexibility and efficiency of newer approaches. Computer science conferences have created disruptive impacts. They have climbed in scientific publication rankings due to their focus on workflows, data properties, and system performance—insights that traditional statistical models often overlook. To stay relevant, statistical

publications should adapt by discussing workflow aspects more integrally and embracing the scientific community's evolving expectations and methodologies. I agree very much with the ideas raised by my fellow panelists. We should have a more streamlined refereeing process that prioritizes specific improvements over extensive rewrites. Statistical journals must take an active role in shaping their place in data science and AI, and engage with new publication platforms and venues to remain relevant and impactful in this rapidly changing field.

Panel III: Advancing Statistical Next-Generation Pipelines and Resources in the Age of AI

Moderator: Peter Song (University of Michigan)

Panel:
- **Wenyi Wang (MD Anderson Cancer Center)**
- **Hulin Wu (University of Texas at Houston)**
- **Qiang Sun (University of Toronto and MBZUAI)**
- **Bin Yu (University of California at Berkeley)**

Discussant: David Donoho (Stanford University)

Wenyi Wang: Let's first discuss how students choose their majors these days, especially with all the buzz around artificial intelligence (AI) in the news. At Texas Medical Center, we offer rigorous statistical training but have few students enrolling, while a nearby medical school sees a surge in students focusing on computational biology, many with a background in statistics. However, these students may not receive as much advanced statistical training due to the extensive training in biological research in their program. This leads to a high imbalance in training and overexposure to large language models. A direct consequence of this imbalance is that AI, particularly in cancer research, is receiving significant attention and funding, making it central to scientific advancement. Yet, statisticians skilled in AI may still struggle to secure funding, as grants addressing specific biological questions often favor those trained in computational biology. Statistics are crucial for rigorous scientific discovery. I am concerned that if we don't prioritize statistical training, science as a whole could suffer. We need stronger leadership training in statistics and increased efforts to attract more talent to the field. Statisticians need a united voice, and perhaps even rebranding the field, to reinforce the idea that statistics is a fundamental part of science, especially as AI continues to grow. Here, I would like to mention Stats-Up AI Alliance, a new initiative designed to strengthen the role of statisticians in an AI-driven world, underscoring the need for statisticians to evolve and collaborate as technology advances.

Hulin Wu: I'd like to share my thoughts on an important question: 'Do we want to expand statistics training to include data science and AI technologies?' Let's start by examining the core identity of statistics, which is

often defined through keywords like data, mathematics, and uncertainty. To me, statistics stands on four main pillars: its theoretical foundation rooted in mathematics, the principles and methods we use to address random variation, the computational implementation borrowed from computer science, and the application, which requires domain knowledge. As we consider branching into fields like AI and data science, I think it's crucial to ask how much of our unique focus on randomness and uncertainty should remain at the core of what we do. Data science, for instance, covers a lot of ground beyond traditional statistics—everything from data collection and cleaning to prediction and engineering. These steps often don't involve the same level of uncertainty that's central to statistics. So, I wonder: How much should we extend into these nonrandom areas, which align more closely with computer technology? At the University of Texas Health Science Center at Houston, where I work, we faced this question head-on. We have a large PhD program, and we recently changed the degree title to 'PhD in Data Science' to better reflect the field's direction. We introduced new courses and certificates in data science as part of this shift. But with this expansion, we're constantly asking ourselves: how far should we go without losing the essence of what makes statistics unique? To keep this balance, we recognize the need for more faculty with expertise in computer science and data science. In conclusion, I believe there's a real challenge in balancing our expansion into data science and AI while preserving the distinct identity and principles of statistics. Any decision to broaden the scope of statistics must be made thoughtfully, considering both the degree of expansion and the resources available to support it.

Qiang Sun: I would like to talk about how statistical training should evolve in this era of data science and AI. The essence of statistics has always been closely tied to mathematics and the concept of uncertainty. The question is, how should it adapt while holding on to its core identity in the face of rapid technological advancements? Real-world relevance is now key. Statistics is shifting toward production, where our work must translate into products that genuinely benefit science and society. This shift means statisticians need a broader skill set, including system engineering and the ability to work across disciplines. I envision a bold reimagining of the curriculum that includes deep learning, AI, and essential engineering skills. I think we should create specialized tracks within statistical training to meet different interests, from theoretical to applied data science. To make this happen, we need a more flexible and inclusive approach to recruiting faculty and students. Open-source data sets from various sectors would also be a huge asset to drive innovation, and there's an urgent need for substantial investments in computing infrastructure to support advanced research. I also believe statisticians need to showcase their work more effectively, similar to the practices observed in the machine learning community. Publishing on diverse platforms and promoting findings beyond traditional statistical outlets can increase our impact. In summary, my vision for statistics is one that embraces change, equips future professionals with both theoretical and practical skills, and fosters a collaborative spirit to keep the field at the forefront in this data-driven age.

Bin Yu: I believe that machine learning (ML) and data science (DS) are not just essential to, but at the forefront

of modern statistics (Stat). There's a significant overlap among these fields, and we should work toward a unified, inclusive approach in our academic courses and research topics. Statistics departments need to actively integrate ML and DS into their curricula and research. The goal of Stat/ML/DS education should be to ensure our graduates can achieve computationally reproducible and scientifically replicable data-driven results in solving domain-specific problems. The best way to reach this goal is through a growth model of learning—or, simply, learning how to learn—by teaching the full data science life cycle with an emphasis on reproducibility, ethics, leadership, and team culture. Modern curriculum should evolve to take on a data science life-cycle perspective, extending beyond algorithms and modeling to reach data-driven conclusions that solve real-world problems. This includes everything from formulating domain problems, accessing or collecting data, cleaning data, modeling, and finally communicating results to the appropriate audience. Machine learning and AI skills, like deep learning and reinforcement learning, are as important as those in traditional statistical modeling. Another focus is on scientific reproducibility through quantitative thinking and a broad approach to uncertainty quantification that includes uncertainties from data cleaning and model choices. My former student Karl Kumbier and I introduced the framework of veridical data science (VDS) in our 2020 *PNAS* paper (Yu & Kumbier, 2020). VDS moves beyond the traditional 'true-model' approach and structures data analysis around the data science life cycle, guided by the principles of predictability, computability, and stability (PCS). PCS builds on and integrates best practices from both statistics and machine learning, providing a framework for developing responsible data science pipelines and tools. This VDS approach is also featured in our new MIT Press book, *Veridical Data Science: The Practice of Responsible Data Analysis and Decision Making*, coauthored by me and Rebecca Barter (Yu & Barter, 2024). The book, available for free online at vdsbook.com, serves as a primary or secondary textbook for upper division undergraduate and entry-level graduate courses in Stat/ML/DS. It's also a solid introduction to data science for domain scientists and data science practitioners, making it a valuable resource for education and professional development in the age of AI.

David Donoho: I'm impressed by the progress at the Texas Medical Center; it's a great example of successful adaptation that others could look to as a model. Like many on the panel, I believe that the field of statistics should continue to broaden to align with AI and data science, but it's essential that we do so without straying from our core values. Looking back at the changes in academic courses over the past decade, I see a clear shift from traditional subjects to more contemporary topics like machine learning. This reflects a transformative time for educational curricula across departments in mathematics and computer science. In our discussion, many emphasized the importance of creating a learning environment that encourages students to embrace new tools and skills, such as computational reproducibility and AI technologies. I also suggest that we strengthen shared resources—things like databases of case studies, documentation, and code repositories—to improve access to data and foster collaboration within our community. In conclusion, let's put forward proactive efforts

to reshape statistics education to meet the demands of our rapidly changing world, while staying true to the fundamental principles of our discipline.

Disclosure Statement

Xihong Lin, Tianxi Cai, David Donoho, Haoda Fu, Tracy Ke, Jiashun Jin, Xiao-Li Meng, Annie Qu, Chengchun Shi, Peter Song, Qiang Sun, Wenyi Wang, Hulin Wu, Bin Yu, Heping Zhang, Tian Zheng, Harrison Zhou, Jin Zhou, Hongtu Zhu, and Ji Zhu have no financial or nonfinancial disclosures to share for this interview.

References

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In B. Krishnapuram & M. Shah (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Donoho, D. (2024). Data science at the singularity. *Harvard Data Science Review*, *6*(1). https://doi.org/10.1162/99608f92.b91339ef

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *The Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016218223

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1007/BF00116037

Sutton, R. S. (2019). *The bitter lesson*. http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses**.** *The Annals of Mathematical Statistics*, *10*(4), 299–326. https://doi.org/10.1214/aoms/1177732144

Yu, B., & Barter, R. L. (2024). *Veridical data science: The practice of responsible data analysis and decision making*. MIT Press.

Yu, B., & Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy of Sciences*, *117*(8), 3920–3929. https://doi.org/10.1073/pnas.1901326117