



## Roundtable on Data Science Postsecondary Education

Meeting #10 - March 29, 2019

The 10th Roundtable on Data Science Postsecondary Education was held on March 29, 2019, at the Arnold and Mabel Beckman Center of the National Academies of Sciences, Engineering, and Medicine in Irvine, California. Stakeholders from data science education programs, government agencies, professional societies, foundations, and industry convened to discuss common challenges in establishing, maintaining, and evolving partnerships in data science between academia and industry; and to learn about ongoing programs at academic institutions and research groups around the United States. This Roundtable Highlights summarizes the presentations and discussions that took place during the meeting. The opinions presented are those of the individual participants and do not necessarily reflect the views of the National Academies or the sponsors. Watch meeting videos or download presentations at [nas.edu/DSERT](https://nas.edu/DSERT).

Eric Kolaczyk, Boston University, welcomed Roundtable participants and noted that although partnerships between industry and academia have existed for years, such collaborations are now occurring at a different scale and with a new intensity, owing in part to the emergence of data science. Academia–industry partnerships enable students to integrate data science skills to address real-world problems. Students also gain insight into the industry workforce and potential career opportunities. And members of industry can experiment with minimal investment, tapping into new developments from academia and identifying prospective hires. Challenges to developing successful partnerships include initiating interactions, maintaining support from institutions, aligning expectations, and navigating issues of data sharing and intellectual property (IP). Roundtable speakers and participants discussed best practices to create effective academia–industry collaborations around data science research and education.

### OVERVIEW OF ACADEMIA–INDUSTRY PARTNERSHIPS

Lise Getoor, University of California, Santa Cruz

Getoor commented that data science presents a unique opportunity for new models of engagement to address challenges in academia and industry. Existing models of academia–industry collaboration include sponsored research, summer internships, capstone projects, visiting researcher status, and formal industrial membership programs.

She explained that there is no one-size-fits-all model for academia–industry partnerships; it is important to develop a shared vision around building a thriving data science education and research community that spans academia and industry, with students at the center. The industry “ecosystem” includes “heavy-hitters” in data science (such as Google, Amazon, Microsoft, IBM, Facebook), start-ups, and new adopters, each with different needs and opportunities. Styles of collaboration (e.g., to educate, share expertise, and/or collaborate on research), expectations, and timelines differ both between industry and academia and across companies. The needs of and opportunities within the academic ecosystem vary based on the institution’s ranking, location, and major disciplines. Cultural differences among data science domains can also be a consideration; for example, the tradition of project-based work that can align well with industry expectations is more common in statistics than in computer science and mathematics, in her experience.

Getoor provided a brief overview of the [Data Science D<sup>3</sup> \(Data, Discovery, and Decisions\) Research Center](#) at the University of California, Santa Cruz. It focuses on academia–industry collaborations around richly structured sociobehavioral data and uses probabilistic programming language to develop templates for sociotechnical systems. The research center follows the National Science Foundation’s (NSF’s) Industry–University Cooperative Research model, which provides a template for addressing IP issues. Industry benefits from the fresh perspectives and research that emerge from partnerships like these, and students benefit from opportunities to work in teams and conduct research with data for real-world problems.

### **PANEL ON MECHANISMS FOR ENGAGING AND FOSTERING INDUSTRY PARTNERSHIPS**

**Adam Causgrove and Rebecca Nugent, Carnegie Mellon University (CMU)**

Causgrove is a corporate relations officer at CMU, where he advocates specifically on behalf of the departments in CMU’s Dietrich College of Humanities and Social Sciences. He and Nugent discussed the value of corporate relations officers, particularly for taking a holistic approach to supporting and sustaining academia–industry partnerships. Corporate relations officers highlight the diverse opportunities available to potential industry collaborators as well as the diverse students at CMU in the hopes that companies will choose to engage in long-term partnerships with any and all of CMU’s colleges.

Causgrove described seven channels through which industry can engage with CMU: student engagement, sponsored research, faculty engagement, professional education, licensing and technology transfer, start-ups, and co-location. Student-centric interactions are particularly popular with industry partners, and engagement is tailored to remain mutually beneficial for CMU and for the companies over time. He mentioned that more than 200 institutions are members of the [Network of Academic Corporate Relations Officers](#), which performs benchmarking, develops best practices for building relationships with industry, and offers resources for institutions that wish to establish corporate partnerships.

Nugent explained that CMU is formalizing an institution-wide Corporate Affiliated Projects (CAP) program. In the CAP program, local, national, and global industry partners work with faculty to scope real-world problems for collaborations with top-tier undergraduate-, masters-, and Ph.D.-level students

and advising faculty. In particular, Dietrich College hosts a Statistics and Data Science Corporate Capstone program, which is focused on experiential learning and tied to a semester-long elective course. This program arose in response to two trends: the recent job market strongly pulled students toward industry careers, and summer internship opportunities are too competitive and restrictive for students (particularly those with summer visa constraints). Meetings occur both in-person and virtually, the experience concludes with student presentations, and both students and faculty receive financial incentives to participate.

Nugent noted the value of collaborating across disciplines, with attention to aligning logistics, project goals, and educational project agreements. The Statistics and Data Science Corporate Capstone program is governed by CMU’s [Educational Project Agreement](#), which includes language to define the relationship, nondisclosure terms, policies for data sharing, and the project cost and scope. This agreement protects the IP of the students and faculty. To begin building a network with industry, she suggested that faculty engage with their institutions’ career centers to organize annual flagship events that draw potential partners to campus at low stakes.

### **Mehran Sahami, Stanford University**

Sahami explained that Stanford’s academia–industry research collaborations in computer science often focus on innovations in artificial intelligence (AI), data science, human–computer interaction, computer science theory, security, graphics, systems, and biocomputation. He provided an overview of data science and AI collaborations at Stanford including the [Stanford AI Laboratory](#), which is a research laboratory and university-wide affiliated program (e.g., statistics, bioengineering, medicine) focused on machine learning, vision, natural language processing, and genomics. Common features of effective academia–industry engagement include formal and informal interactions among the company, faculty, and students; continuous two-way communication; facilitated access to research; and recruitment.

Many of Stanford’s collaborations are housed in the [Computer Forum](#), which is the university’s industrial liaison program. The Computer Forum brings together industry (more than 100 affiliate companies who each pay an annual membership fee of \$21,000) and computer science and electrical engineering faculty and students for both research and recruiting

purposes. The Computer Forum also hosts conferences, workshops, and symposia and gives financial support to the computer science and electrical engineering departments. Once a faculty liaison is assigned to a member company, mutual talks and visits occur, potential research collaborations are identified, and the company decides if it would like to participate in a visiting scholar program to embed one of its researchers in a Stanford research laboratory. Stanford's [Recruiting Program](#), which is part of the Computer Forum, hosts information sessions, on-campus interviews, career fairs, career workshops, company tours, office hours, and networking events.

Sahami also described a Stanford course with corporate engagement. Companies present a high-level problem for which they need a solution, and participating students do a 2-quarter project to explore that area. The cost for each company to participate is \$75,000, and there are more companies that want to participate than there are student teams available each year.

#### **Michael Franklin, University of Chicago and formerly University of California, Berkeley**

Franklin highlighted the University of California, Berkeley's, success in creating multi-faculty projects that engage industry. For example, the [Berkeley Algorithms, Machines, and People Laboratory \(AMPLab\)](#), a big data research center, built the open source Berkeley Data Analytics Stack. AMPLab, a collaborative project, began in 2011 and concluded in 2016, resulting in 34 new faculty, several products, and 4 start-ups. A true public-private partnership, 50 percent of the funding for AMPLab came from NSF, the Defense Advanced Research Projects Agency, the Department of Energy, and the Department of Homeland Security; and 50 percent came from 40 industry partners. AMPLab nurtured its relationship with industry collaborators through twice-yearly retreats, during which faculty received feedback on project directions and students received feedback on research ideas. As part of its outreach and training initiatives, AMPLab also hosted [AMPCamp](#), a big data boot camp.

Franklin explained that building open source software is a valuable way for academia to collaborate with industry. However, a system cannot simply be built and passed on; a community has to be constructed and remain engaged (see [D. Patterson, "How to Build a Bad Research Center"](#)). For example, AMPLab students created a meet-up group for Apache Spark,

which now has more than 500,000 members. He believes that AMPLab's approach was successful because its commitment to producing open source software and publishing vigorously nearly eliminated IP issues and fostered benefits for both industry and academia. Industry secured early access to ideas and plans, recruiting opportunities, and membership in a neutral community. Students accessed early adopters (and sometimes data), advice and mentorship, internship and job opportunities, and practiced communicating their ideas. Faculty participated in a collaborative, flexible, diverse, and impactful platform; gained novel feedback; and received industry funding to augment federal grants.

The University of Chicago, however, is only newly involved in industry partnerships. Challenges to establishing these relationships include companies' limited perspectives about the value of academic research, companies' lawyers becoming involved too early in the process, and increased university competition for the attention of "enlightened" companies (e.g., Amazon, Google, Microsoft). Additionally, Franklin continued, administrators at some universities maintain outdated perspectives about IP and real-world engagement and fail to reward their faculty for industry collaborations. And some faculty underestimate the value of collaboration. To overcome these challenges, he suggested that institutions exploit local campus strengths and reach beyond a single department, as well as identify and exploit regional advantages where there is a concentration of universities, industrial strengths, and unique research assets (e.g., national laboratories). He wondered if NSF could play a role in convening academia-industry partnerships, as its Computer and Information Science and Engineering division has already facilitated successful programs with several industry partners.

## **PANEL DISCUSSION**

### **Establishing Partnerships**

Nugent suggested that academic institutions dedicate time to develop a framework and educate industry about the potential benefits of partnership. Causgrove added that Dietrich College has coordinated with the other six colleges at CMU to ensure that all industry partners receive the same educational agreement—an especially important feature for faculty and companies new to partnerships. Victoria Stodden, University of Illinois, Urbana-Champaign, observed that because academic research is distinct from industrial research (in terms of problems and incentives), it is crucial to understand how the two can reinforce

one another. She agreed that NSF could prompt such conversations and promote resource sharing. Franklin noted that although many complexities need to be addressed before partnerships can be established, he emphasized that a spectrum of research exists (as opposed to there being a distinction between academic and industrial research). Sahami added that academia–industry collaborations are responsible for much of the progress in deep learning; furthermore, more faculty could be inclined to leave academia for industry if siloes between academic and industrial research persist. Mark Tygert, Facebook Artificial Intelligence Research, suggested that participants read the work of Yann LeCun as evidence of productive exchanges between academia and industry.

### Increasing Incentives

Charles Isbell, Georgia Institute of Technology, wondered how to change the culture of academia so that faculty are rewarded for engaging in partnerships. Sahami suggested that junior faculty structure partnerships around potential publications but noted that they sometimes avoid industry collaboration for fear that their Ph.D. students will leave academia for industry jobs. Franklin commented that faculty have to broaden their perspectives of promotion and reward systems (and then educate administrators)—especially in the evolving areas of computer science and data science, in which many definitions of success exist. Nugent said that CMU faculty receive summer research funding as a reward for helping with partnerships.

### Tracking and Replicating Success

Nicholas Horton, Amherst College, asked how the panelists’ institutions have tracked their students’ progress and wondered if alumni serve as allies for these industry partnerships. Nugent said that CMU’s Corporate Capstone program is not mature enough yet to assess the feedback loop, but, anecdotally, students are talking about the program at career fairs and recent alumni are promoting the program to their supervisors. Causgrove added that a number of senior-level alumni relationships have also been leveraged. Sahami reiterated that the key to successful partnerships is maintaining relationships over time. Kathy McKeown, Columbia University, asked how to replicate these programs at scale, especially given the substantial amount of money companies contribute to participate. Franklin replied that although replicating AMPLab has proven more difficult than anticipated, he still believes that it is possible. He wondered if industry could peruse NSF’s pipeline of research proposals to prompt partnerships, and Nugent suggested that universities focus on engaging local companies.

## PANEL ON NATIONAL PERSPECTIVES ON ACADEMIA–INDUSTRY COORDINATION

### Ben Zorn, Microsoft and leader of the Computing Community Consortium (CCC) interim report on “Evolving Academia/Industry Relations in Computing Research”

Zorn described the mission of CCC (a standing committee of the Computing Research Association [CRA]) as to “catalyze the computing research community and enable the pursuit of innovative, high-impact research.” A 2017 CRA survey showed that computer science enrollment at the undergraduate level has more than quadrupled during the past 10 years, which makes it difficult for faculty to teach and maintain close relationships with students in large classes. He added that computing technology influences nearly all aspects of humans’ lives; thus, interesting research challenges and rich opportunities for collaboration between computer science and other disciplines (e.g., transportation, health sciences, and biology) exist.

The CCC Industry Working Group was established in 2018 to better understand academia–industry relations. Its [interim report](#) builds on the CCC’s 2015 report “[The Future of Computing Research: Industry–Academic Collaborations](#).” Anecdotal evidence in the interim report revealed a significant increase in faculty joint appointments in certain research areas, which could affect a university’s culture and mission negatively (e.g., impact on research agenda, conflicts of interest and IP issues, decreased faculty participation on committees for admission and hiring, and decreased mentoring and face time with students). Because some joint appointments could have an indefinite duration, academic institutions might have to develop novel arrangements to cover 50 percent of each participating faculty member’s time (or, in some cases, 80 percent), Zorn explained. He suggested the implementation of contracts as one way to ensure that students remain the priority of the faculty. Many positive outcomes of this type of engagement also exist. These experiences meet industry’s increased demand for talent in an era ripe with access to data and computing capabilities. Faculty and graduate students have the opportunity to participate in ambitious and impactful research and to access increased resources and salary.

CCC’s goal is to preserve the positive aspects of these academia–industry partnerships while understanding and mitigating risks. CCC is interested in [feedback](#) on this interim report; it hopes to expand data gathering, understand best practices of current faculty–stu-

dent arrangements, and document novel company approaches to deepening academic engagement.

### **Chaitan Baru, University California, San Diego (UCSD)**

Baru observed that computer science and data science are optimal areas for collaboration with industry. During the past few years, NSF has facilitated a number of such interactions—for example, [NSF BIGDATA](#), [NSF/Intel Partnership on Foundational Microarchitecture Research](#), [NSF Campus Cyberinfrastructure](#), [NSF Program on Fairness in AI](#). He asserted that hands-on experience is essential for future data scientists and cited programs at North Carolina State University, Harvey Mudd College, and UCSD as exemplars. He suggested that all data science curricula and faculty competencies should align with this vision of “clinical practice” to remain competitive.

Baru described the value of completing an “industry postdoc”—an immersive, practical experience with government agencies, nonprofits, or large or small companies (from Internet giants to start-ups). This experience could occur immediately after the Ph.D. is completed (in order to become better qualified for data science faculty jobs) or after the receipt of a job offer. A variety of modalities exists to fund such an experience (e.g., two-way between the agency and industry or 3-way among the agency, industry, and university), and it should be governed by a mentorship plan that includes standards for compliance.

An example of an implementation vehicle for academia–industry collaboration is [NSF’s Grant Opportunities for Academic Liaison with Industry \(GOALI\)](#). There are currently 300 GOALI awards, only 2 percent of which are in computer science. In the future, Baru hopes that an NSF GOALI program will be created with net new funds and with programs for industry postdocs and industry sabbaticals. Baru concluded by noting that many opportunities exist for academia to collaborate with industry on technological innovation if the right engagement mechanisms are identified.

### **Rachel Levy, Mathematical Association of America (MAA)**

Levy described the mission of MAA as “to advance the understanding of mathematics and its impact on the world.” MAA provides guidelines for departmental reviews and experiential learning-based instruction, and it strives for mathematics to cross disciplines

so that all people view themselves as mathematics “doers.”

Levy shared examples of three MAA programs that relate to data science: (1) [StatPREP](#), which provides resources, workshops, and webinars for faculty on how to bring the modern tools and methods of data science to elementary statistics courses; (2) [PICMath](#), which prepares mathematical sciences students for industry careers through a semester-long course on industry research problems as well as provides training, resources, and support for faculty teaching that course; and (3) [Big Math Network](#), which helps mathematics faculty, via the [Big Jobs Guide](#), advise students interested in industry careers.

While tracking students who earned degrees in mathematics to better understand their job placement, MAA found that their job titles are rarely “mathematician.” Levy observed that mathematicians do not always have a presence in academia–industry partnerships, despite their high level of interest. She suggested that industry could help MAA understand how to create meaningful experiences—building more partnerships, staying connected with mathematics graduates who accept jobs in industry, and creating challenges and competitions with broad participation that integrate data science—that would build competencies for future hires.

## **PANEL DISCUSSION**

### **Balancing Faculty Responsibilities with Industry Experiences**

McKeown reiterated the benefits of faculty joint appointments: faculty have the opportunity to work with interesting industry data and problems, to understand what students will experience when they enter industry, and to establish relationships that could lead to funding opportunities. Nugent noted that faculty who remain on campus and train the Ph.D. students whose advisors are unavailable need to be supported. Industry could sponsor faculty lines at universities to help alleviate this burden, she continued. Mark Green, University of California, Los Angeles, observed that no standards exist to protect students who have invested in the expertise of advisors who become unavailable, and it is unclear what body would have the credibility to suggest them. He appreciated the value of joint appointments but wondered if there is a better process.

Emily Fox, University of Washington, said that her institution recently conducted a survey of Ph.D. students’ perspectives on advising relationships: some

students found it beneficial to have their advisors on leave and working in industry (e.g., increased access to resources), while more students thought that the advisors' decreased availability had a negative impact on the cohesiveness of their Ph.D. cohorts. Fox noted that while the long-term benefit to Ph.D. students is immeasurable, faculty joint appointments take a substantial toll on Ph.D. students working on their dissertations. Green suggested that, to begin to address some of the concerns about faculty joint appointments and academia–industry partnerships, the mathematics community could compile a list of interesting problems that came from industry and led to important research. He also emphasized the need to use data to understand the capacity of the economy to absorb students being trained in Ph.D. programs. Levy noted that this conversation should be expanded to include VITAL faculty (an acronym for visitors, instructors, teaching assistants, adjuncts, and lecturers) as well as industry partnerships with faculty and students at 2-year colleges.

### Expanding Opportunities for Students

Alfred Hero III, University of Michigan, mentioned that the thriving economy in southeast Michigan has enabled the [Michigan Institute for Data Science](#) to be successful in securing industry partnerships. However, because the competition is intense and industry partners often require exclusive nondisclosure agreements, universities run the risk of being limited to partnering with only one company. He suggested that universities engage more with national laboratories, which provide experiential learning on interesting problems without the competition. Deb Agarwal, Lawrence Berkeley National Laboratory, said that the data science community is doing a disservice to students if it continues to focus only on partnerships with companies instead of including government agencies, nongovernmental organizations, and national laboratories. She emphasized that national laboratories have an abundance of opportunities for students to work for the common good on unclassified research related to problems of national interest, generally without IP issues.

Catherine Brooks, University of Arizona, mentioned that her institution has developed a taskforce to identify synergies across the university to better present itself as a unified whole to industry partners. She explained that universities need to be more nimble and less siloed. Kolaczyk added that it is important to propagate lessons from experiential learning at the Ph.D. level across degree levels and across industries.

## A NEW MODEL FOR ACADEMIA–INDUSTRY PARTNERSHIPS

Gary King, Harvard University (via webcast)

King described a political science innovation that addresses the problem of data access for university researchers, motivated by the mission of social science to understand and solve problems that affect human society. King observed that the social sciences have more data than ever before but a smaller fraction of the data that exists in the world than ever before—most of it being tied up in private companies. One goal of King's research is to understand how to incentivize private companies to release data for research that creates public good, without harming themselves.

King is working with Facebook to facilitate studies of the effect of social media on elections and democracy. This data-intensive research is funded by eight ideologically diverse charitable foundations that agreed to pool their funds and let one group of academics decide how to allocate grants. He asked Facebook for full access to its people, products, data, and platforms as well as freedom to publish without prepublication approval. Because Facebook would not agree to both of these terms for any one researcher, King created two groups of researchers: (1) a commission of distinguished academics at [Social Science One](#), an organization he created with Nate Persily at Stanford, who have signed nondisclosure agreements, have complete access to Facebook data, and have agreed not to publish; and (2) a group of outside academics who apply for limited data access and have complete academic freedom (i.e., no prepublication approval) to publish. Facebook, the foundations, and Social Science One agreed on the scope of the project, the commission identified relevant Facebook data sets and issued a request for proposals, and the outside academics applied for access to those data. There are three data sets to which access is now being provided: CrowdTangle, a collection of Facebook's political advertisements, and all of the public URLs shared on Facebook. The project will create its own surveys and will make arrangements with the American National Election Survey, the British National Election Survey, and other large academic surveys to include a question that asks respondents to share their Facebook data with the researchers. The outside academics will follow institutional review board processes and engage in a merit peer review and an ethical peer review, and the final decisions will be made by the commission. Facebook is building a privacy-preserving computer infrastructure.

The timeline for this innovative project has been extended and has included challenges such as dozens of legal agreements. When researchers receive data access (as opposed to data), the academic research model changes from one of individual responsibility to one of collective responsibility. King's goal is to convey to companies and the public that data are an asset to create social good and solve the world's problems, while preserving privacy.

## **PANEL ON INDUSTRY ACTIVITIES AND EXPERIENCES FROM ACADEMIC PARTNERSHIPS**

**Mike Willardson, Facebook**

Willardson described Facebook's mission as to "give people the power to build community and bring the world closer together. People use Facebook to stay connected with friends and family, to discover what is going on in the world, and to share and express what matters to them." As of 2018, Facebook had 35,000 employees and 2.32 billion monthly active users. Willardson provided an overview of the research activities within Facebook's Research Operations and Academic Relations division. The [research activities](#) vary by subject matter; for example, IP is an important concern for research in augmented reality/virtual reality because it is used in commercial products. Facebook believes strongly in building community through open source technology, and investing in open source increases employee retention and recruitment.

Willardson described several innovative Facebook partnerships. The [Open Compute Project](#) democratizes hardware by bringing industry and universities together to build products. This mechanism works well for multiple industry partners because there are no exclusive rights, and everyone benefits equally. The [Telecom Infra Project](#) is a collaborative effort to build and deploy telecommunications network infrastructure. Facebook also engages with faculty and students through fellowship programs, emerging scholar awards, research awards, research collaborations, and visiting researcher and postdoctoral positions. He added that Facebook establishes broad master agreements with universities to cultivate long-term relationships.

**Mary Ellen Sullivan, MassMutual**

Sullivan explained that MassMutual operates for the benefit of its members and participating policy holders by helping people secure their futures and protect their loved ones. MassMutual has 7,500 employees

and 9,000 nationwide advisors. MassMutual employs 100 data scientists in 4 data science domains—risk and product, operations, finance investments, and marketing and sales—to enable data-driven decision making throughout the enterprise.

Sullivan explained that academia–industry partnership is essential at MassMutual. The company supports science, technology, engineering, and mathematics curricula and programs; engages with local faculty; co-sponsors community education and events; engages with student groups; invests in training and development programs; and collaborates on research initiatives with university partners. Smith College, Mount Holyoke, and the University of Massachusetts Amherst each have partnerships with MassMutual, and the University of Vermont will be the company's next collaborator. MassMutual works with university administration, faculty, and student groups to ensure that programs are working effectively and offering mutual benefits. In 2014, MassMutual launched the [Data Science Development program](#) and will launch a [Data Engineering Development program](#) in summer 2019. Each cohort of the Data Science Development Program has 4-8 participants, 80 percent of whom are women. Both programs offer hands-on training and mentorship, full-time employment on an innovative and fast-paced team, and tuition sponsorship for either a master's degree or a certificate from a local university. In January 2018, MassMutual hosted a [Women in Data Science Conference](#), and it hosts monthly data science meetings in Boston, [Data Days for Good](#), and hackathons.

**Peter Norvig, Google**

Norvig said that one of Google's most significant responsibilities is to help grow the field of data science, starting at the K-12 level by developing curriculum and educating teachers. Google supports [Girls Who Code](#), as well as groups within historically black colleges and universities to develop and co-teach classes. Google's own educational materials (some are co-developed with Coursera or Kaggle) are available through massive open online courses. Google is reviewing its guidelines for data sharing and is promoting academia–industry collaborations that will develop responsible and productive researchers by hiring interns, welcoming visiting faculty, and offering faculty joint appointments. Norvig emphasized that when a faculty member decides to leave academia for a career in industry, that move should be viewed as a new opportunity (not a failure). Likewise,

Google staff are encouraged and supported to co-advise students and to teach in the classroom or online.

### **Daniel Marcu, Amazon**

Marcu noted that members of industry and academia alike should be making efforts to enhance their communication and collaboration. Amazon has a variety of collaborative engagement models and a significant research breadth (e.g., hardware, economics, sustainability, logistics, avionics, robotics). Students can participate in 3-month internships or full-time post-doctoral opportunities as well as apply for research grants and Amazon Web Services credits. Faculty can apply for academic grants, secure Amazon Web Services resources and data, and attend Tech Talk Series and academic conferences. The [Amazon Scholars](#) program offers deeper levels of engagement by enabling professors to work on Amazon's large-scale, high-impact technical challenges without leaving their academic institutions. Amazon Community Programs include a graduate research symposium (which pairs student researchers with Amazon's scientists to exchange new innovations and research concepts), scientific meeting sponsorships, and an internal academic advisory council.

When developing partnerships with industry, Marcu suggested that faculty need to understand the potential partner, consider the best-suited model of engagement, and formulate interesting proposals. Administrators could aid in the process by simplifying engagement models. Inhibitors to success include faculty members who dictate terms to the partners and write ineffective proposals as well as administrators who treat industry engagements as one-off activities. Marcu believes standardized agreements could accelerate collaboration.

## **PANEL DISCUSSION**

### **Engaging Students**

Chris Mentzel, Gordon and Betty Moore Foundation, asked the panelists how often their companies engage with disciplines that intersect with data science. Norvig acknowledged that the majority of Google's interactions are with computer scientists and said that it can be difficult to advertise for and evaluate proposals from other fields without appropriate expertise on staff. Marcu said that Amazon engages frequently with economists, computer scientists, data scientists, and machine learning experts. Willardson noted that Facebook engages often with data scientists who have expertise in artificial intelligence, machine learning, connectivity research, and natural language processing, and Sullivan commented that

MassMutual's engagement extends beyond the discipline of data science.

Duncan Temple Lang, University of California, Davis, asked the panelists what skills students need to be prepared for industry careers. Levy suggested [Kaggle](#) as a useful tool for mathematics Ph.D.s who want to move to industry. Sullivan said that MassMutual emphasizes skills that are essential for business but rarely developed at the undergraduate level, such as leading, giving and getting feedback, and tailoring presentations to different audiences. MassMutual began a partnership with EdX and is establishing requirements around a series of self-paced online courses to help reinforce these skills. Norvig agreed that these skills are crucial, especially the ability to work effectively in teams and to give meticulous attention to detail. Marcu noted that it would be beneficial for students to understand that academic research is not inherently superior to industry research. Nugent and Kolaczyk suggested that members of academia and industry avoid referring to these skills as "soft skills." Not only is it offensive to the fields that teach these skills, but such language also causes students to drastically underestimate how important those skills are and how difficult they are to learn.

### **Navigating Two Cultures**

In response to a question from Causgrove about formalizing academia–industry partnerships, Marcu said that although many conversations are happening at different levels across academia and industry, it can be difficult to bridge the communication gap and begin to move forward with effective partnerships. Baru noted that it is easier to partner with companies that understand the culture of academia, and he suggested that those companies help others in industry to better understand the research ethos. Willardson agreed that sharing best practices throughout industry would improve consistency. Setting the context and determining the value proposition before entering into partnership is also effective, he continued. Norvig mentioned that there are different measures of successful partnerships—professors need to publish, while industry teams are recognized for research even if it leads to failure. Tygert mentioned an agreement between Facebook and the University of California, Berkeley, to share students. Instead of negotiating separate agreements, Google, Amazon, and others signed on to this agreement. Hero said that the flow of students and faculty has moved away from academia and toward industry during the past 5 years; he wondered how to reinforce positive relationships between industry and academia and reverse this imbalance. Levy wondered what mechanisms would motivate industry employees to embrace teaching

or training opportunities. Antonio Ortega, University of Southern California, asked about strategies to attract junior faculty to partnerships. Sullivan said that MassMutual's Data Engineering Development program has an academic advisory board with a more junior-level faculty member, and Marcu noted that the number of opportunities in general for junior faculty has increased.

Sahami asked if companies have policies for the length of visiting faculty terms. Sullivan replied that the faculty coming to MassMutual are only joining an academic advisory board or teaching one-off in-house workshops and that academic institutions welcome that level of cross-pollination. Willardson said that Facebook defines a limited term for visiting faculty, and Norvig said that although Google supports freedom of choice, it recognizes that there can be negative repercussions from extended faculty appointments and tries to maintain good relationships with partnering departments.

James Frew, University of California, Santa Barbara, asked about impediments (beyond IP issues) to these partnerships. Willardson said that both partners must be willing to accept some level of calculated risk in order for the partnership to be successful. At MassMutual, the issue is less about risk and more about workforce: Because MassMutual is building pipelines for people to enter its organization, it can be challenging to keep pace with changing skills and relevant curricula. Marcu said that the biggest hindrance is the lack of well-established models of collaboration.

### Sharing Data in Partnership

Noting the growing trend to provide artifacts alongside publications (e.g., the data and code that support a paper's claims), Stodden inquired about policies for sharing artifacts that emerge from collaborative work. Marcu said that this trend presents an opportunity for industry, not a barrier to participation in partnerships. Willardson explained that the subject matter will determine whether Facebook pursues sponsored research agreements (e.g., user data can only be shared in a controlled environment and with prepublication review, so such research is unlikely to be part of these agreements). Facebook is not trying to control outcomes in this case; rather, it is trying to prevent the inadvertent dissemination of confidential information. Norvig suggested that industry provide funding for open source journals and noted that increased partnership among nonprofits, academia, and industry is needed to address issues of data ownership and proprietary publishers. The University of California, for example, recently stopped paying for use of Elsevier. Mark Krzysko, U.S.

Department of Defense, emphasized that sharing and consuming data is complex in part because of challenges with access and dissemination and a lack of clear policies. Norvig said that Google employees have access to internal data, while grant recipients do not. To establish mutually beneficial partnerships, he said that industry needs to make more relevant non-proprietary data sets available and help pose more germane problems. Sullivan said that MassMutual will never share clients' confidential data. Other publicly available data, however, are used for research (e.g., for health and longevity studies, which can be used to provide information to customers).

Baru noted the success of NSF's [Computer Science for All](#) initiative; however, part of the curriculum has languished because teachers did not have access to data sets. It would be helpful if industry partners would contribute data (real or synthetic) for teachers to use. Zorn said that it is important to find the right technology that will empower companies to share data by preventing unauthorized access and highlighting mutually beneficial opportunities of data sharing. Sahami suggested a new model for data sharing in which third-party public institutions are leveraged to socialize the associated risk instead of having either the company or the researcher assume the risk.

## BREAKOUT GROUP DISCUSSIONS

Following the presentations and open discussions, Roundtable participants divided into three groups to create sketches of [Ten Simple Rules](#)<sup>1</sup> for Creating a Successful Academia–Industry Collaboration at the levels of undergraduate, masters, and Ph.D. education. These sketches represent collections of diverse ideas and are not meant to be read as consensus viewpoints. A representative from each group summarized the discussions among the breakout group members as follows:

Hunter Glanz, California Polytechnic State University, presented the following suggestions for effective collaborations between industry and undergraduate students: (1) Keep curriculum current and exploit curricular flexibility; (2) offer experiential learning opportunities early and often; (3) ensure that both parties continually benefit from the interactions; (4) offer capstone experiences; (5) promote early comprehensive experiences (starting with an open-ended problem and working through to the communication of findings) in which students have to make choices; (6) provide multiple points of inclusive entry for

---

<sup>1</sup> Inspired by PLOS's Ten Simple Rules series.

data science learners; (7) educate both parties on data ethics; (8) develop a mutual understanding of unique cultures and environments; (9) provide genuine and varied data sources in a consistent manner; (10) create a reproducible, transferrable data science best practices kit; and (11) promote classroom and company visits.

Kolaczyk highlighted the following suggestions for successful partnerships between industry and masters-level students: (1) Take a holistic approach to training, rather than teaching topics in separate silos; (2) build skills in communication and team interaction; (3) create opportunities for repeated practice; (4) expose students to industry in multiple ways and at many levels; (5) encourage humility and reduce anxiety among faculty and students; (6) become an active listener and learn to use vocabulary that is conducive to collaboration; (7) nurture academia–industry relationships; (8) define collaborative proj-

ects through an iterative process, with both parties vested; (9) own the collaboration on both sides; and (10) lay the intellectual groundwork before involving lawyers.

Nina Mishra, Amazon, shared her breakout group’s discussion of considerations for fruitful collaborations between industry and Ph.D. students: (1) Consider creating a Ph.D. in data science; (2) encourage students to do multiple data science internships; (3) create a consortium of industry collaborators who contribute data and problems; (4) ensure that all parties agree on a project and its duration before it begins; (5) create prolonged internship opportunities; (6) encourage open source and open science; (7) identify potential conflicts of interest ahead of time; (8) formally include internship work in the thesis; (9) avoid letting industry drive what happens to students; and (10) maintain a high bar for dissertation work and graduation.

---

**ABOUT THE ROUNDTABLE:** The Roundtable on Data Science Postsecondary Education is supported by the Gordon and Betty Moore Foundation, the National Institutes of Health, the National Academy of Sciences W.K. Kellogg Foundation Fund, the Association for Computing Machinery, the American Statistical Association, and the Mathematical Association of America. Within the National Academies, this roundtable is organized by the Committee on Applied and Theoretical Statistics in conjunction with the Board on Mathematical Sciences and Analytics, the Computer Science and Telecommunications Board, and the Board on Science Education. Roundtable meetings take place approximately four times per year. Please address any questions or comments to Ben Wender at [bwender@nas.edu](mailto:bwender@nas.edu).

**DISCLAIMER:** This meeting recap was prepared by the National Academies of Sciences, Engineering, and Medicine as an informal record of issues that were discussed during the Roundtable on Data Science Postsecondary Education at its 10th meeting on March 29, 2019. Any views expressed in this publication are those of the participants and do not necessarily reflect the views of the sponsors or the National Academies.

**ROUNDTABLE MEMBERS PRESENT:** Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; Deb Agarwal, Lawrence Berkeley National Laboratory; Jeffery Brock, Yale University; Emily Fox, University of Washington; James Frew, University of California, Santa Barbara; Lise Getoor, University of California, Santa Cruz; Mark Green, University of California, Los Angeles; Alfred Hero III, University of Michigan; Nicholas Horton, Amherst College; Charles Isbell, Georgia Institute of Technology; Mark Krzysko, U.S. Department of Defense; Rachel Levy, Mathematical Association of America; Chris Mentzel, Gordon and Betty Moore Foundation; Nina Mishra, Amazon; Deborah Nolan, University of California, Berkeley; Peter Norvig, Google; Antonio Ortega, University of Southern California; Mehran Sahami, Stanford University; Victoria Stodden, University of Illinois, Urbana-Champaign; Duncan Temple Lang, University of California, Davis; and Mark Tygert, Facebook Artificial Intelligence Research. **GUESTS PRESENT:** Chaitan Baru, University of California, San Diego; Catherine Brooks, University of Arizona; Adam Causgrove, Carnegie Mellon University; Michael Franklin, University of Chicago; Hunter Glanz, California Polytechnic State University; Gary King, Harvard University; Daniel Marcu, Amazon; Rebecca Nugent, Carnegie Mellon University; Mary Ellen Sullivan, MassMutual; Mike Willardson, Facebook; and Ben Zorn, Microsoft. **NATIONAL ACADEMIES’ STAFF PRESENT:** Selam Araia, Linda Casola (via webcast), Chris Fu (via webcast), Tyler Kloefkorn (via webcast), and Ben Wender.

---

## Division on Engineering and Physical Sciences

*The National Academies of*  
SCIENCES • ENGINEERING • MEDICINE

The nation turns to the National Academies of Sciences, Engineering, and Medicine for independent, objective advice on issues that affect people’s lives worldwide.

[www.national-academies.org](http://www.national-academies.org)