Research Statement, Danielle Li

The success of modern firms depends on their ability to embody new ideas in commercial products, and to acquire the resources necessary to support their innovative capacity. To excel at either, managers must effectively evaluate projects and people.

This challenge is often cast as a problem of missing information: entrepreneurs must decide which product features to develop without knowing how well success among early users will translate into the larger market; human resource managers must decide which candidates to hire without observing their aptitude for the role. Yet, in practice, managers are increasingly inundated with information: granular data on customer behavior; advice from investors, employees, and consultants; resumes, test scores, and social media accounts; their own intuition or experience. These sources of information can be conflicting, and they may also be irrelevant. The difficulty of decision-making in modern organizations lies not in overcoming a lack of information, but in curating a myriad of information: what to trust, what to prioritize, what to ignore.

My research focuses on "decision-making technologies," the institutional and technical arrangements by which organizations manage and consume information. At the US National Institutes of Health (NIH), for instance, program managers allocate billions of dollars in funding through decentralized peer review committees, a technology that leverages the expertise and idiosyncrasies of tens of thousands of external advisers to shape a national research agenda. Meanwhile, across a range of industries, firms increasingly use machine learning algorithms to screen job candidates, a shift that has the potential to reshape hiring, the primary process by which firms access human capital—and applicants access economic opportunity. While every organization employs some kind of decision-making technology, relatively little is known about their impact on productivity and the allocation of resources and opportunity.

Answering these questions is difficult. To say that a firm could improve its performance by doing something else—investing in more exploratory projects, relying more on algorithmic recommendations, for instance—requires making inferences about unobserved states of the world. Yet because many firm-level choices are both endogenous and highly consequential, researchers cannot make these inferences by comparing firms, and firms are often unwilling to randomize. In such cases, where observation is unreliable and experimentation is infeasible, my work generates convincing evidence by combining deep institutional knowledge with modern econometrics, novel measurement, and a connection to theory.

I begin by engaging with context. Through collaborations with practitioners, I unearth sources of variation that lead to credible identification strategies, and obtain access to administrative data that reveal an organization's options in addition to its choices. In many cases, I import modern tools from computer science and biomedicine to develop new measures—now adopted by other researchers—that more richly characterize innovative choices. When direct observation is impossible, I apply microeconomic theory to generate empirically testable predictions that distinguish between multiple models of firm behavior. By combining these elements, my research stands out is its ability to convincingly render and interpret organization-level counterfactuals.

My work spans two domains: investments in research and investments in people. With this focus, I take a unifying view of key topics in innovation and personnel economics, namely that the ability to develop ideas and talent form the twin foundations for how organizations create and deliver value.

(1) *Investing in ideas*: Because the value of ideas is so skewed, the success of R&D intensive organizations depends crucially on how well they evaluate research opportunities, not simply on how well they execute a given path. I explore the factors that shape organizations' research priorities, and document their impact on the innovations we get, and the ones we do not.
(2) *Investing in people*: Managing human capital requires firms to make inferences about difficult counterfactuals: how well would someone do if given a new opportunity? My work studies how firms can more effectively learn about workers' potential, and considers the impact of different HR policies on productivity and access to opportunity.

In both streams, my work ultimately raises questions about how the "evidence" in "evidence-based decision-making" is produced. In my ongoing research, I focus on how emerging factors such as algorithm design and the growing adoption of monitoring technologies shape the quality of data-driven decision-making.

1. Evaluating and investing in ideas

Assessing the value of ideas is difficult: ideas may take decades to find their use, and this use may be far from the original intention or may be developed by a different party than the originator. Pfizer and Moderna's m-RNA based vaccines for Covid-19, for instance, both built on decades of failed attempts to develop a vaccine for HIV/AIDS, led by other firms. This difficulty of tracing and appropriating the value of R&D investments creates unique challenges for organizations focused on innovation.

My work in this area follows the value chain of biomedical innovation from public investments in foundational science to private investments in commercial development. I begin by considering how grant funders can more effectively use information to identify promising scientific ideas. Next, I trace public investments in science into the private-sector, documenting their impact on the development of drugs and other medical products. Finally, I consider the research priorities of private firms, focusing on their willingness to explore new research paths.

1.1. **Assessing project quality**

In **Expertise vs. Bias in Evaluation: Evidence from the NIH** (*American Economic Journal: Applied Economics*, 2017; recipient of AEJ Best Paper Award), I examine the extent to which decision-markers should trust advice from potentially biased experts. As the world's largest funder of biomedical research, the NIH's research priorities shape global medical innovation, and the majority of NIH's funding is allocated through peer review. NIH funders face a tension familiar to decision-makers in many settings: they must balance the desire to learn about a project's potential from expert advisers against the concern that those advisers may be biased. Despite the ubiquity of this challenge, my paper was the first to separately identify the role that bias and expertise play in shaping an important investment decision.

Intuitively, if grant reviewers draw on their expertise in scoring applications, then strong applicants should benefit from being evaluated by closely related reviewers and weak applicants should be hurt. If reviewers are simply biased, then related applicants should be more (or

less) likely to be funded regardless of quality. Expertise increases the slope of the relationship between scores and quality, whereas bias changes its level.

Implementing this test requires a measure of quality for all applications, even those that are rejected. I provide this by observing that scientists often publish the research underlying their grant applications even when the application itself is rejected. I use text-analysis to find subsequent publications by the same scientist on the same topic as her initial application. Finally, I exploit the organizational structure of NIH review committees to compare applicants who are quasi-experimentally linked to more or less influential members of the committee.

Using data on 100,000 grant applications, I show that reviewers are biased in favor of applicants whose work relates to their own, but also better informed about their quality. On net, I find that the benefits of expertise dominate the costs of bias. This result therefore cautions organizations that well-meaning attempts to limit conflicts of interest can reduce the quality of investments by reducing the quality of information available to decision-makers. Rather than seeking advice from less knowledgeable sources, organizations may benefit more from investing in efforts to de-bias their evaluators or to better align their incentives.

The results described above raise the question of whether organizations can do anything except simply accept the biases of the experts they consult. Are there other sources of information that decision-makers can turn to? In work with Leila Agha (Dartmouth), **Big Names or Big Ideas: Evidence from the NIH** (*Science*, 2015), I consider the value of experts in a world that is also increasingly data rich. Instead of seeking advice from potentially biased reviewers, the NIH could instead rely on harder metrics such as publication or citation records. Using data on 140,000 NIH grants funded between 1980 and 2008, we provide the first large-scale evaluation of the relationship between peer review scores and subsequent grant performance. We find that reviewers have positive "value-added": their scores are positively correlated with grant outcomes (as measured by associated publications, citations, and patents), even after controlling for an applicant's publication, work, and grant history. A review process that relies only on quantitative metrics would therefore have less predictive power than one that is able to effectively combine objective and subjective forms of evaluation.

Evaluating ideas is demanding process, one that requires reviewers to synthesize a great deal of technical knowledge across many different projects. It is hard work that is done by busy people. In "**Internal Deadlines, Drug Approvals, and Safety Problems**," (*American Economic Review: Insights*, 2020), Lauren Cohen (HBS), Umit Gurun (University of Texas) and I examine the demands of regulatory assessment. Using an international dataset, we show that regulators approve a disproportionate share of drugs at the end of the year, end of the month, and before country-specific holidays (e.g. before Thanksgiving in the US and before Lunar New Year in China, but not vice versa). We introduce a model in which regulators rush to complete their tasks before salient calendar benchmarks, leading to lower quality review. Consistent with this model, we show that drugs approved during approval surges are associated with significantly more adverse effects, including hospitalization and death. These results, which persist across countries even as their official regulatory policies differ, emphasize the importance of developing

organizational policies that focus not only on formal rules, but also address the influence that behavioral patterns can have on decision-making.

1.2. **Valuing R&D investments**

Organizations that invest in scientific ideas do so in the hopes that these ideas will have real-world impacts. The next part of my research develops new ways to measure the value of ideas and applies them toward understanding the returns to R&D investment.

The value of scientific ideas is both uncontested and, in practice, difficult to quantify. This is because foundational knowledge, by definition, can inform discovery in diffuse ways, making it difficult to link ideas to the innovations they enable. In "**The Applied Value of Basic Science**," (*Science*, 2017), Pierre Azoulay, Bhaven Sampat and I develop a new way of assessing the commercial value of public research, using grant acknowledgements and citations between patents and academic publications. This approach relies on data to reveal relationships between ideas and products over time and across disease areas, allowing us to capture the unanticipated diffusion of scientific knowledge through "open science" information channels.

Our results highlight the importance of this methodological innovation. Over 30% of NIH grants produce publications that are eventually cited by private-sector patents, but less than half of these links occur among patents with disease applications in the same area as the original grant. Traditional approaches that focus on same area outcomes and direct patent production would therefore vastly understate the commercial relevance of basic science.

In "**Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules**" (*Review of Economic Studies*, 2018) Pierre Azoulay (MIT), Josh Graff-Zivin (UCSD) and Bhaven Sampat (Columbia) and I build on the measures developed in our *Science* paper to provide the first quasi-experimental evidence on the causal returns to public research investments in biomedical science.

Despite its importance, credible estimates of the return to public research investments have eluded researchers for decades. My work addresses three key challenges that had previously hindered research in this area. First, a credible estimate of R&D returns must address the possibility that funding for an area may be related to its scientific potential or its medical demand. Next, while private R&D is targeted toward specific applications, public R&D investments are made with the opposite goal: to maximize potential spillovers. As documented in my previous work, the prevalence of spillovers makes it difficult to trace the influence of public investments. Finally, when tracing the impact of private funding, the value of expanding public funding depends on the extent to which increases in public spending reduce (crowd out) or spur (crowd in) private investments.

We address concerns about endogeneity using a regression discontinuity design. At the NIH, grant applications are funded in order of their score until the budget is exhausted. We define "windfall" funding as the amount a research area receives from applications near this cutoff, in excess of what would have been expected due to chance, and examine the relation between windfall public funding for a research area, and subsequent private-sector patenting in that area.

To accurately measure patenting outcomes, we develop a method that uses patent similarity to identify all private-sector patents in areas potentially influenced by NIH funding, regardless of whether a patent itself cites NIH-funded work. This approach allows me to examine whether increases in NIH funding expand the total amount of private-sector patenting in a research area, accounting for the possibility that public funds may crowd out (or in) private investments.

My results show that NIH funding increases net private-sector patenting. For patents related to FDA-approved drugs, a dollar of NIH funding leads to $1.4 to $2.7 in drug sales, a return that suggests the private and social returns to public investments in science are very high.

1.3. **Identifying barriers to innovation**

The value of public investments in science depend in large part on the development decisions of private-sector firms: do they make the most of this foundational knowledge? I next examine the factors that shape firms' R&D priorities and their implications for the types of medical innovation that are brought to market.

In "**Missing Novelty in Drug Development**," (*Review of Financial Studies*, 2021), Joshua Krieger (HBS), Dimitris Papanikolaou (Northwestern), and I provide evidence that risk aversion leads pharmaceutical firms to underinvest in novel drug candidates. We introduce a new measure of novelty based on a drug candidate's chemical structure. Using this measure, we show that novel drug candidates—ones that are molecularly distinct from prior candidates—have higher expected revenue and higher stock-market valuations for their underlying patents, but are also less likely to reach FDA approval.

To understand how firms prioritize risk and reward, we examine how they spend a marginal dollar. We obtain variation in cashflows via the introduction of prescription drug coverage for elderly Americans. This policy disproportionately expanded revenues for firms with more drugs targeting the elderly, and those with longer remaining market exclusivity on their drugs. To identify the effect of cash net of differences in demand or product lifecycle, we compare firms with the same overall elderly focus and patent age, but that differ in how their remaining patent life is distributed across drugs different shares of elderly patients.

We find that firms respond to increased cashflows by developing more drugs and—importantly—more novel drugs across all research areas, including areas that did not experience a demand shock from the policy. These results, which hold even for large public firms, stand in contrast to classic models in which firms invest in projects on the basis of their expected returns alone. Rather, the desire to avoid cash shortfalls tomorrow appears to lead firms to invest more conservatively today—a result that is consistent with our descriptive finding that over 15% of new drug candidates are nearly identical to previously developed drugs.

My work on drug novelty highlights the importance of creating incentives for exploration. In "**Insurance Design and Pharmaceutical Innovation**," (*American Economic Review: Insights*, forthcoming), Leila Agha (Dartmouth), Soomi Kim (MIT) and I examine how insurance design can be used as a policy tool to encourage medical innovation.

In 2012, insurance companies began moving from a system that provided coverage for all FDA-approved drugs to one that covers only a preferred set. This practice, known as maintaining a "closed formulary," has since become standard, with almost 900 drugs excluded as of 2020.

We study the impact of closed formulary policies on upstream R&D decisions. We show that exclusions substantially reduce the profitability of targeted drugs, especially in areas with more existing drugs and higher prescription volume, such as diabetes and cardiovascular health. Because these areas had traditionally been highly profitable, exclusion policies generated a dramatic shift in R&D incentives for pharmaceutical firms.

Correspondingly, pharmaceutical investments fell markedly in drug classes at high predicted risk of facing exclusions following the introduction of closed formularies. R&D declined the most in research areas that appear the most "incremental": late entrants to crowded disease classes building on older science. Taken together, these results show that *private* sector policies can play an important role in raising the relative returns to investing in exploratory drug candidates.

In work with Jiro Kondo (McGill) and Dimitris Papanikolaou (Northwestern), "**Trust, Collaboration, and Economic Growth**," (*Management Science*, 2020), I continue investigating the impact that organizational frictions can have on firms' ability to innovate. In this paper, we develop a tractable model that links organizational trust at the micro level to innovation, investment, and productivity growth at the macro level. Innovators generate ideas but are inefficient at implementing them on their own, for example, because they lack access to capital or business expertise. Firms can therefore help innovators develop their ideas more efficiently, but cannot commit to compensating them appropriately. When trust is high, firms anticipate fruitful future collaborations, and can thus credibly commit to not expropriating inventors, leading to more efficient production. Our model qualitatively replicates the empirical relation between measures of trust and various economic outcomes. At the firm level, we document strong correlations between measures of trust gathered from employee reviews, and the rate of high impact patenting in those same firms (controlling for R&D expenditures, assets, and profitability). We find similar relationships at the country level.

1.4. **Future work: Learning from experimentation**

My ongoing work explores the value of information generated through experimentation.

In "**Exploration and Spillovers from Failure**," joint with Alex Frankel (University of Chicago), Josh Krieger (HBS) and Dimitris Papanikolaou (Northwestern), I study the value of learning, particularly from failure. Failures make up the "dark matter" of innovation—much if not most R&D is directed toward projects that never reach some minimum threshold of observability—and consequently little is known about their value. We begin with a model in which firms have better information about incremental projects today, but novel projects generate more information about related projects tomorrow. In our model, firms' incentives to engage in exploration depend crucially on the value that they place on learning from novel projects that fail.

We examine the value of exploration in drug development, a setting in which failure is observable and where novelty can be measured using chemical structure, as pioneered in my previous work. We show that firms are less likely to invest in novel drug candidates but that, conditional on being approved, novel drugs generate more revenue. Our model provides a new way to interpret the classic intuition that novel projects are "higher risk, higher return." Rather than being an inherent trait, we argue that this observed distribution is a consequence of selection: if firms value learning, they should be willing to invest in novel drug candidates even when their expected revenues are low. The fact that this is not the case suggests that firms place less value on future learning. Consistent with this interpretation, we show that firms set higher revenue thresholds for developing novel drugs when they face strong research competition. In such cases, we show that the informational spillovers associated with exploration are more likely to accrue to rival firms, leading the original firm to further discount this source of value. Our results therefore highlight a new dynamic channel linking competition and incentives for exploration.

In another on-going project, "**Learning versus Persuasion in Clinical Trial Design**" with Pierre Azoulay (MIT), Alessandro Bonatti (MIT), and Jennifer Kao (UCLA), I examine firms' incentives to design experiments strategically. In the R&D process, data from experiments serve a dual purpose: they inform the firm about a project's quality, and they allow firms to communicate information about quality to investors, regulators, and consumers. Firms may design different experiments depending on whether their goal is to learn about quality or to persuade others of it. Start-ups wanting to "fail fast" may design experiments to look for bad news while start-ups looking to raise their next round of funding may design experiments to look for good news.

In medicine, firms have a great deal of discretion in how they design clinical trials. They may, for example, create a more difficult test by comparing their drug to the current standard of care, or they may make it easier by comparing to a placebo. In our preliminary work, we document substantial variation in design choices for similar types of trials. Moving forward, we will relate these decisions to a drug's progression through the lens of a Bayesian persuasion model.

These projects highlight the importance of two often overlooked sources of information in innovative settings: information from failed projects, and information generated through a firm's choice of experiment. By demonstrating ways to observe and account for these sources of value, my ongoing research contributes to my broader goal of improving how firms learn from experimentation.

2. Evaluating and investing in people

An organization's innovative capabilities derive, ultimately, from its ability to attract, develop, and retain talented people. My second major stream of work applies my interest in the design of decision-making technologies toward understanding how organizations can more effectively and equitably make hiring and promotion decisions. Here I pay particular attention to two key sources of tension: the value of human versus algorithmic expertise, and the value of current performance versus future potential.

2.1. **Valuing algorithmic and human expertise**

My work on peer review in grant funding highlights the value of expert opinion, but does not consider how it should be combined with hard performance metrics. This question has become particularly important in the context of hiring, given the increasing adoption of machine learning and other types of quantitative evaluation in the recruiting process. In "**Discretion in Hiring**" (*Quarterly Journal of Economics*, 2018), Lisa Kahn (Rochester), Mitch Hoffman (Toronto), and I study how information from job testing scores should be used alongside sources of soft information such as interviewer assessment.

Deciding how to weigh various signals of quality is both an information processing and organizational design problem. When relying on interviews to make hiring decisions, firms effectively delegate some authority to their recruiters because it is difficult to directly verify an applicant's interview performance. The introduction of job testing therefore gives firms the option to limit recruiter discretion and rely more on hard information.

Should firms do this? When recruiters are unbiased and well informed, the answer is unambiguously *no*: discretion allows firms to take advantage of recruiters' private information. In practice, we show that the answer is *yes*: firms can improve hiring decisions by limiting discretion and placing more weight on verifiable information.

Using data from the introduction of a machine-learning based job testing technology, we show that recruiters frequently make exceptions to test recommendations by hiring workers with low test scores when others with high test scores are available. Yet, when faced with identical applicant pools, recruiters who make more exceptions end up with workers who are more likely to quit or be fired. In fact, recruiters who simply follow test recommendations have the best outcomes. This result is consistent with a model in which exceptions are driven by bias or poor judgement. Indeed, our findings show that there are circumstances when firms are better off eliminating discretion rather than continuing to rely exclusively on their recruiters' often-flawed judgement.

Despite the documented value of algorithmic recommendations, there is growing concern that automated approaches to hiring may codify existing biases and restrict access to opportunity. In "**Hiring as Exploration**," (*NBER Working Paper #27736*, 2021), Lindsey Raymond (MIT), Peter Bergman (Columbia), and I show how algorithms can be designed to improve the quality of job-interview decisions while also increasing demographic representation. The central insight in our paper is that most algorithms treat hiring as a static prediction problem when it is better understood as a dynamic learning problem. To identify the best workers over time, firms should engage in both "exploitation" (selecting applicants from groups with proven track records) and "exploration" (selecting applicants from lesser-known groups, in order to learn). Yet despite the importance of exploration, modern hiring algorithms are designed solely to identify applicants from historically successful groups.

In this paper, we build a new type of hiring algorithm that values exploration. Our contextual bandit model selects candidates based on the upper confidence bound associated with our estimates of their expected quality: that is, if two candidates have the same expected quality,

we select the applicant with the noisier estimate. This approach improves learning by favoring candidates with the greater statistical upside.

We use data from professional services recruiting within a Fortune 100 firm to compare our model to the choices of human recruiters as well as hiring models based on supervised learning. Using a variety of econometric tools, we show that incorporating exploration would double the share of Black and Hispanic candidates selected for an interview, while also doubling the firm's overall hiring yield. The same is not true for traditional algorithms, which generate similar gains in hiring yield but select far fewer minority applicants. Together, these findings show that while there need not be an equity-efficiency tradeoff when it comes to expanding diversity, algorithmic design plays a key role in determining whether these Pareto gains can be realized.

2.2. **Valuing current performance versus future potential**

The work above focuses on how firms use information to achieve a single goal: select the "best" worker. My next set of papers consider the challenges that arise when firms balance multiple goals.

In work with with Alan Benson (Minnesota) and Kelly Shue (Yale), "**Promotions and the Peter Principle**" (*Quarterly Journal of Economics*, 2019), I study how firms make promotion decisions. In many firms, promotions serve a dual purpose: they assign workers to new roles and they reward workers for performing well in their current roles. Yet when management skills differ from those of frontline work, promoting workers on the basis of current performance may result in incompetent managers: this is the assertion of the so-called "Peter Principle."

Our paper provides the first direct test of the Peter Principle, using transaction-level data on 50,000 sales workers across 200 firms. Sales is a classic setting in which the worker and management-level skills differ: direct sales positions require confidence and often attracts competitive personalities, while sales management positions require data analysis skills and the ability to coach others. We measure a worker's performance as the revenue they generate and a manger's quality as her value-added to the sales of her subordinates.

The Peter Principle predicts that firms essentially "discriminate" in favor of workers with strong sales performance by promoting them even if they have lower managerial potential. Using an IV strategy based in unrelated variation in the availability of managerial vacancies, we find evidence that this is the case: marginally promoted top sales performers had substantially worse managerial outcomes than marginally promoted workers with weaker sales. Our counterfactual simulations suggest that firms could improve subordinate performance by up to 30% if they focused solely on maximizing managerial quality in their promotion decisions. In showing this, we identify another observable source of information—a sales worker's collaboration experience—that is positively related to managerial performance but overlooked in promotion decisions. This means that firms are either mis-weighting available sources of information or that they place a high value on the incentive or transparency benefits of promoting based on demonstrated job performance.

A key lesson from my work on the Peter Principle is that firms need to find effective ways of predicting workers' potential, not just their current performance. In ongoing work with

the same coauthor team, "**'Potential' and the Gender Promotion Gap**" (working paper, 2021), I show that attempts to predict this counterfactual—how well someone *would* perform, *if* given the opportunity—contribute to gender bias in promotions. Using data on potential and performance rankings of management-track employees, we find that women are consistently judged as having lower leadership potential than men, despite receiving higher performance ratings in their current roles.

This gender gap in potential does not appear justified: women subsequently receive higher ratings of their future performance, relative to male colleagues with the same initial potential scores. Yet even when women are promoted and outperform their male peers, their subsequent potential ratings remain lower. Firms appear to persistently underestimate the potential of their female employees.

Closing this potential gap is difficult. We show that potential ratings, though biased, are informative; doing away with them would increase the equity of promotion decisions, but reduce their quality. Rather, firms would obtain better outcomes by giving higher potential scores to high-performing women. Our findings highlight the importance of de-biasing, rather than eliminating, assessments of potential.

2.3. **Future work: Decision-making with selectively collected evidence**

People and algorithms both form judgements based on the patterns they observe in the world. These patterns, however, are often shaped by selection processes that can be subject to implicit or strategic biases: managers decide which workers or projects to invest in, therefore generating data on the realized potential of some but not others.

In "**Evidence from Police Body Cameras**," Alex Frankel (University of Chicago) and I study the impact of police body-worn cameras (BWCs) on criminal justice outcomes. While BWCs were largely adopted in response to concerns about police misconduct, the footage they yield is far more commonly used as evidence in criminal proceedings when, for instance, drug possession is captured on camera or when the legality of a search is in question.

We show that evidence from BWCs appears to be selective collected. Officers are more likely to capture footage in certain situations: in higher-crime districts, in response to more serious crimes, or for incidents involving younger or Black individuals. This variation in data collection appears to shape conviction outcomes. The introduction of BWCs differentially increases convictions for more heavily surveilled populations, even though, post introduction, the returns to footage (in terms of convictions) are lower for these groups. These preliminary results suggest that heterogeneity in the collection of data, driven by officer preferences, leads to some groups being inefficiently surveilled more than others.

This project introduces a new empirical setting into my work, one in which the use of monitoring technologies can be clearly tied to highly consequential evaluations at the individual level. Going forward, I am interested in broadly exploring the processes that govern how data is transformed into evidence, and its implications for the quality of decisions that organizations make.