The Science of Science

Dashun Wang and Albert-László Barabási

Part 4: Outlook

This is the submitted version.
The printed version[1] is published by Cambridge University Press in 2021.

[1] Wang, Dashun, and Albert-László Barabási. 2021. *The Science of Science*. Cambridge University Press. doi:10.1017/9781108610834.

In the previous three parts, we aimed to offer an overview of the current body of knowledge that the science of science has offered us in the past few decades. Yet, we could not do justice to the full breath of the subject. In this last part, we aim to offer a brief overview of the field's emerging frontiers, briefly discussing several areas that are just gaining traction in the community, and that have the potential to offer new, promising developments in the coming years. Unlike our approach in Parts I–III, here we do not aim to offer a comprehensive coverage, but rather to introduce some interesting problems posed by the existing research, offer some representative examples, suggest new opportunities, and imagine where they might lead us.

We will see how understanding the doings of science may change how science is done—how knowledge is discovered, hypotheses are raised, and experiments are prioritized—and what these changes may imply for individual scientists. We will consider the coming age of artificial intelligence, think through how it might impact science, and illustrate how human and machine can work together to achieve speed and efficiency that neither human nor machine can achieve alone. We will also examine how the science of science can go beyond its current attempts at correcting potential biases and how to generate causal insights with actionable policy implications.

Chapter 4.1

Can science be accelerated?

In the mid-18th century, the steam engine jump-started the Industrial Revolution, affecting most aspects of daily life and providing countries that benefitted from it with a new pathway to power and prosperity. The steam engine emerged from the somewhat counterintuitive yet profound idea that energy in one form—heat—can be converted to another—motion. While some see the steam engine as one of the most revolutionary ideas in science, it was arguably also the most overlooked [1]. Indeed, ever since we knew how to use fire to boil water, we've been quite familiar with that annoying sound the kettle lid makes as it vibrates when water reaches a rolling boil. Heat was routinely converted to motion in front of the millions, but for centuries no one seemed to recognize its practical implications.

The possibility that breakthrough ideas like the steam engine hover just under our noses, points to one of the most fruitful futures for the science of science. If our knowledge about knowledge grows in breadth and quality, will it enable researchers and decision-makers to reshape our discipline, "identify areas in need of reexamination, reweight former certainties, and point out new paths that cut across revealed assumptions, heuristics, and disciplinary boundaries" [2]?

Machines have aided the scientific process for decades. Will they be able to take the next step, and help us automatically identify promising new discoveries and technologies? If so, it could drastically accelerate the advancement of science. Indeed, scientists have been relying on robot-driven laboratory instruments to screen for drugs and to sequence genomes. But, humans are still responsible for forming hypotheses, designing experiments, and drawing conclusions. What if a machine could be responsible for the *entire* scientific process—formulating a hypothesis, designing and running the experiment, analyzing data, and deciding which experiment to run next—all without human intervention? The idea may sound like a plot from a futuristic sci-fi novel, but that sci-fi scenario has, in fact, already happened. Indeed, a decade ago, back in 2009, a robotic system made a new scientific discovery with virtually no human intellectual input [3].

4.1.1 Close the loop

Figure 4.1.1 shows "Adam," our trusty robot scientist, whose training ground was baker's yeast, an organism frequently used to model more complex life systems. While yeast is one of the best studied organisms, the function of 10 to 15 percent of its roughly 6,000 genes remain unknown. Adam's mission was to shed light on the role of some of these mystery genes. Adam was armed with a model of yeast metabolism and a database of the genes and proteins necessary for metabolism in other species. Then it was set loose, with the role of the supervising scientists being limited to periodically add laboratory consumables and to remove waste.

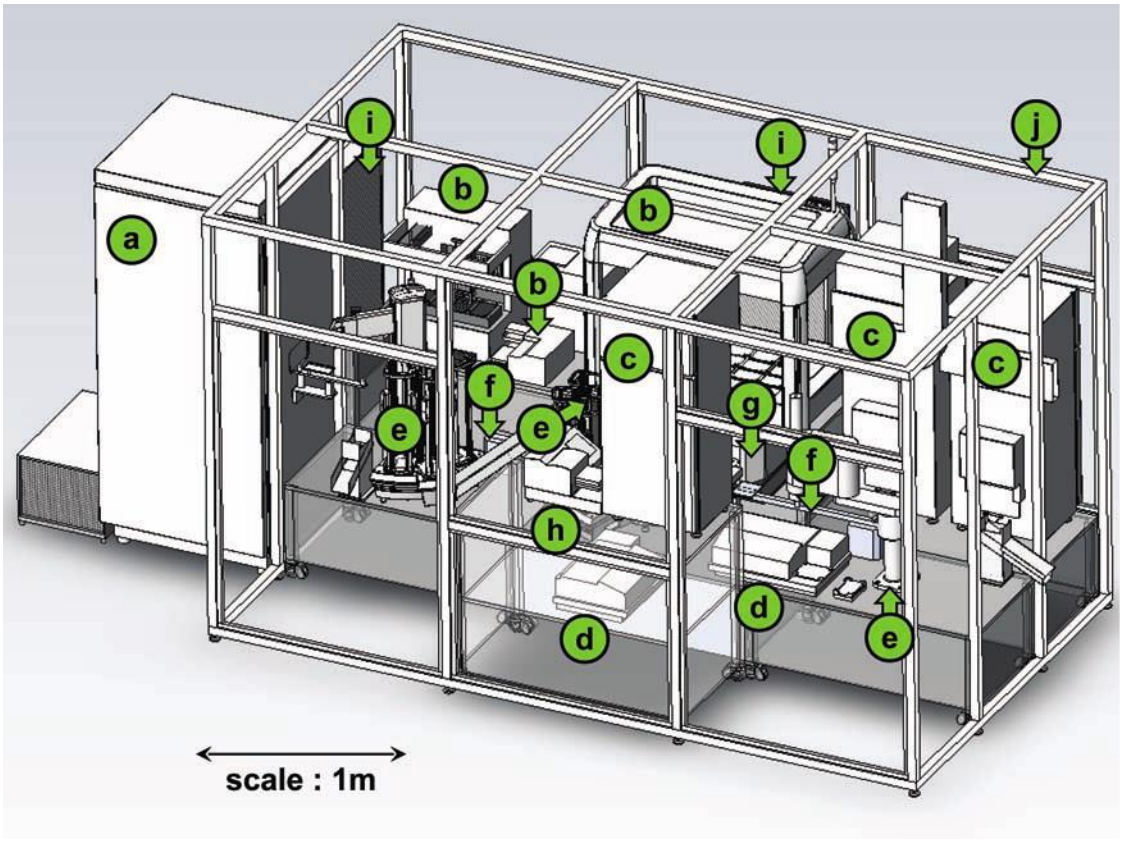

Figure 4.1.1 **The robot scientist Adam.** Adam differs from other complex laboratory systems in the individual design of the experiments to test hypotheses and the utilization of complex internal cycles. Adam has the following components: a, an automated 20°C freezer; b, three liquid handlers c, three automated +30°C incubators; d, two automated plate readers; e, three robot arms; f, two automated plate slides; g, an automated plate centrifuge; h, an automated plate washer; i, two high-efficiency particulate air filters; and j, a rigid transparent plastic enclosure. It also has two bar-code readers, seven cameras, 20 environment sensors, and four personal computers, as well as software, allowing it to design and initiate over a thousand new strains and defined-growth-medium experiments each day. After King *et al.* [3].

Adam sought out gaps in the metabolism model, aiming to uncover "orphan' enzymes, which haven't been linked to any parent genes. After selecting a desirable orphan, Adam scoured the database for similar enzymes in other organisms, along with their corresponding genes. Using this information, Adam then hypothesized that similar genes in the yeast genome may encode the orphan enzyme and began to test this hypothesis.

It did so by performing basic operations: It selected specified yeast strains from a library held in a freezer, inoculated these strains into microtiter plate wells containing rich medium, measured their growth curves, harvested cells from each well, inoculated these cells into wells containing defined media, and measured the growth curves on the specified media. These operations are awfully similar to the tasks performed by a lab assistant, but they are executed robotically.

Adam is a good lab assistant, but what truly makes this machine so extraordinary is its ability to "close the loop," acting as a scientist would. After analyzing the data and running follow-up experiments, it then designed and initiated over a thousand *new* experiments. When all was said and done, Adam formulated and tested 20 hypotheses relating to genes that encode 13 different orphan enzymes. The weight of the experimental evidence for these hypotheses varied, but Adam's experiments confirmed 12 novel hypotheses.

To test Adam's findings, researchers examined the scientific literature on the 20 genes investigated. They found strong empirical evidence supporting six of Adam's 12 hypotheses. In other words, 6 of Adam's 12 findings were already reported in the literature, so technically they were not new. But they were new to Adam, because it had an incomplete bioinformatics database, hence was unaware that the literature has already confirmed these six hypotheses. In other words, Adam arrived at these six findings independently.

Most importantly, Adam discovered three genes which together coded for an orphan enzyme. This finding represents new knowledge that did not yet exist in the literature. And when researchers conducted the experiment by hand, they confirmed Adam's findings. The implication of this is stunning: *A machine, acting alone, created new scientific knowledge*.

We can raise some fair criticisms about Adam, particularly regarding the novelty of its findings. Although the scientific knowledge "discovered" by Adam wasn't trivial, it was implicit in the formulation of the problem, so its novelty is, arguably, modest at best. But the true value of Adam is not about what it can do *today*, but what it may be able to achieve *tomorrow*.

As a “scientist,” Adam has several distinctive advantages. First, it doesn’t sleep. As long as it is plugged into a power outlet, it will unceasingly, doggedly putter away in its pursuit of new knowledge. Second, this kind of “scientist” scales, easily replicable into many different copies. Third, the engine that powers Adam—including both its software and hardware—is doubling in efficiency every year. The human brain is not.

Which means Adam is only the beginning. Computers already play a key role in helping scientists store, manipulate, and analyze data. New capabilities like those offered by Adam, however, are rapidly extending the reach of computers from analysis to the formulation of hypotheses [4]. As computational tools become more powerful, they will play an increasingly important role in the genesis of scientific knowledge. They will enable automated, high-volume hypothesis generation to guide high-throughput experimentation. These experiments will likely advance a wide range of domains, from biomedicine to chemistry to physics, and even to the social sciences [4]. Indeed, as computational tools efficiently synthesize new concepts and relationships from the existing pool of knowledge, they will be able to usefully expand that pool by generating new hypotheses and drawing new conclusions [4].

These advances raise an important next question: How do we generate the most fruitful hypotheses in order to more efficiently advance science?

4.1.2 The next experiment

To improve our ability to discover new, fruitful hypotheses, we need to develop a deeper understanding of how scientists explore the knowledge frontier, and what types of exploration—or exploitation—tend to be the most fruitful. Indeed, merely increasing the pool of concepts and relationships between them typically lead to a large number of low-quality hypotheses. Instead, we need to be discerning. One example of how to hone-in on valuable discoveries is the Swanson hypothesis [5]. Swanson posits that if concepts A and B are studied in one literature, and B and C in another, then the link between A and C may be worth exploring. Using this approach, Swanson hypothesized that fish oil could lessen the symptoms of Raynaud’s blood disorder and that magnesium deficits are linked to migraine headaches, both of which were later validated [5].

Recent attempts at applying computational tools to massive corpora of scientific texts and databases of experimental results have substantially improved our ability to trace the dynamic frontier of knowledge [6, 7]. Analyzing the abstracts of millions of biomedical papers published from 1983 to 2008, researchers identified chemicals jointly studied in a paper, represented specific research problems as links between various scientific entities, and organized them in a knowledge graph [7]. This graph allowed them to infer the typical strategy used by scientists to explore a novel chemical relationship.

For example, Fig. 4.1.2 shows that scientists have the tendency to explore the neighborhood of *prominent* chemicals. This paints a picture of a "crowded frontier" [8], where multiple researchers focus their investigations on a very congested neighborhood of the discoverable space, rather than exploring the space of the unknown more broadly.

While these prominent chemicals may warrant more investigations, this example suggests that there could be more optimal ways to explore the map of knowledge. For example, the model estimates that, overall, the optimal strategy for uncovering 50% of the graph can be nearly ten times more efficient than the random strategy, which tests all edges with equal probability. This illustrates that a deeper understanding of how science evolves could help us accelerate the discipline's growth, allowing us to strategically choose the best next experiments.

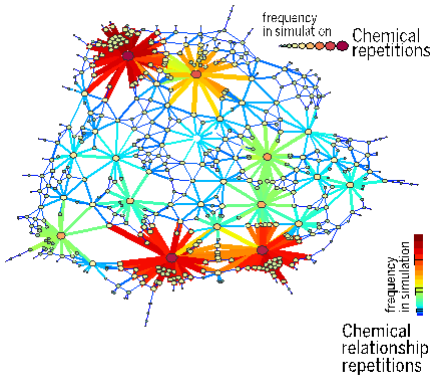

Figure 4.1.2 **Choosing experiments to accelerate collective discovery.** The actual, estimated search process illustrated on a hypothetical graph of chemical relationships. Here each node is a chemical and links indicate the pair of chemicals examined together in a paper, representing publishable chemical relationships. The graph shows simulation results averaged from 500 runs of that search strategy. The strategy swarms around a few "important," highly connected chemicals, whereas more optimal strategies estimated are much more even and less likely to "follow the crowd" in their search across the space of scientific possibilities. Adapted from [7].

4.1.3 New challenges

"White spaces" offer both opportunity and peril. There is, of course, untapped potential in exploring fresh connections. But the underlying dangers are in the "file-drawer problem," i.e. results that never got published, which is driven by scientists' preference for publishing positive results [9, 10] and statistical findings that exceed field-specific thresholds (e.g. p-value < 0.05) [11, 12]. It is possible, therefore, that many gaps in the literature have been explored by previous generations of scientists, but, because they didn't find interpretable results, they never reported their efforts [2].

If we want to accelerate scientific progress, these negative results have positive value. Indeed, instead of being discarded, negative results should be saved, shared, compiled and analyzed. This would not only offer a less biased view of the knowledge landscape—helping us separate opportunity from peril—it would also improve the reproducibility of findings, since we increase the credibility of positive results when we place them in the context of negative ones. This is already done in clinical trials, where results must be published at https://clinicaltrials.gov/, independent of the outcome. Indeed, biochemical journals require investigators to register early phase 1 clinical trials, as failed attempts are of critical importance to public health. For this reason, the top journals have all agreed that they won't publish the findings of phase 3 trials if their earlier phase 1 results—whether positive or negative—weren't reported first [13].

As of 2019, we're already seeing promising indications that other areas of science are moving in this direction. For example, the "preregistration revolution" [14] encourages scientists to specify their research plans in advance, before they gather any data. This practice originated in psychology, but it's being adopted by other disciplines as well. If journals and funding agencies provide the right nudges, preregistered studies may one day become the rule rather than exception.

Attempting to optimize the way we explore the knowledge landscape presents a second challenge: The design of incentive structures which encourage individual risk-taking. Whereas scientists share the overall objective of exploring the space of unknowns, individually they may hold objectives that are not optimized for science as a whole. Sociologists of science have long hypothesized that researchers' choices are shaped by an "essential tension" between productive tradition and risky innovation [15, 16]. Scientists who adhere to a research tradition often appear productive by publishing a steady stream of contributions that advance a focused research agenda. But a focused agenda may limit a researcher's ability to sense and seize opportunities for staking out new ideas. Indeed, although an innovative publication may lead to higher

impact than a conservative one, high-risk innovation strategies are rare, because the potential reward does not compensate for the risk of failing to publish at all [2].

Scientific awards and accolades offer a possible way to encourage risk-taking, functioning as key incentives for taking risks to produce innovative research. Funding agencies can also help. By proactively sponsoring risky projects that test truly unexplored hypotheses, we may see more scientists venturing off well-worn research paths. But this is often easier said than done. Indeed, measurements show that the allocation of biomedical funding across topics and disciplines in the United States is better correlated to previous allocations than to the actual need, defined by the burden of diseases [17]. For example, US spends about 55% of its biomedical research funding on genetic approaches, despite the fact that genome-based variation can explain only about 15-30% of disease causation [18]. Much less funding is devoted to environmental effects and diet, despite these being responsible for the lion's share of the disease burden experienced by the population [18], from diabetes to heart disease. These findings highlight a systemic misalignment between U.S. health needs and research investment, casting doubt on the degree to which funding agencies—which are often run by scientists embedded in established paradigms—can successfully influence the evolution of science without additional oversight, incentives, and feedback.

To manage risk, scientists can learn from Wall Street, which invests in portfolios rather than individual stocks. While no one can really predict which experiment will work and which won't, we can spread the risk across a portfolio of experiments. Doing so would take some of the pressure off of individual research endeavors and individual scientists. For example, policymakers could design institutions that cultivate intelligent risk-taking by shifting evaluation from the individual to the group, as practiced at Bell Labs. They could also fund promising people rather than projects, an approach successfully implemented by the Howard Hughes Medical Institute. With the right incentives and institutions, researchers can choose experiments that benefit not just themselves, but also science and society more broadly [7].

*

Of course, we face challenges in realizing this vision, but the potential benefit of doing so is enormous. Eighty percent of the most transformative drugs from the past 25 years can be traced to basic scientific discoveries [19]. Shockingly, though, these discoveries were published on average *31 years* before the resulting drugs achieved FDA approval. The decades-long gap between a basic discovery to a FDA approval highlights just how vital the science of science could be in fundamentally transforming the practice

of science and its policy. What if we could identify if a new discovery could lead to a new drug at time of its publication? What if we could short cut the pathway to arrive at new technologies and applications faster?

Indeed, as we discuss next, with the rise of artificial intelligence, some of these goals no longer seem far-fetched. In fact, the resulting advances may even change the very meaning of doing science.

Chapter 4.2

Artificial Intelligence

"What just happened?"

The sentiment circling the hallways of the CASP conference on December 2, 2018 was one of puzzlement. CASP, short for the Critical Assessment of Structure Prediction, is a biannual competition aimed at predicting the 3D structure of proteins—large, complex molecules essential for sustaining life. Predicting a protein's shape is crucial for understanding its role within a cell, as well as diagnosing and treating diseases caused by misfolded proteins, such as Alzheimer's, Parkinson's, Huntington's and cystic fibrosis [20]. But how proteins fold their long chains of amino acids into a compact 3D shape remains one of the most important unsolved problems in biology.

Established in 1994, CASP is the Kentucky Derby of protein folding. Every two years, leading groups in the field convene to "horserace" their best methods, establishing a new benchmark for the entire field. Then the researchers return to their labs, study each other's methods, refine and develop their own approaches, only to reconvene and race again in another two years.

At the 2018 conference, two things were unusual. First, there had been "unprecedented progress in the ability of computational methods to predict protein structure," as the organizers put it. To put that progress in perspective, roughly two competitions worth of improvement had been achieved in one. Second, this giant leap was not achieved by any of the scientists in the field. The winning team was a complete stranger to the community.

What happened at the 2018 CASP competition was merely one instance out of many in past few years, where artificial intelligence (AI) systematically outperformed human experts in a large variety of domains. These advances have led to the consensus that the ongoing AI revolution will change almost every line of work, creating enormous social and economic opportunities, and just as many challenges [21]. As the society is preparing for the moment when AI may outperform and even *replace* human doctors, drivers,

soldiers, and bankers, we must ask: How will AI impact science? And, what do these changes mean for scientists?

4.2.1 What's new with this wave of AI?

The technology that underlies the current AI revolution is called *deep learning*, or more specifically, deep neural networks. While there are many things AI experts have yet to agree on—including whether we should call the field "artificial intelligence" or "machine learning"—there is a consensus, in and out of academia, that this really *is* the next big thing.

A defining feature of deep learning is that it *really* works. Since 2012, it has beaten existing machine-learning techniques in more areas than we can keep track of. These advances have surely changed such typical computer science areas as image [22-25] and speech recognition [26-28], question answering [29] and language translation [30, 31]. But deep neural networks have also shattered records in such far-flung domains as predicting the activities of drugs [32], analyzing particle accelerator data [33, 34], reconstructing brain circuits [35], and predicting gene mutations and expression [36, 37].

Most importantly, many of these advances were not incremental, but represent jumps in performance. When in 2012, deep learning debuted at the ImageNet Challenge, the premier annual competition for recognizing objects in images, it almost *halved* the state-of-art error rate. Since then, deep learning algorithms are rapidly approaching human-level performance. In some cases, like strategy games like Go or shogi (Japanese chess), and multi-player video games that emphasize collaborations, or bluffing at Texas hold'em tables, they even surpass the performance of human experts. At the CASP conference in 2018, deep learning added one more accolade to its growing resume of super-human performances: It beat all *scientists* at predicting the 3D structure of proteins.

In simple terms, artificial intelligence helps us find patterns or structures in data that are implicit and probabilistic rather than explicit. These are the type of patterns which are easy for humans to find (i.e. where the cat is in a picture) but were traditionally difficult for computers. More precisely, it used to be difficult for humans to translate such tasks to computers. With AI, the machines have developed an uncanny way to do this translation themselves.

Although AI is popping up everywhere, major recent advances all hinge on a single approach: Supervised Learning, where algorithms are supplied with only two sets of information: a large amount of

input, or “training data,” and clear instructions (“labels”) for sorting the input. For example, if the goal is to identify spam emails, you supply the algorithms with millions of emails and tell it which ones are spam and which are not. The algorithm then sifts through the data to determine what kinds of emails tend to be spam. Later, when shown a new email, it tells you if it looks “spammy” based on what it has already seen.

The magic of deep learning lies in its ability to figure out the best way of representing the data without human input. That’s thanks to its many intermediate layers which each offer a way to represent and transform the data based on the labels. With sufficient number of layers, the system becomes very good at discovering even the most intricate structures or patterns hidden in the data. More notably, it can discover these patterns all by itself. We can think of the different layers of deep neural networks as having the flexibility of tuning millions of knobs. As long as the system is given *enough data* with *clear directions*, it can tune all the knobs automatically and figure out the best way to represent the data.

What’s so different about AI this time around? After all, more than 20 years ago, I.B.M.’s chess-playing program, Deep Blue, beat Garry Kasparov, back then the world champion. In the past AI was meticulous, but it lacked intelligence. Deep Blue defeated Mr. Kasparov because it could evaluate 200 million positions per second, allowing it to anticipate which move was most likely to lead to victory. This type of AI fails at more complex games like Go or the protein folding problem, where it can’t process all the possibilities.

Deep learning, on the other hand, has been wildly successful in these arenas. In 2016, AlphaGo, created by researchers at DeepMind defeated the world Go champion Lee Sedol over five matches. It didn’t win by evaluating every possible move. Instead, it studied games of Go completed by human players, learning what kinds of moves tend to lead to victory or defeat.

But why learn from us, when the system could learn from itself? Here is where it gets *really* interesting. Merely one year after AlphaGo’s victory over humans, DeepMind introduced AlphaZero [38], which has no prior knowledge or data input beyond the rules of the game. In other words, it truly started from scratch, teaching itself by repeatedly playing against itself. AlphaZero not only mastered Go, it also mastered chess and shogi, defeating every human player and computer program.

Most importantly, because AlphaZero didn’t learn from human games, it doesn’t play like a human. It’s more like an alien, showing a kind of intuition and insight which grandmasters had never seen before. Ke Jie, the world champion of Go, remarked that AI plays “like a god.” Indeed, it didn’t rely on human

knowledge to discover its intricate, elegant solutions. And AlphaZero managed this feat at super-human speed: After a mere four hours of chess training and eight hours of Go training, its performance exceeded the best existing programs.

Think about these numbers again. We gave an AI algorithm the rules of humanity's richest and most studied games, leave it with only the rules and a board, and let it work out strategies by itself. It starts by making all sorts of stupid mistakes, like all beginners do. But, by the time you come back and check on it later that day, it became the best player there has ever been.

If deep learning can beat humanity at its own board games, finding previously unimagined solutions to complex problems, how will it impact science, a discipline dedicated to advancing creative innovation?

4.2.2 The impact of AI on science

There are two major avenues through which AI could affect the way we do science. One is similar to what Google has done to the Internet: AI will drastically improve access to information, optimizing the various aspects of science, from information access to the automation of many processes scientists now perform. This is the utopic version, as most scientists would welcome the automation of the routine tasks, allowing us to focus on the creative process. The other avenue will be more like to what AlphaGo has done to games of Go: AI systems could offer high-speed, creative solutions to complex problems. In a dystopic world, AI could one day replace us, scientists, moving science forward with a speed and accuracy unimaginable to us today.

Organizing information

Artificial intelligence already powers many areas of modern society. Every time you type a search query into Google, AI scours the web, guessing what you really want. When you open the Facebook app, AI determines which friend's update you see first. When you shop on Amazon, AI offers you products that you may also like, even though these items were never on your shopping list. AI is also increasingly present in our devices. When you hold up your smartphone to snap a photo, AI circles in on faces and adjusts the focus to best capture them. When you address your "personal assistant"—Siri, Alexa, or Cortana—you're relying on AI to transcribe your commands into text.

What aspects of science can be augmented by this kind of AI? To begin with, today there is more literature published than we could ever hope to keep up with. Could AI identify and personalize what papers we should read? Can it cohesively summarize the text of these papers, extracting the key findings relevant to us, creating a newsletter-style digest of the key advances in the field? These new capabilities will help researchers expand the depth and quality of knowledge they acquire, as well as help identify new research possibilities.

For decision-makers in science AI could offer a more comprehensive "horizon scanning" capability, suggesting areas for strategic investment, and identifying ideas and even assembling teams that could lead to transformative science. Publishers may also use deep learning to identify which referees to seek for a manuscript, or to automatically identify apparent flaws and inconsistencies in a manuscript, avoiding the need to bother human reviewers.

Some of these applications may seem far-fetched, especially if we hope to achieve the high level of accuracy and reliability that scientists and decision-makers require. But the reality is, although technology has substantially re-shaped human society in the past two decades, technologies that could facilitate scientific processes have languished. If you're unconvinced, just take a look at the grant submission websites of the National Science Foundation, or at the ScholarOne manuscript systems which handle most of the site's editorial functionalities—they resemble fossil websites abandoned since the dot com boom.

Solving scientific problems

Will AI one day help us pose and solve fundamental scientific problems? By integrating diverse sources of information that no individual scientist can master, can AI systems help scientists come up with more creative and better solutions faster? Can it also suggest new hypotheses, or even new areas to explore?

We've already seen some encouraging early progress in this area. For instance, researchers applied deep learning to medical diagnosis and developed an algorithm to classify a wide range of retinal pathologies, obtaining the same degree of accuracy as human experts [39]. In another example, an AI algorithm trained to classify images of skin lesions as benign or malignant achieved the accuracy of board-certified dermatologists [40]. And in emergency rooms, deep learning can now help us to decide whether a patient's CT scan shows signs of a stroke [41]. The new algorithm flagged these telltale signals at a level of accuracy comparable to medical experts'—but, crucially, it did so 150 times faster.

And of course, there's AlphaFold, the deep learning system that filled CASP attendees with awe. In the CASP contest, competing teams were given the linear sequence of amino acids for 90 proteins. The 3D shape of these proteins was known but had not yet been published. Teams then computed how the proteins would fold. By sifting through past known protein folding patterns, the predictions by AlphaFold were, on average, more accurate than any of its 97 competitors.

These successful uses of AI technology possess the two critical ingredients for deep learning: a large amount of training data and a clear way to classify it. For example, to detect skin cancers, researchers fed the algorithm with millions of images of skin lesions, telling it which ones are benign and which are malignant. Because the algorithm didn't go through the same training as dermatologists do, it may not see the same patterns that dermatologists are trained to see. This means, the AI system may also recognize patterns that have so far eluded us.

What scientific areas would most benefit from these advances? It may be helpful to think about this question in terms of the two critical ingredients for deep learning—copious data and clear perimeters for sorting it. This suggests that scientific areas that may more directly benefit from AI technology are those that are narrow enough, so that we can provide an algorithm with clear sorting strategies, but also deep enough that, by looking at *all* the data—which no scientist could ever do—could allow the AI to arrive at new results.

But most importantly, although machines are rapidly improving their efficiency and accuracy, the most exciting future of science belongs to neither humans or machines alone, but to a strategic partnership between the two.

4.2.3 Artificial and Human Intelligence

Let's think again about what happened in the AlphaFold case. Scientists using a new technology but without expertise or training in the specific scientific domain, were able to outperform the entire community of experts relying on traditional technologies. This example raises an important question: what if we pair the latest technology *with* researchers' subject matter expertise?

A critical area of future science of science research concerns the integration of artificial intelligence so that machines and minds can work together. Ideally, AI will broaden a scientist's perspective in a way that human collaborators can't, which has far-reaching implications for science.

A recent example comes to mind. Hoping to remedy a present-day challenge in science known as the "reproducibility crisis," researchers used deep learning to uncover patterns in the narrative of scientific papers which signal strong and weak scientific findings. In 2015, the "Reproducibility Project: Psychology" (RPP) manually tested the replicability of 100 papers from top psychology journals by using the exact procedures implemented in the original studies, finding that 61 of 100 papers failed their replication test [42]. Since then, studies in psychology, economics, finance, and medicine have reported similar cases for papers that fail to reproduce [43-46].

In response, researchers combined artificial and human intelligence to estimate replicability [47]. Using data on 96 studies that underwent rigorous manual replication tests, they trained a neural network to estimate a paper's likelihood of replicability and tested the model's generalizability on 249 out-of-sample studies. The results are fascinating: The model achieved an average Area Under the Curve (AUC) of 0.72, indicating that its predictions were significantly better than chance. To put this result in context relative to the prognosticative information provided by expert reviewers, researchers trained a new AI model that used only reviewer metrics using the same data and training procedures, finding that the reviewer metrics-only model is significantly less accurate than the narrative-only model (an AUC of 0.68). These results suggest that the AI relies on diagnostic information not captured by expert reviewers. Indeed, although the statistics reported in the paper are typically used to evaluate its merits, the accuracy of the AI shows that the narrative text actually holds more previously unexplored explanatory power. Most importantly, combining the information from the narrative model and the reviewer metrics model—in other words, combining the insights of machines and humans—yields a new AI model with the highest accuracy (AUC = 0.74).

Analyses of the mechanisms behind the model's predictive power indicate that conspicuous factors—such as word or persuasion phrase frequencies, writing style, discipline, journal, authorship, or topics—do not explain the results. Rather, the AI system uses an intricate network of linguistic relationships to predict replicability. And while words outnumber statistics in scientific papers by orders of magnitude, a paper's text has so far been largely unexploited in the science of science. Algorithms can now take advantage of the full text of papers, to detect new patterns and weak scientific findings that human experts may miss.

This example highlights a novel, and potentially formidable, human-machine partnership. Indeed, while machines can consume and digest more information than humans can, the current AI applications all belong to the category of "narrow AI," in that they can only tackle a specifically defined problem. In this

respect, current AI systems are much like washing machines. They can wash any piece of clothing you throw their way, but they wouldn't know what to do with your dishes. For that, you need to build another narrow machine called a dishwasher. Similarly, we can build AI systems that are extremely good at protein folding, but which can do almost nothing else. By contrast, humans have the ability to learn, extrapolate, and think creatively in ways that machines can't.

The physics Nobel laureate Frank Wilczek famously predicted that, in 100 years, the best physicist will be a machine. Advances like AlphFold offer credibility to this prediction. But Wilczek's prediction also simplifies a complex picture: science is not only about solving well defined problems. The most admired scientists are often those that propose new problems and identify new areas of study. Those that realize that the tools and the knowledge have advanced to the point that new discoveries are possible to break through from fallow grounds. It took humans, therefore, to realize that time is ripe to enter these new areas, and to address the challenges they present. It took humans to realize that the data and tools have matured enough that we can move forward successfully. That is, science is not only about problem solving. It is also about intuition, and ability to spot new frontiers, courage to go there, and leadership.

AI has made fabulous advances in solving problems posed by humans. It could even formulate new hypotheses within the realm of an existing body of knowledge and paradigm. Will AI ever get to the point to detect the need for a new theory, like evolution or quantum mechanics, and pursue it doggedly? As of now, there are no indications on the horizon that AI is capable of that, and many AI experts doubt it could ever be possible [48]. Hence, for now machines do not yet claim potential ownership over the future of science. Rather, our most exciting future discoveries require a strategic partnership between humans and machines. Indeed, if we assign tasks based on the abilities of each partner, scientists working in tandem with machines will potentially increase the rate of scientific progress dramatically, mitigate human blind spots, and in the process, revolutionize the practice of science.

*

There is, however, an important "but." A major drawback with the current wave of AI is that it's a black-box. Sure, it works really well, but no one knows why—which could be a big problem, especially in science. To see why, consider Amazon's experience of using AI to pick their future employees. Since 2014,

Amazon has been building computer programs to review job applicants' resumes. The company's experimental AI tool rated job candidates by one to five stars—much like shoppers do when they rate products on Amazon. At first glance, it looked to be an HR holy grail. You give it 100 resumes, and it'll spit out the top five. But soon, the company realized that its new algorithm systematically penalized female candidates. The program was trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period, most of which came from men. So, the system quickly taught itself to prefer male candidates. It penalized resumes that included the word "women" and downgraded graduates of two all-women's colleges [49].

The moral of this story is not that AI can't do its job right. After all, the system did exactly what it was trained to do. Humans asked it to look at millions of past resumes, both rejects and hires, and to use this information to spot future hires. What Amazon's debacle illustrates is that, as our tools grow in accuracy and sophistication, they'll amplify and help perpetuate whatever biases we humans already have. Which means that as the science of science advances, there will be an increased need to understand biases and causal relationships in the tools and metrics our community builds.

Chapter 4.3

Bias and Causality in Science

Science of science research relies on publications and citations as its primary data sources, an approach that has important implications.

First, the explored insights and findings are limited to ideas successful enough to merit publication. Yet, most ideas fail, sometimes spectacularly. As we lack data on failures, our current understanding of how science works has multiple blind-spots. Arguably, by focusing only on successful research, we perpetuate systematic biases against failure.

Second, for most of this book, the success outcomes we've used rely on citation counts. While this bias toward citations is reflective of the current landscape of the field, it highlights the need to go beyond citations as the only "currency" of science.

Third, the data-driven nature of the science of science research indicates that most studies remain observational. While such descriptive studies can reveal strong associations between events and outcomes, to understand whether a specific event "causes" a specific outcome requires us to go beyond observational studies and systematically assess causality.

Next, we will discuss in more detail these important directions for the science of science. Let's begin with a simple question: How big of a problem is our ignorance of failure?

4.3.1 Failure

In World War II, the British military had access to an armor material that could make its planes bullet-proof. Yet, this new armor was heavy, hence it could be only deployed to cover some, but not all, parts of a plane without compromising its flight range or maneuverability. Hence the designers faced an important question: Which part of their planes should be armored first?

The Allied forces decided to take a data-driven approach. They looked at returning B-29 bombers and they marked every place where they had taken fire. Once they gathered the data, the decision seemed simple:

Just apply the armor to areas most often pocked with bullet holes (Fig. 4.3.1). As they moved forward with the plan, Abraham Wald, a statistician in the research group, stepped in. He explained to the Defense Department that the correct strategy should be doing exactly the opposite: applying armor to areas where bullet holes *weren't* recorded. After all, all the data comes from planes that *successfully* returned to the base. A fuselage that looks like Swiss cheese isn't a real concern if it made a round-trip journey. Instead, parts that were missing bullet holes, corresponding to engines, culprit, and other critical parts, are what need the extra protection—as those planes never made it back.

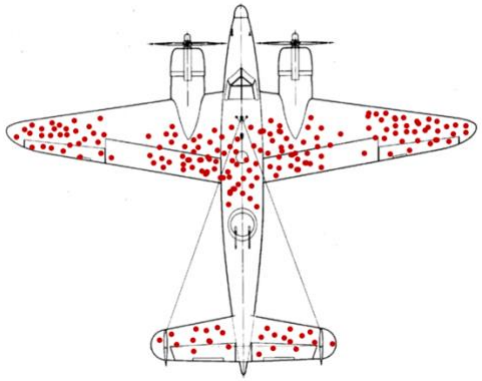

Figure 4.3.1 **Survivorship bias.** Illustration of hypothetical damage pattern on a WW2 bomber. The damaged portions of returning planes show locations where they can take a hit and still return home safely; those hit in other places do not survive. (Image from Wikipdia under CC BY-SA 4.0)

This is a wonderful example where an initial conclusion had to be completely reversed because it's based on data containing only successful samples. Similar biases abound in science: the literature tends to focus on the researchers who have successfully raised funds, published in peer-reviewed journals, patented inventions, launched new ventures, and experienced long, productive careers. These successful cases beg an important question: Given that our current understanding of science has been derived almost exclusively from successful stories, are we certain that the conclusions we've come to don't require major corrections?

Failures in science remain underexplored mainly because it's difficult to collect ground-truth information, accurately tracing failed ideas, individuals, and teams. That situation could be remedied, however, by exploring new data sources and combining them with existing ones. For example, since 2001, patent applications filed at the US Patent and Trademark Office (USPTO) have been published within 18 months of their priority date, regardless of whether they're granted or not. By tracing all applications filed, researchers can now identify successful ideas that were granted patents by the USPTO, alongside those that

were turned down. Grant application databases, which contain both funded and unfunded grant proposals, represent another rich source of information for deepening our understanding of success and failure in science. A limited number of researchers have gained access to internal grant databases at funding agencies such as the U.S. National Institutes of Health (NIH) [50] and the Netherlands Organization of Scientific Research [51]. When combined with existing publication and citation databases, this data could enable scientists to map out the rich contexts in which success and failure unfolds.

Initial investigations in this direction have offered several counterintuitive insights. Take one of our own studies as an example [50]. We focused on junior scientists whose R01 proposals to the NIH fell just above and below the funding cutoff, allowing us to compare "near-miss" with "near-win" individuals and to examine their longer-term career outcomes. These two groups of junior scientists were essentially the same before the "treatment" but faced a clearly different reality afterward: One group was awarded on average $1.3 million over five years, while the other group was not. How big of a difference did this kind of early-career setback make for junior scientists?

To find out, we followed both groups' subsequent career histories, finding that an early-career setback is indeed of consequence: It significantly increases attrition. Indeed, one near miss predicts more than a 10% chance the researcher would disappear permanently from the NIH system. This rate of attrition is somewhat alarming, because to become an NIH PI, applicants have to have a demonstrated track record and years of training. A single set-back, in other words, can end a career.

Yet, most surprisingly, the data indicates that the near-miss individuals who kept working as scientists systematically *outperformed* the near-winners in the long run—their publications in the next ten years garnered significantly higher impact. This finding is quite striking. Indeed, take two researchers with similar performance who are seeking to advance their careers. The two are virtually twins, except that one has had an early funding failure and the other an early funding success. It was puzzling to discover that it's the one who *failed* who will write the higher-impact papers in the future.

One possible explanation for this finding is a screening mechanism, where the "survivors" of the near-miss group have fixed, advantageous characteristics, giving those who remain in the field better performance on average than their near-win counterparts. But when we accounted for this screening effect, we find that screening alone cannot explain the performance gap. In other words, those who failed but

persevered didn't just start out as better performers; they also became better versions of themselves, reinforcing that "what doesn't kill you makes you stronger."

These results seem especially counter-intuitive given that science is dominated by "rich get richer" dynamics, where success, not failure, brings future success. These findings therefore indicate that failure in science has powerful, opposing effects—hurting some careers, but also unexpectedly strengthening outcomes for others. Just like prior success, prior failure can also act as a marker of a future successful career. That's good news, since scientists encounter failure on a weekly if not daily basis.

More generally, though, this study raises a broader point: While we've achieved some success in understanding success, we may have failed to understand failure. Given that scientists fail more often than they succeed, understanding the when-s, why-s, and how-s of failure, and their consequences, will not only prove essential in our attempts to understand and improve science; they may also substantially further our understanding of the human imagination by revealing the total pipeline of creative activity.

Manuscript rejections boost the impact of the paper.

A scientist who submits a new paper, only to have it rejected time after time, may conclude that the paper is simply not that good, and that even if it ever gets published, it will likely fade away into obscurity. The data shows otherwise: rejections actually boost the impact of paper. We know this thanks to a study that tracked the submission histories of 80,748 scientific articles published in 923 bioscience journals between 2006 and 2008 [52]. The study found that resubmissions are rare: 75% of all papers were published in the journal to which they are first submitted. In other words, scientists are good at gauging where their papers are best suited. But when the authors compared resubmitted papers with those that made it into publications on their first try, they were in for a surprise: papers rejected on the first submission but published on the second were more highly cited within six years of publication than papers accepted immediately in the same journal. How does failure improve impact?

One possibility is that authors are good at assessing the potential impact of their research. Hence, manuscripts that are initially submitted to high-impact journals are intrinsically more 'fit' for citations, even if they are rejected. But this theory cannot fully explain what the researchers observed: After all, resubmissions were more highly cited regardless of whether the resubmissions moved to journals with higher or lower average impact. This suggests another possibility: Feedback from editors and reviewers, and the additional time spent revising the paper for resubmission, make for a better—and more citable—final product. So, if you've had a paper rejected, don't be frustrated with the resubmission process—what doesn't kill you may indeed make your work stronger.

4.3.2 A broader definition of impact

Scientists have had a love-hate relationship with any metric for assessing a paper [53]. Indeed, why would researchers ever rely on proxies, rather than engage directly with the paper? And yet, citations are frequently used to gauge the recognition of a scientist by their peers. As our quantitative understanding of science improves, there is an increasing need to broaden the number and range of performance indicators. Consider the game of Go, where in each step of the game, you ought to ask yourself: What's the right move? The answer there is unambiguous—the right move is the one that will most likely win the game. Yet science lacks a single "right move," but many possible paths are interwoven. As the science of science develops, we are likely to witness the exploration of a multitude of "right moves," a better understanding of how they interact with each other, and how they capture new dimensions of scientific production and reward. These new developments will likely fall within the following three categories.

The first encompasses variants on citations. While metrics will continue to rely on the citation relationships between papers, they will likely go beyond sheer citation counts, leveraging intricate structures within the citation network. The disruption index [54], discussed in Chapter 2.5, offers one example: instead of asking how many citations a paper has received, we can view each of the citations within the context of literature that's relevant to the paper. When papers that cite a given article *also* reference a substantial proportion of that article's references, then the article can be seen as consolidating the current scientific thinking. But, when the citations to an article ignore its intellectual forebears, it's an indication that the paper eclipses its prior literature, suggesting it disrupts its domain with new ideas.

The second category, alternative metrics, or altmetrics, complement the traditional citation-based impact measures. The development of web 2.0 has changed the way research is shared within and outside academia, allowing for new innovative constructs for measuring the broader impact of a scientific work. One of the first altmetrics used for this purpose was a paper's page views. As journals moved to the web, it is possible to count precisely how often a paper was looked at. Similarly, the discussion of a paper on various platforms can also gauge its potential impact. Scientists can calculate such metrics using data from social media, like Facebook and Twitter, blogs, and Wikipedia pages.

Researchers have calculated page views and tweets related to different papers [55, 56]. When comparing these measures with subsequent citations, there is usually a modest correlation at best. The lack of correlation with citations is both good and bad news for altmetrics. On the one hand, this lack of

correlation indicates that altmetrics complement citations counts, by approximating public perception and engagement in a way that citation-based measures do not. As funders often demand a measurable outcome on the broad impact of their spending, such complementary metric could be useful. It is particularly appealing that altmetrics can be calculated shortly after a paper's publication, offering more immediate feedback than accruing citations.

On the other hand, the lack of correlation with traditional citation metrics raises questions about the usefulness of altmetrics in evaluating and predicting scientific impact. Indeed, some altmetrics are prone to self-promotion, gaming, and other mechanisms used to boost a paper's short-term visibility. After all, likes and mentions can be bought, and what's popular online may not match the value system of science. Hence, it remains unclear if and how altmetrics will be incorporated into scientific decision-making. Nevertheless, altmetrics offer a step in the right direction, helping us to diversify the way we track the impact of science beyond science. Which brings us to the third category.

An equally promising direction may be quantifying and incorporating a broader definition of impact, especially the ripples of impact that reach beyond academic science. For example, researchers analyzed the output of NIH research grants, but instead of focusing on publications and citations, they studied patents related to drugs, devices, and other medical technologies in the private sector, linking public research investments to commercial applications [57]. They found that about 10% of NIH grants generated a patent directly, but 30% generate papers that are subsequently cited in patent applications. These results indicate that academic research has a much higher impact on commercial innovation and is a lot more important than what we might have thought. In a related example, researchers measured how patented inventions built on prior scientific inquiries by mapping a network of citations between papers and patents [58]. Their research found that an astonishing 80 percent of published papers could be connected to a future patent. These two examples suggest that scientific advances and marketplace inventions are pervasively and intimately connected. By broadening their definition of impact, these studies also demonstrate the value of scientific research reaches beyond academic science. Although the vast majority of research endeavors take place in an ivory tower, that work meaningfully impacts the development of patented inventions and generates productive private-sector applications.

To be sure, more research is needed to understand what each new metric does, and how to avoid misuse. Future science of science research seeking to extend the definition of the impact will prove to be

critical. For example, understanding the practical value of scientific research—particularly when it is paid for by federal funding agencies—is essential for determining how best to allocate money.

4.3.3 Causality

Does funding someone make that person more productive? We can answer this by collecting large-scale datasets pertaining to individual funding and linking that to subsequent productivity. Let's assume that the data shows a clear relationship between the amount of funding and the number of papers published over the next five years. Can we claim to have definitively answered our question?

Not quite. A variety of other factors both increase the likelihood of getting funding and make people appear more productive later on. For example, it could be that some fields are trendier than others, so people working in these fields are more likely to attract funding *and* also tend to publish more. Institution prestige may also play a role: It may be easier for researchers at Harvard to get funding than their colleagues at lower-tier institutions, but those Harvard professors also tend to be more productive.

We can, however, control for these factors by introducing fix effects in the regression table. The limitation of this approach is that we can only control for factors that are observable and there are many unobservable factors that are just as plausible. Take, for example, what's known as "the grit factor." [59] Grittier researchers may be more likely to attain funding and are also more productive. These types of confounding factors are often observational studies' worst enemies, because they question whether *any* observed correlation implies a causal relationship between the two variables.

Why do we care whether the relationship is causal or not? First, understanding causality is critical for deciding what to do with the insights we glean. For example, if funding indeed results in more publications, then we can increase the amount of research in a certain area by increasing our investment in it. If, however, the observed relationship between funding and publications is entirely driven by the grit factor, then increased funding in an underexplored discipline would have little effect. Instead, we'd have to identify and support the researchers in that discipline who have grit.

Luckily, understanding causality is something science of science researchers can learn from their colleagues in economics. Over the last three decades, micro-economists have developed a series of techniques that can provide fairly reliable answers to some empirical questions. These methods, collectively known as the credibility revolution [60], rely on two basic tricks.

The first trick is utilizing a randomized controlled trial (RCT), an approach originated in medical research. To test if a certain treatment works, people participating in the trial are randomly allocated to either (1) the group receiving the treatment or (2) a control group that receives either no treatment or a placebo. The key element here is that the intervention is assigned randomly, ensuring that the treatment and control group are indistinguishable. If there are any detectable differences between the two groups after the trial, they must be the outcome of the treatment alone, and not any other factors. Let's look at an example.

We learned in Part 3 that teams now dominate the production of high-impact science. But we know very little about how scientific collaborations form in the first place. There is empirical evidence that new collaborations tend to occur among people who are geographically close to each other. One hypothesis posed to explain this observation is that finding collaborators presents a search problem: It's costly to find the right ones, so we tend to work with people who are in our physical proximity. That theory makes sense, but there could be alternative explanations. Maybe the geographic closeness we see between collaborators is due to the friendships cultivated in a shared workspace. Or maybe it can be explained by the shared research interests that emerge in small departments. A critical test of the search problem hypothesis would be an RCT: If we reduce search costs for some pairs of potential collaborators by facilitating face-to-face interactions but not for others—creating a treatment group and a control group—will we see any differences in collaboration?

That's precisely what researchers did when they organized a symposium for an internal funding opportunity at the Harvard Medical School [61]. They enabled face-to-face interactions during 90-minute idea-sharing sessions for some pairs of scientists who'd been randomly selected from among the 400 attendees. The probability of collaboration increased by 75% for the treated scientists compared with the control group, who participated in the same idea-sharing but did not have a face-to-face session. Since the two groups were randomized, the increased rate of collaboration among the treatment pairs isn't due to other factors like shared interests or pre-existing friendships. Instead, we can attribute the increased likelihood of collaboration to the fact that the pairs in question met face-to-face.

RCTs remain the most reliable method in the world of causal inference. But experimenting on science also represents a significant challenge. Indeed, RCTs that can alter outcomes for individual researchers or institutions—which are mostly supported by tax dollars—are bound to attract criticism and push-back [62]. This is where quasi-experimental approaches will come in handy.

If we want to identify a causal relationship between factors, there's a second trick we can use: Find a randomly occurring event, use it as a cut-off, and examine what happens before and after. This is known as a "natural experiment" or a "quasi-experiment." The example on how early-career setbacks lead to future success, used this trick [50]. Whether junior PIs were just above or below the funding threshold for NIH grants is an exogenous variation to any individual characteristics of the PIs themselves. People with more grit may be more resilient against setbacks, or publish higher-impact papers, but they could not choose to be in either side of this arbitrary threshold. Since being above or below the threshold only affects your chance of funding, hence if that predicts your future career outcome, it must mean that there is a link between funding and career outcome, independent of any other observable or unobservable factors.

If we can identify causal relationships in the studies we conduct, we can be more certain of the results we arrive at. Partly for this reason, the credibility revolution has had a major impact on economics. Indeed, in 2017, around 50% of working papers published by National Bureau of Economics Research (NBER) have the word "identification" in them [63]. But, this greater certainty comes at a price. Answers whose causal relationship can be well identified tend to cover a narrower range of topics. Indeed, one of the most critical lessons in interpreting results from RCTs is that the causal relationship, when it is established, only applies to the randomized population that was studied. A drug treatment that healed 60-years-old patients in a New York hospital won't necessarily work on adolescents in China. Similarly, effects discovered for NIH PIs around the funding threshold don't necessarily generalize to other populations. This means that describing the conditions of the experiment is just as important as describing the results. As such, RCTs and other causal identification methods highlight an essential tension between certainty and generalizability: Small, limited questions can be answered with confidence, but bigger questions are subject to much more uncertainty.

Given the tension between certainty and generalizability, both experimental and observational insights are important in our understanding of how science works. In a call for more experiments on science itself, MIT economist Pierre Azoulay pointed out [62]: "We inherited the current institutions of science from the period just after the Second World War. It would be a fortuitous coincidence if the systems that served us so well in the twentieth century were equally adapted to twenty-first-century needs." Through quasi-experiments and carefully designed experiments, the science of science will hopefully yield causal insights that will have direct policy implications for this new era.

Yet, even within economics, there is a growing consensus that many big questions in society—like the impact of global warming—cannot be answered using super-reliable techniques [64]. But, we cannot afford to ignore these big questions. Therefore, the science of science of the future will benefit from a flourishing ecology of both observational and experimental studies. By engaging in tighter partnerships with experimentalists, science of science researchers will be able to better identify associations discovered from models and large-scale data that have causal insights. Doing so will enrich their relevance in our policy and decision making.

Bibliography

1. Harari, Y.N., *Sapiens: A brief history of humankind*. 2014: Random House.
2. Evans, J.A. and J.G. Foster, *Metaknowledge.* Science, 2011. **331**(6018): p. 721-725.
3. King, R.D., et al., *The automation of science.* Science, 2009. **324**(5923): p. 85-89.
4. Evans, J. and A. Rzhetsky, *Machine science.* Science, 2010. **329**(5990): p. 399-400.
5. Swanson, D.R., *Migraine and magnesium: eleven neglected connections.* Perspectives in biology and medicine, 1988. **31**(4): p. 526-557.
6. Foster, J.G., A. Rzhetsky, and J.A. Evans, *Tradition and innovation in scientists' research strategies.* American Sociological Review, 2015. **80**(5): p. 875-908.
7. Rzhetsky, A., et al., *Choosing experiments to accelerate collective discovery.* Proceedings of the National Academy of Sciences, 2015. **112**(47): p. 14569-14574.
8. Azoulay, P., et al., *Toward a more scientific science.* Science, 2018. **361**(6408): p. 1194-1197.
9. Ioannidis, J.P., *Why most published research findings are false.* PLoS medicine, 2005. **2**(8): p. e124.
10. Greenberg, S.A., *How citation distortions create unfounded authority: analysis of a citation network.* Bmj, 2009. **339**: p. b2680.
11. Gerber, A.S. and N. Malhotra, *Publication bias in empirical sociological research: Do arbitrary significance levels distort published results?* Sociological Methods & Research, 2008. **37**(1): p. 3-30.
12. Benjamin, D.J., et al., *Redefine statistical significance.* Nature Human Behaviour, 2018. **2**(1): p. 6.
13. Efthimiou, O. and S.T. Allison, *Heroism science: Frameworks for an emerging field.* Journal of Humanistic Psychology, 2018. **58**(5): p. 556-570.
14. Nosek, B.A., et al., *The preregistration revolution.* Proceedings of the National Academy of Sciences, 2018. **115**(11): p. 2600-2606.
15. Kuhn, T.S., *The essential tension: Selected studies in scientific tradition and change*. 1977: University of Chicago Press.
16. Bourdieu, P., *The specificity of the scientific field and the social conditions of the progress of reasons.* Social Science Information, 1975. **14**(6): p. 19-47.
17. Yao, L., et al., *Health ROI as a measure of misalignment of biomedical needs and resources.* Nature biotechnology, 2015. **33**(8): p. 807.
18. Willett, W., *Nutritional epidemiology*. Vol. 40. 2012: Oxford university press.
19. Spector, J.M., R.S. Harrison, and M.C. Fishman, *Fundamental science behind today's important medicines.* Science translational medicine, 2018. **10**(438): p. eaaq1787.
20. Senior, A., J. Jumper, and D. Hassabis, *Deep Mind, AlphaFold: Using AI for scientific discovery*.
21. Harari, Y.N., *Reboot for the AI revolution.* Nature News, 2017. **550**(7676): p. 324.
22. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.
23. Farabet, C., et al., *Learning hierarchical features for scene labeling.* IEEE transactions on pattern analysis and machine intelligence, 2012. **35**(8): p. 1915-1929.
24. Tompson, J.J., et al. *Joint training of a convolutional network and a graphical model for human pose estimation*. in *Advances in neural information processing systems*. 2014.
25. Szegedy, C., et al. *Going deeper with convolutions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
26. Mikolov, T., et al. *Strategies for training large scale neural network language models*. in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. 2011. IEEE.
27. Hinton, G., et al., *Deep neural networks for acoustic modeling in speech recognition.* IEEE Signal processing magazine, 2012. **29**.
28. Sainath, T.N., et al. *Deep convolutional neural networks for LVCSR*. in *2013 IEEE international conference on acoustics, speech and signal processing*. 2013. IEEE.
29. Bordes, A., S. Chopra, and J. Weston, *Question answering with subgraph embeddings.* arXiv preprint arXiv:1406.3676, 2014.

30. Jean, S., et al., *On using very large target vocabulary for neural machine translation*. arXiv preprint arXiv:1412.2007, 2014.
31. Sutskever, I., O. Vinyals, and Q.V. Le. *Sequence to sequence learning with neural networks*. in *Advances in neural information processing systems*. 2014.
32. Ma, J., et al., *Deep neural nets as a method for quantitative structure–activity relationships.* Journal of chemical information and modeling, 2015. **55**(2): p. 263-274.
33. Ciodaro, T., et al. *Online particle detection with neural networks based on topological calorimetry information*. in *Journal of physics: conference series*. 2012. IOP Publishing.
34. Kaggle. *Higgs boson machine learning challenge*. 2014; Available from: https://www.kaggle.com/c/higgs-boson.
35. Helmstaedter, M., et al., *Connectomic reconstruction of the inner plexiform layer in the mouse retina.* Nature, 2013. **500**(7461): p. 168.
36. Leung, M.K., et al., *Deep learning of the tissue-regulated splicing code.* Bioinformatics, 2014. **30**(12): p. i121-i129.
37. Xiong, H.Y., et al., *The human splicing code reveals new insights into the genetic determinants of disease.* Science, 2015. **347**(6218): p. 1254806.
38. Silver, D., et al., *A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play.* Science, 2018. **362**(6419): p. 1140-1144.
39. De Fauw, J., et al., *Clinically applicable deep learning for diagnosis and referral in retinal disease.* Nature medicine, 2018. **24**(9): p. 1342.
40. Esteva, A., et al., *Dermatologist-level classification of skin cancer with deep neural networks.* Nature, 2017. **542**(7639): p. 115.
41. Titano, J.J., et al., *Automated deep-neural-network surveillance of cranial images for acute neurologic events.* Nat Med, 2018. **24**(9): p. 1337-1341.
42. Open Science Collaboration, *Estimating the reproducibility of psychological science.* Science, 2015. **349**(6251): p. aac4716.
43. Nosek, B.A. and T.M. Errington, *Reproducibility in cancer biology: Making sense of replications.* Elife, 2017. **6**: p. e23383.
44. Camerer, C.F., et al., *Evaluating replicability of laboratory experiments in economics.* Science, 2016. **351**(6280): p. 1433-1436.
45. Camerer, C.F., et al., *Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015.* Nature Human Behaviour, 2018. **2**(9): p. 637.
46. Chang, A. and P. Li, *Is economics research replicable? Sixty published papers from thirteen journals say'usually not'.* Available at SSRN 2669564, 2015.
47. Wu, Y., Y. Yang, and B. Uzzi, *An Artificial and Human Intelligence Approach to the Replication Problem in Science*. 2019.
48. Tegmark, M., *Life 3.0: Being human in the age of artificial intelligence*. 2017: Knopf.
49. Reuters. *Amazon scraps secret AI recruiting tool that showed bias against women*. 2018; Available from: https://ca.reuters.com/article/technologyNews/idCAKCN1MK08G-OCATC.
50. Wang, Y., B.F. Jones, and D. Wang, *Early-career setback and future career impact.* arXiv preprint arXiv:1903.06958, 2019.
51. Bol, T., M. de Vaan, and A. van de Rijt, *The Matthew effect in science funding.* Proceedings of the National Academy of Sciences, 2018. **115**(19): p. 4887-4890.
52. Calcagno, V., et al., *Flows of research manuscripts among scientific journals reveal hidden submission patterns.* Science, 2012. **338**(6110): p. 1065-1069.
53. Azoulay, P., *Small-team science is beautiful.* Nature, 2019. **566**(7744): p. 330-332.
54. Funk, R.J. and J. Owen-Smith, *A Dynamic Network Measure of Technological Change.* Management Science, 2017. **63**(3): p. 791-817.
55. Haustein, S., et al., *Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature.* Journal of the Association for Information Science and Technology, 2014. **65**(4): p. 656-669.
56. Perneger, T.V., *Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ.* Bmj, 2004. **329**(7465): p. 546-547.
57. Li, D., P. Azoulay, and B.N. Sampat, *The applied value of public investments in biomedical research.* Science, 2017. **356**(6333): p. 78-81.
58. Ahmadpoor, M. and B.F. Jones, *The dual frontier: Patented inventions and prior scientific advance.* Science, 2017. **357**(6351): p. 583-587.

59. Duckworth, A. and A. Duckworth, *Grit: The power of passion and perseverance*. Vol. 234. 2016: Scribner New York, NY.
60. Angrist, J.D. and J.-S. Pischke, *The credibility revolution in empirical economics: How better research design is taking the con out of econometrics.* Journal of economic perspectives, 2010. **24**(2): p. 3-30.
61. Boudreau, K.J., et al., *A field experiment on search costs and the formation of scientific collaborations.* Review of Economics and Statistics, 2017. **99**(4): p. 565-576.
62. Azoulay, P., *Research efficiency: Turn the scientific method on ourselves.* Nature, 2012. **484**(7392): p. 31-32.
63. Kleven, H.J., *Language trends in public economics.* Princeton, NJ, 2018.
64. Ruhm, C.J., *Deaths of despair or drug problems?* 2018, National Bureau of Economic Research.