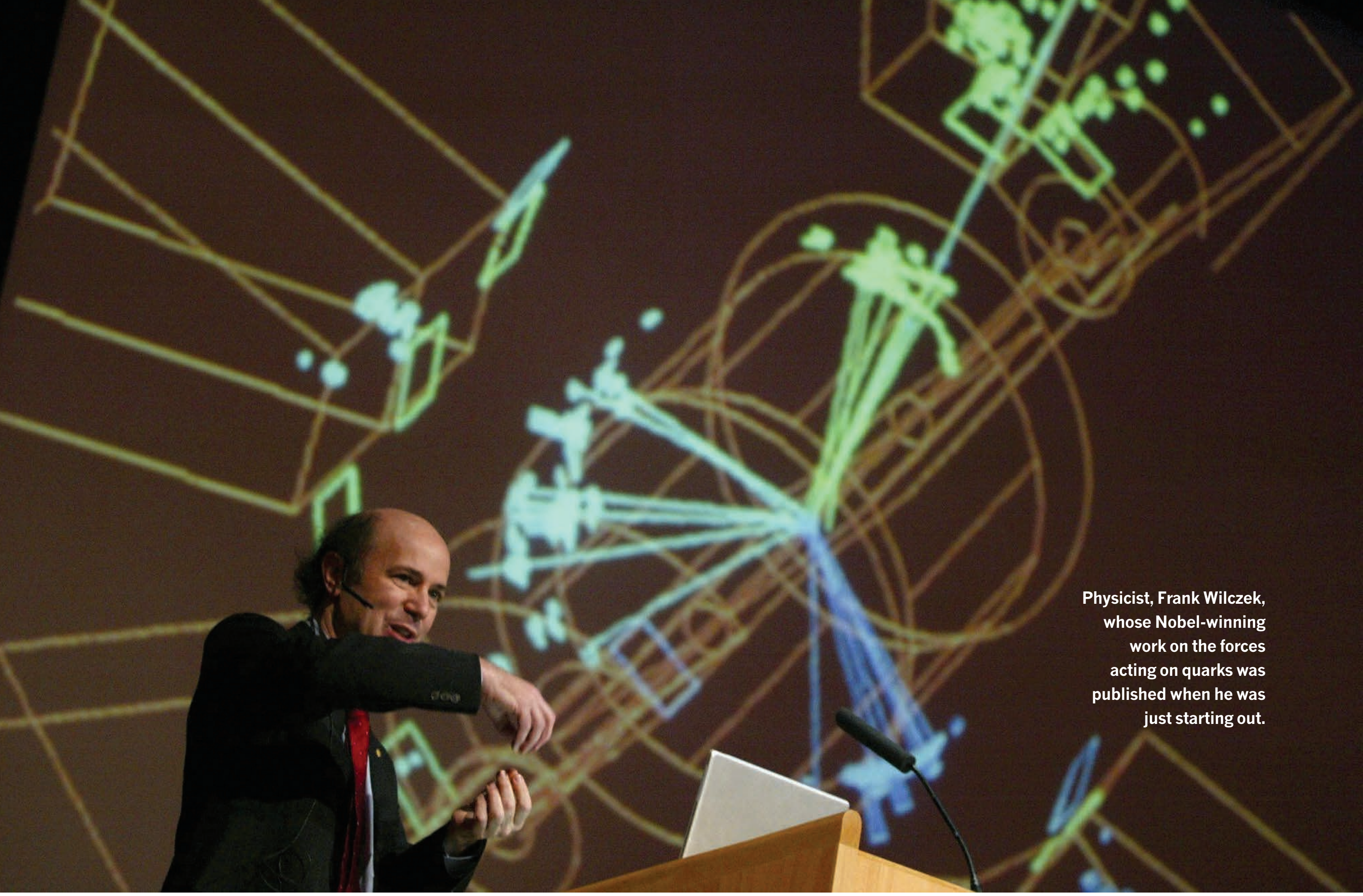

Physicist, Frank Wilczek, whose Nobel-winning work on the forces acting on quarks was published when he was just starting out.

BERTIL ERICSON/AFP/GETTY IMAGES

PREDICTING SCIENTIFIC SUCCESS

Even sophisticated, data-driven models of academic careers have trouble forecasting the highs and lows.

SMRITI MALLAPATY

When Frank Wilczek was a graduate student in his early twenties, he published work on the forces holding quarks together that later won him a Nobel Prize.

At the other end of a career span, John Fenn, a retired analytic chemist in his seventies, developed the award-winning technique for analysing large proteins using mass spectrometry.

From early starters to late bloomers, the timing of a researcher's career high is largely dependent on chance. This was the conclusion of a 2016 study, in which researchers developed a mathematical model to describe publication and citation trends based on the records of thousands of people.

Every piece of work is just as likely to be your highest impact paper as the last, says study co-author Dashun Wang at Northwestern University's Kellogg School of Management in Evanston, Illinois. "To be a successful scientist, you should just keep drawing the lottery and hope for the best."

Sophisticated new models are using vast data sets to help elucidate the process of scientific discovery, and how it will evolve — including at the level of individual careers. As the volume of this information expands, the resulting algorithms and their predictions will improve.

But, in searching for predictable patterns, and a formula for detecting rising research stars, scientists are finding that success is inherently unpredictable, says Daniel Larremore, a computer scientist at the University of Colorado Boulder.

These models are also beginning to reveal the flaws in the research system and point to ways of correcting them. "Through reverse engineering, we can help create a fairer system that nurtures talented people, no matter their ethnicity, gender or location," says Roberta Sinatra, a network and data scientist at the Central European University in Hungary, and first author of the 2016 study.

BETTING ON THE BEST

Researchers have had limited success in finding quantitative and objective ways of predicting a scientist's future performance based on their past merits.

Earlier efforts typically involved statistical checks of single or collected metrics to see how well they correlate with reality. In 2007, for example, Jorge Hirsch, a physicist at the University of California, San Diego, published a paper on the predictive power of a popular measure he had invented for determining the scientific impact of an individual — the *h*-index. Hirsch observed a correlation between a researcher's current and future *h*-index.

Several years later, a group led by computer scientist, Daniel Acuna, now at Syracuse University, developed a formula to estimate an individual's future *h*-index based on several variables, including number of articles, publication in prestigious journals and years since first paper. It accounted for 66% of the variability in the *h*-index of some 3,000 neuroscientists five years later. But some scientists argued that the cumulative nature of the *h*-index overstated its predictability.

Now, mathematicians, network scientists, and physicists are bringing new tools to the challenge. They are creating simple models of the rules of human behaviour, in the same way that the Standard Model explains the existence of the Higgs Boson.

These models exploit rich and accessible long-term data generated about scientists and their scholarly endeavours — from publications and citations, to funding sources, collaborators, mobility, institutional affiliation, ethnicity and gender. But a formula for spotting rising research stars is still elusive. In detecting career trends, the models are also revealing predictive limits.

CHANCE DISCOVERY

Those who study the trajectories of scientific careers had long assumed that researchers were at their most creative early in their careers. Sinatra and Wang's 2016 study proved otherwise. They found that a constant and unique value known as Q, derived from an individual's long-term citation and publication record, could determine the number of citations that their best paper would achieve, but the timing of that paper was anybody's guess. The higher a researcher's Q factor, the higher the impact of their paper.

In a recent study covering a shorter publication window, Wang and Sinatra showed that a career high is typically characterized by a slew of several highly cited papers. "All of an individual's best works tend to happen within that hot streak," says Wang. And while most scientists will experience such a creative burst, it will probably only happen once in their career.

A 2017 study by Larremore also deconstructed the fast-early-peak, slow-slump pattern of productivity. In an analysis of more than 2,000 computer scientists and 200,000 publications, he found that while the researchers' collective publication trajectory followed the rise–fall pattern, it could only explain the productivity of one in every five scientists.

Paper citations don't always follow a reliable pattern either, which makes it difficult to predict career trajectories based on them. Some papers lie dormant for many years before gaining citation traction. A 2015 citation analysis of 22 million articles spanning more than a century found that there are many examples of such 'sleeping beauties'. Among them is a 1955 paper by Eugene Garfield on the utility of a citation index, which caught the research community's attention some half a century later.

First x-ray photograph of a human, in 1895.

While emerging algorithms can potentially anticipate incremental advances in science, such as the observation of gravitational waves, it is beyond their capacity to predict the accidental isolation of penicillin, or the serendipitous discovery of x-rays, as it is beyond the scope of most humans.

"Any kind of model that makes strong bets on the trends of the past is likely to perpetuate the kinds of problems that we have now, without leaving us open to the weird and unexpected innovations that no-one sees coming," says Larremore.

Models of scientists' careers don't need to be good predictors to be useful, says Vincent Traag, a computational social scientist at the Centre for Science and Technology Studies, Leiden University. By allowing researchers to uncover the mechanisms underlying the phenomena they observe — how science itself works — "we can start thinking of how to address questions such as the replicability crisis, publication biases, and inappropriate incentives," says Traag.

Gaps in the publication records of individuals expose the many lost opportunities — from those who have abandoned academia out of a sense of failure, or to raise children, or for unexplained reasons.

"The big piece of the puzzle that is missing is a quantitative understanding of failure," says Wang, who is analysing grant application data from the US National Institutes of Health to capture signals not just of acceptance, but also rejection. "It happens all the time, yet we know so little about it."

When it comes to tracking talent, some traits have little to do with merit. Studies of the *h*-index, for example, have found that women are cited less than men.

"If we put this into an approach that predicts impact, then it would favour men, rather than women," says Sinatra, who is working on developing data-driven measures and models to identify the source and contribution of forms of bias so they can be corrected, and not perpetuated in predictive modelling.

"So much of the past 'success' has been correlated with looking and sounding, well, like me — white, male, native English speaking, past affiliation with Harvard," says Larremore. "There is a danger of reading too much into the patterns of the past." ■

SCIENCE PHOTO LIBRARY; LEWIS HOUGHTON/SCIENCE PHOTO LIBRARY

Algorithms can point to incremental advances, but breakthroughs such as the accidental isolation of penicillin are impossible to predict.