For FEBS Letters

Jim's View: "Playing Billiards with Science"

A friend once told me that it is good to keep an open mind, so long as it is not so open your brains fall out. I sometimes wonder if we are not in danger of this today in biology and in medical research more generally. We are simply awash in data. It seems there are now almost as many kinds of "omics" as there are genes in an animal. Sequences of DNA. Patterns of RNA. Proteomes. Interactomes. Patterns of methylation and the like. The lists go on, the accumulating result of impressive advances in scalable analytical biochemistry and computing power. As these approaches rapidly become standardized and progressively lower in cost, it becomes very seductive for many scientists to collect ever more data with no hypothesis in mind, justified by the apparent virtue of being "unbiased."

As scientists we have always sought to understand the rules that govern the world around us. We have always proceeded by first observing, chronicling the nature of things. We have always developed our biases from such observations, developing them into specific hypotheses which we then test. Traditionally we measure one or very few entities (organisms, molecules, etc.) at a time, intentionally altering only one variable at a time in a carefully controlled, closed system.

What is entirely new is that observations now can involve very large numbers of entities measured simultaneously in different conditions, which are themselves often not controlled (for example, comparing transcriptional patterns among two types of cells or in healthy versus diseased tissue samples). We then use computational approaches to extract statistically significant relationships among the entities in the data sets from which hypotheses can then be formulated.

Wikipedia defines this "discovery science" (also known as discovery-based science) as a "scientific methodology analysis of large volumes of experimental data with the goal of finding new patterns or correlations

leading to hypothesis formation. A variety of methodologies (algorithms) can be used to extract patterns in data science, including artificial intelligence/machine learning." These methods constitute the closely related field of "data science", which is defined as "the interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining."

Certainly, discovery science/data science-based approaches provide a valid means of generating hypotheses, which can then be tested by conventional means. Well formulated and insightful questions based on intuition guided by deep insight will always be the key whatever the methodology. But it is also valid to ask when or even whether the new data science approaches are likely to be as efficient, deep, or (forgive me) beautiful as traditional causal inference.

Billiards

As an example, how would data science go about predicting the patterns of motion of billiard balls? Data tracking the positions and motion of the balls in hundreds, perhaps thousands of games would be entered into a computer. One or another data science method would classify patterns of motion of the balls based on this training set. From these patterns, the motions of individual balls would then be predicted with a sufficient degree of accuracy. Problem solved, right?

At one level yes, and quite efficient indeed. But at a deeper level, would we actually have gained any real understanding of the physics involved? Somehow this "discovery science" approach completely misses discovering Newton's laws of motion. Wouldn't it have been far simpler to track the motion of a single ball and how it changes when it collides with a second ball; and how this varies with the size, speed and direction of the second ball, and so on? Then, with a few direct and elegant experiments in a controlled closed system we would be able to deduce the laws of motion, and they could be forcefully generalized to any number of balls on the table to predict their motions with absolute certainty. Certainty is a key point. Data science methods can at best provide predictions that are statistically

but not absolutely accurate because they result purely from correlations rather than causal inference.

Cause and effect - the unique power of One

How do we go about establishing causality? We are at an intrinsic disadvantage because our brain itself is at its core a correlating machine. If two events occur together frequently enough, we learn to expect one when the other occurs, and naturally take this as evidence that one causes the other. But this is no proof; in fact, distinguishing coincidence from causation is the essence of proof in biochemistry, genetics, and all hard science. To establish causation, we must change only one variable (such as the concentration of an enzyme in a biochemical reconstitution or a single gene in an organism) at a time and compare the result with a control which is otherwise identical. Any resulting change in outcome (rate of reaction, phenotype) must then be caused by the single change that was made. Critically, if even two such variables were changed in a controlled manner simultaneously, we could no longer attribute the change in outcome to one or the other variable, making causal proof impossible with only the correlation remaining.

This is the "power of one." It accounts for every scientific fact in every textbook, and even in this issue of FEBS Letters we see its continuing power in article after article.

Even AI can’t reason Why

Surely, the kind of sophisticated artificial intelligence/machine learning algorithms (for which Silicon Valley is justifiably famous and feared in equal measure) will come to the rescue. But even the world’s leaders in this field have been unable to reverse engineer and figure out how their own algorithms work. As the recipient of the 2011 Turing Award for his work on probabilistic and causal reasoning Judea Pearl (and Dana Mackenzie) have written, “to reach the higher rung of causal inference, in place of ever-more data, machines need a model of the underlying cause and effect”, which is as if to say they need to be provided with the answer in advance.

No doubt someday this will change, but for now it remains a good thing that Isaac Newton watched one apple fall at a time.

NOTE: The section on cause and effect was adapted from my forward to volume 79 of Annual Review of Biochemistry. The quote from Pearl and Mackenzie is from the Wall Street Journal May 19, 2018.

May 27, 2018