

Turing Award Winner, Longtime ASA Member Talks Causal Inference

1 November 2012 12,355 views 9 Comments



Pearl

Longtime ASA member Judea Pearl is the winner of the 2012 Turing Award, the most prestigious award in computer science. Using some of the proceeds from the award, Pearl is setting up a contest to help advance the teaching of causal inference in introductory statistics courses. ASA Executive Director Ron Wasserstein recently interviewed Pearl about the Turing Award and Pearl's interest in promoting instruction of causal inference.

***RW:* Congratulations on receiving the Turing Award! Please tell us a little bit about the work the award committee cited when selecting you and where you hope that work will go in the future.**

JP: The award committee cited my works in probabilistic and causal reasoning that “enabled remarkable advances in the partnership between humans and machines.” The applications they highlighted include “medical diagnosis, homeland security, and genetic counseling to natural language understanding and mapping gene expression data.”

Needless to state, I was not personally involved in these many applications, but they stem indirectly from my earlier work on Bayes networks, a graphical model that was developed in the 1980s by computer scientists and statisticians (including P. Dawid, S. Lauritzen, D. Spiegelhalter, G. Shafer, D. Cox, and N. Wermuth). This development has enabled machines to represent probabilistic information parsimoniously and meaningfully over many variables and to draw its consequences in light of new evidence.

My contribution was to define the mathematical relationships between graphs and probabilities (with A. Paz) and to devise algorithms that compute posterior probabilities swiftly and distributedly. Storage space and update time were the two major hurdles in the 1980s for representing and processing uncertain information by computers.

An interesting anecdote from that period may be of interest here. I think I was the only one then who insisted all along on computing posteriors in a distributed fashion, that is, by passing messages among simple and repetitive computational units, each assigned to a different variable. Why? Because this was the only biologically feasible way we could explain how the human brain deals with uncertainty, say when we read English text or cross a street. My colleagues at that time could not have cared less how humans do things, as long as the algorithm was efficient and correct. But as things turned out, it was this biologically inspired algorithm, called “belief propagation,” that scaled up (as an approximation) in practical applications and eventually enabled computers to

process problems with thousands of variables. The moral of the story is that we should not underestimate what we can learn from fallible humans, even in statistics, where human judgment is often synonymous with bias.

Looking back, everything we learned from Bayes networks turned applicable when we made the transition to causal reasoning, and it did not take long to define a new object, called “causal Bayes network” or “causal diagram,” which encodes not merely conditional independencies, but also effects of outside interventions, like those present in controlled experiments. From here, the road laid opened to “do-calculus”—a symbolic logic for deriving all interventional and observational implications that a given model entails. This led to a complete solution of the infamous “identification problem,” namely, to decide which causal effects are estimable from nonexperimental data and how, given the theoretical assumptions that a researcher encodes in the diagram.

I should also mention the emergence of counterfactual calculus from a semi-deterministic version of Bayes network, with the help of which we were able to unify the graphical and potential-outcome approach of Neyman, Rubin, and Robins, thus forming a symbiotic methodology that benefits from the merits of both paradigms.

One triumph of the symbiotic analysis has been the emergence of lively research activity in nonparametric mediation problems, namely, to estimate the extent to which an effect of treatment on outcome is mediated by a set of variables standing between the two. The importance of this analysis lies, of course, in unveiling the mechanisms—or pathways—of the data-generating process, thus telling us “how nature works.” These sorts of questions were asked decades ago by Fisher and Cochran, but, lacking the tools of graphs and counterfactuals, they could not be addressed until quite recently. I am surprised, therefore, that teachers of statistics (as well as econometrics and other data-intensive empirical sciences) are not rushing to introduce these new tools in their classrooms.

As to the future, I see untapped opportunities in aggregating data from a huge number of different sources, say patient data from hospitals, and coming up with coherent answers to queries about yet unseen environment or subpopulation. I call this task “meta-synthesis,” after realizing that current methods of meta-analysis do little more than average apples and oranges to estimate properties of bananas. What we need is a principled methodology for analyzing differences and commonalities among studies, experimental as well as observational, and pooling relevant information together so as to synthesize a combined estimator for a given research question in a given target subpopulation.

We have begun to look into this challenge through the theory of “transportability” (with E Bareinboim, 2011), and found, not surprisingly, that the inferential machinery of the do-calculus is indispensable when it comes to transferring empirical findings across populations. I have great hopes for this line of research.

***RW:* Do you consider yourself a statistician or a computer scientist, or is that a distinction you even make when you think of yourself and your work?**

JP: There is a lot in common to statisticians and computer scientists, especially those working on machine learning and inference under uncertainty. Both are attempting to make sense of data, and both are going about it in a systematic way. The distinction comes in two dimensions: first, what it means to “make sense of data” and, second, what language we use in our mathematics. For the great majority of statisticians, “making sense” means estimating useful properties of the joint distribution that underlies the observations. For a computer scientist, “making sense” usually means gaining a deeper understanding of the mechanism that generates the data, and such understanding cannot be achieved, even when we have a complete and precise specification of the joint distribution function. Thus, if we take this dividing line seriously, I was a statistician before 1988 and turned computer scientist hereafter, when I made the transition from probabilistic to causal inference.

But the other dividing line is perhaps more fundamental. Computer scientists are extremely sensitive to notation, and extremely careful about making all assumptions explicit in their chosen notation. The reason is simply that a robot is a fairly stupid reasoner, and, if we do not spell out explicitly all assumptions, outrageous conclusions are likely to be derived. Statisticians, on the other hand, are ingenious reasoners and can perfectly manage the analysis while keeping causal assumptions in their heads, without making them explicit in the mathematics. For example, when Fisher invented the randomized experiment, he did not find it necessary to deploy special

notation to distinguish experimental from observational findings. As a result, one would be hard pressed even today to find a mathematical proof that randomized experiments unveil the desired causal effect or, as a more challenging example, a proof that one should not control for a covariate that lies on the pathway between treatment and effect, nor for any proxy of such covariate.

Computer scientists would not have allowed eight decades to pass without developing the mathematical notation and logical machinery needed for producing such proofs. In this sense, I am a computer scientist. And it is not the mathematical proofs per se that I aim to facilitate, but the practical questions that a careful notation can answer, if supported by the appropriate logic. For example, what covariates we should control for or what the testable implications are of a given set of causal assumptions.

In summary, I am a statistician in my aims to interpret data and a computer scientist in the formal tools that I employ toward these aims.

***RW:* Would you please describe for the uninitiated what causal inference is and how your 1988 transition from probabilistic to causal inference came to pass?**

JP: Causal inference is a methodology for answering causal research questions from a combination of data and theoretical assumptions about how the data are generated. Typical causal questions are the following: What is the expected effect of a given treatment (e.g., drug) on a given outcome (e.g., recovery)? Can data prove an employer guilty of hiring discrimination? Would a given patient be alive if he had not taken the drug, knowing that he, in fact, did take the drug and died?

The distinct feature of these sorts of questions is that they cannot be answered from (nonexperimental) frequency data alone, regardless of how many samples are taken; nor can they be expressed in the standard language of statistics, for they cannot be defined in terms of joint densities of observed variables. (Skeptics are invited to write down a mathematical expression for the sentence, “The rooster crow does not cause the sun to rise.”)

These peculiar properties of causal questions, which have rendered them taboo in standard statistics textbooks, are emerging as exciting intellectual challenges in modern causal inference. One such challenge is the necessity of creating a new mathematical language for expressing both the research questions of interest and the theoretical assumptions upon which the answers must depend. The assumptions must be encoded in a new notation, called “causal model,” that is friendly enough to permit scrutiny by researchers and, at the same time, precise enough to permit mathematical derivation of the model’s implications. For example, the model should tell us whether the assumptions have testable implications, what those implications are, whether the queries of interest are estimable from the available data, and, if so, how.

Unknown to the uninitiated, most of these goals have been achieved recently through a friendly and embarrassingly simple symbiosis between graphical and algebraic methods. My transition from probabilistic to causal inference was triggered by the realization that people carry most of their world knowledge through causal, not probabilistic, relationships and that judgments about dependencies and independencies, so critical for the construction of graphical models, emanate from causal considerations—probabilities are just surface decorations. This is shown most vividly in Simpson’s paradox. Why else would an innocent sign reversal of associations be deemed so paradoxical by most people? The only answer we have is that people are incurably predisposed to prefer causal over associational interpretations, and Simpson’s paradox appears paradoxical only because causal interpretations of the data rule out (both intuitively and mathematically) the possibility of sign reversal.

Another factor that prompted my transition was the realization that, unless one is a pollster or weather forecaster, investigators’ main interest lies in causal, not associational, questions. True, cultural taboos and the lack of mathematical notation have inhibited investigators from asking those questions explicitly, forcing them to pose and settle for associational substitutes. The advent of causal calculi now provides a fairly transparent understanding of the assumptions that must be made and the measurements that must be undertaken to answer

such questions, and this enables statisticians to address directly the problems that their customers/users have kept dormant for decades.

***RW:* So, you are setting up a prize to encourage the teaching of basic causal inference in introductory statistics courses. Please tell us about the prize and long-run outcome you hope to stimulate through it.**

JP: We discussed earlier the ongoing excitement in the causal inference community, and all the new problems that can be solved today using modern methods. I would like this excitement to percolate down to the education level. Just think about it. This year, there were 73 papers on causal inference listed in the JSM program. By comparison, there were only 13 such papers in JSM 2002's program.

Though it is a small sample, one cannot deny that these numbers indicate a transformative shift of focus in statistics research. Yet, this shift is hardly seen among statistics educators, and is totally absent from statistics textbooks or even from the pages of *Amstat News*. I have watched this research-education gap widening from the day causality began to come out of her closet, and I can now put my finger on its main causes—statistics instructors are reluctant to teach a topic that tradition (and authority) has branded informal, undefined, anecdotal, and controversial, if not metaphysical. And I do not blame them; I, too, would think twice before standing in front of a class to answer questions which, only a generation ago, caused terrible embarrassment to my own professors when they had to evade them. But times are changing; causality has taken on a new face, both rigorous and friendly, for every statistician to enjoy.

To narrow this gap in statistics education, I have donated part of the Turing Prize money to the American Statistical Association to establish an [annual prize](#) for a person or team that contributes substantially toward introducing causal inference in education. I hope this will stimulate the generation of effective course material, perhaps a video or a 100-page booklet that would convince every statistics instructor that causation is easy (It is!) and he/she too can teach it for fun and profit. The fun comes from showing students how simple mathematical tools can answer questions that Pearson-Fisher-Neyman could not begin to address (e.g., Simpson's paradox, covariate selection, mediation), and the profit comes from the fact we mentioned earlier—that most customers of statistics ask causal, not associational, questions.

Once we get statistics education on the side of history, a renaissance will follow. Causal analysis will become an organic territory of statistics proper, and researchers in statistics-based disciplines will be spared the psychological barriers that have been hampering our generation.