Why Does Software Cost So Much?

featuring Mike Konrad and Robert Stoddard as Interviewed by Suzanne Miller

---

Suzanne Miller: Welcome to the SEI Podcast Series, a production of Carnegie Mellon University Software Engineering Institute. The SEI is a federally funded research and development center operated by Carnegie Mellon University and sponsored by the US Department of Defense. Today's podcast will be available at the SEI website at www.sei.cmu.edu/podcasts.

Today, I am here to welcome my colleagues Robert Stoddard, Mike Konrad. I am Suzanne Miller, principal researcher here. I have worked with these gentlemen for many years. Actually, Mike was my boss at some point. So, I am very excited to talk to them about the work that they have been doing recently in a subject called causal learning, which they have been applying to a topic that we all care about, which is *Why Does Software Cost So Much*?

Before we get started into that discussion, give the rest of the audience that don't know you as well as I do a little bit about why you are here at the SEI, what brought you here, and what is it that brought you to this particular research? Robert, let's start with you.

Robert Stoddard: I have been at the SEI about 13 years after 24 years in industry. I came to the SEI because I really wanted to become a scholar practitioner, work on my Ph.D., which is a lifelong pursuit, and still make use of my industry background. So, I have been working in statistics and probabilistic modeling, and now machine learning and causal learning, and have a great time working with Mike.

Suzanne: Excellent, Mike?

Mike Konrad: I have been at the SEI for about 30 years. *Why did I come to the SEI?* That is long-ago history.

Suzanne: What brought you to this research that you are doing now?

Mike: I have been fascinated by architecture and requirements and how they meet in safety- and -security critical systems and wondering about how to do that better, how to apply what we think we know about design and quality in some kind of systematic way to better achieve good quality results with software. Causal learning seems to be a way to make sense of the soup of the many factors that affect that.

Suzanne: That takes us into the subject of causal learning and how it is different from other kinds of learning methods that apply in other settings, like machine learning, statistically based learning methods. Causal learning actually has more of an intervention kind of aspect to it than some of these other methods, which I think distinguishes it from things pretty clearly. We are not just looking at, *What did we learn from this data?* but [also], *How does this data guide us to make a different policy decision, make a different tactical decision?* To me, that is the fascinating thing about it. Tell us more about what it is and how it is different from these other methods, methodologically and also in terms of its use.

Robert: Just to clarify, the intervention part is what we seek to educate. We want to educate through causal learning what interventions one could take to change outcomes. But the methodology actually can be applied to both experimental data and observational data where we really don't have the ability to run a controlled experiment. In most of our topics, that's the case. We do want to empower the decision-maker to know when to take action and when not to take action.

Suzanne: Some of this is because of the multi-factor nature of these environments that Mike mentioned. When we can't do a controlled experiment, we are left with lots of different factors affecting, and you are trying to help clarify what are the factors that we may want to pay attention to.

Robert: Right, so the causal learning is going to go beyond the correlation and regression and actually let us get a chance to see which factors are really causal in nature versus just correlated. We have a notion of correlation that can be misinterpreted, and it's not really [of a] causal nature.

Suzanne: We have a statement in the statistics world that says correlation does not imply cause [causation], right? We all learned that in our first statistics class. What you are trying to do is sort of help us understand *When does correlation have more meaning and when can we take it farther?* So how do you do that?

Robert: So as our coach--Professor [David] Danks over at CMU's Dietrich College of Humanities and Social Sciences--he is the head of the Department of Philosophy and has all the experts in the causal learning. He basically espouses that correlation is a noisy indicator of

causation. So it is still helpful. We are not saying to disregard correlation, but the causation is where we want to go, and it is related [to] machine learning. I'll let Mike…

Mike: So, a simple example. If you were to look only at simple correlations, it was found [you would find] that in New York City serious crime would go up when ice cream sales would go up. [There might be] a causal explanation for that. [Or] one might instead recognize that there are other reasons why correlation might be there. Maybe it has something to do with summer or heat or warmth outside. That is where you bring criminal and maybe victim in closer proximity, and people buy ice cream.

It turns out that there are statistical tests that you do for correlation that inform you, also, about whether there is in fact a causal relationship between variables. There are those [statistical tests] related to what is called conditional correlation and those related to, say, information theory. Those are ways of actually ascertaining something, which we would call *causality*.

Suzanne: What drove you to start focusing on causality as a topic? Because we have been trying to deal with learning of varying natures, but this is fairly recent in the last five years.

Mike: Can I take that first? OK, let me take that one first. Several years ago we came out with guidance to help organizations on their high-maturity path, but a lot of that guidance was actually based on correlation. What organizations were trying to do—and we were encouraging organizations—was to use results from their inspections or peer reviews and tests to guide when to take some kind of action because maybe quality wasn't good enough or maybe quality was fine.

What we initially were prescribing they do was look for correlational information. But then you have the hidden summer variable thing that I was mentioning. You have other factors that might explain why things are correlated. There was no guarantee that by taking certain interventions that you would make things better. You only were playing into the known correlations and not the issue of summer, so that analogy is clear. It was a very strong motivation to try to understand exactly what was happening, so that we could help organizations make sensible changes, interventions in their project to help guide them towards success.

Robert: Like Mike said, our journey started with predictive modeling for high-maturity organizations. As we studied that, we were jumping into a lot of probabilistic modeling and Bayesian modeling. Then we became big fans of Dr. Judea Pearl. Then in 2015 he [Pearl] came on site to CMU when he was awarded the Dickson Science Prize for CMU. That is when we learned that after a lifelong work in Bayesian probabilistic modeling, he had come to the enlightenment that causal learning was really the end that he should have been pursuing. It goes

beyond just prediction but actually having a model that you know you can [use to] take action and change outcomes.

Suzanne: So all of this research, this is what we do, take all of this learning from different sources and apply it to a real problem. The real problem that you are working on right now is the title of this podcast, why does software cost so much? What is it about causal learning that makes you think that it will help you in answering that question better than we have been able to answer it for the last 30 years? Because, we do not have good answers for that.

Mike: Well, very simply, it is easy to find correlations when it comes to software. It's just that we don't know which ones are actually having this causal relationship. Yes, we have been building software for several decades and we find lots of correlations and therefore it's almost as Bob said earlier, when quoting professor David Danks, correlation is a noisy indicator of causation.

So what do you change? What do you do differently? Do you send people to more training? Is now a good time to re-plan? Do you throw in better staff? Do some mentoring? Do you buy a new tool? Those are the kinds of decisions that need to be made, and right now people are just guided by their experience and correlation, and neither is that great of a guide. We are trying to get a more established, more systematic approach to understanding what changes to make to improve project success.

Robert: I think, additionally, this is enriching our research here at the SEI and, we hope, across more research at the SEI. I firmly believe that this methodology lets us do a better job of filtering out what are truly causal factors versus non-causal factors. My belief is, in most areas of research, up to like 80 percent of the factors that people think are related because the [because of] correlation are not causal. We really think that we are going to help research move forward by focusing on the ones that we believe are causal in nature.

Suzanne: In the past, if you wanted to do root cause analysis (right?), which we have done qualitatively, you ask the five whys and you do things like that. What is it that you do differently? How do you use the rich data sets that are available that give us different kinds of information to go beyond that qualitative sense of, *This is probably the cause*, to something that is more statistically based, more research-based?

Mike: Going back to that example of serious crime, ice cream sales, and summer. If you looked at the two variables of serious crimes and ice cream sales, you [would] find them correlated, but if you conditioned on what season it is, you would tend to find them uncorrelated. That is a signal that summer has some kind of role between ice cream sales on the one hand and serious crime on the other. It is that kind of generalized pattern that we look for, or the algorithms we

use that were developed by individuals like Peter Spirtes, Clark Glymour, also with the Department of Philosophy on campus. These were invented concurrently with Judea Pearl's work that Bob was speaking of earlier, that helped put causal learning on a more systematic setting. Now there are several dozen algorithms. Through the Center for Causal Discovery, which is a joint effort of Carnegie Mellon and University of Pittsburgh and several other universities through NIH funding, these dozens of algorithms provide a basis for any researcher (the research done at the SEI included) that will help—Bob and I contend—provide a more robust and systematic approach towards observational-type data, which is often what we have in software engineering, and identify the causal patterns that seem so important for designing methodologies and analysis algorithms and so forth.

Suzanne: So, another way to explain this is, if I have it right, is that you look at conditions like both of these, ice cream sales and crime increases that occur in summer. So *if we take summer out of the equation, are these then correlated?* What you are doing is a systematic process of elimination of, *OK, here is a correlation. Here is another factor that they are both correlated to. Is this something we can take out?* What you are left with are probable causal factors, although you have to do more analysis to actually establish that, correct?

Robert: Right. The dozens of algorithms that Mike mentioned in that tool, called Tetrad, the algorithms use different approaches to identify that causation. One other simple example of how some of the algorithms work is, *If you have a factor X causal on Y, supposedly, if you run regression, you will give a set of residuals. Then, if you have the two factors change places and run a regression again, you will get another set of residuals. If you compare the two sets residuals, you can conclude which direction the causality is.* There are a number of techniques within these algorithms that help them define the causal links.

Suzanne: Let's pull it back to cost and software. What are you finding when you start looking at that particular problem and all the data that we have? How does causal learning help us get some insights that we may not have had before?

Mike: This is a three-year journey, actually four years, because Bob led the initial exploratory research effort. Now we have a further three years. We are looking at datasets that we have here at the SEI and those with our various collaborators and colleagues. So, for instance, we have looked at the Personal Software Process data and found some reasonable support for a hypothesis that Dr. Bill Nichols (a colleague of Bob and myself) and I have had that there are human traits that manifest themselves in programming in terms of a type of problem-solving characteristics that we might think of as productivity and then also defect proneness. [Those] traits can be modified through learning, through training, but they are also very predictive of what happens.

Behind our current research, we have a kind of theory about how these come together for individuals to create code, choreographed by a plan or a project team in the roles and responsibilities assigned. That is what our research is trying to elicit through looking at, for instance, the Personal Software Process data set. Another data set with another research collaborator, Dr. Sarah Sheard, was looking at the role that complexity has, complexity in the environment or context, including stakeholder relationships, complexity in the architectural solution, complexity in terms of using emerging technology in the solution, and what that does relative to cost. We have been looking systematically at several data sets, and Bob has as well. We have some initial, again early stages, but some initial results that are kind of intriguing and I'll talk about some.

Robert: Yes, so one of the things we are finding on this project is that we need data to do this. There is not like a lot of great data just floating around to use. We have several approaches to deal with that, and that has been spending up a good part of our team's time as we got started to beat the bushes for collaborators that have data and all that. We recognize that we are often going to access a data [set] that is a little bit disjointed, but we know from our coaching from the CMU faculty there are ways to stitch these causal structures together towards a holistic model.

A recent data set I am studying is from a client where they had a survey taken by 30 staff members each week for a year and a half. It had a lot of the factors that I had data-mined out of Watt Humphrey's book on *Leading TSP Teams* that were basically unmeasured but discussed in his book. We have studied those. The interesting thing is we are looking at outcomes [such as] cost, schedule, and quality by these team members and their perception of how it's going on the project for that week and analyzing that data from a correlational standpoint. Anywhere from, like, 16 to 18 of the factors are related, correlated, highly with cost, schedule, quality. When we do a causal structure analysis, one of them was causal on cost, one on quality, and none of them were causal on schedule. That is a pretty dramatic early example of what we are getting out of this.

Suzanne: That is very exciting because I know project managers that I have worked within various client settings are always looking for, *What are the levers?* They talk about, *What are the levers? What are the leverage points?* If we can help them eliminate the ones that are not causal then that allows them to focus. There are a whole lot of implications for that if we have structures that we can trust. That comes out to the next question, *How do you build trust in these causal structures?* There are algorithms that sort of get us to a point where we have some idea, but to socialize these into something someone is going to use, have you guys been thinking about that yet, or is that kind of more in the future for things you need to think about?

Mike: We are doing several things to build confidence in the results that we are getting, and it is an ongoing process. We are looking at a variety of different data sets, often where we know the

provenance and have some confidence in the provenance. We are looking, as Bob mentioned earlier, at overlapping data sets that collectively inform us. No one data set might be in itself complete at capturing causes potential candidate causes of interest. We are using different algorithms that work on very different principles. So, I mentioned information theory earlier as an example, and the other one on the conditional independences within the data set, and those two are relatively different approaches towards teasing out causality. So there's that.

Frankly, we are also building on the history of use of causal techniques and other domains where they have had success. We are basically arguing there is nothing so fundamentally different about this data. All data is different, but there are some similarities in terms of information theory and independences that we are all capitalizing on. Whether we are talking about medical discovery or discovery of software engineering factors that matter in terms of cost, schedule, and quality. Incidentally, on a data set that Sarah Sheard and I have been looking at, we found this concept called cognitive fog, which is where you have a lot of internally inconsistent data and cognitive overload as a factor that predicts what Sarah calls *performance gap*.

This is a kind of quality [that] measures whether you meet the technical requirements on the project. Cognitive fog seems to affect that [technical] performance, which also relates to project success. So we have those two outcomes, project success and performance gap, that are predicted by cognitive fog, which occurs, maybe, about midway in a project. In turn, cognitive fog seems to have, as its cause, number of decision makers. So here we could see as we refine…

Suzanne: So the old adage, *too many cooks spoil the pot* may actually have a research base.

Mike: Yes, yes exactly. So these are early. The datasets are small. We are not yet ready to base any particular policy on this or encourage that. We are finding things which do meet certain familiar sayings or heuristics or truisms. I will just say that.

Robert: Your question about confidence really highlights a second reason we are looking for collaborators. One is to have access to data. The second one is—and this is back to the CMU coaching—once we have a causal result, the number one way you can go increase confidence is if we can get a collaborator in the real world in the field to actually take action based on what our causal structure is saying and see a difference in a result. We are working with collaborators, and we are presenting at the Acquisition Research Symposium to get collaborators who actually would agree to do the second part.

Suzanne: Right, Mike?

Mike: Our collaborator Anandi Hira at the Barry Boehm Center for Systems and Software Engineering [Center for Systems and Software Engineering at the University of Southern

California (Los Angeles)] is where we will both be…actually she will be presenting some of our joint research that she conducted.

Robert: Actually the data she studied, we have taught them how to do causal learning. They have some interesting results on projects, real projects but done by students. They have the ability, in a quicker timeframe, to have the students take [make] change on [to] causal factors and give us the results.

Suzanne: That is nice, so now you've got some short timeline. Many of the projects, certainly in the DOD, are of the much longer time frame. The delayed effect of change is one of the other things that you have to deal with.

So this is pretty early, your first year of a three-year project, but many of our listeners and our viewers get very interested in learning more about the subject that we are talking about. So, have you thought about, sort of, what are some of the ways that people can learn about causal learning in a more general way? Do you have any plans for, kind of, transitioning this out into a larger, either, research community or just in the larger DoD community?

Robert: I am going to let Mike talk about some of the actual technical training that is available now, but I would lead off by saying that I would like the audience to start realizing correlation is not the stopping point. When you hear about research results on NPR in the morning, they always stop at correlation. I am like, *go the extra step. Tell me, what's the causal results* so I can really believe it. Otherwise, I am not really going to change my diet.

I would just like folks to become aware that the causation is important and don't just think correlation is always the answer. The other thing I would mention is we recently posted an SEI Blog post on this topic. At the end of that blog, I think we have like about 15 links to books and tutorials and things like that. That is like one-stop shopping. It is the kind of blog I wished I would have had three and a half years ago when I started my journey studying causal learning.

Mike: I mentioned earlier the Center for Causal Discovery [CCD]. You can find their website either searching within a Google search window or within YouTube. [Search on] "CCD" and maybe add the word "causal" so you don't discover some form of music you didn't realize existed. And so, search for "CCD" and the word "causal" and you will find that there are videotaped sessions of Professor Richard Scheines presenting 13 videos or lectures on causal learning. So, you can learn causal learning. Bob has already mentioned the TETRAD tool, which has those algorithms implemented through a friendly GUI interface. The Center for Causal Discovery also provides an API for those wanting to write programmatic solutions they can do so.

Suzanne: OK. So there already is material for learning this, and what people should look forward to is your reports in the future about how we can apply this in the software engineering community to answer some of our questions.

Mike: There already are some publications out there that just came out through Anandi Hira and others that we will be doing in the next few weeks.

Robert: I would say we have done an extensive [literature] search to see who else has been using causal learning. In the United States right now, it's most heavily used in cancer research. A small pocket of economists are using it, and then a moderate-sized group of psychologists, sociologists are using it for government social policy. From what I can tell, in the engineering and the engineering research side of the house, I think we are precedent setting on that. I really hope that the SEI will be at the top of the FFRDCs for using such a rigorous methodology in research.

Suzanne: Well, as you said, we have access to some software data that is not prevalent within the industry as a whole. So, we have a little bit of a leg up on that, but I am glad to see us using it.

Mike: So, Bob mentioned different kinds of algorithms earlier: for discrete data versus continuous, but there are also [algorithms] for time series data. So, an application there of causal learning for the past few years has been in applying causal learning to FMRI [functional magnetic resonance imaging] data. In particular, brain imagery to see which areas of the brain get activated.

Suzanne: Functional MRI?

Mike: Yes, functional magnetic resonance imaging.

Suzanne: Good, OK. So, lots of different avenues for this. Once you've figured out how to use causal learning to answer some of these questions, what do you think is the next thing that you're going to want to do with this? What's your dream?

Robert: On my own time it will be the stock market.

Suzanne: There's an application of causal learning. You better not be the next one that makes the mortgage industry crash. You're not allowed to do that.

Robert: After this project, we have a restricted research project going on with the Air Force on applying machine and causal learning and understanding the benefits to use them together. I see a number of existing and planned research, both in CERT cybersecurity and in the Software

Solutions Division that would definitely benefit from using this. We have started working a little bit in the technical debt area with Rick Kazman and Ipek [Ipek Ozkaya and Rod Nord].

Suzanne: I was thinking about CERT because I was thinking about all the many ways the security problem is too much data, lots of correlations and not enough understanding of where the causal [exists], which things we can eliminate. So, that could be very fruitful area.

Robert: I think insider threat would definitely be a good area.

Suzanne: Oh yes, that would be one. I want to thank both of you for joining us today. I get to see you in the halls but I don't get as much time to talk about your research as I would like to. So, this is really precious.

Mike: Likewise, it is precious for me and I'm sure Bob as well.

Robert: It's nice to see people other than at the airport.

Suzanne: Yes, don't go there. I do want to let our audience know that the blog post that you mentioned is out there if you search in the SEI blog at insights.sei.cmu.edu and search for S-T-O-D-D-A-R-D, then you should be able to find his blog post. I am sure that you will enjoy it as much as I did. We will provide links to the resources mentioned.

There are a few things we talked about here that we will add into the transcript that may not be in the blog post so that our audience has access to all of that. I will remind everyone that our podcast is available from three sources. You can get it at the SEI website at sei.cmu.edu/podcasts, you can get it at the Carnegie Mellon University's iTunes site, and you can also get it on the SEI's very own YouTube channel. So, you are going to be on YouTube. You can tell your kids.

As always if you have any questions please don't hesitate to contact us at info@sei.cmu.edu. Thank you for viewing.