SOCIETY OF ACTUARIES

PREDICTIVE ANALYTICS AND FUTURISM SECTION

Predictive Analytics and Futurism

ISSUE 20 • MAY 2019

View From the Top—Interviews With Industry Leaders From Huge Reinsurance Companies

By Xiaojie (Jane) Wang and Dave Snell

Page 6

Predictive Analytics and Futurism

Issue 20 • May 2019

Published three times a year by the Predictive Analytics and Futurism Section of the Society of Actuaries.

475 N. Martingale Road, Suite 600
Schaumburg, Ill 60173-2226
Phone: 847.706.3500 Fax: 847.706.3599
www.soa.org

This newsletter is free to section members. Current issues are available on the SOA website (*www.soa.org*).

To join the section, SOA members and non-members can locate a membership form on the Predictive Analytics and Futurism Section webpage at *http://www.soa.org/predictive-analytics-and-futurism/*.

Publication Schedule
Publication Month: August
Articles Due: Contact editor for copy submission date

The digital edition of this newsletter can be found on the section landing page at *https://www.soa.org/sections/pred-analytics-futurism*

From the Editor: Predictive Analytics and Futurism (PAF)—It's Bigger on the Inside!

By Dave Snell

Many PAF Section members are fans of the science fiction series "Dr. Who," the time traveler who uses technology and creativity to save planets, galaxies and sometimes our entire universe. The doctor undergoes many shape changes over the years (convenient when the lead actor is replaced). In fact, the current Dr. Who character is female. Another main character in the series is the TARDIS (Time And Relative Dimension In Space). The TARDIS is the doctor's home, companion and time machine in one. The TARDIS evokes surprise from newcomers as they enter what looks like a British police box (about the size of old stand-alone phone booths) and find inside that it houses infinite rooms and vast spaces. The typical exclamation is that it is "bigger on the inside." The doctor is used to this reaction. One of my favorite episodes is the one where Clara, another new acquaintance, walks in, walks out, walks around it, goes back in and then exclaims, "It's smaller on the outside!" Some fans feel that she was the first to notice that the world is smaller than the TARDIS and not that the TARDIS is simply "bigger on the inside."

> Unencumbered by the baggage of human knowledge, the new AI can learn faster and more effectively!

Predictive analytics and futurism are two topics that have grown in importance over the past several years. One presenter at the recent Predictive Analytics Symposium stated that 90 percent of the data available today was created in the last two years. The amounts and types of data for analysis are increasing at an accelerating rate: YouTube claims that over 1 billion hours of video is being viewed on its site every day. Wearable and embeddable devices like Apple watches and diabetes monitors are streaming data faster than ever previously imagined. The processing power to analyze these data is, likewise, growing at phenomenal rates. Consider that the average smartphone has over a million times more memory and similar speed improvements than the onboard computer (4K of RAM and laughable speed) that put the first human on the moon! The advances in artificial intelligence (AI), which PAF also embraces, have been similarly impressive. The watershed moment, in 1997, when Deep Blue beat World Grandmaster Gary Kasparov in chess, has been eclipsed by the likes of AlphaZero, which has become the world champion in chess, Go, and Shogi; and the training of AlphaZero consisted of merely inputting the rules of the games, as opposed to the tedious inputting of thousands of the best games that humans have played. Unencumbered by the baggage of human knowledge, the new AI can learn faster and more effectively!

Autonomous vehicles are no longer science fiction. They are proving themselves safer than human drivers. The newer concerns regarding them center around the ethical issues. A survey by Massachusetts Institute of Technology researchers surfaced the disturbing conflict that "people thought an autonomous vehicle should protect the greater number of people, even if that meant sacrificing its passengers. But they also said they wouldn't buy an autonomous car programmed to act this way. What does this tell us?"[1] Perhaps actuaries can be involved in the formulation of legislation to ensure ethical vehicle actions.

I am honored to be a judge of the Actuarial Speculative Fiction contest, and the entries this year (results should be available by our next issue) included stories where actuaries, as experts on human life and health with the ability to analyze and understand vast amounts of data analytics, become the people who reverse global warming and save millions of lives. This is a phenomenal time to be an actuary who understands and uses predictive

analytics and has the vision and open-mindedness to embrace futurism. This section is truly "bigger on the inside."

Starting next year, the PAF newsletter, like the TARDIS, will become bigger on the inside. We are transitioning to a web-based format, and this year's issues will appear as both paper (as you may be reading now) and an online version that will contain essentially the same content but in a more reader-friendly format. Beginning next year, the online version will not have the same constraints on article size, code availability and other features that section members have expressed concerns about; and it will be the expanded, more feature-rich counterpart of our printed newsletter. This also opens the door for more timely articles, less likely to be locked into print production schedules. Peruse the print version in your hand and then see the counterpart articles on our website. We welcome your feedback and ideas.

- Starting us off in the **Chairperson's Corner**, Eileen Burns tells about exciting new dimensions to PAF resources, including the digital newsletter, a new Delphi study in progress, and the new platform for Jupyter (named for Julia, Python, and R) notebooks that PAF members will be able to download and use directly. She also describes some hot topics reports and ideas and welcomes the 200-plus new members (that still fit easily into our expanding section).

- Next, we have a **"View from the Top."** Here is an excerpt from a fascinating interview that Xiaojie (Jane) Wang and I conducted with two noted industry leaders from very large companies and two leaders from start-ups. Normally, we would have to drastically cut this down to fit into the PAF newsletter; but here we can give you a few highlights in the printed version and a link to the actual keynote session interview, complete with synchronized slides and audio online.

- Michael Niemerg follows this with **All About Them Curves: Ordered Lorenz Curves and Lift Curves**. Michael explains Lorenz curves and Gini gain, which he describes as "great alternatives to life curves." Lorenz curves do not involve binning, and they are far less susceptible to misleading results from small perturbations. He suggests adding them to complement your lift curves to form a better overall measure of risk stratification.

- In **The Growing World of "Off-Balance Sheet" Talent and Workforce Modeling,** Megan Gauer, Nathan Pohle and Priyanka Srivastava team up to explain the benefits and challenges associated with the changing mix of workers across time zones, priorities and employment status. They point out that 40 percent of U.S. employees fall under alternative staffing arrangements, and the global trend is similarly moving toward freelancers and the self-employed.

- Sarah Abigail follows this with a summary of **Artificial Intelligence and the Actuarial Profession**, an SOA Annual Meeting session that PAF and the General Insurance sections jointly sponsored. The burgeoning interest in AI is evidenced by her statement, "The SOA provided the largest room available at the convention hall, however it was filled to capacity, with people standing in the back, and more attendees unable to get in." Read it and get on board this AI tsunami about to hit the insurance industry.

- Another trio of authors—Zhen Hui Trinh, Ben Johnson and Eileen Burns—write about **Interactive Plotting**. They describe the dangers of too-busy plots as "about as helpful to your reader as donning a track jacket when confronted with Snowmageddon." They also describe the increased benefits and opportunities present with online graphics (Did I mention enough times that we are transitioning to online?) and some tools such as Plotly and Billboarder. They also provide a handy table with links to several interactive plotting tools.

- Wrapping up the print issue, Jeff Heaton gives us an excellent introduction to Cython, a way to address the speed issues of Python runtime performance. In **Getting C++ Performance From Python With Cython**, he covers more than just speed issues. He also describes Cython as a better way to protect your intellectual property of source code. And, of course, Jeff provides links to his online code examples so you can test out the language with code that already works!

Just like AI and predictive analytics and futurism, the PAF section has become smaller on the outside than the inside. The link to our online version is right here: *http://bit.ly/2UTNbfY*. Come on inside! Experience a new dimension and perhaps some time travel to the future of our profession. ■

Dave Snell, FALU, FLMI, ASA, MAAA, CLU, ChFC, ARA, ACS, MCP, teaches AI Machine Learning at Maryville University in St. Louis. He can be reached at *dave@ActuariesAndTechnology.com.*

ENDNOTES

1 The MIT study is summarized in the online interview written by Morgan Meaker, a British journalist, who spoke with Jean-François Bonnefon, Azim Shariff and Iyad Rahwan, MIT researchers who designed an online quiz called The Moral Machine. *https://medium.com/s/story/how-should-self-driving-cars-choose-who-not-to-kill-442f2a5a1b59*

Chairperson's Corner

By Eileen S. Burns

Welcome to our first digital (and print) newsletter! During our January council call, Beth Bernardi gave us a demonstration of what we could expect on the digital platform. We're all excited for the launch and hope that it enables broader reach of our section's content.

Our 2019 calendar is filling up fast. We hope to announce a couple of webcasts soon; recruitment for meeting sessions is underway; and our two current research initiatives are getting closer to completion: the Delphi study began recruiting participants in December, and the research on validation of predictive models has gotten a first draft to the Project Oversight Group. This latter research is scheduled to be published in the early spring and will be presented at the Life and Annuity Symposium.

We are moving forward with our two member-driven initiatives as well. We've selected GitHub as the platform for the Jupyter Notebook contest, and we will be looking into what education on GitHub we can provide as a resource to the contest participants. Our Hack-a-thon planning team picked the data set that will serve as the basis for the event and will continue coordinating with the planning committee for the Predictive Analytics Symposium.

> Last year was the fourth year in a row that we added more than 200 members to our roster.

As part of an effort to raise engagement on our council calls, which are open to both council members and friends of the council, Stuart Klugman, FSA, CERA, SOA sr. staff fellow, Education, joined our January call. Going forward, we'll invite guests to discuss hot topics and ideas for collaboration with other sections. Stuart helped start this effort by sharing details of Society of Actuaries (SOA) educational initiatives in predictive analytics for us to get a better picture of what is being offered, how our members can volunteer to help, and where our section can best add value outside of those offerings. Our next guest on that topic will be Courtney Nashan, SOA director, Employer Relations and Opportunites, who will discuss the SOA's broader initiatives in predictive analytics. Once we have a complete picture of what's out there, my hope is that we'll be able to use our platforms to share that information with our members via a newsletter article, a spot on our web page, or posts on our LinkedIn page.

Finally, I'd like to welcome all of our new members! Last year was the fourth year in a row that we added more than 200 members to our roster. We hope that all our members enjoy and value what we plan to bring you this year! ■

Eileen S. Burns, FSA, MAAA, is a principal and consulting actuary with Milliman. She can be contacted at *eileen.burns@milliman.com.*

View From the Top—Interviews With Industry Leaders From Huge Reinsurance Companies

By Xiaojie (Jane) Wang and Dave Snell

The 2018 Predictive Analytics Symposium was another successful Society of Actuaries (SOA) and Predictive Analytics and Futurism (PAF) event. Feedback from the attendees told us that one of the highlights was the keynote panel presentation, with JJ Carol, senior vice president of SwissRe America Holding Corporation and head of the SwissRe New Solution Group, and Greig Woodring, retired president and CEO of RGA Reinsurance Group of America, Inc. and current chairman of Wamberg Genomic Advisors. JJ and Greig answered a series of questions that we posed to them and then answered several more from the audience. The session was so successful that we decided to transcribe a portion here for you.

These insurance visionaries' subset of advice is interesting and valuable. You may listen to the entire interview (that is, our questions, plus many from the audience) at *http://bit.ly/2Pv3bPE.*

Xiaojie: *Albert Einstein said that a person who never made mistakes never tried anything new. We all know that success is not built on successes but built on failures. So, can both of you, JJ and Greig, share the most important lessons you have learned from implementation of predictive analytic strategy and projects in your organization?*

Greig: I would say the most important thing that I have learned over time is that it is easy to get excited about the sexy new technology and the whizbang application that you are doing, but you need to bring the culture of the organization along with it to some extent. In some cases, that cultural change in adoption of a new technology, such as predictive analytics, is as important or more important to the success of the program as the actual application use that's developed, and that's the hardest thing to do sometimes. Paying attention to that at the outset, paying attention to getting everybody on board, and pulling in the same direction is the most important lesson that I could convey about how to get these things done.

JJ: Greig, I am glad you said that because I absolutely agree that it's critically important. Xiaojie, you mentioned that success is built off of failure. I have a little bit of a pet peeve about the word failure because to me it connotates in some way that you then quit somehow or you stopped doing something, but in most cases I think it's really about learning and pivoting and iterating, over time, which is, I think, what is intended by those statements. I wanted to share just a little story about one of the early predictive models that we started working on and some of the lessons we learned along the way. This project was started by one of the business leaders in our company, which is a great place for any project to start. I think it's important to have a business problem and then figure out how the predictive analytics can help solve that problem. The problem is the gentleman who started the project left the company and so there was no longer a champion for the project, and it stalled out. So, the first lesson is that it is really critical to have a business champion for your project. The second thing that we learned along the way is that we were trying to decide if this is something that we build ourselves, or do we work with a vendor to help create this, and I know many of you are thinking about this along the way as well. The challenges that we ran into included having little experience working with vendors in this particular case … maybe it was just the team that was working with those vendors. We didn't learn the lessons from others and we didn't really think carefully as we were embarking in that journey, about the agreement that we wrote with the vendor. We should have been thinking about things like how do you protect your intellectual property, what is your exit strategy … some of those important things ... making sure that we are thinking of those ahead. Kind of tied to that, there was a gentleman yesterday who talked about it being OK from a regulatory standpoint go into a research and development sandbox and learn about things, which is exactly what we did; but then the mistake we made was that we also didn't talk to our legal team at the same time, and so once the data scientists and actuaries and underwriters had what they felt was a really cool, good idea and a model that worked, it then went to the legal team and they were excited to bring this out to market but couldn't because we were then delayed by having to go through the regulator; so my third lesson is bring the legal team on early so you can have that interaction along the way.

Dave: *Why are we calling predictive analytics a disrupter? How is it really different from technology advances in the past?*

JJ: Well, maybe I'll start with another story. To me, it goes back to your early comments about the human elements of insurance. I was at dinner with the head of a very prominent distributor who told me a story that this is actually the reason why his business was having trouble. It's about the ease of use and the change in customer expectations that really make this different than any other technology historically, and I think that it reminds us that

we really need to think about these problems from an outside-in perspective and think customer first.

Greig: I'd say a lot of these things blend into one another. You talk about predictive analytics and artificial intelligence almost in the same breath at times. To me, those are things in and of themselves as much as they are techniques that are going to permeate almost everything we do; so, I think the disruption is because it is broader than other technologies that have been narrower in the past. You can use predictive analytics in ways that have never been done before. You can use AI in ways that have never been done before. They will profoundly change the ways that we look at not only our own business but the future of our business. That's maybe a little bit of the reason and the rationale of why it has been considered a bigger disruptor than prior technologies have been.

Xiaojie: *I think you both talked about big companies and small start-ups. I think a few months ago, I saw the statistic that 50 percent of the 10 biggest companies in the world have been replaced by newer companies over the last 10 years. Fifty percent of them are new companies. As our industry is very, very old, we should be very careful about our strategy regarding predictive analytics. How do you optimize predictive analytics strategy?*

Greig: I think we have touched on a number of things already. Failure was mentioned at the beginning. A lot of times failures can lead to success. Being willing to pick up the threads of something that looks like a failure and find out what worked in it is often the key to the success of the next iteration of this process. Sometimes you need to give up, and sometimes you are better off not giving up but trying to find another way to do it. Having an entrepreneurial mindset is something that is a blessing and a curse. You can keep trying things for too long; or you can make something when everyone else has given up on it. It is not always easy to see a year or two in advance which of those is going to be the pathway you are on.

JJ: Maybe I'll talk a little bit about the role of the actuary and take a specific example. As actuaries, we are responsible for managing risk, and one of the tools that we use in that is monitoring. If you wanted to optimize a predictive analytics tool that you're working with, let us take the example of predictive underwriting and the role of monitoring. In that case, what you're trying to do here is that you're really trying to optimize two things. You're trying to optimize how many people and how quickly they get through the process of underwriting, and at the same time you're balancing that with trying to optimize the mortality impact. So, you have these two levers that

you're looking at as you are monitoring this business. Just to get really concrete, what are you to do? You are looking at the data. You're looking for things like misrepresentation. You're looking for straight-through processing rates and what those look like. And then, it's not enough to just monitor but as your optimizing that you're looking for ways to change and adapt. How do you find new data sources that can plug holes if you have significant misrepresentation coming through compared to what you expected? And what's the new data source or a new model that you can deploy there as you're looking for efficiencies and straight-through processing? How are you analyzing the questions that you're asking and why you're asking those questions? How are you creating efficiencies in the process? And so I think it really does come down to the basics and taking a deeper look at what you're doing and what your core objectives are.

Greig: Yes, I think it's very important to understand exactly what you're trying to do. I think one of the biggest problems that life insurers have had is when they focused on speed and forgotten that there's fraud, there's deception, and there's antiselection in the marketplace that can really bite you pretty badly if you're only focused on speed.

Dave: *Both of you talked a lot about processing vast amounts of data; but processing vast amounts of data has a cost to it as well. There are human resources, computer resources, hardware and software and storage, and all of that. Are you basically saying that, in the future, the size of the company is what is going to determine success?*

JJ: I should think—Greig, you said it earlier—that some of the smaller companies are some of the more innovative ones. It's more about culture than anything else.

Greig: Yes, JJ, I would agree with that. On the other hand, there does come a point where you have so much data that you become a powerhouse. It's impressive the amount of data an Amazon or a Google possesses, and it's hard to see competing with them for consumer preferences, if you were head-up doing something that they were interested to do as well. So, you have to play both sides of this. You have to be quick on your feet. At the same time, you have to be able to use the power when you have it of big data sources.

JJ: Just to build on that, it is a lot of times about the partnerships that you built. You don't have to do it all yourself.

Xiaojie: *Greig, you mentioned about the different types of data. We all know that data is increasing at exponential speed, and nowadays you can note there are a lot of nontraditional data sources such as credit data, social media data, wearable devices data and the genetic data you mentioned before are showing great promise for our industry. What are some of the new types of data we have been encountering, and what can we expect in the near future?*

Greig: One of the good innovators at RGA is a person named Tim Rozar. He is fond of saying that whoever has the data has the power; and I think there is a lot of truth to that. You should be, as organizations, greedy for data. In fact, you probably have a lot of data that you don't even use. What's really interesting is not only the new sources of data but new ways to look at old data that you have. I have spent some time talking to CEOs of organizations, and one of the things they have been saying for the last 10 years is that in the life insurance industry we tend to cash checks and pay death claims and have very little contact with our policyholders or know very little about them. Then, if you ask them, "What are you doing about that," they really don't have an answer. They waved their hands a little bit, but with a few exceptions, very few companies know anything about their inforce; and that's not hard to fix. If you really wanted to fix it. There are a couple organizations that have done so. They tend to be somewhat unique organizations, but they have done so, and the amount of data and the types of data they get from their policyholders are eye-opening. If companies were aware of what they could gather just from their inforce data, I think they would be surprised. Then, you have to figure out what to do with it. It's one thing to know something about your inforce customers; but then you have to figure out can you cross-sell to them, can you try to improve customer persistency through special offers, or discounts, or who knows what? There are a lot of possibilities, but until you have the data to use, you don't really understand them.

JJ: Greig, maybe I could build on one potentially specific example within looking at your inforce and engagement with your inforce that has to do with this concept of being a risk adviser to our customers. Xiaojie, you were asking about new kinds of data. I think it was Chris Steno who said that 90 percent of the data that exists in the world today has been created in the last two years. I am thinking of one very specific massive increase in data. If you think about diabetics who have been testing their blood twice a day for their blood sugar. Those are two discrete points in time during the day that give them a reading of what their blood sugar looks like; but you have absolutely no insight as a diabetic as to what your blood sugar looks like in those interim time periods. In the advances of technology that are beginning to collect more of a continuous stream of data, those individuals can identify a bigger picture—a more accurate, continuous picture of what their risks look like—and connect that to their behaviors and their actions in a more informed way. And so is there a way as an industry that we can support our own clients, our customers, in understanding their own health and taking action against that to live a longer, healthier life, which is, of course, in our interest as well.

Dave: *Many of the conferences that I go to are not of this type, and they are stressing things like incremental process improvement. Yet, in many respects, AI and predictive analytics are not an incremental*

type of process improvement. It's a quantum leap in many cases. What advice would you want to share to keep an insurance company from becoming obsolete as a result of these huge changes from AI and predictive analytics?

Greig: Well, to keep from becoming obsolete, you need people who are willing to embrace that change. A lot of insurance companies at the top ranks, like most of industry at the top ranks, are heavily populated by baby boomers, who are short-termers at this point and who are not really thrilled about changing in any transformational way. As an organization, though, you need to find ways to do that; and you need to find ways to bring in new ideas—younger people with a longer horizon and longer thought process, because if you don't, you will become one of the casualties of the 50 percent of companies that fall off every 10 years, and that's what you need to do. We talked a lot about different things that would help you do that in terms of doing experiments and keeping your ideas fresh by keeping your feet moving as opposed to standing still, but it's up to companies and the kind of culture that they develop in order to make that happen.

JJ: Ultimately, it's your customers who decide whether you become obsolete or not, so I guess my advice is think customer first. Think outside-in, and don't let your strategy be driven by your technology but that your strategy drives your technology. Sometimes it's staying away from the shiny object and it's focusing on what's your purpose? What is it? Who are you at your core as a company? What are you trying to achieve? Then it's tying that strategy to what your customer expectations are. Simple things like being focused on what your complaints are coming in; and fix those. I mean, there are things that you can do, and yes, you're using predictive analytics, but don't just think about your own goals from of a risk perspective. Think about your customer and how you reaching to them. As actuaries, I think we have a responsibility to be connected throughout our organization, which we often are not very good at on sometimes as an industry, though I think we focus a little bit too much on our own risk protections and not quite enough on the end customer, so I think that needs to continue to evolve.

This was a small subset of the complete interview and panel presentation. Once again, you can listen to the entire interview (our questions, plus many from the audience) at *bit.ly/2UOdyyT* (Session 40). ■

Xiaojie (Jane) Wang, FSA, CERA, MAAA, is corporate vice president at New York Life Insurance Company in New York City. She can be reached at *xiaojie_wang@newyorklife.com.*

Dave Snell, FALU, FLMI, ASA, MAAA, CLU, ChFC, ARA, ACS, MCP, teaches AI Machine Learning at Maryville University in St. Louis. He can be reached at *dave@ActuariesAndTechnology.com.*

All About Them Curves: Ordered Lorenz Curves and Lift Curves

By Michael Niemerg

Often, models are crafted to optimize some objective function. However, in the real world, the quality of a model is multidimensional and can't really be summarized in a single metric. One of the goals in any thorough model validation process is the evaluation of a model with an entire array of diagnostics. The more perspectives we can judge a model from, the better sense we can get of what the model is truly achieving.

Two very closely related validation measures I've found myself employing continually more often are Lorenz curves and Gini gain. Both give a sense of how well a model is able to stratify risk in the sense of rank ordering. The Lorenz curve is a visualization of this stratification, while the Gini gain is a way to transform such a visualization into a single summary statistic.

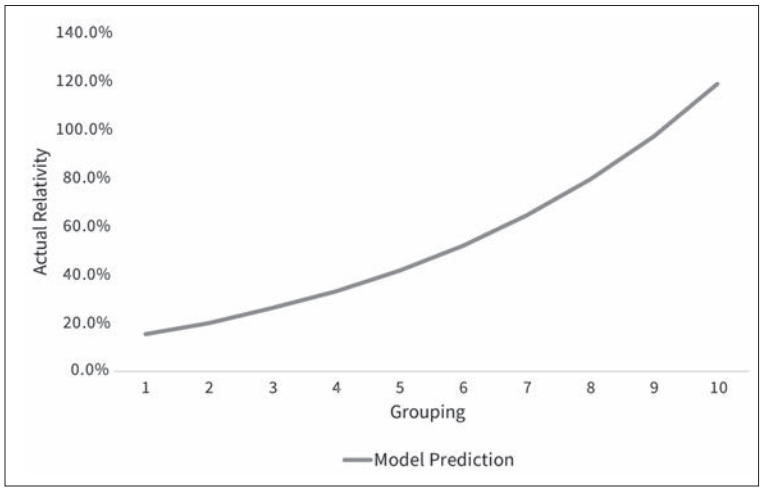
These two model validation diagnostics are great alternatives and complements to lift curves, which are another commonly utilized visualization technique used to measure model stratification. The term "lift curve" has multiple meanings, so I'm going to define precisely what I mean.

The best way to define a lift curve is to explain how it is created. To start, you need three elements for each observation in your dataset: the predicted value of an outcome (coming from either a manual rate, mortality table or any other type of predictive model), the actual value of that outcome, and a grouping by which to bin the observations. The chosen grouping can vary; for example, quantiles or prediction ranges are often used.

For consistency and concreteness, I will call predicted outcomes "claim costs." From these claim costs, I will also frequently refer to as "actual relativities" and "predicted relativities"—where both of these values are expressed as ratios relative to some baseline expectation. Using health rating as an example, we may reference a manual rate as our expected value. Then, if the actual experience of one member was $500 per member per month (PMPM) and their manual rate was $400, the actual relativity is 1.2 ($500/$400). If we were trying to build a predictive model to adjust this manual rate, and for this member the predicted outcome was $600, the predicted relativity would be 1.5. In certain situations, using these relativities instead of the claim costs themselves can be simpler. For instance, if we want to see how much a single variable from our rating formula stratifies risk while holding everything else constant, we can use the baseline in the relativity to account for all the variables in the rating process.

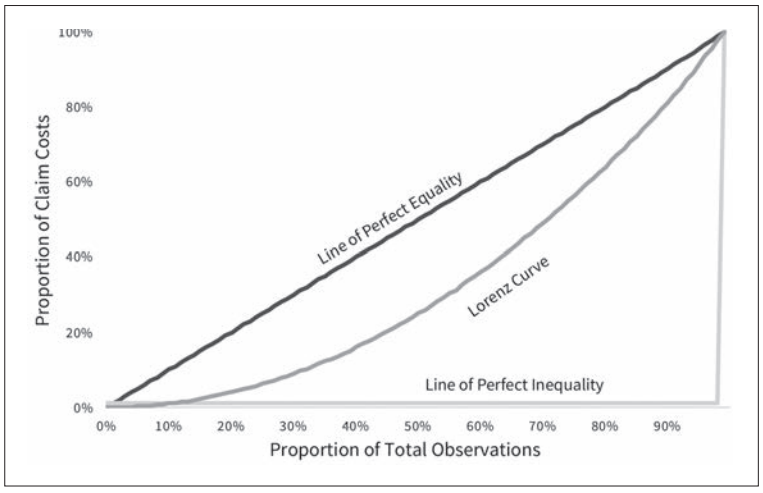
Now, with all the data in hand, we create our graphic to display our lift curve. Taking our observations sorted into groupings, we calculate the actual relativity for each grouping and plot the results (with predicted relativities on the x-axis and actual relativities on the y-axis). Figure 1 offers an illustrative example. For the example in Figure 1, as the grouping increases, so does the actual relativity. If we made this example more concrete and said that the grouping represented predicted relativities, then this chart would be demonstrating that as our predicted relativity increases, the actual relativity also increases (exactly what we were hoping for).

Figure 1
An Illustrative Lift Curve



Now, let's define Lorenz curves and Gini gain. When one typically thinks of Lorenz curves, the immediate association is with economics, since Lorenz curves have a storied history of being used to measure income inequality. In a typical construction of Lorenz curves, the function measures the cumulative proportion of individuals on the x-axis (I'm going to instead use the more neutral lingo described earlier and call these observations) and the cumulative proportion of income on the y-axis where the observations are ordered on income (this will be replaced by our claim costs). The basic idea is to measure how unequal the distribution is. See Figure 2 for an example. The line of perfect equality is what we would expect if we had a perfectly equal distribution (everyone has the exact same claim cost) and the line of perfect inequality is the line we would get if we had a perfectly uneven distribution (one observation generates 100 percent of the claim costs).

Figure 2
An Illustrative Lorenz Curve



What I'm going to talk about from this point on are ordered Lorenz curves, which are very similar conceptually to Lorenz curves, but now we use a predicted relativity to order our observations. Once ordered in this fashion, we graph the cumulative portion of actual claim costs and the cumulative proportion of observations based on this ordering. In building our curve this way, we are able to see whether our predicted relativity is distributing actual claim costs unevenly (which is desirable—ideally, our predictions have some value beyond pure randomness). For instance, suppose we were using an ordered Lorenz curve to test the impact of a new manual rating adjustment factor where a 1.00 signifies no adjustment to the manual rate. Here, we would hope that those observations earning a score below 1.00 would be associated with a disproportionally small amount of losses and that those with an adjustment greater than 1.00 would be associated with a disproportionately high share of losses. This would result in an ordered Lorenz curve with a bowed shape, similar to that in Figure 2.

The Gini statistic allows us a succinct way to summarize a Lorenz curve or ordered Lorenz curve with a single metric. It is equal to the total area between the Lorenz curve and the line of perfect equality divided by the total area between the Lorenz curve and the line of perfect inequality (since the total area under the line of perfect equality will be equal to 0.5, this is also equivalent to two times the area between the Lorenz curve and the line of perfect equality). Having this single summary metric is nice when values are either close and the Lorenz curves are hard to visually distinguish from the line of equality or if you want to summarize an entire array of model validation metrics in a single table to compare them simultaneously.

What are lift curves and ordered Lorenz curves achieving for us? Remember that our end goal is to create a visualization that helps us see risk stratification in a meaningful way. With the alternative approach of plotting every single data point (comparing predicted versus actual), the result would be a line bouncing all over the place, because insurance claim costs are highly volatile. By way of contrast, lift curves and ordered Lorenz curves are both employing a form of smoothing to make results visually interpretable. An ordered Lorenz curve smooths out the variance in actual results between different ranges by showing a cumulative value (any single observation will only contribute a small amount to the cumulative distribution resulting in a curve that appears smooth), whereas a lift curve is doing that with its grouping.

One thing that needs to be kept in mind is that by its very nature, smoothing removes granularity to make underlying trends more visible. One caution when using lift curves here is, therefore, that the break points in a lift curve are arbitrary, and slightly different break points can result in massively different looking curves. Compare Figure 3 (Pg. 12) and note how different these two models' results are. Now look closer—Figure 3(a) is based on the same model (I pulled a sleight of hand). The only difference is that the range on the x-axis in Figure 3(b) is shifted slightly—by a mere 0.01. This slight difference in binning created a massively different looking result. This isn't even a contrived example. I was able to create these graphs using data from a real project and without much experimentation. Moreover, the model represents several hundred thousand lives, so it isn't just a consequence of using a small sample. Most lines of business in insurance are volatile, and this variability impacts the lift curves.

This underscores the fact that you can't focus too much on small perturbations in the lines of lift curves.

Figure 3
Lift Curve Comparison for Slightly Different Ranges of Lift Curves

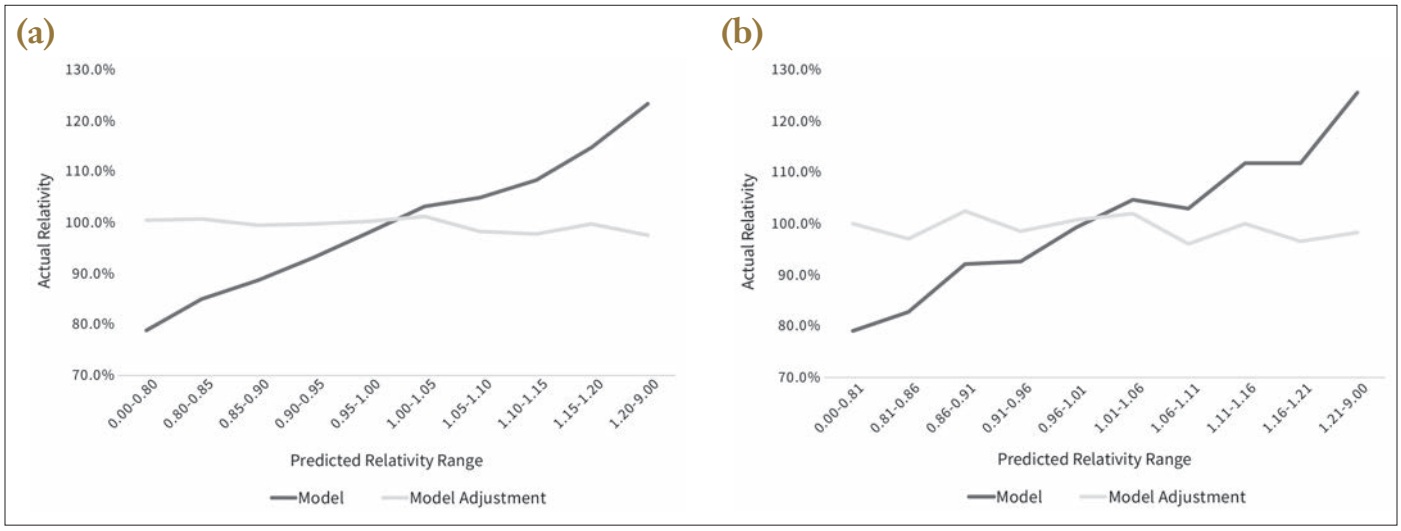
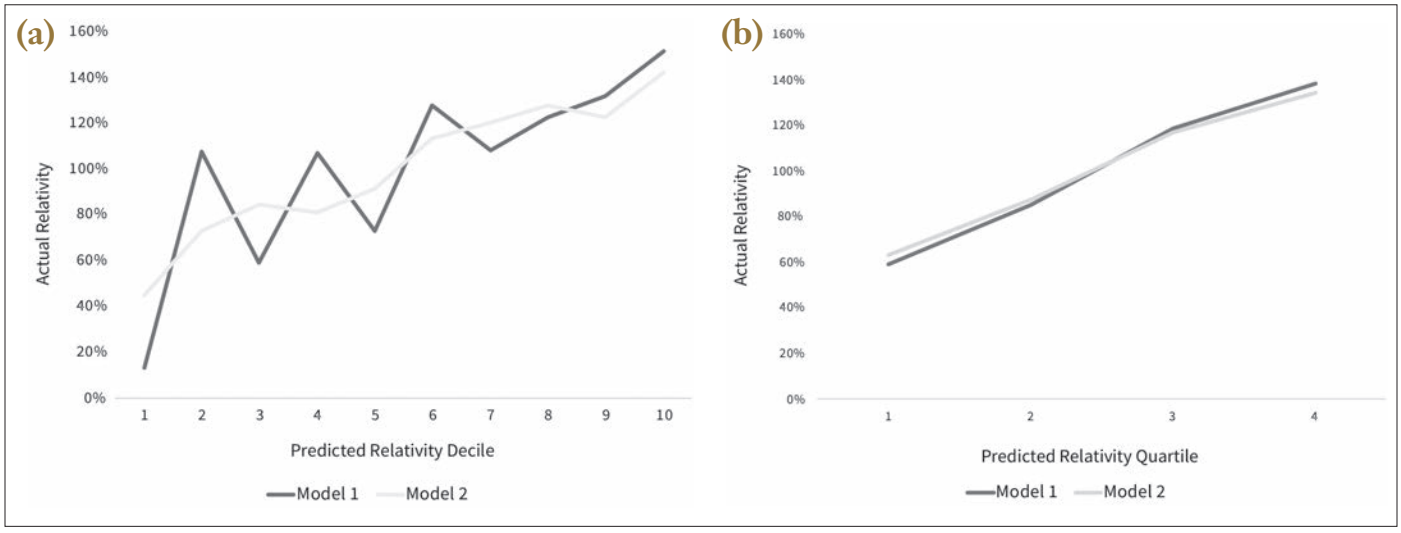


Figure 4
Lift Curve Comparison With Two Different Bin Sizes: Fine (a) and Coarse (b)



Although lift curves are great visual tools, they can be over-interpreted because the binning will always be somewhat arbitrary. For high-variance events such as health care claims or mortality, these small perturbations in the predicted model results can move the predicted value for an observation with a large claim from one bin to the next, and this can result in a significantly different lift curve even if the prediction itself isn't meaningfully different. For instance, a single million-dollar claimaint may have a predicted relativity of 0.899 in one iteration of a predictive model and 0.902 in another version of a predictive model. Because of this, a Lorenz curve is much more invariant than a lift curve to these small perturbations of the scores since it does not include any binning.

For another example, see Figure 4, where I compare two different models (Model 1 and Model 2) against one another. The difference between the two figures here is that I am simply using different bin sizes for my predicted relativities (deciles and quartiles). Although both figures show consistency in the sense that Model 1 stratifies risk better in both figures, the smoothing in Figure 4(b) masks the fact that there appears to be greater volatility in the values for Model 1. However, in this instance, the Lorenz curve may not save us because it also

LIFT CURVES VS. ORDERED LORENZ CURVES

LIFT CURVES

+ Easier to tell how predicted values change as the predicted relativity changes
+ Simple interpretation
+ Easier-to-visualize model bias
- Hard to summarize with a single metric
- Hard to visually distinguish differences between two competing models when the changes in predictive power are marginal
- Small changes in how groupings are determined can correspond to large visual changes

ORDERED LORENZ CURVES

+ Shape is more stable when large claims are present
+ Gini gain summarizes risk stratification in a single statistic that is easily compared across models
+ Not subject to arbitrary groupings
- Hard to tell if the model predictions are biased without using other validation metrics
- Less intuitive and harder to explain

won't obviously show the higher volatility of Model 1 when it is visualized either.

Another word of caution: Although both lift curves and ordered Lorenz curves do a great job of displaying the rank ordering of a model, they don't tell you whether your model is getting you close to your target. For an illustrative example, see Figure 5. If you compare the predicted relativity versus the target for the five observations here, you see a large deviance. The model significantly underpredicts low target values and significantly overpredicts very high model values. However, merely comparing rank orders would suggest a quality predictive model. The lowest actual relativity value corresponds to the lowest predicted relativity, and the highest actual relativity corresponds to the highest predicted relativity.

Figure 5
An Example of a Model That Rank Orders Well but Whose Predicted Values are Biased

Observation	Predicted Relativity	Actual Relativity
1	20%	100%
2	50%	120%
3	120%	140%
4	200%	160%
5	500%	250%

For a lift curve, it is easy to add a quick modification to make this discrepancy easy to visualize. Simply include another line in the graphic that includes the actual values after adjusting for the impact of the predicted relativity (as shown in Figure 3(b) and 4(a)). If the model is unbiased, its predictions will adjust the results to be close to a 100 percent relativity (there will always be some noise as well, so pay attention to sample size). Other validation metrics, such as mean absolute error and root mean-squared error, can be useful here as well.

I use lift curves daily; by pointing out these weaknesses of lift curves, I am not hating on them. However, they can be fickle, so it's important to always interpret them with caution. These pathologies of lift curves are intuitive when you think about them, but it is easy to get careless. That's why I like using Lorenz curves as additional side information. They respond differently to changes in predicted relativities, and when you calculate the Gini gain from them, you are able to summarize the stratification of the model in a single statistic that isn't hampered by the subjectivity of visual interpretation.

In summary, Lorenz curves and Gini gain are good alternatives and complements to lift curves. Together, they form a dynamic combination of ways to measure risk stratification. Don't use just one of them. Use all of them. ■

Michael Niemerg, FSA, MAAA, is a predictive modeling manager for Milliman IntelliScript. He can be contacted at *Michael.Niemerg@milliman.com.*

REFERENCE

Frees, Edward (Jed) W., Glenn Meyers, and A. David, Cummings. 2014. Insurance Ratemaking and a Gini Index. *Journal of Risk and Insurance* 81, no. 2:335–366.

The Growing World of "Off-Balance Sheet" Talent and Workforce Modeling

By Megan Gauer, Nathan Pohle and Priyanka Srivastava

There are a growing number of options for organizations to engage with talent in the global workforce, including the United States. As new generations enter the workforce, there are changes in preferred ways of working, shifts in priorities and demand for more flexibility and independence. Continued trends in working across time zones, along with enhancements in technology, such as improved remote meeting spaces, are helping to enable this shift.

- In the United Kingdom, 50 percent of the working population will be self-employed within the next five years.[1]
- In the United States, more than 40 percent of employees are employed under alternative staffing arrangements, and this number is steadily rising.[2]
- Across the United States, Europe and India, approximately 77 million members of the workforce already identify as freelancers.[3]

There is a spectrum and range to these new groups of talent and workforce options available to companies. These include traditional off-balance sheet options, such as consultants and contractors, but also include emerging options such as gig workers—think Uber drivers or talent assigned to a specific task and also an option to crowdsource talent.

These new options for talent present an opportunity for companies to create a competitive advantage and an opportunity to tap into new ideas and ways of working like never before. As a result, the impacts of this change are not confined to only these new sources of talent but also will impact the roles and responsibilities of employees already within organizations, because the nature of their work will change as well. In this new future of work, all of these resources—both on- and off-balance sheet—will come together to form a new optimal and, if done right, more effective and efficient talent model.

IMPLICATIONS FOR ORGANIZATIONS AND ACTUARIAL FUNCTIONS

The implications of this transition on organizations and their actuarial functions are multidimensional, and the way professional work will be distributed will be different. This fundamentally redefines how work can be organized and accomplished. Instead of limiting the development of work product to only within an organization or an internal team, one can access the global brain and a much wider range of talent by using a variety of available options.

Implication Example 1

Organizations across the globe, insurance companies and the actuarial profession included have started to experiment with crowdsourcing and platforms such as Kaggle (a predictive modeling and analytics crowdsourcing competition platform). Actuarial teams have participated in numerous Kaggle competitions. For example, the 2017 Data Science Bowl, which asked participants if they could improve lung cancer detection, offered a $1,000,000 prize. and several teams with actuaries placed in the top 25 percent.[4] The Society of Actuaries (SOA) has launched a Kaggle Involvement Program, which is described by the SOA as "an opportunity for actuaries to showcase their predictive modeling skills through data science competitions. Competitors are challenged to produce the best models for predicting and describing the datasets uploaded by companies and users."

Implication Example 2

Busy financial reporting time periods, such as year-end and quarter-end, can provide a strain on the on-balance sheet talent model for insurance companies. Plus, regulatory and compliance activities may not be the work activity where companies want to focus their actuarial time. A growing opportunity has been for companies to use talented off-balance sheet actuarial gig workers during these traditionally busy times, allowing more-experienced on-balance sheet workers to focus on results and decision-making

Implication Example 3

A third implication and trend has taken a different format and structure than implication examples 1 or 2. An increasing option has been for companies to assemble cross-functional teams and/or task forces to work together in new, cross-functional ways, enabled by changes in the talent mix. A cross-functional team can bring together resources such as actuaries, business unit representatives, IT and data scientists to create an optimal combination of talent to solve business problems. This type of structure is different than the crowdsourcing option in implication example 1, which typically offers a prize of sorts and is more temporary in nature. For example, perhaps an organization needs a new predictive modeling solution to solve a key business challenge, such as how to better predict staffing

fluctuations based on sales fluctuations. The company could strategically tap into this cross-functional team to generate a solution.

These broad changes and implications create opportunities to reimagine how the on-balance sheet team can be structured and utilized. As the distribution of work changes, current resources will likely have more flexibility to move around within the organization, develop new skills and shift priorities to more strategic parts of the actuarial function. This presents an exciting opportunity for the actuarial workforce to provide even more value to employers.

BENEFITS

There are many benefits to this emerging optimal workforce, and in this article, we discuss three possible benefits.

Benefit 1: Diversity of Thought

A more dynamic workforce can allow companies to attract different types of talent. Organizations will have access to a wider pool of talent to address their most complex problems. This represents a fundamental change from the past, where options were possibly more limited when a new problem arises, either due to time and resource constraints or alike thinking due to echo chamber effects. Under the changing model, companies would have access to the global brain to tackle their challenges and ideas.

Benefit 2: Improved Employee Satisfaction and a Better Work Product

Another advantage of a more flexible talent model is the ability to perform work more cheaply and more quickly. Crowdsourcing and other talent levers present a potentially cheaper option than traditional options. An award offered through a competition can be less than what a full-time resource would cost but is enough to attract talent. Additionally, those same resources are typically not bogged down by business-as-usual activities, leaving them with more time to dedicate to the more important and interesting aspects of a specific task.

Benefit 3: Independence of Thought

Not only do gig workers and off-balance sheet talent provide a more diverse perspective, but they also can provide more objective and independent thinking. These resources are

not constrained by any engrained company culture or echo chambers.

CHALLENGES

The complexity of these various talent levers and timing elevates the importance of workforce planning. Here, predictive modeling can be used to model staffing needs, with inputs including sales, pipeline and seasonality of work. This presents another opportunity for actuaries.

Confidentiality and privacy of data becomes harder to control, but there are measures to mitigate this risk, including the use of nondisclosure agreements to legally protect the organization.

There is also a more tactical, operational challenge of tapping into these new markets. It represents a shift in how organizations plan for workforce need. Because this new set of resources is new to the human resource function as well, it may present challenges in the form of the mechanics of where to start, who to contact and how to identify which types of workers are a best fit for a task.

CLOSING

Even with the challenges presented, when these new types of resources are used optimally, organizations can better meet their business objectives and demands than they can with their current, possibly overburdened, resources. Organizations around the globe recognize this, with 70 percent of companies believing a different mix of talent will be required in the future.[5] However, this is still new to many organizations, with only 16 percent of organizations having a well-defined strategy to deal with a hybrid workforce.[6] Therefore, a large gap will exist, as companies start defining the talent strategy for the future and the optimal talent models, including off-balance sheet talent such as gig workers and the crowd. The time to start planning is now, because the next generation of talent is here. ■

Megan Gauer, ASA, MAAA, is a consulting actuary with experience in the life insurance industry. She can be reached at *megauer@deloitte.com.*

Nathan Pohle, FSA, CERA, MAAA, is a consulting actuary with experience in the life insurance and sports industries. He can be reached at *npohle@deloitte.com.*

Priyanka Srivastava specializes in the area of risk management and applications of predictive analytics for insurance and annuity products. She can be reached at *prisrivastava@deloitte.com.*

ENDNOTES

1 Leighton, Patricia. Future Working: The Rise of Europe's Independent Professionals. *European Forum of Independent Professionals http://www.efip.org/sites/default/files/Future_Working_Full_Report.pdf* (Accessed March 5, 2019).

2 Murray, Patty. Contingent Workforce: Size, Characteristics, Earnings and Benefits. *Letter from the U.S. Government Accountability Office*, April 20, 2015, *https://www.gao.gov/assets/670/669899.pdf* (Accessed Dec. 7, 2018).

3 Matthews, Ben. Freelance Statistics: The Freelance Economy in Numbers. January 8, 2017. *https://benrmatthews.com* (Accessed March 5, 2019).

4 Society of Actuaries. Kaggle Involvement Program. *https://www.soa.org/predictive-analytics/kaggle-program* (Accessed March 5, 2019).

5 Deloitte. Deloitte Study: 73 Percent Report C-Suite Isn't Working Together Despite Need for Increased Collaboration on Human Capital Challenges. April 3, 2018. *https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-study-report-c-suite-isnt-working-together.html* (Accessed March 5, 2019).

6 *https://www2.deloitte.com/content/dam/Deloitte/de/Documents/human-capital/2018-Global-Human-Capital-Trends-The-workforce-ecosystem.pdf*

Artificial Intelligence and the Actuarial Profession

By Sarah Abigail

At the 2018 Society of Actuaries (SOA) Annual Meeting & Exhibit in Nashville, I was invited to coordinate and moderate a session on Artificial Intelligence (AI). The session was hosted by the Predictive Analytics & Futurism Section as well as the SOA General Insurance Section. The SOA provided the largest room available at the convention hall; however, it was filled to capacity, with people standing in the back, and more attendees unable to get in. This article is a synopsis of what was discussed.

It is clear that actuaries are excited about AI and AI's impact on the future of the actuarial profession. Technology is on track to provide more advancements in the next 10 years than we have seen in the last 50 years. Right now, tech companies are investing hundreds of billions of dollars to develop quantum computer chips. It is a race with limitless possibilities that will scale AI platforms in ways we cannot fully comprehend or imagine. What we know for sure is that AI systems to help actuaries with the decision-making process are being developed, and some of those systems are up and running today.

Shankar Vaidyanathan, the founder of Noonum; Martin Snow, FSA, MAAA, the vice president and chief development officer of Atidot; and Gaurav Gupta, the founder of QuaEra Insights, are actively involved in bridging the gap between machine learning (ML) technologies and the insurance industry. In our session, we answered some of the popular questions about AI, including how actuaries can start utilizing machine learning technologies that leverage AI.

After a brief intro about the origins of AI, Vaidyanathan shared how he built an ML tool that can read through millions of pages of financial data to find useful observations and esoteric patterns. Large amounts of data from multiple sources and of various forms may normally take a team of analysts years to read through, but the ML tool can search it and decipher necessary information in a matter of seconds.

With a handful of companies providing ML tools that can help actuaries with the decision-making process, Snow explained how the business side of the AI system works. The most important step an actuary needs to make is to ask the right questions. If a slightly vague question is asked, the software may give a vague set of answers. With ML solutions, there are a lot of people behind the scenes preparing data, designing customized models, engineering software to fit those models, analyzing results and validating those results, along with the team of executives who work with the client and want to make sure that the results are actionable.

One fun example that Snow shared was how his ML tool helped an insurer find customers who needed more insurance. Snow shared the strategy behind how the model was set up and how his ML software curated a list of underinsured policyholders who were most likely to buy more insurance given an extensive list of predictive variables. He shared that ML solutions are helping insurers discover a deeper understanding of their customers, which results in a reduction of balance-sheet risk, and improved margins.

Gupta was the final speaker, and he provided some deep insights on the relationship between the actuarial profession and the AI world. With so few actuaries involved in AI, Gaurav explained that there is a bigger hurdle for insurers to capitalize on AI in the way other industries have. Currently, ML can help insurance companies with automation, from underwriting to claims. Gupta's ML software helped an insurer by using online behavior data to find new low-risk customers. He also shared how ML helped another insurer with correcting a mortality table using 10 times more variables than a typical predictive analytics tool was able to handle.

To get started, Gupta recommended beginning with a small problem that can be solved in three to five months. Both Snow and Gupta recommended using internal data and making sure the answers will be valuable and actionable. The critical element, in Gupta's opinion, is choosing the right team. For the best outcome, he suggested bringing together actuaries, data scientists, and ML experts to develop the right questions.

Engaging actuaries in the AI world is one of my missions. In the next decade, the actuarial profession will heavily rely on AI functions. It is inevitable. AI is world-altering, and the flood gates are all open. The shape and direction of AI developments in the actuarial profession will rely on a handful of actuaries who pioneer it. Actuaries who are willing to get involved today will become ambassadors of the profession to the AI world. ■

Sarah Abigail is the cofounder at Ironbound Consulting Group on Wall Street in New York City. She can be reached at *sarah@ironbcg.com.*

Interactive Plotting

By Zhen Hui Trinh, Ben Johnson and Eileen Burns

A picture is worth a thousand words. Does an idiom exist to better describe what advanced data visualization methods aim to accomplish? As data scientists, our purpose for using graphics is to reduce massive amounts of information into bite-sized deliverables to be easily consumed by our intended audience. Depending on the audience and the type of deliverable—be it a lengthy report, a quick-and-dirty email or perhaps an elaborate slide deck—the sort of visuals we use will inevitably vary. Invariable is the need to convey as complete an analysis as possible without overcomplicating the visuals. Producing a too-busy plot is about as helpful to your reader as donning a track jacket when confronted with Snowmageddon.

As mediums for publishing the results of our analysis move online, we gain a number of advantages that paper reports cannot provide. The ability to incorporate interactive plots within our output is a great convenience, encoding additional information into graphics by utilizing mouse clicks, hovers, highlights and more. That's why we call them "interactive plots," but the interactivity does not stop there. Not only can a reader interact with a plot, but that plot may interact with other visuals. You can design a report so that clicking on one plot could trigger a chain of events that alters the entire display of the document. While paper formats may force us to split plots into several facets to avoid overcrowding, interactive plots enable us to communicate an equivalent amount of information in a much more compact space. Acknowledging the irony of touting the advantages of interactive plots within a static publication, we provide a link in the next section to a GitHub repository that includes code for producing said interactive plots.

Several interactive plotting packages have been developed far past infancy and into what we feel comfortable calling adolescence. Others also feel that there remains functionality to be desired in the available packages and have decided to delay adoption of interactive plots until packages have further matured. This is exactly how our team first felt when we considered migrating from the standard, static visuals. As a precursor to implementing interactive plots in our online web applications, we've compared the capabilities of some of the more popular interactive plotting packages. We build our web applications in Shiny, so we've focused on graphics packages in R, though we note there is similar functionality available in Python. Although the existing packages may not yet have everything you want, they still bring a lot to the table.

During the search, our team identified seven packages for consideration, building off the work of a former teammate, Mandy Zhuang, in 2017. (Thanks Mandy!) Among these, plotly seemed the most popular option within the R and Python communities. However, RStudio's ggvis package seemed a strong contender to our team, due to its similarity to ggplot. And, of course, we were compelled to investigate googleVis with its association to the tech giant. The list also included rCharts, rbokeh, highcharter and billboarder, with friends and colleagues in similar fields highly recommending the latter two. There are several packages not included in our investigation that provide other types of useful interactive graphics, such as maps, time series and tables.

Rather than provide a comparison of all of the packages across each relevant dimension (which would take more than the space available for an article), we're sharing a detailed comparison of two packages: plotly and billboarder. We built these packages into a Shiny app that we used to illustrate the differences below. We've also shared the Shiny app via GitHub and will continue to develop it as we investigate further: *https://github.com/milliman/SOA_PAF_Section_Newsletter_Plotly_vs_Billboarder.*

PLOTLY VERSUS BILLBOARDER

Zoom

A zoom-by-dragging property is enabled for both plotly and billboarder. Plotly allows for point selection, while billboarder allows for range selection. In other words, with plotly, the range of the y-axis can be changed after zooming, but with billboarder, the range of y-axis is fixed. As shown in Figures 1 to 4, we zoomed into the dots representing setosa in the Iris data set with both plotly and billboarder.

Figure 1
Before Zooming in Plotly

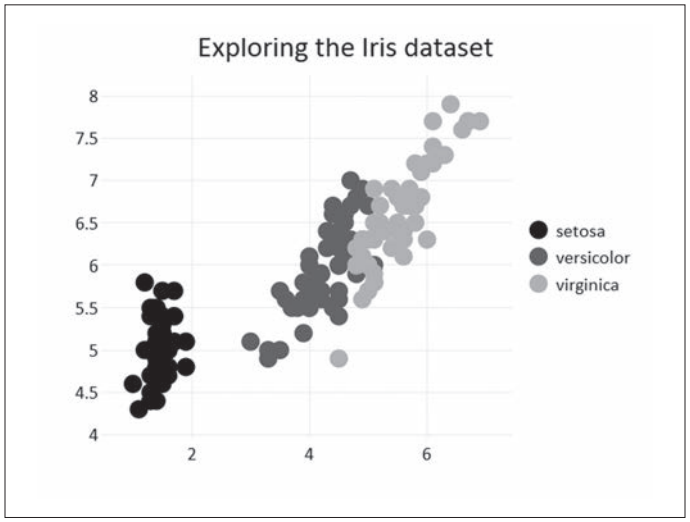


Figure 2
After Zooming in Plotly

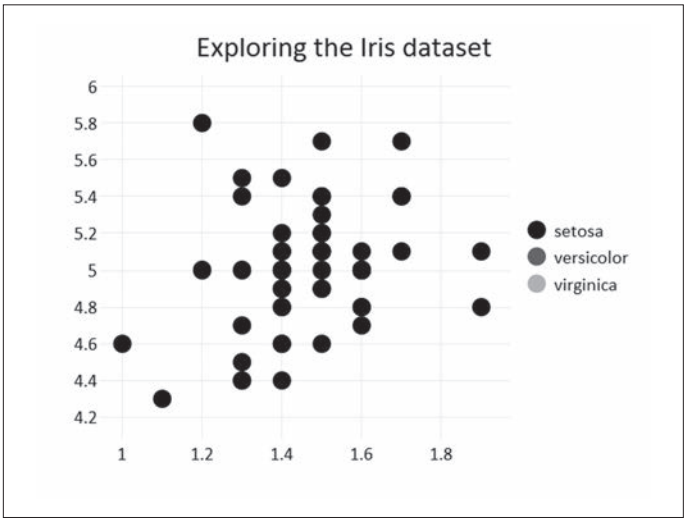


Figure 3
Before Zooming in Billboarder

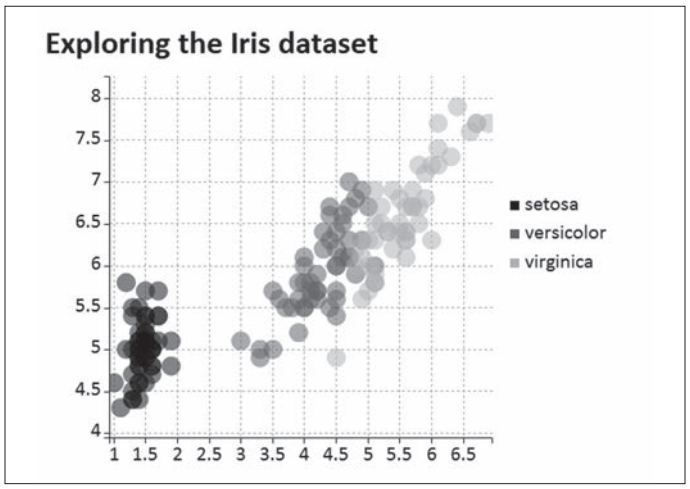


Figure 4
After Zooming in Billboarder

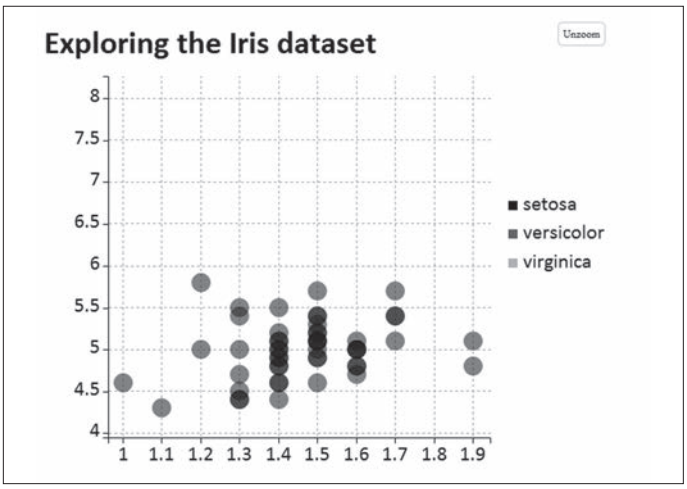


Figure 5
Default Hover Information in Plotly

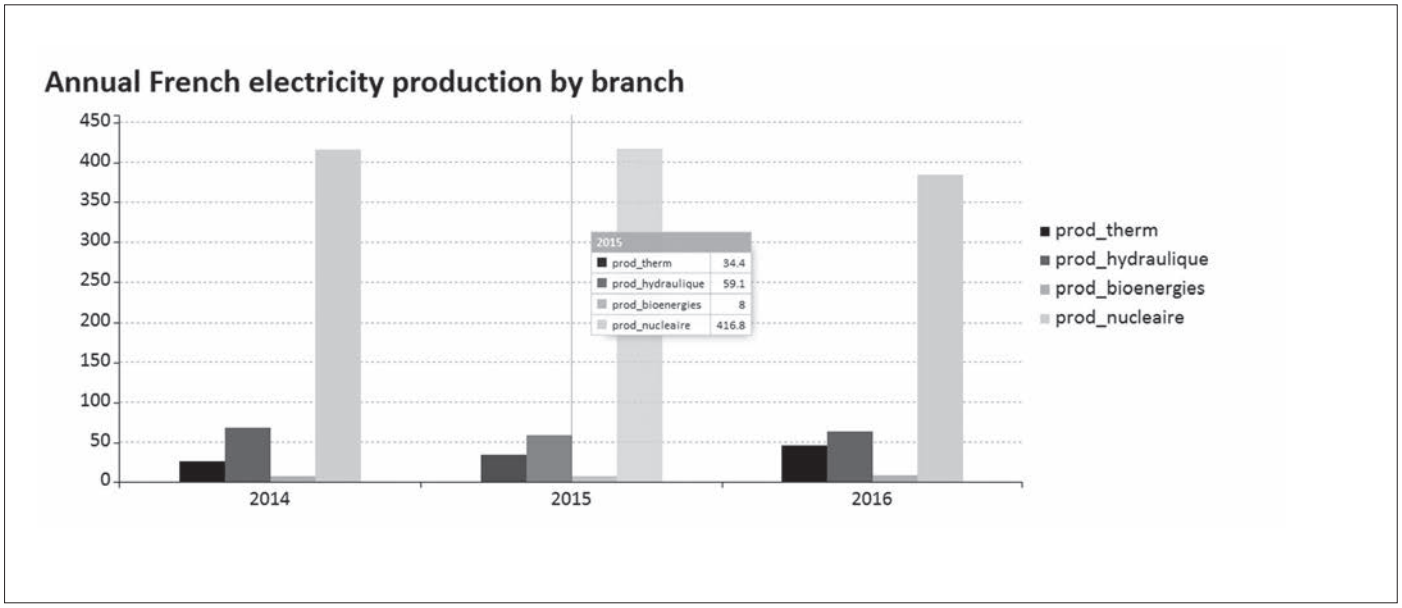


Figure 6
Default Hover Information in Billboarder

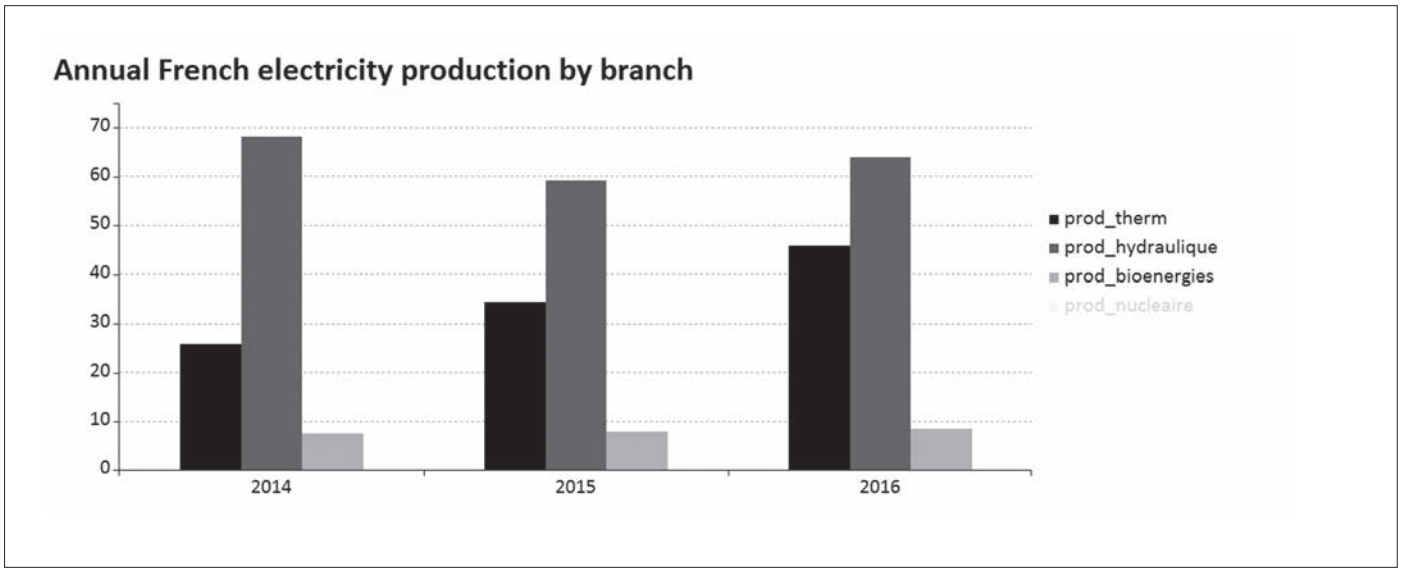
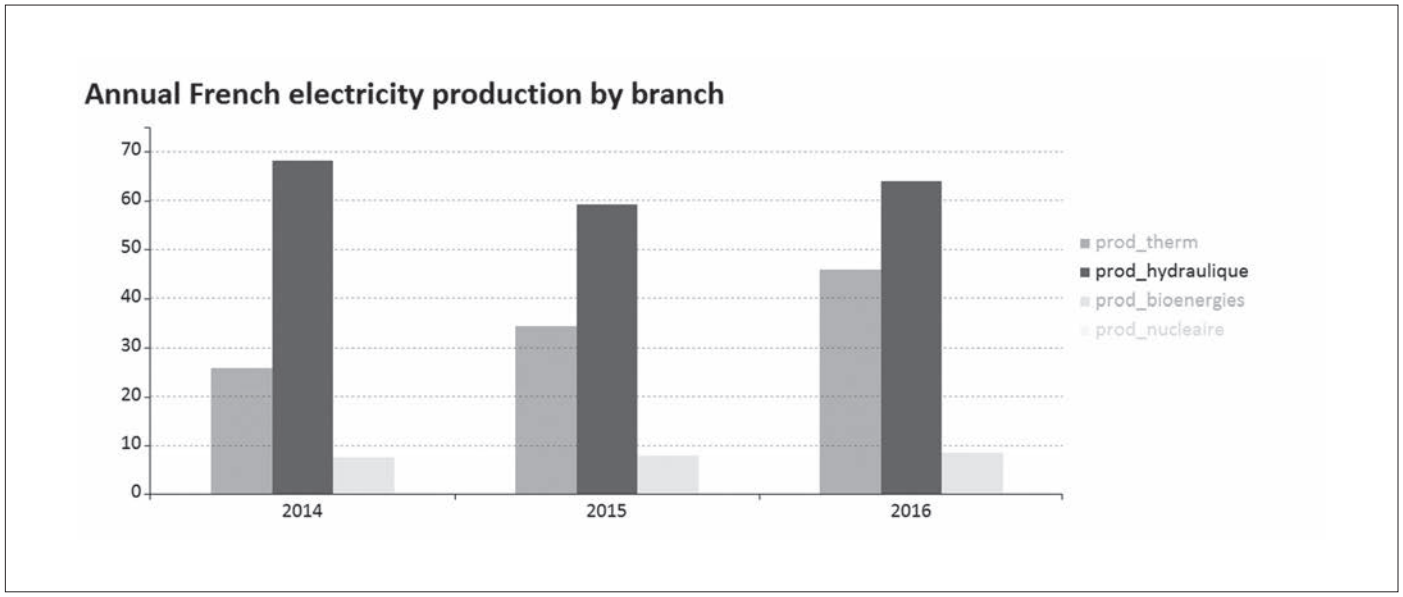


Figure 7
Before Hovering Over the Legend in Billboarder



Hover Information

For plotly, the default display format is: (x value, y value) series label, but there is a lot of freedom to design how you want to display information for a point on hover. Billboarder has a similar display for scatter plots but has an enhanced default option for bar plots and line charts. The x-axis value of the selected data point is displayed on the first line of the box, and the series label and y-value are displayed on the second line of the box with a vertical line in between. This feature can add value because all the y-values for a specific x-value are displayed at the same time.

Legend

For both plotly and billboarder, when a data series in the legend is clicked, the data points of the series will disappear in the graph. This will also lead to changes to the axes—readjusting the range of x- and y-axes to produce a better-scaled plot. For instance, in Figure 5, prod_nucleaire has an exceptionally large value compared to other variables, so the y-axis limit is set to 450 in the plots. After unclicking its legend, as seen in Figures 6 and 7, we see that the y-axis limit is now set to 70.

Of course, they will reappear after you click on their label in the legend again. One extra advantage of using billboarder is that when you hover your mouse over the legend, the data points of that data series will be highlighted in the graph and those of the other groups will become transparent in the background. This feature could be very useful if there are multiple series to compare and you only wish to focus on one. Figures 6 and 7 demonstrate how the bar for prod_hydraulique is highlighted after we hover the mouse over its legend. Plotly does not have this feature.

Display of Large Numbers

In plotly, large numbers are automatically scaled and labeled with k (for values in the thousands), m (millions), b (billions), etc. This feature is very helpful if the values you are displaying are very large. In billboarder, this is not a default

Table 1
Web Resources

Package	Examples	GitHub
plotly	https://plot.ly/r/	https://github.com/ropensci/plotly
billboarder	https://cran.r-project.org/web/packages/billboarder/vignettes/billboarder-intro.html	https://github.com/dreamRs/billboarder
highcharter	http://jkunst.com/highcharter/	https://github.com/jbkunst/highcharter
ggvis	http://ggvis.rstudio.com/ggvis-basics.html	https://github.com/rstudio/ggvis
googleVis	https://cran.r-project.org/web/packages/googleVis/vignettes/googleVis_examples.html	https://github.com/mages/googleVis
rCharts	https://ramnathv.github.io/rCharts	https://github.com/ramnathv/rCharts
rbokeh	http://hafen.github.io/rbokeh	https://github.com/bokeh/rbokeh

setting, but you can make it display the same way as plotly does manually.

ADDITIONAL POINTS OF COMPARISON AND FURTHER RESOURCES

Some packages have commercial or home websites with helpful information and examples, and all have at least a place to see the package in action. Other dimensions you may consider when selecting a package include the following.

1. **Accessibility.** Most are on the Comprehensive R Archive Network (CRAN), enabling easy installation directly from R, though rCharts must be installed from GitHub using the devtools package.
2. **Usability.** All of the packages have a GitHub repository where users can track suggestions and edits to get an idea for how well-used and maintained the package is. Plotly, highcharter, googleVis and rbokeh appear to be the best maintained.
3. **Flexibility.** Packages have varying levels of flexibility for altering basic plot components such as label size.
4. **Cost.** Most packages are free, but some like plotly[2,3] and highcharter have limits. This may be a concern if you wish to enable the users of your Shiny app to download plots after they have adapted them. For example, plotly has a limit to the number of downloads allowed per day, and highcharter charges for commercial use.
5. **Cross-functionality.** All of the packages we looked into work well with Shiny (important for our use case), and some also work with knitr. Depending on your use case, this may qualify or disqualify a given package.

Given our own use case, we're leaning toward using billboarder in our web applications, but we will continue to watch for developments in this space by using the resources we share in Table 1. For any of you who are interested in incorporating interactive plots into your products, we encourage you to consider your own needs and decide which package is best for you. ■

Zhen Hui Trinh is a data scientist for Milliman. She can be contacted at *zhenhui.trinh@milliman.com*

Ben Johnson is an actuarial data scientist for Milliman Financial Risk Management LLC. He can be contacted at *ben.johnson@milliman.com.*

Eileen S. Burns, FSA, MAAA, is a principal and consulting actuary with Milliman. She can be contacted at *eileen.burns@milliman.com.*

ENDNOTES

1 The html widgets for R website shows more about what's available; *http://www.htmlwidgets.org/showcase_plotly.html.*

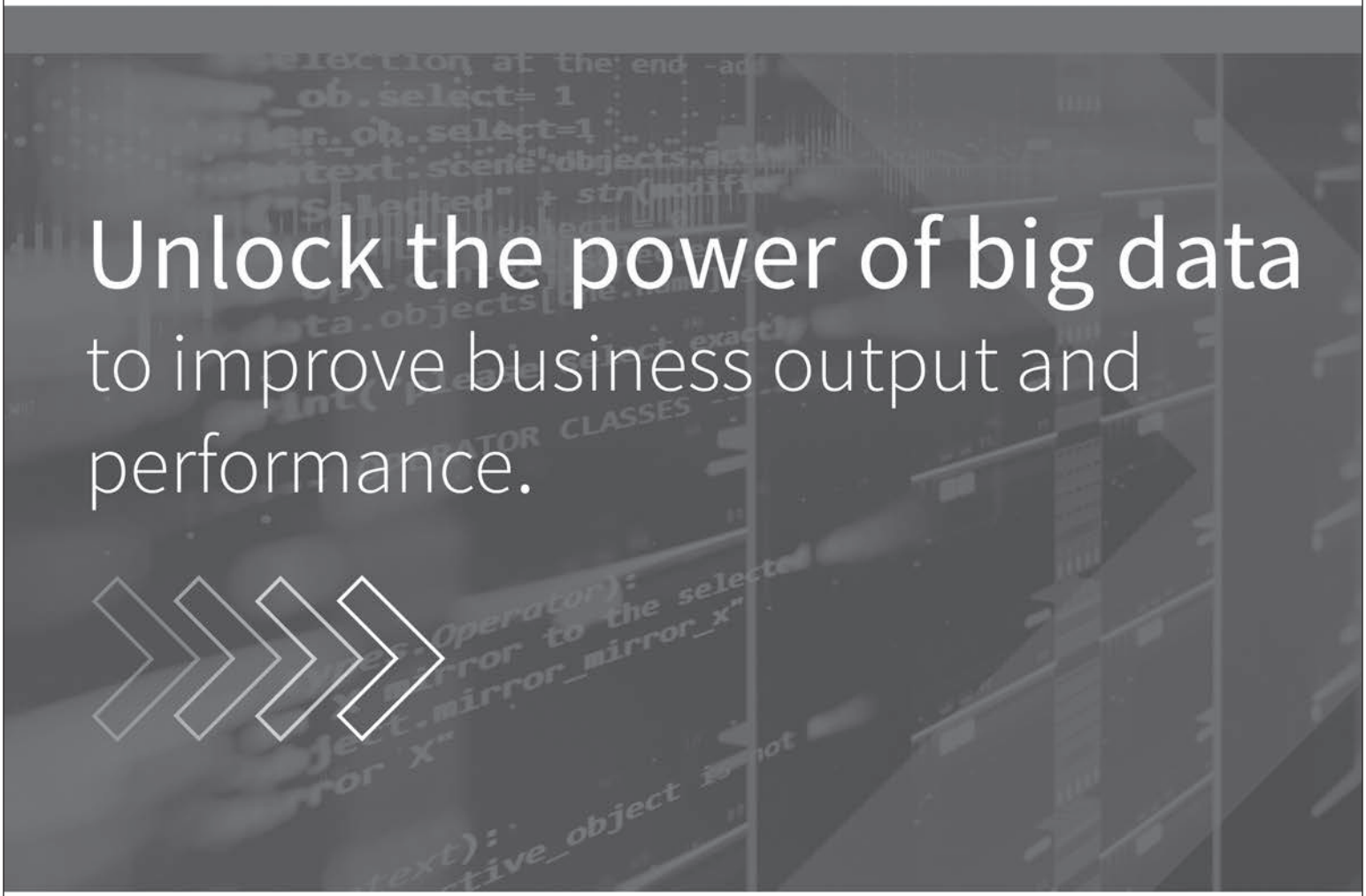
2 Plotly. Chart Studio Cloud. *https://plot.ly/products/cloud* (Accessed Jan. 29, 2019).

3 Plotly. Chart Studio API Rate Limits. *https://help.plot.ly/api-rate-limits/#plotly-api-rate-limits* (Accessed Jan. 29, 2019).

2019 PREDICTIVE ANALYTICS SYMPOSIUM

September 19–20
Philadelphia, PA

Save the date



Unlock the power of big data

to improve business output and performance.

SOCIETY OF ACTUARIES®

Getting C++ Performance From Python With Cython

By Jeff Heaton

Python programs have a wealth of packages that can increase runtime performance. Packages such as Numpy, Scipy, TensorFlow and PySpark are just a few of those available to optimize your program's performance. Normally, when your Python code makes extensive use of loops that process mathematical equations, performance can suffer. Heavily nested mathematical oriented loops and multidimension arrays are normally the domain where compiled low-level languages such as C++ are best suited. However, by making use of a Python package called Cython, you can achieve performance close to that of C++ in Python.

> Cython works by transforming Python into C code. The output from Cython is literally a .C file.

Performance is not the only reason to consider using Cython. You can give the compiled binary produced by Cython to external users of your application, allowing some degree of protection of the intellectual property (IP). Cython works by transforming Python code into C code. The output from Cython is literally a .C file that you must compile with a C compiler. The code generated by compiling a C program to an executable is very difficult to decompile back into C, let alone further back into Python. This makes Cython an effective tool to protect IP contained in your Python source files. While no software protection scheme is perfect, it is much more difficult to reverse engineer compiled C code than higher-level languages, such as R and Python.

Because the output of Cython is C-source code, you must have a C compiler installed to make use of Cython. On Macintosh and Linux, this is easy. Both of these two platforms have open-source C compilers available. Windows is a bit more complex. For Windows, you will need one of the Visual C++ compilers or one of the open-source compilers. Docker is also an option for Windows, because it allows emulation of standard UNIX environments.

CYTHON FOR EXTENSIONS

This article will demonstrate several different ways to use Cython. You can use Cython with either Python 2.x or 3.x; however, the examples contained in this article will make use of Python 3.x. As of the time I wrote this article, the latest version of Python was Python 3.7. I also used the Anaconda release of Python on a Macintosh computer and the GCC C++ compiler version 8.2. Other environments should work; however, small modifications may be necessary.

Each of these examples can be found on my GitHub repository.[1] If performance is the goal in your use of Cython, then you will likely be using Cython to create a Python extension. This allows you to compile part of your Python program to a compiled Cython extension that many of your Python programs can use. The code inside of this extension will be compiled to C and be very efficient. You can import this extension into your Python script just like a regular package.

As an example, Listing 1 shows an extension I created that will calculate the standard deviation of a population.

Listing 1
Calculate the Standard Deviation of a Population (Calculate.pyx)

```
import math

def sdev(lst):
    # Mean
    sum = 0
    for x in lst:
        sum += x

    mean = sum / len(lst)

    # Standard deviation
    sum = 0
    for x in lst:
        delta = x - mean
        sum += delta ** 2

    return math.sqrt(sum / len(lst))
```

While Python contains built-in support for calculating the standard deviation of a population or sample, the above code shows how this can be done from simple algebraic operations. This is exactly the type of code that Cython can speed up. In this form, the above code is standard Python and could be used with

or without Cython. Even when compiling from pure Python, Cython gives considerable speed improvements.

You should save the above code to a filename such as "Calculate.pyx." The PYX extension designates this code as Cython. You must now compile this Python code to C code and then to a Python extension. This is accomplished by a Python build script that is often named "setup.py" and is shown in Listing 2.

Listing 2
Build the Extension (setup.py)

```
from distutils.core import setup
from Cython.Build import cythonize

setup(
  ext_modules = cythonize("Calculate.pyx")
)
```

You should execute the build script to actually compile the Cython extension with the following command:

```
python setup.py build_ext --inplace
```

The "build_ext" parameter indicates that you are building an extension. The "inplace" parameter designates that the extension should be placed in the current folder, as opposed to copied to a system directory. Once this program is run, it will create a .SO file under Mac/Linux or a DLL under Windows. This file is the actual Cython extension. Listing 3 shows how to actually make use of this extension.

Listing 3
Test the Shared Object (test.py)

```
import Calculate as c

print(c.sdev([1,2,3,4,5]))
```

This program simply imports the calculation Cython extension and then computes the standard deviation of the set [1, 2, 3, 4, 5]. The test script is executed with the following Python command:

```
python test.py
```

This will output the standard deviation of the vector.

CYTHON FOR STANDALONE EXECUTABLES

It is also possible to produce a standalone executable with Cython. This executable will be a .EXE file in a Windows operating system or executable file on Linux or Mac (these operating systems do not have a specific extension for executable files). To demonstrate this, we will use a classic Python "Hello World" program, as shown in Listing 4.

Listing 4:
Calculate the Standard Deviation of a Population (HelloWorld.pyx)

```
print("Hello World")
```

Unfortunately, the commands to build a standalone executable are a bit more complex than the previous example. The first step is to invoke Cython and convert the HelloWorld.pyx file to HelloWorld.c.

```
cython --embed -o HelloWorld.c HelloWorld.pyx
```

The next step is to compile the HelloWorld.c to a standalone file. This will require the use of your C++ compiler. The command that I used for GCC is as follows:

```
gcc -Os -I /Users/jheaton/miniconda3/
include/python3.6m/ -o HelloWorldEXE Hel-
loWorld.c -L/Users/jheaton/miniconda3/
lib -lpython3.6m -lpthread -lm -lutil -ldl
```

There are two important paths that are provided to GCC. The first is the path to the include files needed to compile. This includes Python.h. The second path is the location of the Python libraries that are needed to compile your Cython extension. This includes the primary Python library, named "python3.6m"; however, additional libraries should also be specified with more "-l" arguments.

The resulting file is executable from the command line.

CALLING PYTHON PACKAGES FROM CYTHON PROGRAMS

Python programs make use of a variety of packages. Often predictive modeling programs will use Scipy, Numpy, Scikit-Learn, TensorFlow and potentially many others. If you are going to make use of these packages, it is necessary for your Cython program to have access to them. These packages are dynamically linked, and your Cython program will not run without them. You will need each of them to be present to run the stand alone Cython executable. You can look up the file location of any Python package with a single Python command. For example, to find the location of Numpy, you would use the following command:

```
python -c 'import numpy; print(numpy.__file__)'
```

NEXT STEPS

This article provided a brief introduction to Cython. Using these techniques alone, you can considerably increase the speed of Python programs that need to use loops for their calculations. However, this is only the beginning. Cython also adds extensions to the Python programming language that you can use to further enhance the performance of your Python code. Such extensions include static typing and multithreading. ■

Jeff Heaton, Ph.D., is a vice president and data scientist for RGA Reinsurance Company in Chesterfield, Missouri. He can be reached at *jheaton@rgare.com.*

ENDNOTE

1 Heaton, Jeff. GitHub. https://github.com/jeffheaton/present/tree/master/SOA/paf-cython (Accessed March 5, 2019).

SOCIETY OF ACTUARIES®

475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
p: 847.706.3500 f: 847.706.3599
w: www.soa.org

NONPROFIT
ORGANIZATION
U.S. POSTAGE
PAID
SAINT JOSEPH, MI
PERMIT NO. 263