SOCIETY OF ACTUARIES

PREDICTIVE ANALYTICS AND FUTURISM SECTION

Predictive Analytics and Futurism

ISSUE 22 • DECEMBER 2019

Actuarial Superjobs: Evolving Roles Demand an Integrated Skillset

By James Hardington Dunseth, Tony Johnson and Adam Cloe

Page 8

Predictive Analytics and Futurism

Issue 22 • December 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/.

This publication is provided for informational and educational purposes only. Neither the Society of Actuaries nor the respective authors' employers make any endorsement, representation or guarantee with regard to any content, and disclaim any liability in connection with the use or misuse of any information provided herein. This publication should not be construed as professional or financial advice. Statements of fact and opinions expressed herein are those of the individual authors and are not necessarily those of the Society of Actuaries or the respective authors' employers.

Copyright © 2019 Society of Actuaries. All rights reserved.

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

2019 SECTION LEADERSHIP

Officers
Eileen Burns, FSA, MAAA, Chairperson
Cassie He, FSA, MAAA, Vice Chairperson
Michael Niemerg, FSA, MAAA, Secretary/Treasurer

Council Members
Dorothy Andrews, ASA, MAAA
Joy Chen, ASA, CERA
Garfield Francis, ASA
Nathan Pohle, FSA, CERA, MAAA
Dave Snell, ASA, MAAA
Xiaojie (Jane) Wang, FSA, CERA

Newsletter Editor
Dave Snell, ASA, MAAA
dave@actuariesandtechnology.com

Program Committee Coordinators
Dorothy Andrews, ASA, MAAA
2019 Valuation Actuary Symposium Coordinator

Ricky Trachtman, FSA, MAAA
2019 Life & Annuity Symposium Coordinator

Anders Larson, FSA, MAAA
2019 Health Meeting Coordinator

Eileen Burns, FSA, MAAA
2019 SOA Annual Meeting & Exhibit Coordinator

Dave Snell, ASA, MAAA
Xiaojie (Jane) Wang, FSA, CERA
2019 Predictive Analytics Symposium

SOA Staff
Beth Bernardi, Staff Partner
bbernardi@soa.org

Jessica Schuh, Section Specialist
jlschuh@soa.org

Julia Anderson Bauer, Publications Manager
jandersonbauer@soa.org

Sam Phillips, Staff Editor
sphillips@soa.org

Erin Pierce, Senior Graphic Designer
epierce@soa.org

From the Editor: PAF Newsletter—The Source Library for Your Predictive Analytics Needs—For You, and by You!

By Dave Snell

This September, we had another successful Predictive Analytics Symposium for the Society of Actuaries (SOA); and I was honored to welcome about 200 actuaries, data scientists, and insurance professionals to a highly popular assortment of sessions directed specifically for actuaries (often taught by actuaries). My co-chair, Xiaojie (Jane) Wang will carry this onwards next year for you.

One of the most frequently asked questions, by far, from the attendees was along the lines of the following: "Where can I go to get more great information like this on predictive analytics? I want something that will help me navigate the learning curve; but I do not have the time to wade through highly academic literature that might require semesters or years of background reading; and I want something more useful than the pablum and hyperbole of a news media article."

Fortunately, the answer was an easy and satisfying one: In addition to the various SOA-sponsored meetings, courses, and symposia, go to the *SOA.org* website, click on Professional Sections, then Predictive Analytics & Futurism, then Newsletter, and then scroll down to the PAF Article Compilation Excel workbook, and download it. Nick Hanewinckel has been keeping this current for us and it has become a wonderful reference source for actuaries and kindred spirits.

The workbook contains over 200 article summaries that you can search through by issue, date, title, author, actuarial specialty, or even the text of the summary. These range from statistical significance, through Delphi studies, feature selection, supervised and unsupervised methods, generalized linear models (GLMs), neural networks, genetic algorithms, etc., with understandable descriptions of the concepts and practical actuarial examples of the usage. For example, neural networks articles (a very hot topic in machine learning) range from the basics, through generative adversarial networks (GANs), and Neuroevolution of Augmenting Topologies (NEAT). The articles are often accompanied by links to code (in Python, R, Julia, RePast, Cython, Excel, …) that you can download and run and modify for your own applications. When you find an article of interest, you can hyperlink directly to it. And for those of you (or us) that are not even sure what technique you want to learn about, there are plenty of useful guidance articles. Mary Pat Campbell and Michael Niemerg have crafted very useful evaluations of courses and books in past issues to help you plan your own career map. And Nathan Pohle continues the helpful recommendations in a digital article later this month (see note on format change below).

Personally, as editor of the printed edition for the past 11 years, I feel like I may have benefited the most from this treasure trove of knowledge. In the graduate artificial intelligence (AI) machine learning (ML) course I teach, I frequently provide links to supplementary perspectives and explanations of the concepts I am teaching. I know that my students will not be intimidated by reams of formulas to better understand a current technique. Also, I can comfortably recommend (or even assign) these articles because they have been vetted and tested and applied by peers in the actuarial community. This editorial cannot individually thank the many authors (over 50 of you) who have contributed your time and expertise; but I want to give a big thanks to you collectively for authoring such interesting and enlightening articles.

Starting in early 2020, the PAF newsletter will discontinue paper issues, and move on to an all-digital publication format. The digital edition will published every other month, so you

won't have to wait as long between issues. We hope this results in an ability to provide even more value to you.

This issue adds to the collection:

- **Chairperson's Corner: Change is constant—by Eileen S. Burns**

  - Eileen gives us a whirlwind summary of the last three years of our section. Reading through her article you can appreciate why it is the fastest growing section of the Society of Actuaries. We accomplished an impressive collection of new initiatives and started several more … too many to summarize here, so read her Chairperson's Corner and feel proud to be part of PAF!

- **Actuarial Superjobs: Evolving Roles Demand an Integrated Skillset—by James Hardington Dunseth, Tony Johnson, and Adam Cloe**

  - While many current occupations, even some in the actuarial sector, are in danger of elimination by AI and ML, a new opportunity exists for those actuaries who embrace these technologies—a chance for a "Superjob." Read what these attractive and lucrative jobs might be like, and how you can better prepare for them.

- **Big Data—You've Rocked My World!—by Dorothy Andrews**

  - Contrary to what you might infer from the article title, Dorothy provides counterpoint from some of our favorite data science skeptics (Cathy O'Neil, author of *Weapons of Math Destruction* is a favorite of mine) and suggests Big Data sometimes leads to sloppy analysis and that "a truly outstanding data scientist knows how to put 'science' in the phrase data science." She raises important issues of correlation without causation.

- **Autoencoders for Anomaly Detection—by Jeff Heaton**

  - Autoencoder neural networks are cool! I like them a lot. They can perform dimensional reduction and many other tasks. Jeff presents them in a clear manner and then goes on to show an intriguing application where an autoencoder can alert us to events out of the ordinary, such as a network penetration attempt.

- **An Ever-Welcome Warning Against Big Data Hype: A Review of *Big Data, Big Dupe* by Stephen Few—By Mary Pat Campbell**

  - Mary Pat reviews a book that debunks some of the hype we read about using quantity at the expense of quality. She also points out some unrealistic outcomes when data scientists lack business knowledge. "Profitability is driven by policy year!" "Great, we'll go back in time and write more 2015 business." She sums it up as, "Go out and make sense of your data!"

All these articles are readable by an actuary without extensive background in predictive analytics (even by me!); yet they help extend our knowledge in a world where continual learning is mandatory for prolonged success.

Enjoy your PAF library! ■

Dave Snell, FALU, FLMI, ASA, MAAA, CLU, ChFC, ARA, ACS, MCP teaches AI Machine Learning at Maryville University, where he was designated Outstanding Adjunct Faculty for the 2018–2019 Academic Year He can be reached at *dave@ActuariesAndTechnology.com*

SOCIETY OF ACTUARIES®

SOA Job Center

The Premier Source For Your Actuarial Career

Thinking about taking the next step in your actuarial career? Check out the SOA Job Center and get access to benefits such as a free resume review, career resources and more.

Learn more at **Jobs.SOA.org**

Chairperson's Corner: Change is Constant

By Eileen S. Burns

No futurist or predictive modeler could have foreseen all of the changes I've observed between 2016—the year I joined the Predictive Analytics and Futurism (PAF) Section Council—and 2019, the year I leave, even if we narrowed the field to me, the PAF Section, the Society of Actuaries (SOA) and actuarial interests in predictive analytics and futurism.

Personally, it's been a journey! I went from knowing next to nothing about SOA sections to leading the fastest growing section, writing articles, doing research, planning meetings, presenting internationally, brainstorming the future of work with other section chairs and SOA staff, and trying to find ways to add additional value for our members. And that's just the volunteering.

Over the same three years, my job has taken two big leaps. The first was thanks to a new hire in my practice at Milliman who was very technology focused. My team had been using predictive analytics to do experience analysis: cleaning data, building models to answer questions, delivering reports in Microsoft Word. Thanks to this new hire's gentle push, we began looking into alternative ways of sharing our analyses and wound up developing a software product—using our good friends R and Shiny.

At the same time, industry regulations evolved to demand insurers begin to leverage their data better and govern their data and models more effectively. Watching that evolution made me optimistic that one day soon insurers will have all of their data in one place, relatively cleaned, and there for the taking by predictive modelers. This view ultimately led me to my second leap, joining a different team within Milliman, where I might one day help those predictive modelers to use that data where it sits. Building that first software product had given me a taste for product management, exactly what is needed to enable those predictive modelers.

Over the same horizon, the PAF Section has accomplished a lot, and set even more into motion. This year we sponsored six sessions at the SOA's Annual Meeting & Exhibit In Toronto in October, a 50 percent increase over the two prior years. We have

twice as many members. We attended the third Predictive Analytics Symposium (PAS) in September in Philadelphia. We've closed two competitions—the Jupyter Notebook competition and the hackathon at the PAS. We've finished one research project (see my article later in this issue) and are midway through a second. Our podcasts are maybe the one thing that hasn't changed: They're still the most downloaded podcasts from the SOA.

Within the SOA, we're offering a predictive analytics (PA) certificate, newsletters are digital and even the exams have changed! There's now a PA exam on the associate of the Society of Actuaries (ASA) curriculum, and more changes are in the works for the fellow of the Society of Actuaries (FSA) exams. Not to mention the focus on predictive analytics in the research agenda.

Further abroad in the insurance industry, insurers are no more sure that PA is the way of the future by my read, but they're all at least investigating it. Some think actuaries should be doing it and some don't. Actuaries are brilliantly capable but will always be a small blip on the screen in the field of predictive analytics. That's the reason behind the last two initiatives our section is starting under my watch. We are inspiring our members out into the world to learn from and share with others advancing this art.

The first initiative is to encourage actuaries interested in predictive analytics and futurism to create local networks to share ideas and push each other forward. These local networks don't need to be made up solely of actuaries; the intent is to connect people with similar interests. We've recruited a few volunteers for a pilot and hope to kick off a few events in the coming months to demonstrate the idea and encourage more.

The second initiative is crafting journey maps for actuaries interested in becoming more adept in the field of predictive analytics. We're in the process of putting out a call for volunteers to share their experiences with online or in-person learning modules. After consolidating input from these volunteers, we'll compile a list of recommended resources based on your current level of experience and desired analytical programming language.

Where will we be three years from now? I will wait to find out because I certainly can't extrapolate from here. ■

Eileen S. Burns, FSA, MAAA, is a principal and director of product development in the Life Technology Solutions practice with Milliman. She can be reached at *eileen.burns@milliman.com.*

Actuarial Superjobs: Evolving Roles Demand an Integrated Skillset

By James Hardington Dunseth, Tony Johnson and Adam Cloe

Editor's note: As used here, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about *for a detailed description of Deloitte's legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.*

Rapidly evolving technology, the arrival of artificial intelligence (AI), a tsunami of data, an explosion in contingent work and diversity/generational change are all trends that are rapidly transforming professions and causing organizations to holistically redesign their workforce and how work is executed. The actuarial profession is not immune to this transition. In fact, the actuarial profession has already begun to utilize automation tools to complete a number of tasks historically executed manually by actuaries. Over time, and with some human guidance, those machines have proven reliable, not to mention they *actually* perform some actuarial tasks faster and with fewer errors than humans. More than 40 percent of respondents to the 2019 Deloitte Global Human Capital Trends Survey indicated that their organization has already employed automation "extensively or across multiple functions."[1] Despite these reports, many individuals continue to fear robotics and artificial intelligence as they associate automation with job loss. However, as organizations begin to test these emerging technologies, the fear of the unknown should quickly dissipate once the potential benefits are realized. The result on *our* profession? An impending shift in the way actuaries spend their time, as they are augmented by technology, toward more strategic roles and responsibilities.

This is not the first time the actuarial profession has adapted to a dynamic workplace and disruptive technologies. Writing exams without a calculator, using a physical mortality table to price a product or calculating reserves using nothing but a freshly sharpened pencil, battery-powered TI30 calculator and a grid-lined notebook were once realities of being an actuary. Although the transitions were slow and sometimes difficult, it is clear that artificial intelligence, automation and cognitive technologies are not the profession's first disruption. From the introduction of desktop computers, to laptops and, most recently, cloud computing, actuaries have demonstrated the ability to adapt and to continue providing enterprisewide value in the face of change and uncertainty. Perhaps in 10 years, actuaries will view the professionals of today as archaic for insisting on gathering, preparing and manipulating data themselves.

As organizations search for the next competitive edge, early adapters are poised to become market leaders aided by automation and cognitive technology solutions. These emerging technologies do not pose a threat to the profession; rather, organizations hope to use them to facilitate interactions *with* actuaries, saving employees time and allowing them to maximize the value they contribute to organizational performance. However, the organization is not the only beneficiary. Given that most tasks subject to automation are rules-based, procedural and mundane, actuaries will be able to devote additional time to tasks that cater to their strengths, give them energy and passion and improve job satisfaction. Actuaries should welcome the opportunity to devote their day to cognitive activities, that is, those tasks that are well outside the current scope of most artificial intelligence technology. For instance, many valuation actuaries today spend a majority of the quarter-end production time creating the first draft of earnings results. An actuary aided by artificial intelligence would have results much earlier and could instead spend their time digging into variances, communicating with other business units to explain trends and presenting results to management. With ever-changing regulatory environments, consumer demand for increasingly sophisticated financial products and incessant pressure to evolve from competitors, organizations need their actuaries to ascend into these higher-value roles.

Deloitte's 2019 Global Human Capital Trends report categorizes the jobs of the future as standard jobs, hybrid jobs and

superjobs.[2] With the help of an increasingly diverse skillset and cutting-edge technologies, standard actuarial roles, marked by repeatable tasks, standardized processes and specific skillsets, can transform into hybrid jobs. Emerging hybrid jobs require a diverse portfolio of skills ranging from technical competencies to uniquely human soft skills such as communication, creativity and inquisitiveness; however, hybrid jobs are only a transitional step toward what the report calls superjobs. These roles are increasingly analytic, with a simultaneous focus on integration, opportunity and communication. Actuarial superjobs will consolidate several traditional roles, enter new territories and incorporate lifelong learning, demanding a new breed of actuary. Consider the following examples of possible superjobs for the actuary of the future:

- **Artificial intelligence expert.** Combines deep AI experience and actuarial content knowledge to design and update AI robotics solutions to help with assumption setting, source-of-earning analysis, explaining reserve movements, etc.

- **Machine interpreter.** Reviews and interprets machine output, communicating to nontechnical parties as needed to drive strategic action.

- **Superforecaster.** Combines big data and AI technologies with human oversight and actuarial judgment to forecast scenarios across business functions that may impact organizational performance.

- **In-force management influencer.** Works with psychologists to combine data analysis, predictive analytics and behavioral science to positively influence healthy policyholder behavior, reduce unintended actions from insureds and optimize profits for the organization.

- **Algorithm auditor.** Combines strategic thinking, professional judgment and machine learning methods to validate current actuarial models and future AI-based actuarial calculation engines.

Consider an actuary in the role of an algorithm auditor. Armed with a thorough understanding of AI technologies, specifically as they relate to the intersection of machine algorithms, human ethics and actuarial professionalism, the algorithm auditor transcends departments to ensure the organizations' solutions are free of unintentional biases and in compliance with regulation. This superjob requires strategic thinking, professional judgment and knowledge of machine learning methods to validate actuarial models and the design and output of AI algorithms. The algorithm auditor will have an important voice in decisions regarding automation as it impacts the enterprise risk profile. As a result, this role requires organizationwide relationships and is marked by increased visibility in the insurance and risk management industry. Moreover, as more of the work traditionally given to entry-level actuaries is transferred to machines, the role of algorithm auditor will help junior staff learn the ins and outs of the business.

Superjobs are an example of how automation and cognitive solutions offer immense value, not in eliminating actuarial jobs, but augmenting them.

Superjobs are an example of how automation and cognitive solutions offer immense value, not in eliminating actuarial jobs, but augmenting them. Further, with organizations' use of robotics and cognitive technologies, today's enterprises are more machine-powered and data-driven than ever. In addition, enduring interpersonal skills such as problem solving, communication, interpretation and design skills will likely become more valuable. As the profession searches for ways to define actuaries as much more than data scientists specializing in insurance, the gap between outcome-driven superjobs and the fixed-task roles of today will likely become increasingly evident. Thus, the question becomes, how can actuaries bridge the gap between current roles and hybrid jobs, and eventually hybrid jobs and superjobs?

Today's actuary already possesses a breadth of expertise encompassing a deep understanding of risk management, predictive analytics and various financial products, but these skills alone will not be sufficient to stay competitive in a rapidly evolving workforce. The half-life of professional skills has shrunk considerably, indicating the actuary of the future will need to participate in lifelong learning, outside the world of credentialing exams.[3] To further muddy the waters, traditional actuarial skills have become more common, creating the need for a professional with highly integrated quantitative knowledge and the soft skills that undeniably benefit actuarial work, such as creativity, inquisitiveness, fellowship and other skills not replicated by machines. To facilitate the transition to superjobs, it is important for organizations to invest in resources that support their employees. This should include investing in robust training programs and continued learning curriculums and the implementation of a flexible actuarial talent model, which will vary significantly from the rigid, hierarchal structure many organizations have traditionally used.

Actuaries need to change, evolve and disrupt themselves to be able to create the superjobs that will inevitably help separate the actuary of today from the actuary of tomorrow. Those that do not risk being disrupted by other workforce segments that are pushing the envelope of what outcomes they can achieve. Organizations and individuals both have a critical responsibility to drive this transition so that, together, they may both enter the future of work as leaders creating unprecedented value. ■

Tony Johnson, ASA, MAAA, is a senior manager with experience in the financial services industry, including the life insurance industry. He can be reached at *tojohnson@deloitte.com*.

James Dunseth, ASA, is a consulting actuary with experience in the life and health insurance industries. He can be reached at *jdunseth@deloitte.com*.

Adam Cloe is an actuarial analyst with experience in the life insurance industry. He can be reached at *acloe@deloitte.com*.

ENDNOTES

1 Volini, Erica, Indranil Roy, Jeff Schwartz, Mauren Hauptmann, Yves Van Durme, Brad Denny and Josh Bersin. 2019. From Jobs to Superjobs. In *Leading the Social Enterprise: Reinvent With a Human Focus*. 2019 Deloitte Global Human Capital Trends. *https://www2.deloitte.com/content/dam/insights/us/collections/HC-Trends2019/DI_HC-Trends-2019.pdf*.

2 Ibid.

3 Deloitte. Meet Your Future Workforce. Quartz, Jan. 8, 2019. *https://qz.com/1123158/meet-your-future-workforce/*. (Accessed Sept. 13, 2019).

SOCIETY OF ACTUARIES®

MARK YOUR CALENDAR FOR KEY 2020 SOA EVENTS

ReFocus Conference
March 1–4 | Las Vegas, NV

Life & Annuity Symposium
May 4–5 | Saint Louis, MO

Health Meeting
June 8–10 | Chicago, IL

Valuation Actuary Symposium
Aug. 31–Sept. 1 | New Orleans, LA

Annual Meeting & Exhibit
Oct. 25–28 | Seattle, WA

For an updated listing of professional development opportunities,
visit ***SOA.org/Calendar.***

Big Data—You've Rocked My World!

By Dorothy Andrews

We can no longer be sure of what we believe thanks to big data. The advent of big data has caused researchers to question the soundness of age-old sampling techniques and scientific methods,[1] the chain of custody and stability of big data,[2] the necessity for its neatness[3] and the need for causation over correlation.[4] The following is my appreciation for these revelations.

In statistics, a metric is "unbiased" if its mathematical expectation based on a sample equals its population equivalent. If this relationship holds, then there is no need to calculate the metric on the entire population in order to make inferences about the population. The sample metric will do. Historically, it has been cheaper to run statistical tests on samples rather than on entire populations, as Cukier and Mayer-Schoenberger[5] would agree. Depending on the experiment, however, samples are still preferred to running experiments on entire populations. Fast forward to the age of big data and big data analytics and we are seeing many analyses being performed on populations. In these cases, there is no need to question "unbiasedness" in the results because the sample is the population, N = All.[6] One would think the scientific community would be more excited than they are about being able to use populations over samples but that does not seem to be the case. Big data is "destabilizing" their models and systems,[7] forcing them to develop new approaches to solving problems before there are willing or able to do so.

> Many statistical texts exist that prescribe techniques for handling "messy data."

There other considerations that raise concerns over the future applicability of results even from population sized datasets. A major concern is that "real life" unlikely reflects all the variations the future may hold and, in those cases, simulating data to anticipate future scenarios is an often employed technique. This poses risks, which are significant, for interpreting the results, but such discussion is beyond the scope of this writing.

Plantin et al. point to a compromise in the "chain of custody" of data, which they call the "control zone."[8] They express concern for the integrity of data collected by individuals lacking "traditional scientific credentials"[9] and affiliations with respected institutions. (One must wonder how they feel about Nate Silver basing his election predictions, in part, on Yahoo polls.) Despite their views sounding a bit elitist, it is important to apply sound data collection and governance controls to minimize errors and biases in results to ensure results are stable over time.

The well-known 80/20 rule as applied to model building means about 80 percent of the effort in building a model is spent on cleaning and scrubbing the data and about 20 percent of the effort is spent building model code and results, and validating the results. Many statistical texts exist that prescribe techniques for handling "messy data." For example, when data is missing in a field in a record, one technique is to estimate the missing value with the mean, median or mode of existing values of the field to prevent having to exclude the record from the calculation of a metric. Cukier and Mayer-Schoenberger[10] highlight Google Flu Trends (GFT)[11] as an example where messy data was tolerated because the dataset was big. The GFT was a flu-tracking system, grounded in big data, that was meant to predict influenza outbreaks. Its predictions outnumbered those of the Centers for Disease Control and Prevention (CDC) for all the wrong reasons. The researchers point to Google's willingness to tolerate some "messiness"[12] in their big data because of their view that volume trumps messiness in detecting patterns in data. In effect, Google was saying all the messiness will be lost in the decimal places.

All Google really did was swap causation for correlation, according to the researchers. There is a lot of that going around, according to Barrowman,[13] and not for the betterment of analytics. He talks about one of my heroes in statistics, Ronald A. Fisher, father of modern-day statistics and father-in-law to my other statistical hero, George E.P. Box. Barrowman points out Fisher was skeptical of the data and the research done linking smoking to lung cancer,

believing such linkage was spurious at best. To be clear, Fisher questioned the correlational basis being advanced to support the link, not the possibility of the linkage. Fisher spent a significant portion of his career ferreting out "spurious correlations," creating a "cottage industry,"[14] according to Barrowman. Tyler Vigen has profited from this industry, writing books and creating websites on the subject. This writer employs his website[15] when presenting on predictive modeling issues involving correlations. It is particularly amusing to find a spurious correlation relevant to the audience, like how the number of lawyers in Iowa is positively corelated with the number of days of sunshine in the state. It is a spurious relationship, but highly amusing.

Barrowman[16] provides insights to some excellent tools useful in explaining causal relationships, such as path analysis, structural equation modeling, counterfactual analysis, instrumental variables analysis and directed acyclic graphs that no data scientist should live without. His discussion on selection bias is particularly relevant. In modeling, it is just as important to understand the data that is excluded from a sample as it is to understand the data that is included. Statistically significant is an aberration of the data included in the modeling dataset. This aberration is directly related to selection bias.

O'Neil[17] has made it her mission to become an evangelist of big data skepticism and West[18] is indeed a disciple. He sees the need for the data scientists-in-training to attend Sunday services to soak in the gospel of data skepticism preached from the pulpit by Her Holiness O'Neil. West[19] is critical of upcoming data scientists being too focused on techniques and not paying enough attention to the social and ethical implications of the results of their analyses. I am sold on O'Neil's teachings, too. In *On Being a Data Skeptic*, O'Neil defines a skeptic as "someone who maintains a consistently inquisitive attitude toward facts, opinions, or (especially) beliefs stated as facts. A skeptic asks questions when confronted with a claim that has been taken for granted."[20] Further, she exclaims a truly outstanding data scientist knows how to put "science" in the phrase data science. Many in academia feel the art of designing balanced and unbiased modeling datasets is being lost because data scientists are blinded by their quest for statistical significance. The result of this blindness is modelers who forget about what can go wrong with their models and where they can fail. The model does not speak the truth if it is founded on a poorly designed dataset.

O'Neil[21] gives some significant insights regarding data blindness, which she has termed "The Measurement Addiction." This addiction problem creates four hinderances to skepticism. The first hinderance is an addiction to metrics since they are grounded in mathematics, which is perceived as hard, objective, logical, axiomatic and trustworthy. Non-skeptics are unlikely to question the appropriateness of metrics used to assess a model

because of this perception. The second hinderance is an over focus on numbers and not enough on behavior. It is important not to confuse correlation with causation. Causation is the root of all behavior, not a p-value. The third hinderance is incorrectly framing the problem. It is important to have your model peer reviewed as a check that the correct mathematics has been applied to the problem to minimize model risk. Modeling assumptions should be kept at a minimum to prevent biasing the range of results. Finally, the fourth hinderance is ignoring perverse incentives. Models naturally beg for gaming because they cannot account for all possible contingencies modeled phenomenon respond to. This is an area O'Neil says is sorely ignored by modelers. The models most susceptible to gaming are those that heavily utilize proxy variables and assumptions. Proxy assumptions are often used to model missing data. Where data is missing, it is worth the effort to have data corrected at its source before modeling. Campbell's law summarizes the impact of proxies quite poignantly. It states, "The more any [proxy] quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."[22] This statement generalizes with the removal of the word "social." The message to actuaries, especially those working as or with data scientists, is it is important to identify and resolve weakness in data, big or small, to increase its value and reliability to analytical modeling.

The last bit of gospel O'Neil would say is critical to recognize is the wider cultural consequences of models. The Problem of Externalities is the modeler's problem, according to O'Neil. In her view, modelers have a responsibility to ensure the external effects of their models are benign, that the positive effects outweigh the negative effects, or be subject to the heavy hand of government regulation. ■

Dorothy L. Andrews, ASA, MAAA, CSPA, is the chief behavioral data scientist for Insurance Strategies Consulting LLC. She can be reached at *dorothylandrews@msn.com*.

ENDNOTES

1 Cukier, Kenneth, and Viktor Mayer-Schoenberger. 2013. The Rise of Big Data: How It's Changing the Way We Think About the World. *Foreign Affairs* 92, no. 3:28–40; Tufekci, Zeynep. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. Presented at the Eighth International Conference on Weblogs and Social Media (ICWSM), June 2–4, 2014, Ann Arbor, Michigan.

2 Plantin, Jean-Christophe, Carl Lagoze, Paul Edwards and Christian Sandvig. 2017. Big Data is Not About Size: When Data Transforms Scholarship. In Ouvrir, Partager, Réutiliser: Regards critiques sur les données numériques, 128–48. Paris: Éditions de la Maison des Sciences de l'Homme.

3 Cukier and Mayer-Schoenberger, The Rise of Big Data.

4 Barrowman, Nick. 2014. Correlation, Causation, and Confusion. *The New Atlantis* Summer/Fall:23–44; Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343, no. 6176:1203–05.

5 Cukier and Mayer-Schoenberger, The Rise of Big Data.

6 Ibid., 30.

7 Plantin, et al., Big Data is Not About Size, 14.

8 Ibid., 6.

9 Ibid., 10

10 Cukier and Mayer-Schoenberger, The Rise of Big Data.

11 Lazer, et al., The Parable of Google Flu.

12 Cukier and Mayer-Schoenberger, The Rise of Big Data, 33

13 Barrowman, Correlation, Causation, and Confusion.

14 Barrowman, Correlation, Causation, and Confusion, 25.

15 *www.spuriouscorrelations.com*.

16 Barrowman, Correlation, Causation, and Confusion.

17 O'Neil, Cathy. The Rise of Big Data, Big Brother. *Mathbabe: Exploring and Venting About Quantitative Issues* (blog). May 2, 2013. *https://mathbabe.org/2013/05/02/the-rise-of-big-data-big-brother*.

18 West, Jevin. Calling B.S. in the Age of Data Science Euphoria. Presented at National Academies of Sciences, Engineering, and Medicine's Integrating Ethics and Privacy Concerns into Data Science Education meeting and webcast Dec. 8, 2017, Washington, D.C. *https://vimeo.com/250857594?ref=em-share*.

19 Ibid.

20 O'Neil, Cathy. 2014. *On Being a Data Skeptic*. Sebastopol, California: O'Reilly Media, Inc., 1.

21 Ibid.

22 Campbell, Donald. 1979. Assessing the Impact of Planned Social Change. *Evaluation and Program Planning* 2, no. 1:49. *https://doi.org/10.1016/0149-7189(79)90048-X*.

Autoencoders for Anomaly Detection

By Jeff Heaton

In data science, anomaly detection is the identification of unusual items, events or observations that raise suspicions by differing significantly from previously seen data. Typically, the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions. Anomaly detection can also be particularly useful to determine how suited a model trained on a particular dataset is at handling a new dataset. This suitability detection is the focus of this article.

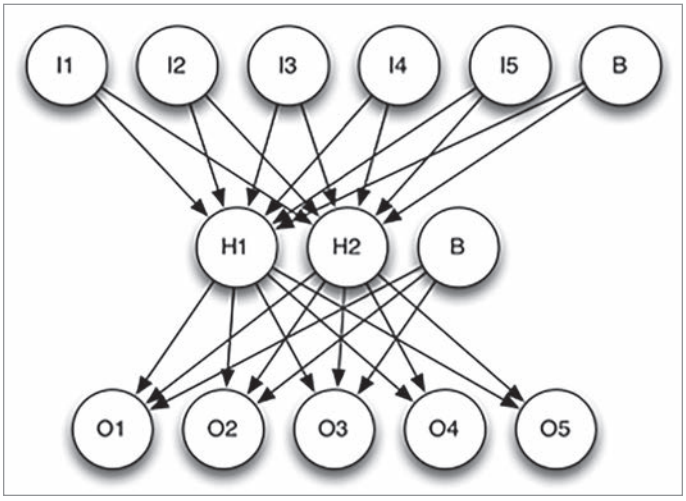
INTRODUCTION TO AUTOENCODERS

An autoencoder is a type of neural network that has the same number of input neurons as output neurons. The number of input/output neurons you have corresponds to the size of your feature vector after the data source has been encoded. For instance, you might have a single input for continuous and a set of dummy variables for each of your categorical inputs. The autoencoder is trained in a supervised fashion; however, the x (inputs) and y (targets) are the same. It is also important to note that the autoencoder is using the ability of a neural network to perform a multi-output regression. The neural network is learning to directly copy the inputs to the outputs. This structure is seen in Figure 1.

At first glance the autoencoder may not seem that useful. We are training a neural network to simply pass the input through to the output. However, there is always at least one hidden layer with fewer neurons than the input and output layers. These hidden layers teach the neural network to compress the input data. You can think of the connections between the input and hidden layers as learning to be a data decompressor and the connections between the hidden and output layers as learning to be a decompressor. It is common to separate the autoencoder into two neural networks. This way the hidden layer becomes the output layer for the compressor.

Extracting the output from the hidden layer can be thought of as a form of dimension reduction, similar to principal component analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE). Because the neural network shown in Figure 1 contains only two hidden neurons, it would reduce the dimensions from the five input neurons down to two dimensions.

Figure 1
Autoencoder Structure



Despite the fact that an autoencoder is trained like a normal supervised neural network, usually using some variant of back-propagation, this training process is considered unsupervised. This is because no one value from the dataset is the target—all of the values from the dataset are the target value. In this regard, the autoencoder training is unsupervised in the same sense as PCA or t-SNE are not provided with a target. However, unlike PCA or t-SNE, the autoencoding neural network also includes a decoder. T-SNE and PCA both lack a well-defined means of returning to the high-dimension input that they processed. In this regard, an autoencoder shares more with a compression algorithm, such as PKZIP, than a dimensionality reduction algorithm.

You can essentially think of the training process of the autoencoder as creating a compression algorithm optimized to the data you provided. Such domain-specific compression-decompression (codec) algorithms are not uncommon. Portable Net Graphics (PNG) format is a lossless codec for image compression. The Joint Photographic Experts Group (JPEG) format is a lossy codec for image compression. A lossy codec will lose some of the original detail from the source data; a lossless codec maintains absolute data integrity. For images and audio, absolute data integrity is not always required.

This specialization among codecs is what allows an autoencoder to be used for anomaly detection. Early cellphone compression algorithms were designed to compress human voice as effectively as possible and make the best utilization of the very slow

cellular networks. When non-voice sounds, such as music, were compressed with these early voice-centric codecs, the music would clearly sound distorted. These early cell phones were anomaly detectors. They produced very little distortion among the human voice data that they were designed for and very high distortion on all other sounds. The more noise introduced into the signal, the less similar that signal was to the original type of data the codec was designed for. Essentially, the effectiveness of the specialized lossy codec for a particular dataset shows how much of an anomaly the new dataset is when compared to the original dataset the codec was designed for.

Now consider an autoencoder. We create an anomaly detector by training this autoencoder on data that we consider "normal." Overfitting is not that big of a concern, since this is effectively an unsupervised learning; however, a k-fold or similar scheme might be used for early stopping of the neural network training once the out of sample error ceases to improve.

INTRODUCTION TO THE KDD-99 DATASET

The KDD-99 dataset is famous in the security field and almost a "hello world" of intrusion detection systems in machine learning. This dataset was used for the Third International Knowledge Discovery and Data Mining Tools Competition, held in conjunction with the Fifth International Conference on Knowledge Discovery and Data Mining. According to the KDD archive, "The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between 'bad' connections, called intrusions or attacks, and 'good' normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment."[1] This dataset is commonly used for computer security and anomaly detection examples. This is the dataset that I used for this example on autoencoder anomaly detection.

The KDD-99 dataset includes a target that identifies the type of attack or if the transaction was normal. We will not directly use this target in the training. Rather, we will separate the data into normal and attack rows. We will train the neural network on the normal rows. We will then compare the difference between the root mean square error (RMSE) for normal vs. error. This RMSE is the difference between the data before and after the autoencoder compresses and decompresses it. The RMSE effectively measures the amount of noise added by running through the autoencoder. Just to be sure there is no overfitting, we will compare out-of-sample normal to the error rate for normal as well.

ANOMALY DETECTION EXAMPLE

This example is from a college course that I teach on deep learning. I will not reproduce all of the code here.[2] This example is in the Python programming language, using TensorFlow 2.0 for deep-learning support.

The TensorFlow autoencoder neural network is set up by the following lines of code:

```
model = Sequential()
model.add(Dense(25, input_dim=x_normal.shape[1],
  activation='relu'))
model.add(Dense(3, activation='relu'))
model.add(Dense(25, activation='relu'))
model.add(Dense(x_normal.shape[1])) # Multiple
  output neurons
model.compile(loss='mean_squared_error',
  optimizer='adam')
model.fit(x_normal_train,x_normal_
  train,verbose=1,epochs=100)
```

You can see that the number of input neurons and output neurons are the same, specified by the value x_normal.shape[1]. These both correspond to the number of predictors in the feature vector generated from the KDD-99 dataset. There are additionally 25 neurons added before and after the three bottleneck neurons to assist with compression and decompression. The three hidden layer neurons specify the number of dimensions that the autoencoder is reducing the data to.

The results from the experiment are shown below.

```
In-Sample Normal Score (RMSE): 0.30
Out of Sample Normal Score (RMSE): 0.31
Attack Underway Score (RMSE): 0.53
```

The in-sample and out-of-sample normal data RMSE were approximately the same, between 0.30 and 0.31. The attack rows were noticeably higher at an RMSE of 0.53. This is consistent with anomaly detection in that the anomaly data is compressed with more noise than normal data. ■

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

ENDNOTES

1 University of California, Irvine. KDD Cup 99 Data: Abstract. *KDD Archive*. Last modified, Oct. 28, 1999. *http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html*.

2 For the complete Python source code for the example, see Heaton, Jeff. T81-588: Applications of Deep Neural Networks. *GitHub*. Last updated, Aug. 20, 2019. *https://github.com/jeffheaton/t81_558_deep_learning/blob/master/t81_558_class_14_03_anomaly.ipynb*.

REACH OUT TO MORE THAN 30,000 ACTUARIES THROUGH SOA SPONSORSHIP

The SOA offers companies the chance to directly reach actuaries from across practice areas and around the world. Choose from diverse sponsorship options that fit your company's budget and desired audience.

Corporate Sponsorship

An effective and convenient program – one agreement for the year for maximum exposure at the SOA's four major meetings:

- Life and Annuity Symposium
- Health Meeting
- Valuation Actuary Symposium
- Annual Meeting & Exhibit

Thought Leadership Sponsorship

Thought leadership opportunities encourage the spread of ideas through engaging presentations by experts in the field. Interested companies may apply to sponsor a series of two (2) sessions at the following SOA major meetings: Valuation Actuary Symposium and Annual Meeting. Alternatively, you can select SOA webcasts to sponsor, reaching a vast audience at minimal cost.

Event Sponsorship

Employ a targeted approach to reaching actuaries in specialized practice areas at specific meetings. Give your company visibility and exposure to actuaries from around the world with an array of benefits at various investment levels. Choose from new sponsorship offerings.

Customized packages can also be arranged to meet your individual needs.

For more information on sponsorship options in 2020, contact ***sponsorship@soa.org*** or visit ***soa.org/sponsorship.***

SOCIETY OF ACTUARIES®

An Ever-Welcome Warning Against Big Data Hype: A Review of *Big Data, Big Dupe* by Stephen Few

By Mary Pat Campbell

Recently, I had the opportunity to test out a tool intended to provide more useful information about individuals, which I could use for my job. The tool was supposed to have used natural language processing and various artificial intelligence algorithms to appropriately structure and cluster results, so that what one received would be more useful than a simple internet search.

Of course, the first person I searched for was myself.

The result? Underwhelming, especially with the amount of completely wrong information.

My employment dates? Completely wrong for every single position mentioned, of the ones where I actually worked. In addition, it listed organizations and companies with which I've had no associations, ever. Now, of course, other Mary Pat Campbells exist, but the point of these types of tools is that they are supposed to automatically distinguish between the actuary Mary Pat Campbell and the psychotherapist Mary Pat Campbell and all the other Mary Pat Campbells.

I have the daily experience of Amazon ads being served up to me on every web browser page, trying to sell to me the book I just bought. Amazon has 20 years of my browsing and buying data—and this is what they do with it. I have political ads served to me on Twitter that are completely out of line with my politics, even though I have put my political labels right in my profile, and what I tweet and retweet gives it all away. For some reason, JCPenney sends me ads on Facebook indicating it thinks I want to buy a top hat. I have yet to buy a top hat.

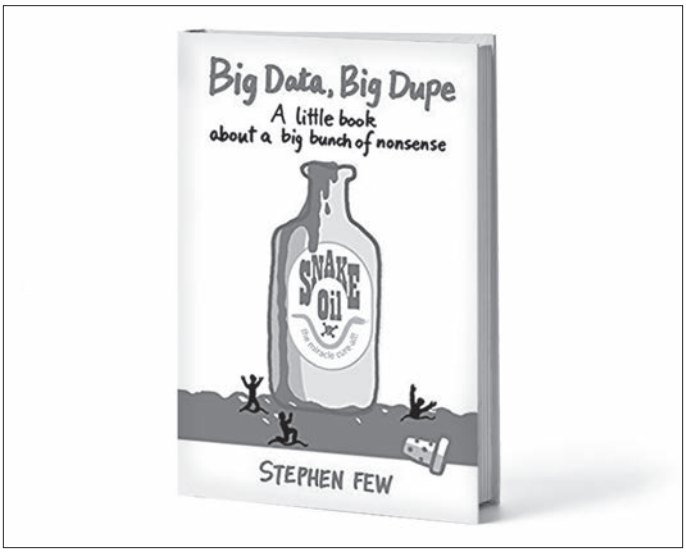
All of these have been courtesy of big data, which not only has been touted as some paradigm-changing aspect of the business environment, but now has been cited as a danger to democracy in Congressional investigations. Are these benefits, and dangers, over-dramatized?

Given I have been deeply unimpressed with the results of big data as it has applied to me, I have obviously bought into the central premise of Stephen Few's book *Big Data, Big Dupe*.[1] Few argues that big data is an ill-defined marketing buzzterm, and he rebuts the hype some have published about this supposedly new phenomenon of the past decade.

The book comes with the subtitle "A little book about a big bunch of nonsense," which is absolutely true—the book is fewer than 100 pages, and there really isn't much in the bulk of the book other than debunking the razzmatazz surrounding big data. He especially likes beating up on the concept of the three V's: volume, velocity and variety. Most of his critique can be found in its original form in blog posts and newsletters at his site.[2] The original entries range from 2009 to 2016, and the book itself was published in 2018.

In particular, the argument regarding both the definition of big data and what qualities are supposed to make it different from the data landscape before takes the form of back-and-forth writing between the proponents of big data as a concept and Few.[3] Some of that is reproduced in the book, and I found it tedious.

Few's book has the problem of most business books: The bulk of the book is of little value if you don't need to be convinced of the premise, and the most important message is in the appendix or epilogue. You can follow the links provided and read (for free) what Few originally wrote before being professionally edited, if you need convincing that big data often doesn't deliver what the proponents say. I am unsure that actuaries need this warning,



given Actuarial Standard of Practice No. 23 on Data Quality.[4] Actuaries have always been alive to the possibilities that data may be of very poor quality, and that simply adding more data does not necessarily make the result better.

> I have heard horror stories from others where generalist data scientists bring back analyses on books of business that were completely useless. . . .

Haranguing against a business fad, while always a popular pastime, is not useful if you're trying to accomplish something. That's where the real value of the book comes in: the epilogue, in which Few advocates an approach to effective data use. In January 2013, he wrote a blog post in which he proposed three S's in response to the three V's: small, slow and sure, which forms the heart of the book's epilogue.[5]

Using Few's own words:

> **Small.** "As data increases in volume, we should keep in mind that only a relatively small amount is useful."
>
> **Slow.** "No matter how fast data is generated and transmitted, the act of data sense-making, which must precede its use, is necessarily a slow process. We must take time to understand information and act upon it wisely."
>
> **Sure:** "Just because you can collect data about something doesn't mean you should. In fact, given all the data that you've already collected, wouldn't it make sense to spend more time making use of it rather than getting wrapped up in the acquisition of more? When you recognize an opportunity to do something useful with data, that's when it becomes sure."[6]

In supporting this "slow data movement," Few provides nine items at the very end of the book for organizations and people to consider when they want to make effective use of their data. These nine items tie to the small, slow and sure items. Few focuses on "data-sensemaking," the entire point of using data in the first place. All the fancy statistical tools and huge volume of data will be useless if we don't understand what we're looking at.

Of the nine items, most of which should be extremely evident to actuaries (example: "3. Information is useful only if it is understood."[7]), the last one has the most impact to us professionally:

> "9. Skilled data sensemakers combine general analytical skills with specific domain knowledge; one without the other is not enough."[8]

I have heard horror stories from others where generalist data scientists bring back analyses on books of business that were completely useless because the analysts didn't understand the nature of the products ("Profitability is driven by policy year!" "Great, we'll go back in time and write more 2015 business"). The situation I've most run into with respect to insurance products and data science is having to operate in a restrictive environment compared to attempting to serve up ads online. Some data, even if available, cannot be legally used. Other data are questionable to use in an industry where reputation with customers is extremely important.

This book may be useful in providing warning to those who do not realize the danger of indiscriminately trying all the computing toys available. At best, some time and effort are wasted if care is not taken. At worst, pricing, underwriting and more could get out of whack if one does not have appropriate knowledge about both the tools and the data the tools are being applied to.

As noted, the book is small, and this reads extremely quickly. If you know of people who need to be warned about what big data can do for their companies, this book would be a good way to warn them—and, as noted above, what they should be doing instead is given at the end. Go out and make sense of your data!

And to JCPenney: No, I don't want a top hat.

Mary Pat Campbell, FSA, MAAA, PRM, is vice president, insurance research at Conning in Hartford, Conn. She can be reached at *marypat.campbell@gmail.com*.

ENDNOTES

1 Few, Stephen. 2018. *Big Data, Big Dupe: A Little Book About a Big Bunch of Nonsense*. El Dorado Hills, CA.: Analytics Press.

2 *http://www.perceptualedge.com/*.

3 For an example, see Few, Stephen. Big Mouths on Big Data. *Visual Business Intelligence* (blog). April 30, 2014. *http://www.perceptualedge.com/blog/?p=1891*.

4 Actuarial Standards Board (ASB). 2016. Actuarial Stand of Practice (ASOP) No. 23: Data Quality. Revised edition. Doc. No. 185. *http://www.actuarialstandardsboard.org/asops/data-quality/*.

5 Few, Stephen. The Slow Data Movement: My Hope for 2013. *Visual Business Intelligence* (blog). Jan. 2, 2013. *http://www.perceptualedge.com/blog/?p=1460*.

6 Ibid.

7 Few, *Big Data, Big Dupe*.

8 Ibid.

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