Industrial Research in Applied Statistics

Taylor Arnold[1]

Introduction

For almost a year, I sat in Washington D.C.'s National airport every Sunday waiting for my flight to Houston. I was 22 years old, with an undergraduate degree in Mathematics, now working in consulting for IBM. I followed physicians in area hospitals each week collecting data. In the evenings, I taught myself R, the statistical programming language, in order to organize and analyze the data so that we could build models to predict the flow of patients through the hospital system. Each week, I found that the theories of mathematics I learned in school were insufficient to address all of the tasks I was assigned. My work involved structuring datasets, making our code run faster, and understanding the specific details of how hospitals function. Data constructed from the social interactions was anything but straightforward.

I decided I needed to learn more and headed to Yale University for a Ph.D. in Statistics. I chose the program because I wanted to develop my mathematics background while blending it with computational statistics. My doctoral dissertation concerned the computational challenges of applying a certain class of statistical models to estimate forms of structural dependence in datasets with a large number of variables. As a result of this study, towards the end of my time in graduate school, I became broadly interested in how methods from exploratory data analysis, data visualization, and statistical learning could be applied to very large datasets. My mentors in graduate school provided me with fundamental skills in statistical computing for structuring and writing efficient code. I felt, however, a substantial disconnect between the data and problems I was working with in an academic setting and the data that statisticians typically work with in industry applications. Most of the work in academic statistics concerns the probabilistic modeling of data, with a particular focus on the estimation of unknown population parameters. In practice, often much more time is spent acquiring, structuring, and visualizing data.

With a desire to learn about the real challenges of applying computational statistics to large, messy data sets, I took a position as a research statistician at Travelers. In that role I applied machine learning algorithms to the task of predicting fraud and the price of future insurance claims. Two years later, I became a senior member of the technical staff at AT&T Labs Research in New York City where I focused on location analytics using cell-phone telemetry data.

In the sections that follow, I explain some of the research questions that I worked on during my time in these two positions and how the skills I learned in graduate school prepared me

---

[1] Taylor Arnold is assistant professor of statistics at the University of Richmond. His email address is tarnold2@richmond.edu.

to address them. I focus on what made these questions, and experiences in general, particularly interesting. I also include a discussion of some critical drawbacks of working in an industry lab. I conclude with my own vision for mutually beneficial partnerships between industry labs and academic researchers.

Travelers Research and Development

My first job out of graduate school was on the Travelers' research and development team responsible for personal automobile insurance. Their large number of messy data sets and unsolved problems that required new innovative approaches made the position especially appealing to me. I found building models within the insurance industry particularly rewarding. I built predictive algorithms that Travelers actually used to make real decisions about what policies to write and how much to charge for them.

While I worked on a number of interesting problems, the majority of my time was spent constructing 'pure premium' models. These models predict, using historic data, the expected amount of money that would be paid out on a particular automobile policy. Actuarial and sales teams incorporate overhead and market-based adjustments to pure premium models in order to arrive at the final rates that are actually charged to consumers for an insurance policy. Several machine learning competitions have featured anonymized datasets with the goal of predicting pure premium values [1, 11]. These competitions, however, obscure the most interesting features of building pricing models. Here, I describe three particularly challenging research questions that underly the construction of premium models within the insurance industry.

The distribution of observed pure premiums makes it difficult to apply many standard statistical techniques without some modifications. Most policies do not receive any claims and have an observed pure premium cost of zero. When claims are made, the amount of money requested has a property known in statistics as a 'heavy tail': a small number of insurance claims require an extremely large payout. These large costs primarily come from extensive medical expenses as a result of automobile accidents. The general distribution of premiums, therefore, should be modeled by a mixed distribution with a discrete mass at zero and a continuous distribution on the positive real numbers. It is possible to split a premium model into separate frequency (the discrete part, predicting whether the policy will have a claim) and severity (the continuous part, predicting how large a claim will be) components. However, splitting the model this way ignores important correlations between frequency and severity. A better alternative is to use a Tweedie distribution, which arises from assuming that the number of claims made on a policy follows a Poisson distribution and the amount of any given claim is distributed with a gamma distribution [12]. Software exists for fitting a generalized linear model where the dependent variable has a Tweedie distribution (Figure 1 shows simulated values from three Tweedie models with varying dispersion parameters) [6]. Interesting research questions arose whenever we wanted to use a new approach or statistical method in our pure premium models. For example, we wanted to incorporate constraints into our models to reduce the number of variables used in the final output. Implementing constrained models required new mathematical derivations and software implementations. Since estimating parameters in the Tweedie model can become numerically unstable, as well as demanding a significant

amount of computational power, these implementations required careful thought and non-trivial extensions of currently available algorithms.

Automobile pure premium models are typically constructed to estimate the cost of insuring a particular automobile. Variables used in this calculation may come from features of the automobile itself (e.g., cost, make, age, and safety features) or from details of the specific policy (e.g., zip code, deductibles, mile driven per year). Some particularly powerful features are also associated directly with the individual drivers on a policy. Examples of predictive driver-level features include credit histories, ages, number of prior claims, and the number of prior traffic violations. The challenge becomes how to summarize drive-level variables at the level of a particular automobile. Should we construct a variable equal to the average age of all drivers? Could we create variables for the minimum and maximum age of all drivers on a policy? Or should we count the number of driver below some age threshold? Any of these new features could be computed for a policy and used in the pricing algorithm. A choice of how to create these aggregated features must be made for dozens of driver-level variables, with the typical trade-offs between variance and bias when including too many or too few correlated variables into a single model. The challenge of summarizing predictive variables at the level of an observed response, a particular example of feature engineering, is a frequent challenge in industry applications. I believe that is one of the single biggest challenges in applied machine learning that is largely overlooked within academic research.

Another important challenge in deriving pure premium models is ensuring that models conform to various government regulations. In the United States, automobile insurance is regulated at the state level, and each of the fifty states has their own set of individual rules. Credit information, for example, is not an allowed predictor variable for pricing policies in Massachusetts. In New Jersey, only a limited number of geographic regions can be defined for pricing and discount purposes. Many states allow insurers to use the age of drivers in pricing models but require that aging can only decrease prices and never increase them. Building models that follow these regulations while retaining most of their predictive properties was a constant challenge within the research and development group at Travelers.

The research problems I encountered at Travelers point to two take aways about graduate education in statistics. We need more statisticians in industry who have the training and interest to conduct original, open-ended research. Many of the most interesting and beneficial projects could not be solved with off-the-shelf statistics tools. They require experience with graduate-level statistical theory as well as general skills in conducting original research. At the same time, we need graduate programs in statistics to include more training in computer science and the empirical social sciences. Computer science and engineering courses can provide skills for writing efficient code to deal with larger datasets, understanding how to implement new estimation algorithms, knowing the principles of building databases, and experience writing and testing code that may be used in production. Social science applications give experience with the techniques and challenges of using data and models to understand human behavior. They also are more likely to explain the political and legal challenges that may underly the collection of data or deployment of empirically trained models.

AT&T Labs Research

In April 2014, I transitioned to the statistics department at AT&T Labs Research. The group has a long history of exceptional work in the field of applied and computational statistics and traces its roots back to the original Bell Labs [7]. Rick Becker was one of the three original authors of the S language, the pre-cursor to the popular R programming language for statistical computing, which was developed at AT&T in the 1980s [4]. Simon Urbanek is one of the small set of core developers of the current R-Project. Chris Volinski and Robert Bell were both on the wining team for the million dollar Netflix movie recommendation competition [5]. A large draw for my move to AT&T was the chance to work with these and other fantastic scholars in the field of computational statistics.

Another motivation for my interest in working at AT&T was the desire to work with extremely large datasets, a continuation of my graduate school research. My world-class colleagues at the labs in a range of fields gave me the opportunity to work collaboratively on new research questions and to keep learning about new areas. My group focused on cell-phone location analytics, which required working with large data sources. Our primary dataset was built from observations known as call detail records, or CDRs. A CDR is generated whenever there is an interaction between a cellphone and cell tower. CDR's can include a cellular voice call, a text message, or the transferring of generic data. With the wide-spread coverage of 4G networks and proliferation of cellphone applications, most cellphones today are involved in a nearly constant stream of CDRs that cover the majority of the day [9]. By associating each cell tower in a CDR with its location, these records make it possible to determine approximately where a device is at any given moment in time (see Figure 2 for an example) [13]. This data has been widely used as legal evidence and was recently employed to assist aid workers helping with the West African Ebola virus epidemic from 2013-2016 [14].

The location analytics data that I worked with was so large that it needed to be distributed over hundreds of machines. Overall, the data I had access to amounted to several petabytes (1000 terabytes) and took days to process even over our large cluster. Given my expertise, I was tasked with building a data pipeline from scratch that ingested the raw CDR records and produced a normalized databased of each observed device's location; a daunting but exciting challenge. In order to work with data stored over a large distributed system, I had to learn two new frameworks (Hadoop and HBase) and learn how to write code for them in a programming language I was not very familiar with (Java). Because data arrived hourly, and needed to be processed immediately, my data pipeline needed to automatically run throughout the day. In applied statistics we are often reminded and taught how to interactively check whether there are potential issues in a data source. With the system I was building, it was important to build in automated tests that would check new data as it came in. This was necessary because there were frequent upstream data issues with the raw CDR files that were being delivered. For example, all of the data from a particular city for 6 hours in a day might go missing due to an internal networking issue. Or, the format of a field would occasionally change and cause some of the code to break. The completed data pipeline opened up many research questions for our team. Quick access to small selections of the corpus (through the distributed database) allowed for exploratory analysis that allowed us to start thinking critically about what the data was able to show. For example,

we found that using the location data was great for detecting movement along highways and public transit routes. It was less useful, however, in the accurate detection of static devices.

Once the location data was cleaned and stored on our research servers, we created tools for modeling and visualizing the data. Mike Kane, Simon Urbanek, and I built a set of tools in R for working with large distributed datasets [3]. These functions focused on being able to process a fixed number of lines of data, allowing for chunk-wise operations on large datasets. Using these tools, we developed a distributed algorithm that allows for applying penalized regression to arbitrarily large datasets. Our work on this problem eventually led to a textbook focused on the computational details of working at scale with large distributed datasets [2]. I also worked on integrating a new spatio-temporal visualization algorithm known as 'nanocubes' into an R package [10]. This allowed for our researchers at AT&T labs to easily explore small subsets of our data within their browsers.

During my time at AT&T labs, I had the chance to develop new software and study approaches for working with extremely large datasets. Doing research at the labs gave me expertise in the modeling and management of large datasets at a scale that would have been nearly impossible to work with in academia. My experience in an industry lab, in short, offered educational opportunities beyond what was available within a formal graduate program.

Drawbacks

Positions in industry labs are not without their own unique issues. For example, in an industry position there is a complete lack of personal ownership over ideas, work, and software. Projects that take months or years of work often result in no tangible outcomes that are seen outside of the company. Business concerns may force researchers to abandon interesting lines of work in favor of other tasks.

I engaged in a wide array of interesting research projects at Travelers and AT&T. Unfortunately, almost none of this work is publicly available. Industry labs typically forbid the publishing of research that uses internal data; without the datasets as examples, most of the methodological innovations made in my work were hard to motivate or even explain.[2] At Travelers we were not even allowed to publish software that we had built. AT&T Labs, with its long tradition in computing, was more willing to allow the publication of software. The two papers I have from my time there both focus on specific software libraries we built. However, even this type of publication is increasingly rare.

Another concern I had while employed within an industry research lab was whether my work was being used in ethical and appropriate ways. Take, for example, the cell phone location analytics projects. All of the applications I directly worked on were either banal internal studies, such as testing network dead spots, or external consulting projects that

---

[2] Prior to my time at AT&T there were more opportunities to publish data-driven research. See the following paper by my colleagues for a great example that illustrates the nature of the internal projects we worked on: [8].

made use of highly aggregated tabulations to show the general movement of people through space for urban planning purposes. However, there was no way for me to stop, or to even be aware of, my code being used for more objectionable applications. These concerns may also extend to all publicly available research. When publishing method papers or open source software there is also no way to ensure that derivative work is being used responsibly. But, at least in the publicly available case the research is not being internally motivated or funded by these applications. Also, I believe that the net benefit of publicly available research generally outweighs the concerns of misuse. The potential for abuse is harder to justify with research that is never made externally available.

After two and half years at AT&T, I left to join the faculty at the University of Richmond. The year prior to my departure, I taught two courses as a part-time lecturer at Yale University. This experience reignited my passion for teaching and convinced me that I wanted to make that a permanent part of my work. I also wanted the opportunity to make more of my research public in order to get external feedback and to see my methods and code made usable in other domains.

Academia and Future Directions

As I have transitioned back into academia, my experience in industry continues to shape my approach to research and teaching. For example, I no longer see a sharp divide between my work as a researcher and an educator. Teaching students how to work with messy, unstructured data makes me reflect on how to turn my individual applied projects into larger-scale methodological frameworks. Similarly, watching students use and struggle with existing software libraries helps me to understand the shortcomings in available tools and how they can be addressed.

One challenge of leaving industry labs has been finding ways to continue my scholarship in large-scale statistical computing without access to industrial datasets. Currently, I have solved this problem by finding publicly available datasets that share common traits of those seen in industrial applications. For example, I have a current project that involves working with the entire corpus of page histories from Wikipedia, which amounts to several terabytes of textual data. The size of the corpus and complexities of the dealing with the data address many of the same challenges I faced working with call detail records at AT&T labs. I am involved in another project that uses computer vision techniques to extract features from video files. While the raw features are associated with individual frames, the predictive modeling tasks I am interested in — such as scene detection and character movement — require building models for sequences of images. The challenges here mirror the issues of aggregating driver-level data to a particular automobile that I faced at Travelers.

My experience in industry has impacted my own teaching philosophy. Across all of my classes, my ultimate goal is to help students develop the skills needed to engage in the ethical and insightful analysis of data. For me this means that I need to teach the entire pipeline of working with data instead of focusing only on probabilistic modeling. In my introductory courses we spend a lot of time talking about how to correctly structure data in a spreadsheet. We also spend several weeks working on how to interpret statistical

visualizations in both written and oral formats. In my courses on data science, students learn how to fetch data through APIs and spend several weeks building interactive websites with Javascript. These experiences improve their ability to present useful data visualizations as well as make them comfortable working in new programming languages and approaching tasks outside their typical comfort zone.

I found my experience in industrial research labs to both rewarding and generally enjoyable. At the same time, I also understand the difficulties of life in an industry lab and appreciate the relative freedoms afforded by a position in the academy. Some of the most influential scholars to my own work have had similar histories that intersect between industry and academic positions, including John Tukey (who split his time between Princeton and AT&T Labs), danah boyd (a researcher at Microsoft with an ongoing position at NYU), Yann LeCun (a computer science professor at NYU and director of research at Facebook), and Hadley Wickham (RStudio and Rice University). These scholars have produced some of the most important work in applied statistics. Hadley Wickham's triptych of papers and associated software for applied data analysis — "A layered grammar of graphics" [15], "Tidy Data" [17], and "The split-apply-combine strategy for data analysis" [16] — have been highly influential, for example, to my own work. I hope to see more direct partnerships where academic faculty can participate in research with industry labs. These exchanges have the benefit of bringing to light many understudied problems in applied statistics. It also provides an external source for critically reviewing the ways data are being used in industry and its potential effects on society as a whole.

References

1. *Allstate Claim Prediction Challenge*. https://www.kaggle.com/c/claim- prediction-challenge. Accessed: 2018-09-30.
2. T. Arnold, M. Kane, and B. Lewis. *A Computational Approach to Statistical Learning*, Chapman and Hall/CRC, 2019.
3. T. Arnold, M. Kane, and S. Urbanek, iotools: High-Performance I/O Tools for R, *The R Journal* **9** (2017), no. 1, 6–13.
4. R. Becker, J. Chambers, and A. Wilks, *The New S language: A Programming Environment for Data Analysis and Graphics*, Wadsworth & Brooks, 1988.
5. R. Bell, J. Bennet, Y. Koren, C. Volinsky, The Million Dollar Programming Prize, *IEEE Spectrum* **46** (2009), no. 5, 28-33.
6. J. Faraway, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman and Hall/CRC, 2016.
7. J. Gertner. *The Idea Factory: Bell Labs and the Great Age of American Innovation*, Penguin, 2012.
8. S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, Identifying Important Places in People's Lives from Cellular Network Data, *Proceedings of the 9th International Conference on Pervasive Computing* (2011), 133–151.
9. O. Järv, R. Ahas, and F Witlox, Understanding Monthly Variability in Human Activity Spaces: A Twelve-Month Study Using Mobile Phone Call Detail Records, *Transportation Research Part C: Emerging Technologies* **38** (2014), 122–135.
10. L. Lins, J. Klosowski, and C. Scheidegger, Nanocubes for real-time exploration of spatiotemporal datasets*, IEEE Trans- actions on Visualization and Computer Graphics* **19** (2013), no. 12, 2456– 2465.
11. *Prudential Life Insurance Assessment*. https://www.kaggle.com/c/prudential- life-insurance-assessment. Accessed: 2018-09-30.
12. G. Smyth and B. Jørgensen, Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling, *ASTIN Bulletin: The Journal of the IAA* **32** (2002), no. 1, 143– 157.
13. H. Wang, F. Calabrese, G. Lorenzo, and C. Ratti, Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records, *Proceedings of the 13th IEEE International Conference on Intelligent Transportation Systems* (2010), 318–323.
14. A. Wesolowski, C. Buckee, L. Bengtsson, E. Wetter, X. Lu, and A. Tatem, Commentary: Containing the Ebola outbreak — The Potential and Challenge of Mobile Network Data, *PLoS Currents* **6** (2014).
15. H. Wickham, A Layered Grammar of Graphics, *Journal of Computational and Graphical Statistics* **19** (2010), no.1, 3–28.
16. H. Wickham, The Split-Apply-Combine Strategy for Data Analysis, *Journal of Statistical Software* 40.1 (2011), no. 1, 1–29.
17. H. Wickham, Tidy Data, *Journal of Statistical Software* **59** (2014), no. 10, 1–23.

Captions

Figure 1 — A histogram of simulated random draws from a Tweedie distribution with for free different dispersion parameters. When modelling insurance premiums, high dispersion values are used to describe policies that incur claims on only is a small percentage of policies.

Figure 2 — Maps show registered cell phone towers (solid dots) in the vicinity of Rochester, New Hampshire. Each path describes an artificial collection of towers that sequentially handle cell phone traffic for a fictional driver commuting from Madbury to Rochester. In the left panel, the driver takes a sequence of smaller roads, Littleworth, Calef Highway, and Gonic Road. The right shows an alternative path that travels by Route 16 (thick grey line).