Sound the Gong

The Laws of the Jungle: Data Science Edition

Contributing Editor Ruobin Gong shared some pieces of advice on cutting a path through the data science jungle:

The winter months of an academic year may appear slow and dormant. But for a thriving discipline like statistics, that cannot be further from the truth. From November to February, many statistics departments are preoccupied with two things: placing their graduating PhD students to academic (and non-academic) positions, and quite often, evaluating job applicants from their peer institutions. This is the time of the year when we all are once more amazed by the stellar, and often imaginative, accomplishments of the next generation. Many candidates for tenure-track statistics faculty positions show strong publication records outside of traditional statistics venues. While most were trained as statisticians, some competitive candidates obtained their doctoral degrees from departments of machine learning and computer science. As prospective employers, we delight in the stimulating consequences of statistics being a core player in data science. As researchers and teachers, on the other hand, we are prompted to examine how well our discipline has been doing in an ever-changing landscape. Is our research making good impact? Are we training our students to make good impact?

The rise of data science reflects the necessity of a scholarly discourse that unites the strengths of many disciplines under the quantitative perspective. A manufactured notion at inception—awkward-fitting and forced into existence—data science has evolved into a natural, broadly accepted, and spirited one over the past decade. During this slow yet steady evolution, its constituent disciplines also found a renewed cadence of research. Many developed perspectives and approaches that were more interdisciplinary than ever before. The cutting-edge topics of data science—privacy, data ethics, digital humanity, and personalized healthcare, to name just a few examples—all require the collective intelligence of statistics, computer science, and the subject-matter experts to put their heads together (and their feet into one another's shoes) to tackle the real questions. Collaboration and competition are two dynamic forces that propel this progress.

Data science is a jungle. It is a vast ecosystem, sustained by the vitality of its members and their symbiotic relationship with one another. A jungle harbors unsurpassed diversity and creativity that comes with it. But a jungle can also be a cruel place, where the livelihood of every being and every species hinges upon a keen sense to survive, to procure resources, and to adapt to the shifting environment. What are the laws of the jungle that is data science? As residents of the data science jungle, what should we do to survive and to thrive? I offer five reflections from a statistician's experience.

First, ***break the mental boundaries***. Traditional mathematical statistics emphasizes scholarly contributions made inside of statistics, that is, the development of theory and methodology that benefit other statisticians and future statistical research. To make impactful contribution in data science, a statistician must look and think beyond disciplinary confines. Long before data science, applied statistical research has been the pioneer in disciplinary boundary breaking. Today, boundary-breaking efforts can and should extend beyond applied research, into the realm of theory and methodology development for problems that stem from non-statistical origins. Get excited about questions regardless of where they come from, so long as they are real, important, intriguing, and amenable to tackling with quantitative evidence.

Second, we should ***play to our strengths***. Know and leverage our disciplinary-specific training in our quest to making an impact. Whether we like to admit it or not, every discipline trains their students into possessing a unique mindset. This mindset, often no less holistic and potent than a world view, encapsulates the epistemology and the wisdom distilled over the history of the field. Elements such as uncertainty quantification, sampling and randomization, probabilistic modeling and regression are among the key wisdoms that embody the spirit of statistics. Taking them to heart is a statistician's rite of passage. Use our wisdom wisely.

Third, ***open mind and open arms***. Newsworthy breakthroughs in machine learning and data science occasionally arrive at our doorstep like a knight with glistering armor. But even in Shakespeare's time, people recognized that not all that glistered was gold. These thoughts may be an instinct of survival or denial; either way, succumbing to them without rational deliberation may incur a loss on our part. In the movie *Arrival*, if our heroine doesn't risk her

life and befriend the scary-looking extraterrestrial beings, how would the Earthlings learn about the magical and powerful world embodied by their teleological language?

An "open mind, open arms" policy calls for disciplines to teach to each other their different vocabularies and practices. Be warned, however, that when disciplinary lines become blurry, an inevitable consequence is that wheels get reinvented, or worse, appropriated. We might feel betrayed when our creation pops up in a different literature, under a new name and without due acknowledgment. (As a not-so-new joke puts it, "machine learning" too often just means logistic regression.) But statisticians are not innocent when it comes to knowledge appropriation either. Indeed, some ideas that are widely believed as quintessentially statistical originated outside of statistics. Let us not be deterred by the perceived risk or harm of intellectual property theft. Good ideas will shine wherever they fulfill a purpose, whether foreseen or otherwise. Let us rejoice in knowing that we came up with them.

Fourth, ***practice productive skepticism***. Statisticians are critical thinkers. We take great care in our own work to lay out all the assumptions and admit earnestly to weaknesses of our solution. In the work of others, we are never shy to point out faults and deficiencies. The skepticism we hold against others and ourselves is a testament to our professional ethics, and is precisely why statistical methods command trust and respect in support of scientific advances.

When wielded appropriately, skepticism keeps us on our toes so that we never grow complacent. It also means that we do things slowly, and sometimes give up on doing anything at all if a perfect solution is deemed beyond reach. Today, publications in machine learning and computer science conferences decorate many statisticians' CVs. In a way, they are badges of honor that attest to the versatility of our contributions. For those who've had their skin in the game, however, we understand too deeply the sacrifice that must be made to partake in fast-paced conference publishing. As soon as the process starts, speed becomes the essence: there is little time to think, and virtually none to practice skepticism.

To join fast publishing may feel like lowering our standards, or even "selling our soul." But to apply the "open mind, open arms" advice here, the publication schism is a cultural one that merely reflects the disciplinary preference at striking the speed-versus-quality tradeoff. In impact-oriented scholarship, a case can be made for faster iterations of research so that good, albeit less-than-impeccable, solutions reach the audience that needs them in a timely manner.

As the old saying goes, don't let Perfect be the enemy of Good.

Finally, ***be kind and be generous*** to the young, the ignorant, and the brave. In the eyes of discipline-specific experts, interdisciplinary work would likely appear foreign. It would not conform to "normal science" or the methods of inquiry that are traditionally agreed-upon. Interdisciplinary work can take unexpected forms, either as a novel application utilizing tools from one discipline to solve the problem of another, or as a synthesis or reconciliation of existing approaches from multiple disciplines. An interdisciplinary contribution might not check all the boxes that a discipline-specific one is expected to. Perhaps a theory cannot be made water-tight in an uncharted territory; a methodology may be employed straight out of the box when its innovation is beside the point; or the substantive findings may not presented in ways that are familiar. The evaluation of such work under discipline-specific lenses might expose problems, and it may be too easy to dismiss it on these grounds even though the virtue of the effort is buried under caveats.

To be clear, I do not advocate for sloppy work disguised as speedy publication or interdisciplinary data science. Rather, my hope is that we prepare ourselves to see the value in a genuine effort to innovate, and to offer constructive—rather than destructive—guidance to those who need it.

In the jungle of data science, aspiring data scientists and early-career scholars who devote their passion to interdisciplinary work need all the help they can get to survive, to grow, and to proliferate statistical thinking. After all, a thriving jungle is a truly amazing place, a paradise in which every sentient being is offered an opportunity to not only live but also reach its full potential. Navigating a fruitful path through the jungle is a career-long project for every data scientist.