White Paper

It Is a Mistake to…. Extrapolate

by John Elder

Founder & CEO

December 2014

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

This article is Part 8 (of 11) of a series by the author on the Top 10 Data Mining Mistakes, drawn from the *Handbook of Statistical Analysis and Data Mining Applications*.

Modeling "connects the dots" between known cases to build up a plausible estimate of what will happen in related, but unseen, locations in data space. Obviously, models – and especially nonlinear ones — are very unreliable outside the bounds of any known data. (Boundary checks are the very minimum protection against "over-answering", as discussed in the next installment.)

But, there are other types of extrapolations that are equally dangerous. We tend to learn too much from our first few experiences with a technique or problem. The hypotheses we form – which our brains are desperate to do to simplify our world – are irrationally hard to dethrone when conflicting data accumulates. Similarly, it is very difficult to "unlearn" things we've come to believe after an upstream error in our process is discovered. (This is not a problem for our obedient and blindingly fast assistant: the computer. It blissfully forgets everything except what it's presented at the moment.) The only antidote to retaining outdated stereotypes about our data seems to be regular communication with colleagues and clients about our work, to uncover and organize the unconscious hypotheses guiding our explorations.[1]

Extrapolating also from small dimensions, *d*, to large is fraught with danger, as intuition gained in low-*d* is useless, if not counter-productive, in high-*d*. (That is, an idea may make sense on a white board, and not work on a many-columned database.) For instance, take the intuitive *Nearest Neighbor* algorithm, where the output value of the closest known point is taken as the answer for a new point. In high-*d*, no point is typically actually close to another; that is, the distances are all very similar and, by a univariate scale, not small. "If the space is close, it's empty, it's not empty; it's not close" is how (Scott, 1992) describes this aspect of the "curse of dimensionality".
(Friedman, 1994) illustrates four properties of high-*d* space:

- Sample sizes yielding the same density increase exponentially with *d*. Radiuses enclosing a given fraction of data are disproportionately large.
- Almost every point is closer to an edge of the sample space than to even the nearest other point.
- Almost every point is an outlier in its own projection.[2]

As our most powerful technique – visualization – and our deep intuition about spatial relationships (in low-*d*) are rendered powerless in high-*d*, researchers are forced to employ much more simplistic tools at the early stages of a problem until the key variables can be identified and the dimensions thereby reduced.

The last extrapolation is philosophical. Most researchers in Data Mining, Machine Learning, Artificial Intelligence, etc., hold the theory of evolution as an inspiration, if not motivating faith. The idea that the awesome complexity observed of life might have self-organized through randomization and indirect optimization can bolster one's belief that something similar might be accomplished in software (and many orders of magnitude

faster). This deep belief can easily survive evidence to the contrary. I have heard many early proponents of *Neural Networks*, for instance, justify their belief that their technique will eventually provide the answer since "that's how the brain works".[3] Others have such faith in their mining algorithm that they concentrate only on obtaining all the raw materials that collectively contain the information about a problem and don't focus sufficiently on creating higher-order features of the raw data. They feed, say, the intensity values of each pixel of an image into an algorithm, in hopes of classifying the image – which is almost surely doomed to fail – instead of calculating higher-order features – such as edges, regions of low variance, or matches to templates – which might give the algorithms a chance.

A better mental model of the power and limitations of Data Mining is small-scale evolution, rather than large-scale. We can observe, for instance, selective breeding of a population of mutts over several generations, to create a specialized breed such as a greyhound. But, it is a bold and unproven hypothesis that one could do so, even with a billion years, beginning instead with pond scum. So why give a model the equivalent? Better to use higher-order features of the raw data; it is well known now that good "feature engineering" can strongly impact the success of one's model. As a rule, use all the domain knowledge and creativity your team can muster to generate a rich set of candidate data features. Data Mining algorithms are strong at sifting through alternative building blocks, but not at coming up with them in the first place.

Figure 1: The Kasparov vs. Deep Blue chess match: a showdown on "thinking"?

Figure 1 depicts the March 25, 1996 cover of *Time* magazine, which provocatively asked: "Can Machines Think? They already do, say scientists. So what (if anything) is special about the human mind?"[4] Magazine covers can perhaps be forgiven for hyperbole; they're crafted to sell copies. But inside, someone who should know better (an MIT Computer Science professor) was quoted as saying, "Of course machines can think. After all, humans are just machines made of meat." This is an extreme version of the "high-AI (artificial intelligence)" view (or perhaps, the "low-human" view). But, anyone who's worked hard with computers knows that the analytic strengths of computers and humans are more complimentary than alike.[5] Humans are vastly superior at tasks like image recognition and speech understanding, which require context and "common sense" or background knowledge to interpret the data, but computers can operate in vast numbers of dimensions – very simply, but with great precision. It's clear to me that the great promise being fulfilled by data mining is to vastly augment the productivity of – but not to replace – skilled human analysts. To believe otherwise – at the extreme, in an eventual "singularity event" in time where humans and machines will merge to create a type of immortal consciousness – is an extrapolation more akin to faith than science.

About the Author

Dr. John Elder, Founder and CEO of Elder Research, leads the largest and most experienced data science consulting firm in the U.S. For 20 years, the team has applied advanced analytics to achieve high ROI for investment, commercial and security clients in fields from text mining and stock selection, to credit scoring and fraud detection. John has Engineering degrees from Rice and the University of Virginia, where he's an adjunct professor. He's authored innovative tools, is a popular keynote speaker, and has chaired International Analytics conferences. Dr. Elder served five years on a panel appointed by President Bush to guide technology for National Security. He has co-authored three books (on data mining, ensemble modeling, and text mining), two of which won Prose "book of the year" awards.

---

[1] This is so critical that, if you don't have a colleague, rent one! A tape-recorder or a dog will even be better than keeping all of your dialog internal.

[2] That is, each point, when projecting itself onto the distribution of other points, thinks of itself as weird… kind of like Junior High.

[3] Though research from the 1990's argues instead that each human neuron (of which there are billions) is more like a supercomputer than a simple potentiometer.

[4] *Time* magazine was reporting on the previous month's first chess match between Gary Kasparov (perhaps the best chess player in history) and "Deep Blue", a specialized IBM chess computer. Kasparov lost the first game – the first time a Grand Master had been beaten by a program – but handily won the full match. Still, that was to be the high-water mark of human chess achievement. A year later, "Deeper Blue" won the re- match, and its likely humans will never reign again. (IBM enjoyed the publicity and didn't risk a requested third match and, of course, computer power has grown by orders of magnitude since then.) Unlike checkers, chess is still nearly infinite enough that computers can't play it perfectly, but they can simply march through a decision tree of possibilities as deep as time allows (routinely to a dozen or more plies, or paired move combinations). Though it seems like a good test of intelligence, the game of chess actually plays well to the strengths of a finite state machine: the world of possibilities is vast, but bounded, the pieces have precise properties, and there is close consensus on many of the game tradeoffs (i.e., a bishop is worth about three times as much as an unadvanced pawn). There are also vast libraries of carefully worked special situations, such as openings, and end-game scenarios, where a computer can play precisely and not err from the known best path. The automation component with the greatest uncertainty is the precise tradeoff to employ between the multiple objectives — such as attack position (strong forward center?), defense strength (take time to castle?), and the pursuit of materiel (capture that pawn?) – that vie for control of the next move. To define this score function by which to sort the leaf nodes of the decision tree, the Deep Blue team employed supervised learning. They took the best role models available (Grand Master matches) and trained on the choices made by the GM's over many thousands of recorded games to discover what parameter values for the move optimizer would best replicate this "gold standard" collection of choices. Lastly, the designers had the luxury of studying many of Kasparov's games, and purportedly devised special anti-Kasparov moves. (Incidentally, Kasparov was refused the chance to study prior Deep Blue games.) Given how "computational" chess is then, it's a wonder any human does well against a machine! But I digress! My original point is that chess skill is a poor metric for "thinking". But I also enjoy that data mining (inductive learning) helped uncover the best tradeoffs.

[5] Perhaps, being from MIT, he's never worked with humans.

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

National Capital Region
2101 Wilson Boulevard
Suite 900
Arlington, VA 22201

855.973.7673

Headquarters
300 W. Main Street
Suite 301
Charlottesville, VA 22903

434.973.7673

Maryland Office
839 Elkridge Landing
Suite 215
Linthicum, MD 21090

855.973.7673