Seven Concepts a Dept. of Defense Program Manager Needs to Successfully Develop Social Science Models: Part 1

Posted on December 17, 2015

***There you go again*. Ronald Reagan [1]**

It was with a mix of deja vu, amusement and resignation that I saw the latest Dept. of Defense (DoD) pronouncements—try here and here —about their intentions to take a very important innovation in machine learning, recurrent neural networks [2], and use this as the centerpiece of a major new machine-human interaction initiative.

It's that word "human" that's setting me off, as when it comes to technical applications, DoD can't ever seem to do "human. Wow, creative new initiative but…been there, done that, and I've seen so many similar things come and go over the years—decades in fact—always with the same result [3]: a big heap of money spent that might as well have been stuffed into [Chinese] fireworks and sent skyward on the Fourth of July.

It's probably getting worse for me now that I'm just a couple hours south of the Beltway and can attend meetings on short notice. There's a pretty consistent script: You start with something ambitious though awkwardly defined—the sort of thing that in academia I would have sent back to a grad student for a re-write—but generally plausible. You've got a bunch of people in a room [4], and some of these are absolutely top in their field and are sincerely trying to be helpful and want to get to a feasible project definition, figure out some appropriate technology, and move everything forward. For an hour or two, things are going pretty well.

But invariably, after a promising start, we head down a very predictable rabbit hole and end up —yet again—at the Mad Hatter's Tea Party. And typically stay there. Curiously, in my experience, this is endemic just to DoD, making it all that more puzzling.

Or not, since the proximate cause of the descent into madness can always be traced to inevitable presence of a cluster of pallid, over-weight men (they're always white men) of late-middle age—the Pillsbury Doughboy look—representing all of the usual suspects of the permanent civilian defense contractor class. Whenever things start looking promising, these dudes start asking the stupidest of questions, exhibiting unbounded cluelessness concerning the topic at hand, and going into long discourses on the impossibility of doing the sorts of research that other parts of the US government have been doing with great success for decades [5], often on precisely the same topic under discussion, and the likes of Amazon, Google and Facebook have as a gadzillion-dollar business model.

So, methinks, what gives? Are these guys really that stupid and sent by their bosses to get them out of the building? If we were ISIS, of course, these folks would be placed at the top of the roster list for suicide missions, but we're not ISIS. So why are they here?

With the intensified exposure to this phenomenon in the past couple of years, I've finally figured it out: the Doughboys are the equivalent of the Communist era minders in Soviet puppet states, and, consistent with the tactics of the Old Left, their entire purpose is to make sure that these meetings remain completely pointless [6] and avoid the disastrous possibility that DoD might, say, spend $10-million on some social science [7] research that would prevent a $100-million mistake or even worse, spend $100-million on research that would prevent one or more $1-trillion mistakes, or, worst of all, develop a sophisticated social science research culture within DoD comparable to that found in numerous other parts of the government, to say nothing of academia and the private sector. No, discretionary DoD money needs to kept where it has been all along, funding mind-boggling levels of contractor fraud, weapons systems that don't work , and 12-figure cost over-runs.

Well, gotta give the Doughboys credit, they've done one heck of a good job for their corporate masters! But that doesn't mean we have to like it.

So in the spirit of the Yule season and as a public service, MouseCorp—which, full disclosure, is not entirely uninterested in having DoD learn how to do research appropriate to the 21st century—will provide guidelines to a small number of critical concepts which individuals trying to manage these new programs might, just might, be able to use to shut up these parasitic bastards [8]. There are more than seven—though the list is still fairly small—so for convenience this will be done in two segments, the first focusing on fairly specific technical concepts, the second, in a week or so, on some more general literatures.

For the sake of exposition, let's assume somehow one or more of these new proposed projects makes it through the preliminary efforts to kill it, and you are managing it, and you quickly figure out that you could get a big boost if you'd incorporate some state-of-the-art social science modeling methods into the project. You've got the Doughboys with their corporate overlords and Gucci-clad lobbyists hamstringing you at every step, trying to make sure the project fails, but you've got some of that anachronistic Greco-Roman Stoic civic virtue thing going, and you'd really like the project to succeed. And since this is DoD, you've got a budget at least an order of magnitude greater than what the National Science Foundation or "the part of the national security community that shall not be named" would have available—granted, fully half of the funding will go to program reviews and PowerPoint slides [9]—so resources are not the issue. Deciding on workable approaches is the issue. So here are some key concepts, with more to

follow:

1. The Forecaster's Quartet

Become familiar with the following:

- Daniel Kahneman. *Thinking Fast and Slow*: 30 years of research which won a Nobel Prize and is a great read, long residing on the business best-seller list [10]
- Nassem Nicholas Taleb. *The Black Swan*.
- Philip Tetlock. *Expert Political Judgment* [11]
- Nate Silver. *The Signal and the Noise*: popular-level antidote to the contention that human behavior is not predictable

And finally at the article length and a more challenging technical level, but in terms of political prediction using formal models, easily the most important work in the past quarter century:

> Michael D. Ward, Brian D. Greenhill, Kristin M. Bakke. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research* 47(4) 363–375 [12]

Like most such paradigm-smashing contributions, they had a very difficult time getting it published.

2. Model specification and the centrality of theory

This comes first in the list of technical terms, since if you don't get the model specified right, everything else is doomed, and the best weapon in your arsenal here is a thorough review of existing theory. The Doughboys hate that—their motto is "A week in the lab saves an hour in the library" since that attitude keeps the meter ticking and the money flowing. DoD projects [13] tend to approach every problem like they were the very first people to ever think about it, whereas in more cases than you'd expect, someone was thinking about it 2,500 years ago and helpfully wrote those ideas down. If not 2,500 years ago, then almost certainly in the past 50 years. A lot of it is garbage but you will discover that the slop dished out by people utterly unfamiliar with existing theory generally isn't very helpful either.

A good theory tells you which haystack you need to look in to find the needle. Once you've got that, you don't want to just pile on more hay. Conversely, specify the model incorrectly, and you're just doing the wrong thing better. [14]

3. Latent dimensions and colinearity

A "latent dimension" is the technical term for what would commonly be called a "generalization," and in statistical terms involves a set of indicators which co-vary. "Economic development" is the standard—and appropriate—example: we know from experience that advanced industrialized economies differ in a large number of ways from developing economies, and while GDP/capita is the most common way of measuring these, plenty of others would work just as well. Famously in the conflict forecasting realm, infant mortality rate. "Democracy", "quality of governance," "globalized economy" and "political instability" are other common relevant latent dimensions in the conflict forecasting literature.

The key point about latent dimensions is once you've got a couple of measures for a dimension—or even a single really reliable measure—adding more variables gives you very little information. In fact, in the linear models commonly used in statistical studies—discussed in the next entry—this becomes counter-productive because of a problem called co-linearity, which plays havoc with the variance of your coefficient estimates.

Latent dimensions are also the reason Achen's "Rule of Three"—discussed next time—is so successful. The Doughboys hate this: their interest is in piling on redundant indicators to drive up the costs and delay the project, and hairball models are a solution, not a problem. Resist.

4. "Error" is part of the process

For starters, in social science models, it's not really "error" in the same sense that we think of "error" in tightly controlled physical processes. You are better off thinking about "error" as "things not included in the model" because we're dealing with open complex processes, not the engine cylinder clearances on a BMW 760. They've not been included because the information is not reliably available, or is not cost-effective [15], or the indicators will actually introduce more error than they reduce, or the process is intrinsically random. [16]

So, it's not "error", it's "everything else". But keeping with convention, we'll keep calling it "error."

5. Accuracy, Sensitivity, Precision and ROC curves

"You can't manage what you can't measure" right? In the bad-old-days, social scientists tended to measure errors using a single linear correlation-based measure called R-squared but

contemporary models for predicting whether or not something will happen are generally evaluated on the aforementioned series of measures that the machine-learning folks have been using for quite some time. Look'em up: Wikipedia has vast resources here. The "ROC curve"—and the most common statistic based on it, the AUC [area-under-curve: Google it]—is particularly challenging to grok because it has an invisible dimension—the change in the threshold at which the model predicts the event will happen—but this is now nearly universally used, so put in the effort.

Most prediction problems relevant to national security concern "rare events" and these present a number of measurement challenges, not all of them resolved. Again, get up to speed on this and know the "gotchas" E.g. it is trivial to generate very high *accuracy* on any rare events problem without producing anything useful for policy purposes. [17]

6. Estimation

All non-trivial models have coefficients [18] and these must be estimated from the data. All estimates have—that word again—errors, or more accurately, variation, and this can also be estimated, though the veracity of that estimate is dependent on the extent to which the characteristics of the data—nowadays the term is usually "data generating process"—corresponds to the assumptions used to derive the estimation method, and as anyone with any experience in the field knows, the Data Fairy is frequently not very kind. All estimation methods can exhibit pathological behaviors when confronted with sufficiently weird data but fortunately for the widely used modeling methods—discussed in the next entry—these are extensively studied and understood.[19] New methods?: you'll be the test case, and these may go very badly. [20]

Historically, social science statistical work was done "in-sample", where the data used to estimate the model and the data used to test it were the same. This leads to "over-fitting" or "fitting the error" and often did not generalize. Contemporary work—and virtually all machine learning work—uses any of a number of more robust "split sample" designs where the "training" and "test" data are separate.

7. Significance testing versus Bayesian approaches

Historically, virtually all social science statistics were done using the "null hypothesis significance testing" approach [21], which is both highly counter-intuitive and very often misinterpreted even by people who should know better, but was the only practical method prior

to the availability of large amounts of computing power and some innovations in estimation methods that only occurred a couple decades ago. Significance testing is gradually being replaced by Bayesian methods [22], which are more likely to provide the information you are actually looking for, and in principle should integrate well with qualitative approach: the buzzword here is "informed priors." The important thing: understand what each approach does and does not do.

Conclusion

Enough for now, and I'm a bit over the word limit already. There's more to come, and I've already assigned too much homework. Still, just with this material alone, you can start the push-back and keep your project from going off the rails: just look those Doughboys straight in the eye and growl "So, feeling lucky, punk?" [23]

Beyond the Snark

Not a lot this time: Wikipedia is generally pretty good on the subject of statistics—well, it is amazingly comprehensive and technically accurate; sometimes the exposition leaves a bit to be desired. Quite a number of people who teach methodology, where a huge amount of labor goes into producing a good set of lecture notes, have posted these on the Web. As has MIT. [http://ocw.mit.edu/courses/sloan-school-of-management/15-075j-statistical-thinking-and-data-analysis-fall-2011/] I've generally used standard terminology here, though in a few areas things get a little confusing because the statistics and machine-learning communities, having developed more or less independently (this is actually quite remarkable but hey, academic silos are built to withstand a *lot* of pressure) not infrequently use different terms for the same concepts. But in general with all of these terms "Google it" will get you lots of information.

My original and somewhat technical exposition on what is wrong with most of the existing approaches: http://7ds.parusanalytics.com/Schrodt.7DS.JPR.May13.pdf. By the way, a few people have interpreted this article as my rejecting quantitative approaches. Far from it, starting with the fact that doing quantitative work is how I keep food on the table. I'm merely saying if you are going to use these increasingly effective methods, do it correctly.

GSA excess: http://newsfeed.time.com/2012/04/18/gsa-scandal-so-what-does-823000-buy-you-in-las-vegas/ And the practical consequences:https://www.washingtonpost.com/politics/clampdown-after-gsa-scandal-puts-some-federal-workers-in-a-pinch/2015/02/08/d8217240-a5a4-11e4-a7c2-

03d37af98440_story.html

Defense contractor-sponsored equivalents to GSA: funny, you don't see any stories about those events. An absence of curiousity which I'm sure is entirely unconnected with the presumably rather costly defense contractor advertisements that keep popping up on my—perhaps not your—Web versions of the *New York Times* and *Washington Post*. "Yes, an F-35…why, just in time for the holidays, a perfect gift for my nephews! I'll take three, my good man, and can you have them wrapped and delivered? Jolly good of you to remind me with that expensive animated advertisement!"

How Rome became with the hegemonic successor to the Hellenistic empires: haven't read it but Mary's Beard new popular history, *SPQR [http://www.amazon.com/SPQR-History-Ancient-Mary-Beard/dp/0871404230]* has gotten a lot of good reviews. There's more to the story than gladiators. Really. And with a level of wealth inequality approaching that of imperial Rome, there's stuff we can learn.

Footnotes

1. With Reagan now subjected to vicious character assassination by the Fox News crowd, I'm going to open the next few entries with Reagan quotes. Judge people by their enemies, eh? Besides, in the current environment, he'd be considered a bit of a lefty, what with all that arms control and raising taxes.

2. I'm not providing many links this time: With every technical phrase in this entry, if I can Google it, you can Google it.

3. And whenever I'm told this, people outside of DoD—never from the inside—suggest I'm cynical because all of the highly successful projects are secret. Well, if they are I'm one important guy, because a huge pile of money has been spent on unclassified foolishness over the years just to distract me from learning about the good stuff. Really, I don't think I'm that important.

4. In the old days, these came with a nice spread of donuts and sandwiches—typically we're doing these things for little or no compensation beyond expenses. But with the combination of the Tea Party trash-and-burn budgetary tactics and that wonderful tax-payer-funded Las Vegas bacchanalia of the GSA, those days are gone. Launder that same tax money through a defense contractor, of course, and the bacchanalia are still absolutely fine: been to a few of those as well,

and other than the experience of seeing the reckless extravagance making me want to throw up, they are splendid exercises, and most of the attendees, staggering about under the burden of unlimited quantities of cheap booze, consider them a professional entitlement. Though I wouldn't be surprised if those things have pushed more than a few folks—at least the caterers—into the Tea Party, albeit with no effect.

Get yourself to one of these defense industry affairs and it's night after night of lobster, champagne and live entertainment in lavishly decorated resort hotel ballrooms, all provided by —smirk, smirk, wink, wink—"company sponsorship." Going to a government-funded conference on finding a cure for Alzheimer's, or teaching kids to read, or preventing rusting bridges from collapsing?: it's gonna be Subway and a diet cola, maybe followed by a pitcher of Miller Lite shared with your buddies at an anonymous sports bar in a strip mall, and remember to bring your wallet, because all this comes out of your own pocket.

Some dumb hick from Abilene, Kansas pointed out the problems with this system back in 1961. Lotta good that did.

5. Seriously, do you think the Federal Reserve Board and the Centers for Disease Control spend their days listening to a gaggle of demented bozos braying about how human behavior is unpredictable?

6. There's a little ritual in these meetings where the program manager earnestly intones a little mantra—I'm not sure whether the original was in Latin or Sanskrit—about their deep responsibility of not wasting public money. Given they know the Doughboys are in the room precisely to insure that public money *is* wasted, it must be hard to do this with a straight face. Though I suppose that why program managers get paid the big bucks. Joke.

7. For the Doughboys, "social science" is an oxymoron, an observation they will share endlessly. In fact they probably have " 'social science' is an oxymoron" tattooed somewhere on their bodies, probably somewhere I'd certainly prefer not to look.

8. Technical term of art: actually, one of my several working titles for this essay was "Seven Under-stated Reflections on When the Hell are You People Going to Learn How to Keep Those Parasitic Bastards from Making Off with My Tax Dollars?" But that doesn't scan particularly well.

9. When the history of the decline of United States hegemony is written in a century or two, the role of the Doughboys will probably deserve at least a footnote. Though the role of the

PowerPoint virus—no, not a virus carried by PowerPoint; PowerPoint *is* a virus—will get a chapter.

That history may be written in English—in New Delhi—rather than Chinese, just as the successor to the Hellenistic empires was not the obvious Mediterranean candidate, the wealthy and commercially savvy Phoenicians from their base in Carthage—but instead a theretofore marginal group of Italians living along the Tiber. The second mouse, as it were.

10. I will assume you can find these on Amazon or, if you prefer not to support a soul-destroying mega-corporation whose business model involves removing every last visage of humanity from the workplace, a local bookshop if you have one. Probably run either by some balding old guy who is likely to engage you in an extended conversation when you really just wanted to buy a book, or someone with cats. Though I'm also actually becoming rather fond of the Barnes and Noble chain, particularly when after seeing the grey hair, they let this old guy to the front of the line for a seat at an overflowing author's event, and at Union Square in New York City no less! I digress…damn old guys who don't know when to stop…well, "communicating", if that's even the relevant concept…

11. We'll deal with his more recent work on superforecasters in the next entry.

12. General link is http://jpr.sagepub.com/content/47/4. Since I still have an adjunct academic appointment with library access, the version I'm seeing isn't paywalled; your results may differ. If you dig a bit,you can usually located non-firewalled versions of widely-cited academic papers, and this would qualify. Or you can pay: none of that payment goes to the authors, of course, as academic publishing doesn't work that way, instead it is a monk-like humble offering to the cause of restricting the flow of knowledge generated through public funding and to further increasing the level of inequality through the support of a tiny oligopoly of rapaciously profitable publishers. Yet again, I digress.

13. And GOP presidential candidates…

14. For some fairly technical reasons, "specification error" in some of the most commonly used models is even worse, since if a variable you've incorrectly put in the model is correlated with variables which are actually causal, the variable will appear to have a stronger effect than it actually has. Though if you are only interested in prediction, this isn't that big a deal. Still, a model that is consistent with a correct theory will almost certainly have better properties than a model that isn't. Which is also why the "data will replace theory" arguments are overly optimistic: a good theory is vastly more useful than any undifferentiated mess of data.

15. Hey, give us poor proles who still actually have to pay taxes a break here, will you?: even in DoD research there should be a concept of "too expensive."

16. Or appears to be, as in chaotic processes such as weather. I'll say a bit about chaotic processes in the next entry; for now suffice it to say these are not the semi-mystical phenomenon some would have you believe, just an unexpected but completely deterministic aspect of a very simple dynamic equation you can easily experiment with in one column of a spreadsheet. Intrinsic randomness could be an essay in its own right…maybe later…

17. But caution, particularly if you've only skimmed Taleb: rare events are not the same thing as black swans.

- Black swan: an event that has a low probability even conditional on other variables
- Rare event: an event that occurs infrequently, but conditional on an appropriate set of variables, does not have a low probability

Using a medical analogy, certain rare forms of cancer appear to be highly correlated with specific rare genetic mutations. Conditioned on those mutations, they are not black swans.

Taleb definitely gets these distinctions, but many of the popularizations of Taleb (who in turn is quite consciously—he profusely acknowledges their work—a popularizer of Kahneman and his collaborators, and Tetlock) miss it.

Also worth noting here—since I ran out of my seven allocated categories—are events which are *too* predictable: these are called "auto-regressive" or "auto-correlated," which simply means that the value of a variable at time *t* is highly correlated with the value at *t-1*. Most human activities have this characteristic—humans, and particularly human institutions, are fairly boring and predictable, except when they aren't. The sorts of sequences one tends to look at in political conflict forecasting are highly autocorrelated except for a small number of highly consequential exceptions which are…rare events. From the perspective of a methodologist, it makes the whole problem rather interesting.

And in a final, really techy, aside, there's a tendency to confuse autocorrelated *variables*and autocorrelated *errors*. The presence of the latter considerably complicates estimation, all the more so when you get both at once. Errors are autocorrelated for the same reason variables are autocorrelated: human behaviors tend not to change much over time, and "errors" are just the factors not included in the model, and quite a few of those involve humans.

18. Google “coefficient” if you aren’t familiar with the term: I can’t begin to count the number of meetings I’ve sat through where we appeared to be operating under an assumption that functioning models would be delivered by, well, maybe Elminster Aumar, High Wizard of the Forgotten Realms?…I dunno, sure the heck wasn’t going to be through any systematic estimation method worthy of discussion.

19. Newer machine learning methods also have a “hyperparameter” issue—the estimators can be configured in a wide variety of different ways, some better than others, and optimizing these using vast amount of machine cycles is another important new research field. Older methods were derived algebraically and generally had a very small number of free parameters.

20. It is remarkably difficult to find new methods that consistently outperform the “obnoxiously effective” old standbys I will discuss in the next entry—conventional and logistic regression, support vector machines, conventional neural networks, and clustering methods. That’s why they are old standbys. That’s also why a genuinely effective new entrant like recurrent neural networks is such a big deal.

21. Usually referred to an “frequentism”, particularly by its detractors. Its supporters call it “statistics.” In some circumstances, frequentism is completely appropriate. But it isn’t universally appropriate and until about 20 years ago, it was treated as such.

22. Look at the research interests of the faculty in almost any university statistics department and you’ll find that most of the younger people are working on Bayesian methods: this is not an instance of random selection.

23. Like so many memorable quotes, that’s not the actual line, which had a rather rambling preamble before finally getting around to: “You’ve gotta ask yourself one question: “Do I feel lucky?” Well, do ya,punk?” Rather as the oft-quoted “Play it again Sam” condensed seven lines of dialog none containing the word “again.” For simplicity, and cognizant of the date, 17-Dec-15, stick with “Han shot first.”

Posted in Methodology, Politics | 1 Comment | Edit