

Still just parametrized models

Eric Blair

2 July 2008

You will recognize Wired as a heavy-paper glossy magazine, owned by the same company (Condé Nast) that owns Ars Technica, Glamour, Modern Bride, Teen Vogue, and Bon Apétit.

A few months ago, it ran an issue whose cover story claimed that everything you knew about environmentalism is wrong, and you should do contrarian things like buying a rugged SUV instead of a hybrid, live somewhere where you'll run your air conditioning 24/7, and so on. A¹ great² many³ people⁴ commented on how ill-founded the recommendations were.

However, there were at least a few points that were kinda true: it *is* generally ecologically cheaper to run the air conditioning than the heater, but there are many places and many houses where you don't have to run either. My brother lives in San Diego and just runs a fan from time to time. San Diego sprawls, but he shouldn't buy a hybrid to get around—he should buy a bike.

This month's issue vends paper via the same revolutionary formula as the antienvironmentalist issue, but gets none of it right at all. In fact, its own examples sometimes support the opposite conclusion. I'm writing a response, despite the *don't feed the trolls* rule, because the authors are actually making very common mistakes, which have been made repeatedly over the last several decades, so this is a springboard to discuss a few other scientific revolutions that didn't happen.

This month's declaration: The End of Science⁵(!!). The guy who wrote about the End of History was maybe right for a year or two—The Lull in History(!!). But Wired's scientific revolution doesn't have half as much going for it. Frankly, the other Condé Nast publications better maintain credibility by just offering 15 NEW HAIR AND SEX TIPS every month.

The basic claim is that having petabytes of data is fundamentally different from having smaller amounts, to the point that the traditional method of developing and testing a model is somehow no longer relevant. In the words of the lead essay⁶, “Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.”

¹<http://putative.typepad.com/putative/2008/06/everything-you.html>

²<http://thinkprogress.org/wonkroom/2008/06/07/wired-ignorant-libertarianism/>

³<http://www.fluffybunnybutts.com/2008/05/wired-sells-out-to-monsanto.html>

⁴<http://gristmill.grist.org/story/2008/5/20/15537/7410>

⁵http://www.wired.com/science/discoveries/magazine/16-07/pb_intro

⁶http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Rather than reading Wired directly, I recommend that you instead read KK's response to Wired⁷, which is much more coherent, although still a bit hyperbolic. It uses the word *pioneer*.

Some examples Back to Wired, it presents a long series of examples where people develop and test models using large data sets. I could write a paragraph about how every last one misses the mark, but I'll just give you three that should give you a sense of how inquiry, data, and models interact and how that differs from the End of Science story.

What they're trying to say above is that we can program our computers to just suck in data and spit out correlations, and that has meaning, and is outside of models. More or less: we give it data, and the computer thinks for us and draws conclusions that are true but beyond our comprehension.

We'll start with an example to show you what we're not talking about:

The best practical example of this is the shotgun gene sequencing by J. Craig Venter. Enabled by high-speed sequencers and supercomputers that statistically analyze the data they produce, Venter went from sequencing individual organisms to sequencing entire ecosystems. In 2003, he started sequencing much of the ocean, retracing the voyage of Captain Cook. And in 2005 he started sequencing the air. In the process, he discovered thousands of previously unknown species of bacteria and other life-forms.

You can ask Wikipedia about shotgun sequencing, and it'll tell you that it is based on the same genetic model everybody else is using, but uses a novel random component to gather a lot more data a lot faster than prior methods could. That is, this is the type of science we call *gathering data*. Where and how to look has always been advised by some human-sensical story, and observers have always striven to let the machinery work and not judge the data while gathering it.

No naturalist feels the need to prove the causal mechanism underlying a new frog before declaring the new frog's existence, so this says nothing about the rise or fall of causal mechanisms. But it is certainly true that methods like these are giving us an order of magnitude more data. We want less restrictive models that will let these reams of data speak for themselves as much as possible.

So let's move on to what happens after the data is gathered: taking action or deriving meaning from data. Any action by Google is taken by tech enthusiasts as divine, so it is naturally a key example (or three):

Google's founding philosophy is that we don't know why this page is better than that one: If [sic] the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required. [...]

This makes two statements: Google doesn't primarily rank quality via content analysis (i.e., computer-reading the page and evaluating the relevance or quality of the words on the page), but it does use a simple, human-comprehensible model relating

⁷http://www.kk.org/thetechnium/archives/2008/06/the_google_way.php

page relevance to incoming links. That is, it doesn't use a literal content model, but it does use another link-based model. We'll see this pattern of passing on one model for a 'looser' model a few times more below. Google's page ranking model even has an underlying causal story, that high quality content causes people to link to the page.

One last example, from this page⁸, about political micro-targeting. I conclude with this one because it's the only one Wired gets right.

As databases grow, fed by more than 450 commercially and privately available data layers as well as firsthand info collected by the campaigns, candidates are able to target voters from ever-smaller niches. Not just blue-collar white males, but married, home-owning white males with a high school diploma and a gun in the household. Not just Indian Americans, but Indian Americans earning more than \$80,000 who recently registered to vote.

This is known as "data mining," and has its formal origins probably thirty years ago, or so. Being from the pre-Wired dark ages, it is very model-dependent: claiming that elements of the data set are correlated to each other is a model, and searching for the best correlation is a series of model tests, often done using the traditional hypothesis tests they forced you to learn in undergrad stats. Your typical data mining textbook includes lots of other models, including overlapping categories, trees, separating hyperplanes, and other very structured forms.

To clarify, think of what a model-free search would really consist of. Think of all the ways that our politicians could handle this data: they could look for the political preferences of people with blood type AB positive, or the product of age cubed times the cube root of sushi consumption for women or the 3.2nd root of sushi consumption for men, or the preferences of people whose house number starts with a 4. Rest assured, they ain't wasting any processor time testing the 3.2nd root of anything, even though it would be as valid to the computer as any other list of numbers. Instead, they set a framework such as a hierarchy of characteristics, and set the computer to find the best such hierarchy.

[Wired gives an example of data mining airline prices. There was a period where airlines sent signals for the purposes of colluding on prices using the cents part of a price, so there really *was* a pattern among tickets whose price has a four in the dimes place (or whatever). I have no idea how that pattern was spotted. U.S. ticket prices are now in whole dollar amounts by law.]

The pattern in the data People marketing products such as toothpaste or politicians love data mining. After all, it's a results-oriented system that only asks how many people purchase the product, not why. Thus, it's perfect for non-causally oriented analysis, and has been for decades before Wired declared it to be a paradigm shift. But every data mining textbook is heavy on causal models. This is not a contradiction: the model is just not where you expect it to be.

Or, consider the field of 'nonparametric statistics'. By that, we mean writing down models that aren't the one- or two-parameter models you get in the back of stats textbooks (Normal, Poisson, Binomial, ...). Instead, a typical procedure defines a bar

⁸http://www.wired.com/science/discoveries/magazine/16-07/pb_vote

chart with maybe a hundred segments, and then estimates the heights of all 100 bars. Great, so this ‘nonparametric’ method has a hundred parameters to fit instead of two.

All of the examples here have a similar flavor: we don’t specify a tight model with only a few parameters, but instead a loose model which may need a million little parts to be specified: instead of a broad regression on a few variables, calculate a different value from scratch for every web page, or every combination of age/race/gender. But that doesn’t mean there’s no model, or that you’ve somehow escaped the paradigm of describing a human-sensible model and then asking the computer to fill in parameters from the data. [Also, it doesn’t save you from the problem that you can fit a loose enough model to any garbage—more on this next time.]

Or consider the case of agent-based modeling, which is a hair’s breadth from simple simulation. This was trendy in the 1980s because of the new study of chaos and all the resultant poster and calendar sales. All the rules of the agents in the simulation or the steps in your chaos model are all very simple and easy to understand, but there’s no way to know what it will do in the end but follow the iterations of the model to their computer-calculated conclusion. We can now repeat everything stated above: the outcome is beyond small-scale parametrization, and causal mechanisms are hard to come by (otherwise we wouldn’t need to follow the simulation along, but could predict the outcome). But on another level, it’s still just a model: simple rules and typically a set of parameters that can be tweaked as desired.

For all of agent-based modeling, chaos theory, nonparametric stats, and whatever the fuck Wired is talking about, some proponents trumpet how their new method is outside of the parametrized small-scale model that we had been contending with since the advent of modern science. But upon further inspection, we find a framework of assumptions that is really just a model pushed behind the curtain, and we find that the final goals of the data search are a set of numbers or relations, which is another way of saying parameters.

One retort is that the framework underlying the computer’s search for parameters for a regression model is a model, but the framework underlying the computer’s search for parameters in these complicated systems is a meta-model, or just a set of rules, or a constraint set, or some other means of avoiding the word *model*. But this is semantics. When we sit down to the computer to fit the old-school models to data or fit the broad meta-heuristic-constraints to data, we do exactly the same thing, albeit with more or less typing.

I was heavily involved in the writing of a book on computational methods for models like these⁹, aimed at treating a range of models as broad as that described here. It opens by declaring that its goal is to *estimate the parameters of a model with data*. Save perhaps for the pure data-gathering exercise, that phrase describes every example here. In every case, we’re assigning a human structure with a finite number of levers, and relegating the computer to finding how to best position the levers. You can ask (and may benefit from asking) the same questions of any study: what is the underlying framework, and its underlying causal story? What parameters are being tweaked and/or output? How do you know when the parameters are good so you can stop searching?

From a philosophy of science perspective, nothing seriously new is happening,

⁹<http://press.princeton.edu/titles/8706.html>

save for an increasing trust when the machine gives us a million parameters instead of two. From a practical perspective, the engineering advances are clearly incremental: a question of distributing computation among PCs, managing databases, and finding ways for us humans to comprehend and take action on all that computer output.