Latest Articles

On Being a Data Skeptic

Data is here, it's growing, and it's powerful.

August 31, 2013

Policeman (source: The British Library)

Skeptic, Not Cynic

I'd like to set something straight right out of the gate. I'm not a data cynic, nor am I urging other people to be. Data is here, it's growing, and it's powerful. I'm not hiding behind the word "skeptic" the way climate change "skeptics" do, when they should call themselves deniers.

Instead, I urge the reader to cultivate their inner skeptic, which I define by the following characteristic behavior. A skeptic is someone who maintains a consistently inquisitive attitude toward facts, opinions, or (especially) beliefs stated as facts. A skeptic asks questions when confronted with a claim that has been taken for granted. That's not to say a skeptic brow-beats someone for their beliefs, but rather that they set up reasonable experiments to test those beliefs. A really excellent skeptic puts the "science" into the term "data science."

In this paper, I'll make the case that the community of data practitioners needs more skepticism, or at least would benefit greatly from it, for the following

I'm charging myself with making a case for data practitioners to engage in active, intelligent, and strategic data skepticism. I'm proposing a middle-of-the-road approach: don't be blindly optimistic, don't be blindly pessimistic. Most of all, don't be awed. Realize there are nuanced considerations and plenty of context and that you don't necessarily have to be a mathematician to understand the issues.

My real goal is to convey that we should strive to do this stuff right, to not waste each other's time, and to not miss business or creative opportunities.

I'll start with a message to the overly sunny data lover. A thoughtful user of data knows that not everything they want to understand is measurable, that not all proxies are reasonable, and that some models have unintended and negative consequences. While it's often true that doing something is better than doing nothing, it's also dangerously easy to assume you've got *the perfect answer* when at best you have a noisy approximation.

dont trust one-liners :)

If you've seen the phrase "if it's not measured, it doesn't exist" one too many times used in a nonironic, unthoughtful way, or even worse if you've *said* that phrase in a moment of triumphant triviality, then I hope I will convince you to cast a skeptical eye on how math and data are used in business.

Now on to a message for the other side of the spectrum: the data science naysayer. It's become relatively easy to dismiss the big data revolution as pure hype or marketing. And to be honest, it sometimes is pure hype and marketing, depending on who's talking and why, so I can appreciate the reaction. But the poseurs are giving something substantial a bad name.

Even so, how substantial is it? And how innovative? When I hear people proclaiming "it's not a science" or "it's just statistics," that's usually followed by a claim that there's nothing new to be gained from the so-called new techniques.

And although a case can be made that it probably isn't a science (except perhaps in the very best and rarest conditions), and although the very best and leading-edge statisticians already practice what can only be described as "big data techniques," that doesn't mean we're not dealing with something worth

we, and presumably they, have already come to rely on. To pick an example from the air, spam filtering has become so good that we are largely shielded from its nuisance, and the historical cries that spam would one day fill our inboxes to their brims have proved completely wrong. Indeed the success stories of big data have become, like air, part of our environment; let's not take them for granted, and let's not underestimate their power, for both good and evil.

It's no surprise that people are ignorant at both extremes of blind faith and dismissive cynicism. This ignorance, although typically fluffy and unsubstantiated, is oftentimes willful, and often fulfills a political agenda.

Further in from the extremes, there's a ton of wishful thinking and blithe indifference when it comes to the power of data and the potential for it going wrong. It's time people educate themselves about what can go wrong and think about what it would take to make things right.

The Audience

Although they're my peeps, I'll acknowledge that we nerds are known for having some blind spots in our understanding of our world, especially when it comes to politics, so the first goal of this paper is to get the nerds in the room to turn on their political radars, as well as to have an honest talk with themselves about what they do and don't understand.

At the same time, the technical aspects of data science are often presented as an impenetrable black box to business folks, intentionally or not. Fancy terminology can seem magical, or mumbo-jumbo can seem hyped, useless, and wasteful. The second goal of this paper is to get nontechnical people to ask more and more probing questions from data folk, get them more involved, and tone down the marketing rhetoric.

Ultimately I'm urging people to find a way to bridge the gap between dialects —marketing or business-speak and math or engineering—so that both sides are talking and both sides are listening. Although cliched, it's still true that communication is the key to aligning agendas and making things work.

be approached.

Trusting Data Too Much

This section of the paper is an effort to update a fine essay written by Susan Webber entitled "Management's Great Addiction: It's time we recognized that we just can't measure everything". It was presciently published in 2006 before the credit crisis and is addressed primarily to finance professionals, but it's as relevant today for big data professionals.

I'd like to bring up her four main concerns when it comes to the interface of business management and numbers and update them slightly to the year 2013 and to the realm of big data.

1) People Get Addicted to Metrics

We believe in math because it's "hard" and because it's believed to be "objective" and because mathematicians are generally considered trustworthy, being known to deal in carefully considered logical arguments based on assumptions and axioms. We're taught that to measure something is to understand it. And finally, we're taught to appear confident and certain in order to appear successful, and to never ask a dumb question.

The trust we have in math engenders a sanitizing effect when we use mathematical models. The mere fact that there's math involved makes a conclusion, faulty or not, seem inevitable and unimpeachable. That's probably a consequence of any private language, whether it's alchemists turning lead into gold or bankers designing credit default swaps.

Once we start seeing mathematical models as trustworthy tools, we get into the habit of measuring things more and more in order to control and understand them. This is not in itself a bad thing. But it can quickly become a form of addiction—especially if we only acknowledge things that are measurable and if we hunger to fit everything into the data box.

Once we get used to the feeling of control that comes along with modeling and

Examples: First, let's give examples of things that are just plain hard to measure: my love for my son, the amount of influence various politicians wield, or the value to a company of having a good (or bad) reputation. How would you measure those things?

Secondly, let's think about how this data-addicted mindset is blind to certain phenomena. If we want to keep something secret, out from under the control of the data people, we only need to keep those things out of the reach of sensors or data-collection agents. Of course, some people have more reason than others to keep themselves hidden, so when the NSA collects data on our citizens, they may well be missing out on the exact people they're trying to catch.

Thirdly, even when we have some measurement of something, it doesn't mean that it's a clean look. Sales data varies from month to month, and sometimes it's predictable and sometimes it isn't. Not all data is actionable and not all "hard" numbers are definitive. But acting decisively when you are in a state of uncertainty is a pretty common outcome, because it's hard to admit when one doesn't have enough information to act.

Nerds: Don't pretend to measure something you can't. The way to avoid this mistake is by avoiding being vague in describing what the model does with the input. If you're supposed to measure income but you are actually using census data to approximate it, for example, then say so.

Be sure you are communicating both the best guess for an answer as well as the error bars, or associated uncertainty. Be creative in the way you compute error bars, and try more than one approach.

Business people: Not everything is measurable, and even when it is, different situations call for different kinds of analysis. The best approach often means more than one. Back up your quantitative approach with a qualitative one. Survey and poll data can be a great supplement to data-driven analysis.

Don't collect numbers for the sake of collection; have a narrative for each datapoint you care about and ask for details of the black boxes you use. What are the inputs and what are the outputs? How is the information being used?

from you.

2) Too Much Focus on Numbers, Not Enough on Behaviors

When modelers can't measure something directly, they use proxies—in fact, it's virtually always true that the model uses proxies. We can't measure someone's interest in a website, but we can measure how many pages they went to and how long they spent on each page, for example. That's usually a pretty good proxy for their interest, but of course there are exceptions.

Note that we wield an enormous amount of power when choosing our proxies; this is when we decide what is and isn't counted as "relevant data." Everything that isn't counted as relevant is then marginalized and rendered invisible to our models.

In general, the proxies vary in strength, and they can be quite weak. Sometimes this is unintentional or circumstantial—doing the best with what you have—and other times it's intentional—a part of a larger, political model.

Because of the sanitizing effect of mathematical modeling, we often interpret the results of data analysis as "objective" when it's of course only as objective as the underlying process and relies in opaque and complex ways on the chosen proxies. The result is a putatively strong, objective measure that is actually neither strong nor objective. This is sometimes referred to as the "garbage in garbage out" problem.

Examples: First, let's talk about the problem of selection bias. Even shining examples of big data success stories like Netflix's movie recommendation system suffer from this, if only because their model of "people" is biased toward people who have the time and interest in rating a bunch of movies online—this is putting aside other modeling problems Netflix has exhibited, such as thinking anyone living in certain neighborhoods dominated by people from Southeast Asia are Bollywood fans, as described by DJ Patil.

In the case of Netflix, we don't have a direct proxy problem, since presumably we can trust each person to offer their actual opinion (or maybe not), but rather it's an interpretation-after-the-fact problem, where we think we've got

Next, we've recently seen a huge amount of effort going into quantifying education. How does one measure something complex and important like high school math teaching? The answer, for now at least—until we start using sensors—is through the proxy of student standardized test scores. There are a slew of proprietary models, being sold for the most part by private education consulting companies, that purport to measure the "value added" by a given teacher through the testing results of their students from year to year.

Note how, right off the bat, we're using a weak proxy to establish the effectiveness of a teacher. We never see how the teachers interact with the students, or whether the students end up inspired or interested in learning more, for example.

How well do these models work? Interestingly, there is no evaluation metric for these models, so it's hard to know directly (we'll address the problem of choosing an evaluation metric below). But we have indirect evidence that these models are quite noisy indeed: teachers who have been given two evaluation scores for the same subject in the same year, for different classes, see a mere 24% correlation between their two scores.

Let's take on a third example. When credit rating agencies gave AAA ratings to crappy mortgage derivatives, they were using extremely weak proxies. Specifically, when new kinds of mortgages like the no-interest, no-job "NINJA" mortgages were being pushed onto people, packaged, and sold, there was of course no historical data on their default rates. The modelers used, as a proxy, historical data on higher-quality mortgages instead. The models failed miserably.

Note this was a politically motivated use of bad models and bad proxies, and in a very real sense we could say that the larger model—that of getting big bonuses and staying in business—did not fail.

Nerds: It's important to communicate what the proxies you use are and what the resulting limitations of your models are to people who will be explaining and using your models.

And be careful about objectivity; it can be tricky. If you're tasked with building a model to decide who to hire, for example, you might find yourself comparing

feedback on their environments compared to the men.
Your model might be tempted to hire the man over the woman next time the two show up, rather than looking into the possibility that the company doesn't treat female employees well. If you think this is an abstract concern, talk to this unemployed black woman who got many more job offers in the insurance industry when posing as a white woman.

In other words, in spite of what Chris Anderson said in his now-famous Wired Magazine article, a) ignoring causation can be a flaw, rather than a feature, b) models and modelers that ignore causation can add to historical problems instead of addressing them, and c) data doesn't speak for itself—it is just a quantitative, pale echo of the events of our society.

Business peeps: Metaphors have no place in data science—the devil is always in the detail, and the more explicitly you understand what your data people are doing and how they use the raw data to come to a conclusion, the more you'll see how people can fall through the cracks in the models, and the more you'll understand what the models are missing.

3) People Frame the Problem Incorrectly

The first stage in doing data science is a translation stage. Namely, we start with a business question and we translate it into a mathematical model. But that translation process is not well-defined: we make lots of choices, sometimes crucial ones. In other words, there's often more than one way to build a model, even something as simple as a measurement. How do we measure a company, for example? By the revenue, the profit, or the number of people it employs? Do we measure its environmental impact? What is considered "progress" in this context?

important to always circle back!

Once we've made a choice, especially if it's considered an important measurement, we often find ourselves optimizing to that progress bar, sometimes without circling back and making sure progress as measured by that *is* actually progress as we truly wish it to be defined.

Another way things could go wrong: the problem could even be relatively well-defined but then the *evaluation* of the solution could be poorly chosen.

Goodhart's! :p

it's typical to see meaningless evaluation metrics applied, especially if there's money to be made with that bad choice.

Examples: First I'll take an example from mine and Rachel Schutt's upcoming O'Reilly book, *Doing Data Science*—specifically what contributor Claudia Perlich, Chief Data Scientist at media6degrees, told us about proxies in the realm of advertisers and click-through rates.

Ideally advertisers would like to know how well their ads work—do they bring in sales that otherwise they wouldn't have seen? It's hard to measure that directly without mind-reading, so they rely on proxies like "did that person buy the product after clicking on the ad?" or, more commonly since the data is so sparse on that question, "did that person click on the ad?" or even "did that person see the ad?".

All of the above might be interesting and useful statistics to track. However, what's really hard is to know is whether the original question is being answered: are the ads getting people to buy stuff who would not have bought that stuff anyway?

Even with formal A/B tests, data is messy—people clear their cookies, say, and are hard to track—and there are small discrepancies between user experiences like delays that depend on which ad they see or whether they see an ad at all, that could be biasing the test. Of course without A/B, there are sometimes outright confounders, which is even worse; for example, people who see ads for perfume on a luxury shopping site are more likely to be perfume purchasers. What's more, the guys at media6degrees recently discovered a world of phantom clicks made by robots that never buy anything and are obviously meaningless.

You might think that the advertisers who learn about the futility of measuring success via click-through-rates would stop doing it. But that's not what you see: the options are limited, habits die hard, and, last but not least, there are lots of bonuses computed via these artificially inflated measurements.

Second, let's consider a problem-framing example that many people know. When the Netflix Prize was won, the winning solution didn't get implemented. It was so complicated that the engineers didn't bother. This was a basic framing

Next, let's give an example of how people hang on to a false sense of precision. In a noisy data environment, we will see people insist on knowing the r^2 out to three decimal points when the error bars are bigger than the absolute value. Which is to say you're in a situation where you don't really even know the sign of the result, never mind a precise answer.

Or, you sometimes see people stuck on the idea of "accuracy" for a rare-event model, when the dumbest model around—assigning probability 0% to everything—is also the most accurate. That's a case when the definition of a "good model" can be itself a mini model.

There's lots at stake here: a data team will often spend weeks if not months working on something that is optimized on a certain definition of accuracy when in fact the goal is to stop losing money. Framing the question well is, in my opinion, the hardest part about being a good data scientist.

Nerds: Are you sure you're addressing the question you've been asked? The default functions a given optimization technique minimizes often ignore the kind of mistake being made—a false negative might be much worse than a false positive, for example—but often in a real-world context, the kind of mistake matters. Is your metric of success really the same as what the business cares about?

Business people: This is often where the politics lie. When we frame a problem we sometimes see sleight of hand with the naming or the evaluation of a project or model. We want to measure progress but it's hard to do that, so instead we measure GDP—why not the quality of life for the median household? Or for the poorest? We want to know how much our work is valued but that's hard, so instead we refer to titles and salary. We want to know who has influence, but instead we look at who has followers, which is skewed toward young people without jobs (i.e., people with less influence). Keep in mind who is benefitting from a poorly defined progress bar or a poorly chosen metric of success.

4) People Ignore Perverse Incentives

Models, especially high-stake models where people's quality of life is on the

no direct gaming, poorly thought-out models, or models with poor evaluation metrics, can still create negative feedback loops.

First let's talk about gaming. It's important to note that it's not always possible to game a model, and the extent to which it *is* possible is a function of the proxies being used and the strength of those proxies.

So for example, the FICO credit score model is pretty good on the spectrum of gamability. We know that to increase our credit score, we should pay our bills on time, for example. In fact most people wouldn't even call that gaming.

Other models are much more easily gamed. For example, going back to credit rating agencies, their models weren't publicly available, but they were described to the banking clients creating mortgage-backed securities. In other words, they were more or less transparent to exactly the people who had an incentive to game them. And game them they did.

Finally, it's well known that high-stakes student testing is perennially gamed via "teaching to the test" methods. See for example the educational testing book by Daniel Koretz, *Measuring Up* (Harvard University Press), for an explanation of the sawtooth pattern you see for student testing as new versions of the test are introduced and slowly gamed.

Nerds: acknowledge the inevitable gaming. Make plans for it. Make your model gaming-aware and make sure proxies are cross-checked. For example, a model that is high impact and widely used, relies on proxies, and is highly transparent is going to be gamed. It's not enough to show that it worked on test data before it was implemented—it has to work in the situation where the people being judged by the model are aware of how it works.

Business people: This gaming feedback loop is known as both Campbell's Law and Goodhart's Law, depending on your background. Be aware that you will distort the very thing you wish to measure by the measurement itself. Don't rely on one distortable metric—use dashboard approaches. And if you can, measure the distortion.

multiple metrics

The Smell Test of Big Data

If I can imagine influence happening in real life, between people, then I can imagine it happening in a social medium. If it doesn't happen in real life, it doesn't magically appear on the Internet.

For example, if LeBron James were to tweet that I should go out and watch some great movie, then I'd do it, because I'd imagine he was there with me in my living room suggesting that I see that movie, and I'd do anything that man said if he were in my living room hanging with me and my boys. But if LeBron James were to tell me to lose weight while we're hanging, then I'd just feel bad and weird. Because nobody has found a magic pill that consistently influences average people to make meaningful long-term changes in their weight—not Oprah, not Dr. Oz, and not the family doctor. Well, maybe if he had a scalpel and removed part of your stomach, but even that approach is still up for debate. The truth is, this is a really hard problem for our entire society which, chances are, won't be solved by simply munging twitter data in a new way. To imply that it can be solved like that is to set up unreasonable expectations.

Bottom line: there's a smell test, and it states that real influence happening inside a social graph isn't magical just because it's mathematically formulated. It is at best an echo of the actual influence exerted in real life. I have yet to see a counterexample to this. Any data scientist going around claiming they're going to transcend this smell test should stop right now, because it adds to the hype and to the noise around big data without adding to the conversation.

On the other hand, having said that, there are cool things you can see with Twitter data—how information spreads, for example, and the fact *that* information spreads. Let's spend our time on seeing how people do stuff more easily via social media, not hoping that people do stuff they would never do via social media.

Trusting Data Too Little

On the other side of the spectrum, we have a different problem, with different and sometimes more tragic consequences: that of *underestimating* the power of models and data.

One of the not-so-tragic consequences of underestimating data and the power

will.

There are more tragic consequences of underestimating data, too. As a first example, look no further than the unemployment rate and the housing crisis, which still lingers even after five years post-crisis. Much of this, arguably, can be traced to poor models of housing prices and of specific kinds of mortgage derivatives, as well as meta-economic modeling about how much and how fast middle class Americans can reasonably take on debt and whether we need regulation in the derivatives market.

But wait, you might say—those were consequences of financial and economic models, which are a bit different from the models being created by modern "big data" practitioners. Yes, I'd agree, but I'd argue that the newer big data models might be even *more* dangerous.

Whereas the old models predicted markets, the new models predict people. That means individuals are to some extent having their environments shaped by the prevalent models, along the lines of what is described in Eli Pariser's excellent book, *The Filter Bubble* (Penguin, 2012).

Pariser correctly describes the low-level and creeping cultural effects of having overly comfortable, overly tailored individual existences on the web. It leads to extremism and a lack of empathy, both bad things.

But there is also a potential to be imprisoned by our online personas. This has come up recently in the context of the NSA's data collection processes as exposed by Edward Snowden, but I'd argue that we have just as much to fear from the everyday startup data scientist who has not much to lose and plenty to gain from fiddling with online data.

1) People Don't Use Math to Estimate Value

There are plenty of ways to get ballpark estimates of models, and I rarely see them used outside of finance. What's the expected amount of data? What's the expected signal in the data? How much data do we have? What are the opportunities to use this model? What is the payoff for those opportunities? What's the scale of this model if it works? What's the probability that the idea

model.

Example: This kind of back-of-the-envelope reasoning can also be used to evaluate business models. What is the market for predicting whether people would like a certain kind of ice cream? How much would people pay for a good ice cream flavor predictor? How much would I pay? How accurate is the data for ice cream preferences, and how much can I get to train my model? How much does that data cost in real time?

Another satisfying application of this line of reasoning: "Hey, we've got 15 people sitting on this conference call, each of whom is probably costing $100/hour, trying to make a decision that's worth $700." Put an end to terribly inefficient meetings.

Nerds: Put this on your resume: it takes a few minutes to figure out how to reckon this way and it's a real skill. Call it "McKinsey on data steroids" and let 'er rip.

Business peeps: Get those nerds into business planning meetings with you—they have something to contribute.

2) Putting the Quant in the Back Room

Quants or data scientists, terms I use interchangeably, are often treated almost like pieces of hardware instead of creative thinkers. This is a mistake, but it's understandable for a few reasons.

First, they speak in a nerd dialect of English, especially when they first emerge from school. This is an obstacle for communication with business people right off the bat.

Second, they often don't have the domain expertise needed to fully participate in business strategy discussions. This is also an obstacle, but it's exaggerated —mathy folks are experts at learning stuff quickly and will pick up domain expertise just as fast as they picked up linear algebra.

Third, they have a seemingly magic power, which makes it easy to pigeonhole

Finally, sometimes businesses don't actually want data people to do meaningful work—they just hired them as ornaments for their business, as marketing instruments to prove they're on the cutting edge of "big data." God forbid if the quants were actually given data from the business in this circumstance, because they'd probably figure out it's a shell game.

Nerds: Ask to be let into strategic discussions. Focus on learning the domain expertise and the vocabulary of the people around you so you can understand and contribute to the discussion. Point out when something people are guessing at can be substantiated with data and elbow grease, and do some mathematical investigations on questions the business cares about to show how that works. If you never get let into a discussion like this, look around and figure out if you're just window dressing, and if so, get yourself a better job where you'll actually learn something.

Business peeps: Invite the nerds into the room, and tell them you're open to suggestions. Spend an extra few minutes on making sure communication lines are clear, and look meaningfully at the nerds, with eyebrows raised expectantly, when questions of approximation or uncertainty come up—they might have useful things to say.

3) Interpreting Skepticism as Negativity

Another reason that quants are largely ignored is that, when they deliver news, it's not always good.

This is a mistake, and here's why. Your data scientist may very well know your business better than you do, assuming you give them a real job to do and that they're competent and engaged. Ignoring them is tantamount to a doctor turning off the heart monitor to avoid bad news about their patient.

I'm not saying that data people aren't risk averse—they sometimes are, and business folk might have to learn to discount bad news from the data team somewhat. But turning off that channel altogether is ignoring good information.

to be realistic.

Business folk: Look, in this drip-feed VC culture, where every board meeting has to be good news or your funding dries up, it's hard to hear that the numbers are bad, or that they don't tell the whole story and they could be bad, or that you're measuring the wrong thing, or that your model is biased, when you're desperate for control.

Even so, look at the long term and trust that your people are invested in the best result for the company.

4) Ignoring Wider Cultural Consequences

There are various ways that models can affect culture, and although many of them are pretty subtle and difficult to measure, others are not very subtle or difficult to measure but fall into the category of "not my problem." In economics this phenomenon is often referred to as "externality," and it's famously difficult to deal with.

For example, how do you make companies pay for the pollution they cause, especially when the effects are difficult to measure and spread out over many people and many years?

Another related possibility is the large-scale feedback loop. For example, it's been well documented that the existence of the system of Pell Grants, which are federal student loans for low-income students, has led to increased college tuitions. In other words, the extent to which Pell Grants has made college more affordable for poor students has been erased by the rising tuition cost.

When we acknowledge that there's a potential for models to cause externalities and large-scale feedback loops, it is tantamount to thinking of the public as a stakeholder in our model design. We recognize the possibility of long-term effects of modeling and our responsibility to make sure those effects are benign, or if they aren't, that their positive effects outweigh their negative effects.

Examples: It is an unfortunate fact that not much research has gone (yet) into

that. This is a field I'd love to see taken up by a data-driven and independent scientific community. And it's a great example of how something may well exist even if it's not being measured.

For now we can speculate, and call for examination, on the cultural effect of various models.

For example, to what extent does public access to average standardized test scores lead to increased residential segregation? In other words, when people choose where to buy a house, they look up the standardized test scores to decide which public schools are good. This means towns with better scores are more desirable, increasing the cost of the houses in those neighborhoods.

This apparent feedback loop adds to the widening equality gap in education and arguably undermines the extent to which it can be honestly called "public education." If you're paying an extra $200K for good schools, then you're essentially pricing out poor people from that resource.

Here's another well-known example: *U.S. News & World Report* publishes a well-known list of top colleges every year. Because of its widespread influence, this promotes cross-college competition for top students and has led to widespread gaming, which some people argue has decreased the quality of education nationwide in pursuit of a good rank.

Next, an up-and-coming model that worries me: online credit scoring, where the modelers include data such as how well you spell and punctuate to decide how much of a credit risk you are. This approach confuses correlation with causation and results in a credit scoring system that gives up your historical behavior data—how promptly you pay your bills—for demographic data—how "other people like you" have historically behaved. What's the feedback loop look like for such models if they become widely used?

Finally, let's go back to the assumption we often make in modeling that "N=ALL", which was brought up by Kenneth Cukier and Viktor Mayer-Schoenberger in their recent *Foreign Affairs* article, The Rise of Big Data. In their article, they argue that part of the power of the big data revolution comes from the fact that—unlike in the past where we had to work with small sample sizes and get only approximate results—nowadays in our world of GPS, tracking, and

people's behavior. But then again, the extent to which the "N=ALL" rule fails is critical to understanding how we impact culture with our models. Who is left out of the data? Whose vote is not counted? Are we starting with a model trained on one population, say in the United States, and deploying it in a totally different population? What is the result?

Nerds: This cultural feedback loop stuff is hard, maybe the hardest part of your job. How do we even devise a test for measuring feedback loops, especially before a model goes into production? And yet, it's not a waste of your time to go through the thought experiment of how models will affect people if they are used on a wide scale.

Business people: Please consider the "losers" of a system as well as the "winners." So when you're bragging about how your online credit score gives some people better offers, make sure to acknowledge it also gives some people worse offers, depending on the person. What criteria you use to decide who to make which offer needs to be considered deeply and shouldn't be a stab in the dark.

The Sniff Test for Big Data

In almost any modeling scenario, there's almost always a predator and a prey. And as the modeler, 99% of the time you're the predator.

In other words, you're usually trying to get people to do something—buy something, pay attention to something, commit to something, contribute their data in one way or another—and you're likely manipulating them as well.

That's not to say you're not giving them anything back in return, but let's be honest—most consumers don't understand the value of their contributions, or even that they've entered into a transaction, so it's unreasonable to say it's a fair deal.

Be clear about who your prey is and what effects you have on them, and come to terms with it. If you can't figure out who the prey is, but you're still somehow making money, think harder.

negative and fearful atmosphere that paralyses and intimidates us, a healthy level of skepticism is good for business and for fruitful creativity.

But of course, that doesn't mean it's easy to create or to maintain.

To look at examples of existing centers of skepticism, we might wish to look at data skepticism as it currently exists in academia. Unfortunately, my impression having attended a recent academic conference on this topic is that there's an unreasonable distance between academics who study data and the culture of data from actual data practitioners.

And while the academics are thinking about the right things—cultural effects of modeling, the difference between code and algorithm—the language barrier and hands-on experience differential is a serious obstacle, at least for now.

That means we need to find a place inside business for skepticism. This is a tough job given the VC culture of startups in which one is constantly under the gun to perform, and it's just as hard in the cover-your-ass corporate culture typical of larger companies.

In other words, I'm calling for finding a place for skepticism, but I have no easy formula for how to achieve it, and I have sympathy for people who have tried and failed. I think the next step is to collect best practices on what works.

Let's end on a positive note with some really good news.

First, there are excellent tools currently being built that should prove extremely helpful in the quest for thoughtful data storytelling, communication, and sharing.

The open source community is hard at work perfecting products such as the IPython Notebook that allows data scientist to not only share code and results with nontechnical people, but to build a narrative explaining their decision processes along the way. The medium-term goal would be to allow that non-technical person to interact and play with the model to gain understanding and intuition, which will go a long way toward creating an environment of candid communication with the business.

although that is not a panacea.

Data is a tool, and like any other tool, it's neither good nor evil; that depends on how it's used. And whether or not any application of data is good or bad doesn't even have to do with whether the data is misapplied: bad guys can do great data analysis, and they can also screw up, just like the good guys (and most guys think they're good guys). But you're more likely to use data effectively, and understand how other people are using it, if you develop a healthy skepticism, which is to say a healthy habit of mind to question—and insist on understanding—the arguments behind the conclusions.

Post topics: Building a data culture