All we want are the facts, ma'am

In these troubled times, we can all appreciate a stand-up guy like Joe Friday, the fictional *Dragnet* detective who is best remembered for saying *"Just the facts, ma'am."* Except he never said that. What he actually said was *"All we want are the facts, ma'am."* That's actually a much better sentiment. Be gracious when you're interviewing---let the interviewees have their say and don't force them towards one topic. When you've gathered all the data, then sort out the facts.

I've never been involved in a detective interview, but occasionally I'm interviewed by a journalist. Some publications use a quaint procedure known as *fact checking*; for example, the *New Yorker* had a wonderful piece by John McPhee in their Feb. 9th issue on their fact-checking process. (My favorite part was about fact-checking the absence of a comma. The phrase *"William Penn's daughter Margaret fished in the Delaware"* implies that Penn *may* have multiple daughters, and thus by Grice's maxim that he *does*. If you put commas around "Margaret," then the name becomes a non-essential appositive, and implies that she is his only daughter.) Journalists have been worried for at least 15 years about the deterioration of fact checking; as publications cut costs, fact-checking can be an early victim.

I recently had a run-in with the fact-checkers for *Wired* magazine. They wrote and asked me:

```
Is it true that at your ETech presentation in March, you said, in a direct homage
to George Box, "All models are wrong, and you don't need them anyway"? Is that
accurate?
```

Great, I thought--*Wired* is a publication with integrity and wants to get the facts right. I wrote back:

```
The quote I used was "essentially all models are wrong, but some are useful".

The point I was making -- and I don't remember the exact words -- was that if the
model is going to be wrong anyway, why not see if you can get the computer to
quickly learn a model from the data, rather than have a human laboriously derive a
model from a lot of thought.
```

In other words, with more data you can use weaker priors (perhaps with non-parametric or semi-parametric models), but you certainly still need models.

I figured they would either use the quote I gave them, paraphrase it, or drop it completely if it didn't fit with the point of the story. But when Chris Anderson's story *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* came out in June 2008, there was a fourth possibility that I hadn't even counted upon: they attributed to me a made-up quote that actually contradicts the reply I gave them:

> *Peter Norvig, Google's research director, offered an update to George Box's maxim: "All models are wrong, and increasingly you can succeed without them."*

To set the record straight: **That's a silly statement, I didn't say it, and I disagree with it.**

The ironic thing is that even the article's author, Chris Anderson, doesn't believe the idea. I saw him later that summer at Google and asked him about the article, and he said "I was going for a reaction." That is, he was being provocative, presenting a caricature of an idea, even though he knew the idea was not really true. That's a mode I expect from other publications, but it's not what I want from *Wired*, and I don't expect *Wired* to make up facts to support their caricature.

Anderson deserves credit for spotting an important development and bringing it to the attention of a wider audience. When he writes "in the era of big data, more isn't just more. More is different.", he has got it just right, in my opinion. But he could have written an article that preserves his key insight while being both provocative *and* true. He correctly noted that the methodology for science is evolving; he cites examples like shotgun sequencing of DNA. Having more data, and more ways to process it, means that we can develop different kinds of theories and models. But that does not mean we throw out the scientific method. It is not "The End of Theory." It it is an important change (or addition) in the methodology and set of tools that are used by science, and perhaps a change in our stereotype of scientific discovery. Today we think of a scientific breakthrough as the brilliant lone scientist, watching an apple drop or sitting in his bathtub, and suddenly having a *Eureka!* moment.

In reality, most discoveries are the result of a series of small steps made by many scientists. In the end, the time is right for someone to synthesize these small steps into a new theory. When the new theory is as pithy as $F = m\ a$, the promoter of the final step can take the lion's share of the credit (even if he points out that he was standing on the shoulders of giants). But if the final step is a complex theory that is not fully known even to the discoverer, but can be approximated by a model derived from massive amounts of data, then things becomes messier--how much of the credit goes to the one who took the final step? How much to the many people who collected the data? How much to the statisticians who helped tune the data models?

Using data and statistics is not a new idea in science, and it is *always* done with respect to a theory. To take a simple example, if you flip a coin 1000 times and it comes up heads 513 times, you could say that you have a model that predicts a 51.3% (or you might say "about 50%") chance that the next flip will be heads. But that prediction depends on theory: the theory that the future will be similar to the past; the theory that the result of the coin flip does not depend on the current exchange rate for gold, nor on whether there are a prime number of people whose middle name is "Alice" in the room, and so on.

In my talk at ETech, all the examples I gave of using data also involved building models based on a theory. For example, you can create a spell corrector from data rather than by the sweat and tears of linguists, but you need a model. The current dominant model assumes that words can be modeled by a Markov sequence, that the distribution of words is stationary over time, and that spelling errors with a small edit distance are more likely than errors with a large edit distance. And so on. We recognize that these models leave out a lot of phenomena, but they are good enough for the task at hand.

I think we are all agreed that models are here to stay, and that it doesn't make sense to talk of doing science (or computer science) without them. There is an important technical point about the use of data for large-scale non-parametric models that I would characterize thusly:

> In complex, messy domains, particularly game-theoretic domains involving unpredictable agents such as human beings, there are no general theories that can be expressed in simple equations like $F = m\ a$ or $E = m\ c^2$. But if you have a dense distribution of data points, it may be appropriate to employ non-parametric density approximation models such as nearest-neighbors or kernel methods rather than parametric models such as low-dimensional linear regression.

But perhaps only a statistician could love that idea. I care deeply about how to adapt methods like support vector machines to very large data sets in a computationally tractable way. But I understand that the casual reader of *Wired* might not share that passion.

However, there is another point that should appeal to all: the idea that it can be difficult to fully understand the implications of a complex model. I think that should have been the centerpiece of Anderson's article (and then he could have quoted me accurately). The great thing about $F = m\ a$ is that it is so simple that we can easily see how to apply it to falling objects on Earth, and then to the orbits of the moon and planets, and then to the flight of spacecraft. But complex models may hold secrets that they are less willing to give up.

There's a famous (well famous in Artificial Intelligence circles, anyways) Zen koan that goes like this:

> In the days when Sussman was a novice, Minsky once came to him as he sat hacking at the PDP-6.
> "What are you doing?", asked Minsky.
> "I am training a randomly wired neural net to play Tic-Tac-Toe," Sussman replied.

> "Why is the net wired randomly?", asked Minsky.
>
> "I do not want it to have any preconceptions of how to play", Sussman said.
>
> Minsky shut his eyes.
>
> "Why do you close your eyes?", Sussman asked his teacher.
>
> "So that the room will be empty."
>
> At that moment, Sussman was enlightened.

Minsky was showing Sussman that the randomly-wired neural net (a complex model if ever there was one) actually *did* have preconceptions; it's just that we don't understand what these preconceptions are. So it is with large data sets. There is plenty to learn from them, both about the domains that the data come from, and about the methodology for learning from data.

But to be clear: the methodology still involves models. Theory has not ended, it is expanding into new forms. Sure, we all love succinct theories like $F = m\ a$. But social science domains and even biology appear to be inherently more complex than physics. Let's stop expecting to find a simple theory, and instead embrace complexity, and use as much data as well as we can to help define (or estimate) the complex models we need for these complex domains.

OK, so *Wired* had an important insight, but missed the real story, and misquoted me. At the time I was willing to shrug off the misquote. But the *New Yorker* article inspired me to aspire higher. Publications can and should have a higher standard for accuracy, and we readers should call them on it when they miss.

So, three cheers for John McPhee, for *The New Yorker*, and for their corps of fact-checkers. I'm proud that my subscription to *The New Yorker* supports the fact-checkers in some small way.

As for *Wired*, I have one closing thought: *All we want are the facts, ma'am.*

Peter Norvig

---

11 Comments **norvig.com** 1 **Login**

♡ **Recommend** 11 Tweet Share **Sort by Best**

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

baesharam • 6 years ago

Someone once said, "Wired is to Science what Playboy is to erotica".

2 ^ | ⌄ • Reply • Share ›

anandventakaraman • 9 years ago

Many years ago 20/20 ran a story on New Zealand TV chastising the practice of whaling in Japan, and how it was a barbaric and uncivilized practice. When I saw the story, I wrote to the editor that it's unfair to make a self-exonerating statement like this singling out a people for reprimanding, especially in a place like New Zealand, famous for its lean cut beef, lamb shanks and its ever-popular delicacy - boiled-alive crayfish (lobsters).

The following week they read my letter out on TV in the feedback section of the program. Only it was paraphrased. Yep - Who'd have even thought that they "paraphrase" your letters, and without clearing with you first! They read "Anand Venkataraman from Massey University writes: 'It's not unfair for the Japanese to be whaling. Eating beef in New Zealand is a bigger problem". My instant reaction was "Yeah, I'm Indian, as you have cleverly induced from my name oh 20/20 wizards, and I don't eat beef, but it's not because I'm Indian that I don't eat beef." In fact, I don't follow any religion! I'm a vegetarian, and a "wanna-be" vegan because I'm a PeTA activist. A series of email exchanges followed, leading eventually to my conviction that no retraction will be published.

On that day I decided that objectivity in journalism is constrained by that subset of objective facts or their projections into the space of what makes for interesting and entertaining news. It is perhaps the germ of my discontent that led to my eventual blog post two years ago on "Google news, air and TCP/IP" (http://www.pandamatak.com/p...

&
--
Anand Venkataraman

3 ^ | ˅ • Reply • Share ›

leidner • 9 years ago

Dear Peter,

As somebody who wrote an undergrad paper on a corpus-based study defining and non-defining relative clauses back in '94, I couldn't agree more on the importance of the comma :-).

Yes, there is a lot of pressure on the newspapers (and by implication on news agencies) to cut expenses, so the decline of fact checking is perhaps not a big surprise. In fact, large part of the decline is due to the declining sales of newspaper advertising, and search engines have a hand in this.

So I think what we all should do (and your position gives you some leverage to affect change, I would think) is to think about new revenue sharing models for newspapers and search engines. Content providers and online advertisers form a symbiosis: no content, not search for content - "Just the facts, [Sir]".

Unfortunately, there's a follow-up problem once this issue is fixed: there is a danger that ever more populist content will appear on the Web, and the more investigative journlism may no longer be appreciated (in terms of

$$$) by enough people to make it worthwhile. Paper newspapers dealt with the problem implicitly, because you can only buy them as a whole. In the online world, the story "Michael Jackson died (74th update)" is separated from more specialized, in-depth piece "Plight of women in Southern Aghanistan". One way is to offer investigative journalism for pay. However, if popular sells, than even that model could in the mid-term be threatened...

Best regards
Jochen

1 • Reply • Share ›

Terranova • 10 years ago

i enjoyed this and learned from it. thanks.

Reply • Share ›

Anonymous • 10 years ago

The techniques you've cited are sometimes called "model free" in the literature. Maybe you were upset at this journalist, and so that's why you glossed over this, but it is troubling to me that your reply ignores this. Of course "free" is technically the wrong word, but there is a legitimate intellectual question of whether "dumb" nearest neighbor models (or their analogues in other areas) are good enough. Is modern science moving away from having sophisticated models? Elementary techniques from machine learning seem to be yielding interesting science results recently:

http://news.cnet.com/robo-s...
http://blog.wired.com/wired...

How much of society, communication, cognition, and science are really just based on using "dumb" nearest neighbor matching, i.e. the most naive model? Or similar theoretically "boring" models? It's fascinating to me.

Reply • Share ›

alagesan • 10 years ago

I have one doubt regarding "free will". Even probability is just a way of quantifying what we don't know by using previous statistical results. Everything seems to be deterministic with cause and effect relationship. I don't think some local randomness can be thought of as free will, like for example at this moment I can choose to either type in this comment or just stare at the monitor. I am not convinced to think this ability to choose as free well. Even the previous example may not be random, as I don't know the causes and so I claim it to be random. So then what exactly is free will? It appears as if free will should be equivalent to randomness.

Reply • Share ›

Popperian • 10 years ago

Popperian · 10 years ago

> essentially all models are wrong, but some are useful

What does that say about your meta-model of models? That it is wrong. I think that some models may be right ("true"), even if we can never know (in the Platonic sense) that they are, i.e. even if we can never have "ultimate" justification for them. (I also think that the condition here holds.)

> But that prediction depends on theory: the theory that the future will be similar to the past; the theory that the result of the coin flip does not depend on the current exchange rate for gold, nor on whether there are a prime number of people whose middle name is "Alice" in the room, and so on.

I accept that prediction always has theoretical underpinnings (I see it as a deductive argument, and only theories can provide strictly universal projectans premises and concepts), but I should note that the particular theories mentioned are in no way necessary for such a prediction to be