asecondmouse
Reflections on social science, politics and education

Instability Forecasting Models: Seven Ethical Considerations

Posted on February 20, 2019

So, welcome, y'all, to the latest bloggy edition on an issue probably relevant to, at best, a couple hundred people, though once again it has been pointed out to me that it is likely to be read by quite a few of them. And in particular, if you are some hapless functionary who has been directed to read this, a few pointers

- "Seven" is just a meme in this blog
- Yes, it is too long: revenge of the nerds. Or something. More generally, for the length you can blame some of your [so-called] colleagues to whom I promised I'd write it
- You can probably skip most of the footnotes. Which aren't, in fact, really footnotes so much as another meme in the blog. Some of them are funny. Or at least that was the original intention
- You can skip Appendix 1, but might want to skim Appendix 2
- ICEWS = DARPA Integrated Conflict Early Warning System; PITF = U.S. multi-agency Political Instability Task Force; ACLED = Armed Conflict Location and Event Data; PRIO = Peace Research Institute Oslo; UCDP = Uppsala [University] Conflict Data Program; DARPA = U.S. Defense Advanced Research Projects Agency; EU JRC = European Commission Joint Research Centre
- Yes, I'm being deliberately vague in a number of places: Chatham House rules at most of the workshops and besides, if you are part of this community you can fill in the gaps and if you aren't, well, maybe you shouldn't have the information [1]

Violating the Bloggers Creed of absolute self-righteous certainty about absolutely everything, I admit that I'm writing this in part because some of the conclusions end up at quite a different place than I would have expected. And there's some inconsistency: I'm still working this through.

Prerequisites out of the way, we shall proceed.

Our topic is instability forecasting models—IFMs—which are data-based quantitative models, originally statistical, now generally using machine learning methods, which forecast the probabilities of various forms of political instability such as war, civil war, mass protests, even coups, at present typically (though not exclusively) at the level of the nation-state and with a time horizon of about two years. The international community developing these models has, in a sense, become the dog that caught the car: We've gone from "forecasting political instability is impossible: you are wasting your time" to "everyone has one of these models" in about, well, seven years.

As I've indicated in Appendix 1—mercifully removed from the main text so that you can skip it—various communities have been at this for a *long* time, certainly around half a century, but things have changed—a lot—in a relatively short period of time. So for purposes of discussion, let's start by stipulating three things:

1. Political forecasting per se is nothing new: any policy which requires a substantial lead time to implement (or, equivalently, which is designed to affect the state of a political system into the future, sometimes, as in the

Marshall Plan or creation of NATO and later the EU, very far into the future) requires some form of forecasting: the technical term (okay, *one* technical term…) is "feedforward." The distinction is we now can do this using systematic, data-driven methods.[2]

2. The difference between now and a decade ago is that these models work and they are being seriously implemented, with major investments, into policy making in both governments and IGOs. They are quite consistently about 80% accurate,[3] against the 50% to 60% accuracy of most human forecasters (aside from a very small number of "superforecasters" who achieve machine-level accuracy). This is for models using public data, but I've seen little evidence that private data substantially changes accuracy, at least at the current levels of aggregation (it is possible that it might at finer levels in both geographical and temporal resolution) [4]. The technology is now mature: in recent workshops I've attended, both the technical presentations and the policy presentations were more or less interchangeable. We know how to do these things, we've got the data, and there is an active process of integrating them into the policy flow: the buzzphrase is "early warning and early action" (EWEA), and the World Bank estimates that even if most interventions fail to prevent conflict, the successes have such a huge payoff that the effort is well worthwhile even from an economic, to say nothing a humanitarian, perspective.
3. In contrast to weather forecasting models—in many ways a good analogy for the development of IFMs—weather doesn't respond to the forecast, whereas political actors might: We have finally hit a point where we need to worry about "reflexive" prediction. Of course, election forecasting has also achieved this status, and consequently is banned in the days or weeks before elections in many democracies. Economic forecasting long ago also passed this point and there is even a widely accepted macroeconomic theory, rational expectations, dealing with it. But potential reflexive effects are quite recent for IFMs.

As of about ten years ago, the position I was taking on IFMs—which is to say, before we had figured out how to create these reliably, though I still take this position with respect to the data going to these—was that our ideal end-point would be something similar to the situation with weather and climate models [5]: an international epistemic community would develop a series of open models that could be used by various stakeholders—governments, IGOs and NGOs—to monitor evolving cases of instability across the planet, and in some instances these alerts would enable early responses to alleviate the conflict—EWEA—or failing that, provide, along the lines of the famine forecasting models, sufficient response to alleviate some of the consequences, notably refugee movements and various other potential conflict spill-over effects. As late as the mid-2000s, that was the model I was advocating.

Today?—I'm far less convinced we should follow this route, for a complex set of reasons both pragmatic and ethical which I still have not fully resolved and reconciled in my own mind, but—progress of sorts—I think I can at least articulate the key dimensions.

1. Government and IGO models are necessarily going to remain secret, for reasons both bureaucratic and practical.

Start with the practical: in the multiple venues I've attended over the past couple of years, which is to say during the period when IFMs have gone from "impossible" to "we're thinking about it" to "here's our model", everyone in official positions has been adamant that their operational models are not going to become public. The question is then whether those outside these organizations, particularly as these models are heavily dependent on NGO and academic data sets, should accept this or push back.

To the degree that this tendency is simply traditional bureaucratic siloing and information hoarding—and there are certainly elements of both going on—the natural instinct would be to push back. However, I've come to accept the argument that there could be some legitimate reasons to keep this information confidential due to the fact that the decisions of governments and IGOs, which can potentially wield resources on the order of billions of dollars, can have substantial reflexive consequences on decisions that could affect the instability itself, in particular

- foreign direct investment and costs of insurance
- knowledge that a conflict is or is not "on the radar" for possible early action
- support for NGO preparations and commitments
- prospects for collective action, discussed below

2. From an academic and NGO perspective, there is a very substantial moral issue in forecasting the outcome of a collective action event.

This is the single most difficult issue in this essay: are there topics, specifically those dealing with collective action, which should be off-limits, at least in the public domain, even for the relatively resource-poor academic and NGO research communities?

The basic issue here is that—at least with the current state of the technology—even if governments and IGOs keep their exact models confidential, the past ten years or so have shown that one can probably fairly easily reverse engineer these except for the private information: at least at this point in time, anyone trying to solve this problem is going to wind up with a model with relatively clear set of methods, data and outcomes, easily duplicated with openly available software and data.[6][7]

So in our ideal world—the hurricane forecasting world—the models are public, and when they converge, the proverbial red lights flash everywhere, and the myriad components of the international system gear up to deal with the impending crisis, and when it happens the early response is far more effective than waiting until the proverbial truck is already halfway over the cliff. And all done by NGOs and academic researchers, without the biases of governments.

Cool. But what if, instead, those predictions *contribute to* the crisis, and in the worst case scenario, cause a crisis that otherwise would not have occurred. For example through individuals reading predictions of impending regime transition, using that information to mobilize collective action, which then fails: we're only at 80% to 85% accuracy as it is, and this is before taking into account possible feedback effects. [8] Hundreds killed, thousands imprisoned, tens of thousands displaced. Uh, bummer.

One can argue, of course, that this is no different that what is already happening with *qualitative* assessments: immediately coming to mind is the Western encouragement of the Hungarian revolt in 1956, the US-supported Bay of Pigs invasion, North Vietnam's support of the Tet Offensive, which destroyed the indigenous South Vietnamese communist forces,[9] and US ambiguity with respect to the Shi'a uprisings following the 1991 Iraq War. And this is only a tiny fraction of such disasters.

But they were all, nonetheless, disasters with huge human costs, and actions which affect collective resistance bring to mind J.R.R. Tolkien's admonition: "Do not meddle in the affairs of wizards, for they are subtle and quick to anger." Is this the sort of thing the NGO and academic research community, however well meaning, should risk?

3. Transparency is nonetheless very important in order to assess limitations and biases of models.

Which is what makes the first two issues so problematic: despite the convergence in the existing models, every model has biases [10] and while the existing IFMs have converged, there is no guarantee that this will continue to be the case as new models are developed which are temporally and/or spatially more specific, or which take on new problems, for example detailed refugee flow models. Furthermore, since the contributions of the academic and NGO communities were vital to moving through the "IFM winter"—see Appendix 1—continuing to have open, non-governmental efforts seems very important.

Two other thoughts related to this

1. Is it possible that the IFM ecosystem has become *too* small because the models are so easy to create? I'm not terribly worried about this because I've seen, in multiple projects, very substantial efforts to explore the possibility that other models exist, and they just don't seem to be there, at least as for the sets of events currently of interest, but one should always be alert to the possibility of what appears to be a technological maturity is a failure of imagination.
2. *Current* trends in commercial data science (as opposed to open source software and academic research) may not be all that useful for IFM development because this is not a "big data" problem: one of the curious things I noted at a recent workshop on IFMs is that deep learning was never mentioned. Though looking forward counterfactually, it is also possible that rare events—where one can envision even more commercial applications than those available in big data—are the next frontier in machine learning/artificial intelligence.

4. Quality is more important than quantity.

Which is to say, this is not a task where throwing gigabytes of digital offal at the problem is going to improve results, and we may be reaching a point where some of the inputs to the models have been deliberately and significantly manipulated because such manipulation is increasingly common. Also there is a danger in focusing on where the data is most available, which tends to be areas where conflict has occurred in the past and state controls are weak. High levels of false positives—notably in some atomic (that is, ICEWS-like) event data sets—are bad and contrary to commonly-held rosy scenarios, duplicate stories aren't a reflection of importance but rather of convenience, urban and other biases.

The so-called web "inversion"—the point where more information on the web is fake than real, which we are either approaching or may have already passed—probably marks the end, alas, of efforts to develop trigger models—the search for anticipatory needles-in-a-haystack in big data—in *contemporary* data, though it is worth noting that a vast collection of texts from prior to the widespread manipulation of electronic news feeds exists (both in the large data aggregators—LexisNexis, Factiva, and ProQuest—and with the source texts held, under unavoidable IP restrictions, by both ICEWS, the University of Illinois Cline Center, the University of Oklahoma TERRIER project and presumably the EU JRC) and these are likely to be extremely valuable resources for developing filters which can distinguish real from fake news. They could also be useful in determining whether, in the past, trigger models are real, rather than a cognitive illusion borne of hindsight—having spent a lot of time searching for these with few results, I'm highly skeptical, but it is an empirical question—but any application of these in the contemporary environment will require far more caution than would have been needed, say, a decade ago.[11]

5. Sustainability of data sources.

It has struck me at a number of recent workshops—and, amen, in my own decidedly checkered experience in trying to sustain near-real-time atomic event data sets—the degree to which event data—structural data being generally solidly funded as national economic and demographic statistics—used in IFM models depends on a large number of small projects without reliable long-term funding sources. There are exceptions—UCDP as far as I understand has long-term commitments from the Swedish government, both PRIO and ACLED have gradually accumulated relatively long-term funding through concerted individual efforts, and to date PITF has provided sustained funding for several data sets, notably Polity IV and less notably the monthly updates of the Global Atrocities Data Set—but far too much data is coming from projects with relatively short-term funding, typically from the US National Science Foundation, where social science grants tend to be just two or three years, with no guarantee of renewal, and grants from foundations which tend to favor shiny new objects over slogging through stuff that just needs to be done to support a diffuse community.

The ethical problem here is the extent to which one can expect researchers to invest in models using data which may not be available in the future, and, conversely, whether the absence of such guarantees is leading the collective research community to spend too much effort in the proverbial search for the keys where the light is best. Despite several efforts over the years, political event data, whether the "atomic" events similar to ICEWS or the "episodic" events similar to ACLED, the Global Terrorism Database, and UCDP, have never attained the privileged status the U.S. NSF has accorded to the continuously-maintained American National Election Survey and General Social Survey, and the user community may just be too small (or politically inept) to justify this. I keep thinking/hoping/imagining that increased automation in ever less expensive hardware environments will bring the cost of some of these projects down to the point where they could be sustained, for example, by a university research center with some form of stable institutional support, but thus far I've clearly underestimated the requirements.

Though hey, it's mostly an issue of money: Mr. and Ms. Gates, Ms. Powell-Jobs, Mr. Buffet and friends, Mr. Soros, y'all looking for projects?

6. Nothing is missing or in error at random: incorrect predictions and missing values carry information.

This is another point where one could debate whether this involves ethics or just professional best-practice—again, don't confine your search for answers to easy methods are easy where you can just download some software—but these decisions can have consequences.

The fact that information relevant to IFMs is not missing at random has been appreciated for some time, and this may be one of the reasons why machine learning methods—where "missing" is just another value—have fairly consistently out-performed statistical models. This does, however, suggest that imputation—now much easier thanks to both software and hardware advances—may not be a very good idea and is potentially an important source of model bias.

There also seems to be an increasing appreciation that incorrect predictions, particularly false positives (that is, a country or region has been predicted to be unstable but is not) may carry important information, specifically about the resilience of local circumstances and institutions. And more generally, those off-diagonal cases—both the false positives and false negatives—are *hugely* important in the modeling effort and should be given far more attention than I'm typically seeing. [12]

A final observation: at what point are we going to get situations where the model is wrong because of policy interventions? [8, again] Or have we already? — that's the gist of the EWEA approach. I am guessing that in most cases these situations will be evident from open news sources, though there may be exceptions where this is due to "quiet diplomacy"—or as likely, quiet allocation of economic resources—and will quite deliberately escape notice.

7. Remember, there are people at the end of all of these.

At a recent workshop, one of the best talks—sorry, Chatham House rules—ended with an impassioned appeal on this point from an individual from a region which, regrettably, has tended to be treated as just another set of data points in far too many studies. To reiterate: IFMs are predicting the behaviors of people, not weather.

I think these tendencies have been further exacerbated by what I've called "statutory bias" [10, again] in both model and data development: the bureaucratic institutions responsible for the development of many of the most sophisticated and well-publicized models are prohibited by law from examining their own countries (or in the case of the EU, set of countries). And the differences can be stark: I recently saw a dashboard with a map of mass killings based on data collected by a European project which, unlike PITF and ICEWS data, included the US: the huge number of cases both in the US and attributable to US-affiliated operations made it almost unrecognizable compared to displays I was familiar with.

This goes further: suppose the massive increase in drug overdose deaths in the US, now at a level exceeding 70,000 *per year*, and as amply documented, the direct result of a deliberate campaign by one of America's wealthiest families, whose philanthropic monuments blot major cities across the land, suppose this had occurred in Nigeria, Tajikistan or Indonesia, might we at the very least be considering that phenomenon a candidate for a new form of state weakness and/or the ability of powerful drug interests to dominate the judicial and legislative process? But we haven't.

On the very positive side, I think we're seeing more balance emerging: I am particularly heartened to see that ECOWAS has been developing a very sophisticated IFM, at least at the level of North American and European efforts, and with its integration with local sources, perhaps superior. With the increasing global availability of the relevant tools, expertise, and, through the cloud, hardware, this will only increase, and while the likes of Google and Facebook have convinced themselves only whites and Asians can write software, [13] individuals in Africa and Latin America know better.

Whew…so where does this leave us? Between some rugged rocks and some uncomfortable hard places, to be sure, or there would have been no reason to write all of this in the first place. Pragmatics aside—well-entrenched and well-funded bureaucracies are going to set their own rules, irrespective of what academics, NGOs and bloggers are advocating—the possibility of developing models (or suites of models) which set off ill-advised collective action concerns me. But so does the possibility of policy guided by opaque models developed with flawed data and techniques, to say nothing of policies guided by "experts" whose actually forecasting prowess is at the level of dart-throwing chimps. And there's the unresolved question of whether there something special about the forecasts of a quantitative model as distinct from those of an op-ed in the *Washington Post* or a letter or anonymous editorial in *The Economist*, again with demonstrably lower accuracy and yet part of the forecasting ecosystem for a century or more. Let the discussion continue.

I'll close with a final personal reflection that didn't seem to fit anywhere else: having been involved in these efforts for forty or so years, it is very poignant for me to see the USA now almost completely out of this game, despite the field having largely been developed in the US. It will presumably remain outside until the end of the Trump administration, and then depending on attitudes in the post-Trump era, rebuilding could be quite laborious given the competition with industry for individuals with the required skill sets though, alternatively, we could see a John Kennedyesque civic republican response by a younger generation committed to rebuilding democratic government and institutions on this side of the Atlantic. In the meantime, as with high speed rail, cashless payments, and universal health care, the field is in good hands in Europe. And for IFMs and cashless payments, Africa.

Footnotes

1. I went to college in a karst area containing numerous limestone caves presenting widely varying levels of technical difficulty. The locations of easy ones where you really had to make an effort—or more commonly, drink—to get yourself into trouble were widely known. The locations of the more difficult were kept confidential among a small group with the skills to explore them safely. Might we be headed in a similar direction in developing forecasting models?—you decide.

Someone about a year ago at one of these IFM workshops—there have been a bunch, to the point where many of the core developers know each other's drink preferences—raised the issue that we don't want forecasts to provide information to the "bad guys." But where to draw the line on this, given that some of the bad guys can presumably reverse engineer the models from the literature, given the technical sophistication we've seen by such groups, e.g. in IEDs and the manipulation of social media. Suddenly the five-year publication lags (and paywalls?) in academic journals becomes a good thing?

2. I finally realized the reason why we haven't had serious research into how to integrate quantitative and qualitative forecasts—this is persistently raised as a problem by government and IGO researchers—is the academics and small research shops like mine have a *really* difficult time finding real experts (as opposed, say, to students or Mech Turkers) who have a genuine interest and knowledge of a topic, as distinct from just going through the motions and providing uninformed speculation. In such circumstances the value added by the qualitative information will be marginal, and consequently we're not doing realistic tests of expert elucidation methods. So by necessity this problem—which is, in fact, quite important—is probably going to have to be solved in the government and IGO shops.

3. I'm using this term informally, as the appropriate metric for "accuracy" on these predictions, which involve rare events, is complicated. Existing IFMs can consistently achieve an AUC of 0.80 to 0.85, rarely going above (or below) that level, which is not quite the same as the conventional meaning of "accuracy" but close enough. There are substantial and increasingly sophisticated discussions within the IFM community on the issue of metrics: we're well aware of the relevant issues.

4. One curious feature of IFMs may be that private data will become important at *short* time horizons but not at longer horizons. This contrasts to the typical forecasting problem where errors increase more or less exponentially as the time horizon increases. In current IFMs, structural indicators (mostly economic, though also historical), which are readily available in public sources, dominate in the long term, whereas event-based conditions may be more important in the short term. E.g. "trigger models"—if these are real, an open question—are probably not relevant in forecasting a large-scale event like Eastern Europe in 1989 or the Arab Spring, but could be very important in forecasting at a time horizon of a few weeks in a specific region.

5. *Science* had a nice article [http://science.sciencemag.org/content/363/6425/342] recently on these models: Despite the key difference of IFMs being potentially reflexive and the fact that that one of our unexplored domains is the short term forecast, some of the approaches used in those models—emphasized in the excerpt below—could clearly be adapted to IFMs

> Weather forecasts from leading numerical weather prediction centers such as the European Centre for Medium-Range Weather Forecasts (ECMWF) and National Oceanic and Atmospheric Administration's (NOAA's) National Centers for Environmental Prediction (NCEP) have also been improving rapidly: A modern 5-day forecast is as accurate as a 1-day forecast was in 1980, and useful forecasts now reach 9 to 10 days into the future (*1*). Predictions have *improved for a wide range of hazardous weather conditions [emphasis added]*, including hurricanes, blizzards, flash floods, hail, and tornadoes, with skill emerging in predictions of seasonal conditions.
>
> …
>
> Because data are unavoidably spatially incomplete and uncertain, the state of the atmosphere at any time cannot be known exactly, producing forecast uncertainties that grow into the future. This sensitivity to initial conditions can never be overcome completely. *But, by running a model over time and continually adjusting it to maintain consistency with incoming data [emphasis added]*, the resulting physically consistent predictions greatly improve on simpler techniques. Such data assimilation, often done using four-dimensional variational minimization, ensemble Kalman filters, or hybridized techniques, has revolutionized forecasting.
>
> Sensitivity to initial conditions limits long-term forecast skill: Details of weather cannot be predicted accurately, even in principle, much beyond 2 weeks. But weather forecasts are not yet strongly constrained by this limit, and the increase in forecast skill has shown no sign of ending. *Sensitivity to initial conditions varies greatly in space and time [emphasis added]*, and an important but largely unsung advance in weather prediction is *the growing ability to quantify the forecast uncertainty [emphasis added]* by using large ensembles of numerical forecasts that each start from slightly different but equally plausible initial states, together with perturbations in model physics.

6. I'm constantly confronted, of course, with the possibility that there are secret models feeding into the policy process that are totally different than those I'm seeing. But I'm skeptical, particularly since in some situations, I'm the only person in the room who has been witness to the *process* by which independent models have been developed, such being the reward, if that's the word, for countless hours of my life frittered away in windowless conference rooms watching PowerPoint™ presentations. All I see is convergence, not just in the end result, but also in the development process.

Consequently if a trove of radically different—as distinct from incrementally different, however much their creators think they are novel—secret models exists, there is a vast and fantastically expensive conspiracy spanning multiple countries creating an elaborate illusion *solely for my benefit*, and frankly, I just don't think I'm that important. I'm sure there are modeling efforts beyond what I'm seeing, but from the glimmers I see of them, they tend to be reinventing wheels and/or using methods that were tried and rejected years or even decades ago, and the expansiveness (and convergence) of known work makes it quite unlikely—granted, not impossible—that there is some fabulously useful set of private data and methodology out there. To the contrary, in general I see the reflections from the classified side as utterly hampered by inexperience, delusional expectations, and doofus managers and consultants who wouldn't make it through the first semester of a graduate social science methodology course and who thus conclude that because something is impossible for them, it is impossible for anyone. Horse cavalry in the 20th century redux: generally not a path with a positive ending.

7. Providing, of course, one wants to: there may be specialized applications where no one has bothered to create public models even though this is technically possible.

8. One of the more frustrating things I have heard, for decades, is a smug observation that if IFMs become successful, the accuracy of our models will decline and consequently we modelers will be very sad. To which I say: bullshit! Almost everyone involved in IFM development is acutely aware of the humanitarian implications of the work, and many have extended field experience in areas experiencing stress due to political instability (which is not, in general, true of the folks making the criticisms, pallid Elois whose lives are spent in seminar rooms, not in the field). To a person, model developers would be ecstatic were the accuracy of their models to drop off because of successful interventions, and this is vastly more important to them than the possibility of Reviewer #2 recommending against publication in a paywalled journal (which, consequently, no one in a policy position will ever read) because the AUC hasn't improved over past efforts.

9. Back in the days when people still talked of these things—the end of the Vietnam War now being almost as distant from today's students than the end of World War I was from my generation—one would encounter a persistent urban legend in DoD operations research—ah, OR…now there's a golden oldie…—circles that somewhere deep in the Pentagon was a secret computer model—by the vague details, presumably one of Jay Forrester's systems dynamics efforts, just a set of difference equations, as the model was frequently attributed to MIT—that precisely predicted every aspect of the Vietnam War and had decision-makers only paid attention to this, we would have won. You know, like "won" in that we'd now be buying shrimp, t-shirts and cheap toys made in Vietnam and it would be a major tourist destination. I digress.

Anyway, I'm pretty sure that in reality dozens of such models were created during the Vietnam War period, and some of them were right some of the time, but, unlike the Elder Wand of the Harry Potter universe, no such omniscient Elder Model existed. This land of legends situation, I would also note, is completely different than where we are with contemporary IFMs: the models, data, methods, and empirical assessments are reasonably open, and there is a high degree of convergence in both the approaches and their effectiveness.

10. I'd identify five major sources of bias in existing event data: some of these affect structural data sets as well, but it is generally use to be aware of these.

1. **Statutory bias,** also discussed under point 7: Due to its funding sources, ICEWS and PITF are prohibited by a post-Vietnam-era law from tracking the behavior of US citizens. Similarly, my understanding is that the EU IFM efforts are limited (either by law or bureaucratic caution) in covering disputes between EU members and internal instability within them. Anecdotally, some NGOs also have been known to back off some monitoring efforts in some regions in deference to funders.
2. **Policy bias**: Far and away the most common application of event data in the US policy community has been crisis forecasting, so most of the effort has done into collecting data on violent (or potentially violent) political conflict. The EU's JRC efforts are more general, and for example have foci on areas where the EU may need to provide disaster relief, but is still strongly focused on areas of concern to the EU.
3. **Urban bias**: This is inherent in the source materials: for example during the Boko Haram violence in Nigeria, a market bombing in the capital Abuja generated about 400 stories; one in the regional capital of Maiduguri would typically generate ten or twenty, and one in the marginal areas near Lake Chad would generate one or two. Similarly, terrorist incidents in Western capitals such as Paris or London generate days of attention where events with far higher casualty rates in the Middle East or Africa typically are covered for just a day.
4. **Media fatigue**: This is the tendency of news organizations to lose interest in on-going conflicts, covering them in detail when they are new but shifting attention even though the level of conflict continues.
5. **English-language bias**: Most of the event data work to date—the EU JRC's multi-language work being a major exception—has been done in English (and occasionally Spanish and Portuguese) and extending beyond this is one of the major opportunities provided by contemporary computationally-intensive methods, including machine translation, inter-language vector transformations, and the use of parallel corpora for rapid dictionary development; IARPA has a new project called BETTER focused on rapid (and low effort) cross-language information extraction which might also help alleviate this.

11. See for example https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

12. Though this is changing, e.g. see Michael Colaresi https://twitter.com/colaresi/status/842291411298996224 on bi-separation plots, which, alas, links to yet-another-frigging paywalled article, but at least the sentiment is there.

13. See https://www.nytimes.com/2019/02/13/magazine/women-coding-computer-programming.html. Google and Facebook have 1% blacks and 3% Hispanics in their technical employees! Microsoft, to its credit, seems to be more enlightened.

Appendix 1: An extraordinarily brief history of how we got here

This will be mostly the ramblings of an old man dredging up fading memories, but it's somewhat important, in these heady days of the apparently sudden success of IFMs, to realize the efforts go *way* back. In fact there's a nice MA thesis to be done here, I suppose in some program in the history of science, on tracking back how the concept of IFMs came about. [A1]

Arguably the concept is firmly established by the time of Leibnitz [], who famously postulated a "mathematical philosophy" wherein

> "[…] if controversies were to arise, there would be no more need of disputation between two philosophers than between two calculators. For it would suffice for them to take their pencils in their hands and to sit down at the abacus, and say to each other (and if they so wish also to a friend called to help): Let us calculate."

I'm too lazy to thoroughly track things during the subsequent three centuries, but Newtonian determinism expressed through equations was in quite the vogue during much of the period—Laplace, famously—and by the 19th century data-based probabilistic inference would gradually develop, along with an ever increasing amount of demographic and economic data, and by the 1920s, we had a well-established, if logically inconsistent, science of frequentist statistical inference. The joint challenges of the Depression and planning requirements of World War II (and Keynesian economic management more generally) led to the incorporation of increasingly sophisticated economic models into policy making in the 1930s and 1940s, while on the political side, reliable public opinion polling was established after some famous missteps, and by the 1950s used for televised real-time election forecasting.

By the time I was in graduate school, Isaac Asimov's *Foundation Trilogy*—an extended fictional work whose plot turns on the failures of a forecasting model—was quite in vogue, and on a more practical level, the political forecasting work of the founder of numerical meteorology, Lewis Fry Richardson—originally done in the 1930s and 1940s then popularized in the early 1970s by Anatol Rapoport and others, and by the establishment of the *Journal of Conflict Resolution*—who in 1939 self-published a monograph titled *Generalized Foreign Politics* where he convinced himself [A2] that the unstable conditions in his arms race models, expressed as differential equations, for the periods 1909-1913 and 1933-1938 successfully predicted the two world wars. Also at this point we saw various "systems dynamics" models, most [in]famously the Club of Rome's fabulously inaccurate *Limits to Growth* model published in 1972, which spawned about ten years of [also very poorly calibrated] similar efforts.

More critically, by the time I was in graduate school, DARPA was funding work on IFMs at a level that kept me employed as a computer programmer rather than teaching discussion sections for introductory international relations classes. These efforts would carry on well into the Reagan administration—at no less a level than the National Security Council, under Richard Beale's leadership of a major event data effort—before finally being abandoned as impractical, particularly on the near-real-time data side,

In terms of the immediate precedents to contemporary IFMs, in the 1990s there were a series of efforts coming primarily coming out of IGOs and NGOs—specifically Kumar Rupesinghe at the NGO International Alert and the late Juergen Dedring within the United Nations (specifically its Office for Research and the Collection of Information)—as well as the late Ted Robert Gurr in the academic world, Vice President Al Gore and various people associated with the US Institute for Peace in the US government, and others far too numerous to mention (again, there's a modestly interesting M.A. thesis here, and there is a very ample paper trail to support it) but again these went nowhere beyond spawning the U.S. State Failures Project, the direct predecessor of PITF, but the SFP's excessively elaborate (expensive, and, ultimately, irreproducible) IFMs initially failed miserably due to a variety of technical flaws.

We then went into a "IFM Winter"—riffing on the "AI Winter" of the late-1980s—in the 2000s where a large number of small projects with generally limited funding continued to work in a professional environment which calls to mind Douglas Adams's classical opening to *Hitchhiker's Guide to the Galaxy*

> Far out in the uncharted backwaters of the unfashionable end of the western spiral arm of the Galaxy lies a small unregarded yellow sun. Orbiting this at a distance of roughly ninety-two million miles is an utterly insignificant little blue green planet whose ape-descended life forms are so amazingly primitive that they still think digital watches are a pretty neat idea.

Yeah, that's about right: during the 2000s IFM work was definitely amazingly primitive and far out in the academically unfashionable end of some uncharted backwaters. But this decade was, in fact, a period of gestation and experimentation, so that by 2010 we had seen, for example, the emergence of the ACLED project under Clionadh Raleigh, years of productive experimentation at PITF under the direction of Jay Ulfelder, the massive investment by DARPA in ICEWS [A3], substantial modeling and data collections effort at PRIO under the directorship of Nils-Petter Gleditsch and substantial expansion of the UCDP datasets. While models in the 1960s and 1970s were confined to a couple dozen variables—including some truly odd ducks, like levels of US hotel chain ownership in countries as a measure of US influence—PITF by 2010 had assembled a core data set containing more than 2500 variables. Even if it really only needed about a dozen of these to get a suite of models with reasonable performance.

All of which meant that the IFM efforts which had generally not been able to produce credible results in the 1990s became—at least for any group with a reasonable level of expertise—almost trivial to produce by the 2010s.[A5] Bringing us into the present.

Appendix footnotes

A1. A colleague recently reported that a journal editor, eviscerating an historical review article no less, required him (presumably because of issues of space, as we all are aware that with electronic publication, space is absolutely at a premium!) to remove all references to articles published prior to 2000. Because we are all aware that everything of importance—even sex, drugs, and rock-and-roll!—was introduced in the 21st century.

A2. I'm one of, I'm guessing, probably a couple dozen people who have actually gone through Richardson's actual papers at Lancaster University (though these were eventually published, and I'd also defer to Oliver Ashford's 1985 biography as the definitive treatment) and Richardson's parameter estimates which lead to the result of instability are, by contemporary standards, a bit dubious and using more straightforward methods actually leads to a conclusion of stability rather than instability. But the thought was correct…

A3. Choucri and Robinson's *Forecasting in International Relations* (1974) is a good review of these efforts in political science, which go back into the mid-1960s. As that volume has probably long been culled from most university libraries, Google brings up this *APSR* review by an obscure assistant professor at Northwestern but, demonstrating as ever the commitment of professional scientific organizations and elite university presses to the Baconian norm of universal access to scientific knowledge, reading it will cost you $25. You can also get a lot from an unpaywalled essay by Choucri still available at MIT.

A4. The ICEWS program involved roughly the annual expenditures of the entire US NSF program in political science. Even if most of this went to either indirect costs or creating PowerPoint™ slides, with yellow type on a green background being among the favored motifs.

A5. As I have repeated on earlier occasions—and no, this is not an urban legend—at the ICEWS kick-off meeting, where the test data and the unbelievably difficult forecasting metrics, approved personally by no less than His Stable Genius Tony Tether, were first released, the social scientists went back to their hotel rooms and on their laptops had estimated models which beat the metrics before the staff of the defense contractors had finished their second round of drinks at happy hour. Much consternation followed, and the restrictions on allowable models and methods became ever more draconian as the program evolved. The IFM efforts of ICEWS—the original purpose of the program—never gained traction despite the success of nearly identical contemporaneous efforts at PITF—though ICEWS lives on, at least for now, as a platform for the production of very credible near-real-time atomic event data.

Appendix 2: Irreducible sources of error

This is included here for two reasons. First, the exposition of a systematic set of reasons as to why IFMs have an accuracy "speed limit"—apparently an out-of-sample AUC in the range of 0.80 to 0.85 at the two-year time horizon for nation-states—and if you try to get past this, in all likelihood you are just over-fitting the model. Second, it takes far too long to go through all of these reasons in a workshop presentation, but they are important.

- Specification error: no model of a complex, open system can contain all of the relevant variables: "McChrystal's hairball" is the now-classic exposition of this.

- Measurement error: with very few exceptions, variables will contain some measurement error. And this presupposing there is even agreement on what the "correct" measurement is in an ideal setting.

- Predictive accuracy is limited by measurement error: for example in the very simplified case of a bivariate regression model, if your measurement reliability is 80%, your accuracy can't be more than 90%. This biases parameter estimates as well as the predictions.

- Quasi-random structural error: Complex and chaotic deterministic systems behave as if they were random under at least some parameter combinations. Chaotic behavior can occur in equations as simple as $x_{t+1} = ax_t^2 + bx_t$.

- Rational randomness such as that predicted by mixed strategies in zero-sum games.

- Arational randomness attributable to free-will: the rule-of-thumb from our rat-running colleagues: "A genetically standardized experimental animal, subjected to carefully controlled stimuli in a laboratory setting, will do whatever it damn pleases."

- Effective policy response: as discussed at several point in the main text, in at least some instances organizations will have taken steps to head off a crisis that would have otherwise occurred, and as IFMs are increasingly incorporated into policy making, this is more likely to occur. It is also the entire point of the exercise.

- The effects of unpredictable natural phenomenon: for example, the 2004 Indian Ocean tsunami dramatically reduced violence in the long-running conflict in Aceh, and on numerous occasions in history important leaders have unexpectedly died (or, as influentially, not died and their effectiveness was gradually diminished).

Tetlock (2013) independently has an almost identical list of the irreducible sources of forecasting error.

Please note that while the 0.80 to 0.85 AUC speed limit has occurred relentlessly in existing IFMs, there is no theoretical reason for this number, and with finer geographical granularity and/or shorter time horizons, this could be smaller, larger, or less consistent across behaviors. For a nice discussion of the predictive speed limit issue in a different context, criminal recidivism, see Science 359:6373 19 Jan 2018, pg. 263; the original research is reported in Science Advances 10.1126/sciadv.aao5580 (2018)

Share this:

Press This Twitter Facebook 4

Reblog Like

Be the first to like this.

Related

Going Feral! Or "So long, and thanks for all the fish…"
In "Higher Education"

Seven lessons the national Democratic Party should draw from the victory of John Bel Edwards [1]
In "Politics"

The legal status of event data
In "Methodology"

This entry was posted in Methodology, Politics. Bookmark the permalink.

asecondmouse

Create a free website or blog at WordPress.com.