Article from

Forecasting and Futurism

Month Year July 2015
Issue Number 11

What I've Learned from the Good Judgment Project

By Mary Pat Campbell

We have often heard of the supposed wisdom of crowds, and the downfall of experts, but as one person noted last year, not all crowds are all that good at predicting:[1]

> *"I read the results of my impromptu experiment as a reminder that crowds are often smart, but they aren't magic. Retellings of Galton's experiment sometimes make it seem like even pools of poorly informed guessers will automatically produce an accurate estimate, but, apparently, that's not true."*

The context of that quote was that the author, Jay Ulfelder, had a cousin who ran an impromptu contest online, asking people how many movies he had watched (in the theater) in the past 13 years. The cousin kept a record of every movie he watched (to remind himself of the perk of being master of his own schedule as a freelance writer).

Forty-five people submitted answers, and the average (the supposed "wisdom of crowds") was way off from the actual answer. However, some of the answerers were close to the true answer.

Jay continues:

> Whatever the reason for this particular failure, though, the results of my experiment also got me thinking again about ways we might improve on the unweighted average as a method of gleaning intelligence from crowds. Unweighted averages are a reasonable strategy when we don't have reliable information about variation in the quality of the individual guesses (see here), but that's not always the case. For example, if Steve's wife or kids had posted answers in this contest, it probably would have been wise to give their guesses more weight on the assumption that they knew better than acquaintances or distant relatives like me.
>
> Figuring out smarter ways to aggregate forecasts is also an area of active experimentation for the Good Judgment Project (GJP), and the results so far are encouraging. The project's core strategy involves discovering who the most accurate forecasters are and leaning more heavily on them. I couldn't do this in Steve's single-shot contest, but GJP gets to see forecasters' track records on large numbers of questions and has been using them to great effect. In the recently-ended Season 3, GJP's "super forecasters" were grouped into teams and encouraged to collaborate, and this approach has proved very effective. In a paper published this spring, GJP has also shown that they can do well with non-linear aggregations derived from a simple statistical model that adjusts for systematic bias in forecasters' judgments. Team GJP's bias-correction model beats not only the unweighted average but also a number of widely-used and more complex nonlinear algorithms.

What is this "Good Judgment Project" and who are their forecasters?

Jay's post happens to have been written at the end of their third season, and I've joined the GJP for the 4th season. While there are details of the current season I can't share, I can explain the background of the project, some of the basics of participation, and, most importantly, what I've learned so far.

HISTORY OF THE GOOD JUDGMENT PROJECT

The Good Judgment Project sprouted out of a number of surprises in the U.S. intelligence community. How could they have been blindsided by so many developments?

Part of the research coming out of those failures was a competition called the IARPA ACE tournament. IARPA stands for Intelligence Advanced Research Projects Activity, providing funding and running projects that are intended to dig into intelligence issues that cross multiple organizations within the U.S. government. According to their own description, IARPA undertakes "high-risk/high-payoff research … [in which] failures are inevitable. Failure is acceptable so long as the failure isn't due to a lack of technical or programmatic integrity and the results are fully documented."[2]

CONTINUED ON **PAGE 18**

The Good Judgment Project feeds into that mission—especially for the individual participant. Failure is a big part of the project—failure in forecasting. But more on that in a bit.

The ACE tournament run by IARPA stands for "Aggregative Contingent Estimation," and it's run under the Office of Anticipating Surprise (man, I'd love to direct that office). It was an attempt to provide better forecasts of geopolitical events. The Good Judgment Project is a spinoff from the project, being run by researchers at University of Pennsylvania and UC-Berkeley. They had put together an approach, in which forecasters were trained and measured that outperformed many of the other ACE tournament participants, and IARPA ACE sponsored them as a part of a four-year research project.

What is interesting is that while the project has discovered "superforecasters" as part of their project, they have also shown effective ways to train people to forecast.[3] The training involves learning how to think probabilistically (which we actuaries should be good at), how to battle cognitive bias (which we may be no better than most people), and in general, how to become more successful in forecasting.

As noted in the article referenced above, there are "clusters" of questions that are attacking higher-level issues from different angles:

> "Within each cluster, we offer numerous specific forecasting questions. For example, within the cluster about European economic and political integration, we asked a question in fall 2014 about whether voters in Scotland would pass the independence referendum, and within the Iran cluster, we have a question currently open that asks when Iran will release Jason Rezaian, the Washington Post's Tehran bureau chief, who has been detained for over five months."

I have seen both questions, one obviously closed (the Scots did not vote for independence) and the other still open. I will not comment on the questions specifically, but about what I've learned about myself and about forecasting in general.

JOINING SEASON 4

I first heard about the Good Judgment Project via an NPR story in April 2014:[4]

> "For the past three years, Elaine Rich and 3,000 other average people have been quietly making probability estimates about everything from Venezuelan gas subsidies to North Korean politics as part of the Good Judgment Project, an experiment put together by three well-known psychologists and some people inside the intelligence community.
>
> "According to one report, the predictions made by the Good Judgment Project are often better even than intelligence analysts with access to classified information, and many of the people involved in the project have been astonished by its success at making accurate predictions.
>
> "When Rich, who is in her 60s, first heard about the experiment, she didn't think she would be especially good at predicting world events. She didn't know a lot about international affairs, and she hadn't taken much math in school.
>
> "But she signed up, got a little training in how to estimate probabilities from the people running the program, and then was given access to a website that listed dozens of carefully worded questions on events of interest to the intelligence community, along with a place for her to enter her numerical estimate of their likelihood. …
>
> "She's in the top 1 percent of the 3,000 forecasters now involved in the experiment, which means she has been classified as a superforecaster, someone who is extremely accurate when predicting stuff like: Will there be a significant attack on Israeli territory before May 10, 2014?
>
> "In fact, Tetlock and his team have even engineered ways to significantly improve the wisdom of the crowd—all of which greatly surprised Jason Matheny,

one of the people in the intelligence community who got the experiment started.

"'They've shown that you can significantly improve the accuracy of geopolitical forecasts, compared to methods that had been the state of the art before this project started,' he said.

"What's so challenging about all of this is the idea that you can get very accurate predictions about geopolitical events without access to secret information. In addition, access to classified information doesn't automatically and necessarily give you an edge over a smart group of average citizens doing Google searches from their kitchen tables."

At the end of the article, I noticed they were going to start recruiting people for the fourth round in the research. All the prior forecasters who wanted to continue would do so, but there would be a new crop of people coming in. I preregistered and then qualified by taking an online quiz touching on a variety of geopolitical subjects (most news junkies can easily answer these) as well as some reasoning items.

After being accepted in the fourth season, I got some training, involving some big themes in putting together a forecast and in improving one's performance. I always have access to these materials if I need to review the concepts, but I knew several of these just due to my own readings on cognitive biases.

One of the most important things I learned, though, was how I'd be scored.

HOW TO EVALUATE A FORECAST … AFTER THE FACT

One of the most important parts of the project is that forecasts get a score for their forecast accuracy after the fact. What's used is a Brier score, originally developed by Glenn W. Brier in 1950. The GJP uses the full Brier score, originally developed by Glenn W. Brier in 1950,[5] which works for a wider variety of questions than the yes/no example given below. However, for the purposes of illustration, I'm going to use the simplest formulation of this score.

The simplest type of forecasting question is forecasting the probability that a specific event will happen. The outcome will be yes/no, and you're putting a probability on the "yes" occurring.

To give a really simple question: "Will it snow >1 inch in North Salem, N.Y. on March 1, 2015?"

Let's pretend I forecast this for the five days preceding March 1:

Date	Forecast Probability
24 Feb 2015	50%
25 Feb 2015	50%
26 Feb 2015	60%
27 Feb 2015	75%
28 Feb 2015	95%

At the end I will calculate a Brier score, which depends on whether I came close to the actual result, 0 = it didn't happen and 1 = it did.

It just so happened we got over an inch of snow on March 1.

If I had prescience, I would have predicted 100 percent for each day. That will be one extreme.

If I had the opposite of prescience, I would have predicted 0 percent for each day. That will be the other extreme.

The basic Brier score formula is:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_N)^2$$

Where N is the number of days in the forecast period, f_t is the forecast percentage on day t, and o_N is the ultimate outcome.

This score was originally developed for weather forecasts, where one would make a prediction of probability of rain for each day—each day would have one forecast. The GJP is

CONTINUED ON **PAGE 20**

looking at something different—because this is about events possibly developing over time, one would want to see forecasts coalescing and changing over time. One would hope it gets closer to correct.

If I had perfect prescience, the Brier score would be 0, and if the perfect opposite, the result is 1. So the lower the Brier score, the better.

Date	Forecast Probability	Brier score for day
24 Feb 2015	50%	0.25
25 Feb 2015	50%	0.25
26 Feb 2015	60%	0.16
27 Feb 2015	75%	0.0625
28 Feb 2015	95%	0.0025
	OVERALL BRIER SCORE	0.145

In my example, I did not do too poorly. The Brier score for going with a 50/50 guess is 0.25, so one would compare against that. The Brier score used by the GJP is not exactly as I did above, because it needed to be adaptable to multiple-choice answers, and not merely yes/no. The specific details are not important.

One important thing to note: because of the squaring of the difference of the probability and the actual outcome, one is penalized for being far from the mark. If you way underpredict the chances of an event, you get heavily penalized; and if you overpredict the chances, you get extremely penalized. Deviating in the wrong direction from the 50/50 mark hits you very hard, so one must be spare with predictions of 5 percent or 95 percent probabilities for any event.

But the point is that once a question is closed, and all the GJP questions are of a finite period and do get resolved (more on that in a bit), one can look at how one did. More importantly, one can look at one's own rank among forecasters within a small group.

THE IMPORTANCE OF FEEDBACK IN FORECASTING

I have forecasted a few dozen questions that have closed thus far, having made more than 100 individual forecasts (one can change the forecast for any specific open question once a day). I am grouped with several other people, but that's only for comparison purposes right now. Unlike some of the prior seasons, we new forecasters are not being grouped to collaborate on forecasts. Yet.

In addition, unlike prior seasons in which forecasters were not expected to explain their reasoning behind their forecasts, we are encouraged (and given extra points) for flagging key comments about our reasoning, and also checking off categories of activities we did to make the forecast (such as adjusting a forecast for the passage of time—getting closer to a deadline may make the event more or less likely to be fulfilled.)

Finally, I can look at the closed questions and see how they were resolved, and look at my entire forecasting history for the question. I have found this the most valuable portion of the project for me. We are encouraged to write post-mortem comments for ourselves (which we can also look at later), and in doing these post-mortems I have discovered the following:

REGULAR UPDATING IS GOOD

Part of the reason my scores were bad on some questions was because I did not revisit my forecasts often enough. I do not have time to check every day, but I do make sure I look at all my forecasts at least once a week.

TRY MORE!

Originally, I stuck to areas where I understood the issues better (or so I thought), and I'm coming to realize that I'm losing some valuable points thereby. Most of us aren't experts in all the topics being covered, and just doing a few Internet searches can do enough to get one off the 50/50 line for a forecast.

This one I'm still having trouble with.

REMEMBER THAT NOT ALL TIME SERIES (OR QUESTIONS) ARE EQUAL

One question I messed up on was because I forgot how volatile certain time series can be. Some of the questions asked are based on financial markets indicators, and not all of them develop smoothly.

In particular, I have to be careful of "threshold" questions—some of the finance questions are whether a particular financial indicator goes above or falls below a particular threshold within a time period. That's a very different question from whether it will still be above that threshold at the end of the period.

I had forgotten that certain things could jump drastically on news, within a few hours even, and though the particular item I was following did settle back to "normal" areas, the threshold had been crossed. And my Brier score was hit.

STAY AWAY FROM FUZZY QUESTIONS

Most of the questions they've been presenting to forecasters are very clear: will a certain event occur by a certain date? Whether that event occurred is usually clear to all.

However, they've started getting into "fuzzier" questions, and I have found some of what is being done with those frustrating.

I understand why they're doing this—the really important intelligence would tend to be of a fuzzy nature. They are also trying to be fair—there are responses and clarifications. One can request a clarification while a question is still open. After questions close, you can provide feedback as to whether you agree with how they resolved a question.

That said, I have limited time. I do not need extra frustration in my life and enough fuzzy questions in my day job. Maybe, if the GJP continues past the 4th season, I'll get the comfort to work on those fuzzy problems. But right now, I want to stick to items that are more clear.

GJP'S OWN FINDINGS

It does take a while for academic research to get published, but some of the results from the GJP has appeared in journals. The most recent journal article I can find is from January 2015:[6]

> This article extends psychological methods and concepts into a domain that is as profoundly consequential as it is poorly understood: intelligence analysis. We report findings from a geopolitical forecasting tournament that assessed the accuracy of more than 150,000 forecasts of 743 participants on 199 events occurring over two years. Participants were above average in intelligence and political knowledge relative to the general population. Individual differences in performance emerged, and forecasting skills were surprisingly consistent over time. Key predictors were: (a) dispositional variables of cognitive ability, political knowledge, and open-mindedness; (b) situational variables of training in probabilistic reasoning and participation in collaborative teams that shared information and discussed rationales (Mellers, Ungar, et al., 2014); and (c) behavioral variables of deliberation time and frequency of belief updating. We developed a profile of the best forecasters; they were better at inductive reasoning, pattern detection, cognitive flexibility, and open-mindedness. They had greater understanding of geopolitics, training in probabilistic reasoning, and opportunities to succeed in cognitively enriched team environments. Last but not least, they viewed forecasting as a skill that required deliberate practice, sustained effort, and constant monitoring of current affairs.

I think that abstract is accessible to the non-academic, but let's look at the media coverage of this:[7]

> "'Most people would expect to find domain experts doing well in their domain,' says Nick Hare, one of the super-forecasters (informally, they go by 'supers') whose performance in the project landed him an invitation to the Good Judgment Project's annual summer conference. But, in fact, 'there are people who are good at all domains'—outperforming even specialists. And they could hold the key to reconfiguring the way intelligence services think about making predictions in the future.

CONTINUED ON **PAGE 22**

WHAT I'VE LEARNED … | FROM PAGE 21

> "So, what makes Hare such a good forecaster? His success, he says, comes down not to knowledge but his capacity for 'active, open-minded thinking': applying the scientific method to look rigorously at data, rather than seeking to impose a given narrative on a situation."

I think this is really key. The point is to consider possibilities that might not accord with what you expect. In my own case, I'm looking at the feedback to try to improve, and I'm thinking of using these approaches in forecasts in my own job in insurance research.

Too often we may settle on an answer or forecast too quickly, based on our biases. The following:

- Actively seeking out information disconfirming our "gut instinct";
- Taking notes on our reasoning, to be referred to later;
- Regularly revisiting our predictions; and
- Conducting a post-mortem of the reasoning and process once an outcome is known;

are all great techniques I've learned (or re-learned, the hard way) by participating in the Good Judgment Project. I hope it continues for a fifth season, so I can continue to improve … and perhaps some of y'all will join me!

Mary Pat Campbell

Mary Pat Campbell, FSA, MAAA, PRM, is VP, Insurance Research at Conning in Hartford, Conn. She can be reached at marypat.campbell@gmail.com.

RELATED ARTICLES FROM SOA NEWSLETTERS

Wolzenski, Ben. "Predictably Irrational, by Dan Ariely The Hidden Forces that Shape our Decisions". Forecasting & Futurism Newsletter, December 2014. *https://soa.org/Library/Newsletters/Forecasting-Futurism/2014/december/ffn-2014-iss10.pdf*

Campbell, Mary Pat. "Know Thyself and Others". The Stepping Stone, May 2011. *https://www.soa.org/library/newsletters/stepping-stone/2011/may/stp-2011-iss42-campbell.aspx*

And maybe y'all can find more F&F articles that are related. ▼

ENDNOTES

1. Jay Ulfelder: Crowds Aren't Magic. May 20, 2014. http://goodjudgmentproject.com/blog/?p=215
2. About IARPA, accessed 2 March 2015. http://www.iarpa.gov/index.php/about-iarpa
3. "The Good Judgment Project: Improving intelligence estimates to support decision-makers", CHIPS Magazine, Jan – Mar 2015. http://www.doncio.navy.mil/chips/ArticleDetails.aspx?ID=5976
4. Spiegel, Alex. "So You Think You're Smarter than a CIA Agent" NPR parallels blog, April 2, 2014. http://www.npr.org/blogs/parallels/2014/04/02/297839429/-so-you-think-youre-smarter-than-a-cia-agent?utm_source=digg&utm_medium=email
5. Brier Score. Wikipedia. Accessed 2 March 2015. http://en.wikipedia.org/wiki/Brier_score
6. Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015, January 12). The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics. Journal of Experimental Psychology: Applied. Advance online publication. http://dx.doi.org/10.1037/xap0000040
7. Burton, Tara Isabella. "Could you be a 'super-forecaster'?". BBC Future. 20 January 2015. http://www.bbc.com/future/story/20150120-are-you-a-super-forecaster