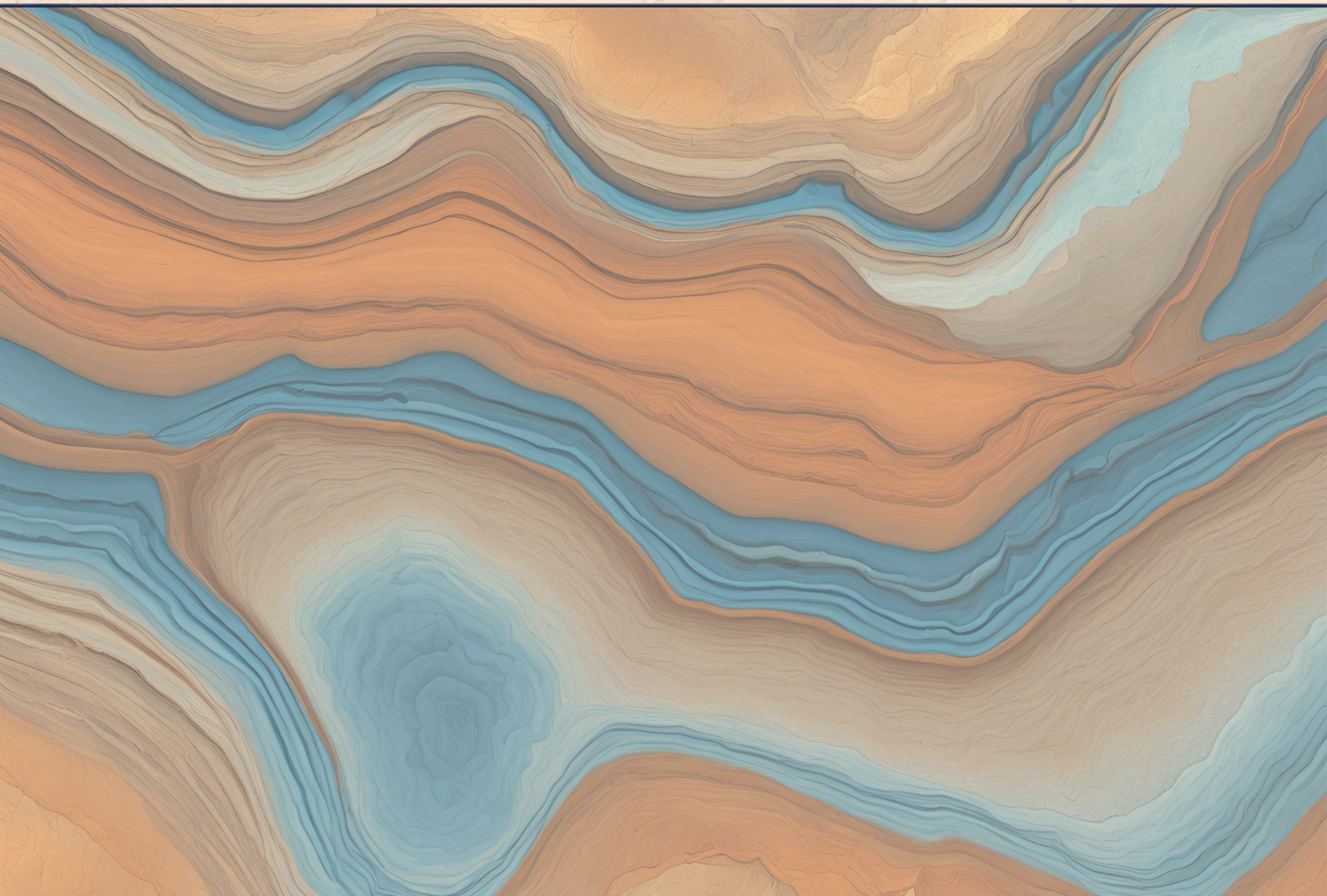
Comparing Top Forecasters to Domain Experts

Reviewing the evidence

An Arb Research Report

March 2022



The hypotheses

The superforecasting phenomenon — that certain teams of forecasters are better than other prediction mechanisms (large crowds, simple statistical prediction rules, etc) seems sound. But serious interest in superforecasting stems from the reported triumph of forecaster *generalists* over non-forecaster *experts*. (Another version says that they also outperform analysts with classified information.) So distinguish some claims:

1) Forecasters > public
2) Forecasters > simple models
3) Forecasters > experts
    a) Forecasters > experts with classified info
    b) Averaged forecasters > experts
    c) Aggregated forecasters > experts

Is (3) true? We review studies comparing the performance of trained forecasters and experts in some domain. We also tried to cover prediction markets vs experts.

Summary

We are fairly pessimistic (but uncertain):

- We think that claim (1) is true with 99% confidence[1] and that claim (2) is true with 95% confidence. But surprisingly few studies compare experts to generalists (3). The analysis quality and transparency leave much to be desired. The best study found that forecasters and health professionals performed similarly. In other studies, experts had goals besides accuracy, or there were too few to aggregate well.

- 3a) There's a common misconception that superforecasters outperformed intelligence analysts by "30%". Instead: Goldstein et al showed that [EDIT: the Good Judgment Project's best-performing aggregation method][2] outperformed the intelligence community, but this was partly due to the different aggregation technique used (the GJP weighting algorithm seems to perform better than prediction markets, given the apparently low volumes of the ICPM market). The forecaster *prediction market* performed about as well as the intelligence analyst prediction market; in general, prediction pools outperform prediction markets under current conditions (e.g. subsidies, volume, incentives, demographics). [85%]

[1] This is almost a trivial claim, since forecasters are by definition more interested in current affairs than average, and much more interested in epistemics than average. So we'd select for the subset of "the public" who should outperform simply through increased effort, even if all the practice and formal calibration training did nothing, which it probably doesn't.
[2] Previously this section said "superforecasters"; after discussion, it seems more prudent to say "the Good Judgment Project's best-performing aggregation method". See this comment for details.

- 3b) In the same study, the forecaster average was notably *worse* than the intelligence community.

- 3c) Ideally, we would pit a crowd of forecasters against a crowd of experts. Only an unpublished extension of Sell et al. manages this; it found a small (~3%) forecaster advantage.

- The bar may be low. Expertise, plus basic forecasting training and active willingness to forecast regularly, were enough to be on par with the best forecasters. [33%][3]

- In more complex domains, like ML, there could be significant returns to expertise. So it might be better to broaden focus from *generalist forecasters* to *competent ML pros who are excited about forecasting*. [40%]

[3] Our exact probability hinges on what's considered low and on how good trained Hypermind forecasters are. This is less obvious than it seems: in a similar tournament, the CSET Foretell Top Forecasters were not uniformly good; a team including Misha finished with a 4x better relative Brier score than the "top forecasters team." Further, our priors are mixed: (a) common sense makes us favor experts, (b) but common sense also somewhat favors expert forecasters, (c) Tetlock's work on expert political judgment pushes us away from politics experts, and finally (d) we have first-hand experience about superforecasting being real.

Table of studies

<table>
<tr><th></th><th>Comparison</th><th>Result</th><th>Notes</th></tr>
<tr><td colspan="4">Geopolitics</td></tr>
<tr><td>Goldstein et al (2015)</td><td>US Intelligence Community Prediction Market (ICPM)

Good Judgement Project (GJP): an average, vs prediction market (PM), vs best method (selected post hoc)[4]. Participants are rewarded for accuracy. ICPM was low stakes: play-money[5], while GJP participants “were paid a small honorarium for their active participation”.</td><td>Equal performance for expert and forecaster prediction markets. The best aggregation method was notably better than ICPM. The best method was selected post hoc among 20, but several of the other methods performed within 2% of the best.

Mean of means of daily Brier scores[6] (MMBD)<table><tr><td>Goldstein et al (2015)</td><td>MMBD</td><td>95% CI</td></tr><tr><td>ICPM</td><td>.23</td><td>(.19, .27)</td></tr><tr><td>GJP (avg)</td><td>.32</td><td>(.29, .35)***</td></tr><tr><td>GJP (PM)</td><td>.21</td><td>(.17, .26)</td></tr></table></td><td>Unpublished document used to justify the famous “Supers are 30% better than the CIA” claim.

The most direct comparison between forecasters (GJP PM) and experts (ICPM) finds similar performance (insignificant diff).
Prediction markets seem worse than super-aggregating opinion pools (see Appendix A); this study itself shows a large gap between GJP (PM) and GJP (best).</td></tr>
</table>

[4] All Surveys Logit ”takes the most recent forecasts from a selection of individuals in GJP’s survey elicitation condition, weights them based on a forecaster’s historical accuracy, expertise, and psychometric profile, and then extremizes the aggregate forecast (towards 1 or 0) using an optimized extremization coefficient.” Note that this method was selected *post hoc*, which raises the question of multiple comparisons; the authors respond that “several other GJP methods were of similar accuracy (<2% difference in accuracy).”

[5] There is some inconclusive research comparing real- and play-money: Servan-Schreiber et al. (2004) find no significant difference for predicting NFL (American football); Rosenbloom & Notz (2006) find that in non-sports events, real-money markets are more accurate and that they are comparably accurate for sports markets; and Slamka et al. (2008) finds real- and play-money prediction markets comparable for UEFA (soccer).

[6] MMBD is not a proper scoring rule (one incentivizing truthful reporting). If a question has a chance of resolving early (e.g., all questions of the form “will X occur by date?”), the rule incentivizes forecasters to report higher probabilities for such outcomes. This could have affected GJP (avg and best) predictors, who were rewarded for it; but should have not affected ICPM and GJP (PM), as these used the Logarithmic Market Scoring Rule. See Sempere & Lawsen (2021) for details.

<table>
<tr>
<td></td>
<td>N=139 geopolitical questions</td>
<td>
<table>
<tr><td>GJP (best)</td><td>.15</td><td>(.10, .21)***</td></tr>
</table>

Mean Percentage of Days Directionally Accurate (MPDDA)
<table>
<tr><td></td><td>MPDDA</td><td>95% CI</td></tr>
<tr><td>ICPM</td><td>81.58</td><td>(76.63, 86.54)</td></tr>
<tr><td>GJP (avg)</td><td>79.74</td><td>(73.82, 85.66)</td></tr>
<tr><td>GJP (PM)</td><td>83.45</td><td>(78.83, 88.06)</td></tr>
<tr><td>GJP (best)</td><td>88.20</td><td>(83.87, 92.5)***</td></tr>
</table>
** p < .001 vs ICPM.
</td>
<td></td>
</tr>
<tr>
<td>Stastny & Lehner (2018)</td>
<td>Qualitative forecasts from intelligence reports. Seasoned professional analysts produced[7]:
• initial personal probabilities,
• probabilities imputed in the reports,
• imputed[8] probabilities in light of current events,
• updated personal</td>
<td>Mean absolute error of ICPM was better (p<.001) than in the reports. Moreover, the initial forecasts by seasoned intelligence analysts were better (p<.05) than the forecasts imputed by them from the reports. Note that Initial forecasts were almost as good as ICPM forecasts.

Mean absolute error
<table>
<tr><td></td><td>All q.</td><td>Non-fuzzy q.</td><td>Fuzzy q.</td></tr>
<tr><td>Initial</td><td>0.317</td><td>—</td><td>—</td></tr>
<tr><td>Imputed</td><td>0.416</td><td>0.412</td><td>0.427</td></tr>
<tr><td>ICPM</td><td>0.302</td><td>0.305</td><td>0.3</td></tr>
</table>
</td>
<td>Mandel (2019) critiques the study. Table 1 is illuminating:

Mean Brier scores
<table>
<tr><td></td><td>all</td><td>non-f.</td></tr>
<tr><td>Initial personal</td><td>.194</td><td>.200</td></tr>
<tr><td>Imputed</td><td>.252</td><td>.254</td></tr>
<tr><td>Imputed upd.</td><td>.238</td><td>.243</td></tr>
</table>
</td>
</tr>
</table>

[7] Note that our understanding is that these were not averaged. On average there have been ~2.5 imputation predictions per report.
[8] Unclear if imputers did a reasonable job of separating their personal views from their imputations. Mandel (2019) notes that Person correlation between mean Brier scores for personal and imputed forecasts is very high, r(3)=.98, p=.005. Imputers average Brier scores ranged from .145 to .362 suggesting that traditional analysis' apparent accuracy depends on whether interpreters are better or worse forecasters. Lehner and Stastny (2019) responded. We don't take a stance on their dispute.

<table>
<tr>
<td></td>
<td>probabilities

Aforementioned ICPM

N=99 geopolitical questions, 28 of which had a “fuzzy” resolution criteria</td>
<td>Initial and imputed probabilities were compared to ICPM probabilities selected on the days on which the readers submitted their initial and imputed probabilities.

Due to the posting delay, ICPM had information not available to the report authors. However, longer posting delays would decrease ICPM's advantage.

Both ICPM probabilities and imputed estimates were poorly calibrated: with Calibration Indexes of .047 and .097 respectively (much higher than .025, .014, and .016 from other studies).</td>
<td>
<table>
<tr><td>Personal upd.</td><td>.150</td><td>.158</td></tr>
<tr><td>ICPM</td><td>.188</td><td>.195</td></tr>
</table>

Updated personal forecasts did better than ICPM (p=.087). Data suggests that seasoned analysts performed comparably to the prediction market. Note that their initial average Brier scores ranged from .145 to .362 so there is room for selection.

(See fn 8 for whether we can conclude anything about the quality of intelligence reports.)
</td>
</tr>
<tr>
<td>Kajdasz et al. (2014)</td>
<td>ICPM v. InTrade v. “10 best IC experts we could identify on each topic”

N=10 geopolitical questions[9]

N=”152 individual forecasts from the ICPM, InTrade, and individual IC experts over approximately matching topics and time horizons.”</td>
<td>Note that the three groups answered different questions ("approximately matching topics")

The market prices provided significantly[10] more accurate forecasts than experts. No statistical difference[11] in accuracy between the ICPM and InTrade.

Brier score summary statistics
<table>
<tr><td></td><td>$n_{forecasts}$</td><td>mean</td><td>std</td></tr>
<tr><td>ICPM</td><td>48</td><td>.0746</td><td>.13336</td></tr>
</table>
</td>
<td>Different number of $n_{forecasts}$ is confusing (as it suggests that prediction from groups might not have been well balanced; it would have been better if every forecast of IC SME was matched with forecasts from ICPM and InTrade on the same day and on the same time horizon).[12]

ICPM’s .075 Brier is 3x lower than its average across many questions reported in Goldstein et al. (2015). And InTrade is at .0366, which suggests that traders were rarely (if ever) predicting confidently, and so were rarely on the</td>
</tr>
</table>

[9] “We replicated some of these markets in the ICPM, or identified closely analogous predictions if they existed, so that direct comparisons between the two prediction markets could be made over time.”

[10] They report $F_{A\ comp}$ (1, 149) = 19.85, p<.01, $\hat{W}^2$ ψ = 0.1095.

[11] They report F(1, 149) = 1.33, n.s.

[12] Authors write: “We repeatedly collected forecasts from our markets and our experts to sample various time horizons, ranging from very near-term forecasts to as long as 4 months before a subject was resolved. All told, we collected 152 individual forecasts from the ICPM, InTrade, and individual IC experts over approximately matching topics and time horizons.”

<table>
<tr>
<td></td>
<td></td>
<td>
<table>
<tr><td>InTrade</td><td>50</td><td>.0366</td><td>.0634</td></tr>
<tr><td>IC SME</td><td>54</td><td>.1895</td><td>.2529</td></tr>
</table>
</td>
<td>wrong side of maybe.[13]</td>
</tr>
<tr>
<td>Beadle (2022); summary here</td>
<td>465,673 predictions over 3 years

1,375 participants

150 resolved questions, 240 in total

“The average time perspective in the FFI tournament was 521 days, i.e. four times as long as the questions in GJP.”</td>
<td>Unfortunately FFI-superforecasters were selected and evaluated on the 150 same questions, which makes regression toward the mean much more likely.[14]

“The standardised Brier scores of FFI superforecasters (0.36) were almost perfectly similar to that of the initial forecasts of superforecasters in GJP (0.37)”. “Note that GJP forecasters improved their scores after updating. However, the FFI forecasters could not update on their predictions”.

“Based on the first 150 questions, the average Brier score of the participants in FFI's tournament is 0.52 (SD: 0.11).” No better than predicting 50% on all questions.

“Moreover, even though regular forecasters in the FFI tournament were worse at prediction than GJP forecasters overall (probably due to not updating, training or grouping), the relative accuracy of FFI's superforecasters compared to regular forecasters (–0.06), and to defence researchers with access to classified information (–0.1) was strikingly similar.”</td>
<td>“In 2017 the Norwegian Research Defence Establishment (FFI) ran a forecasting tournament intended to investigate if GJP’s findings in the ACE tournament would replicate on questions with a longer time horizon, and in a Nordic context.”

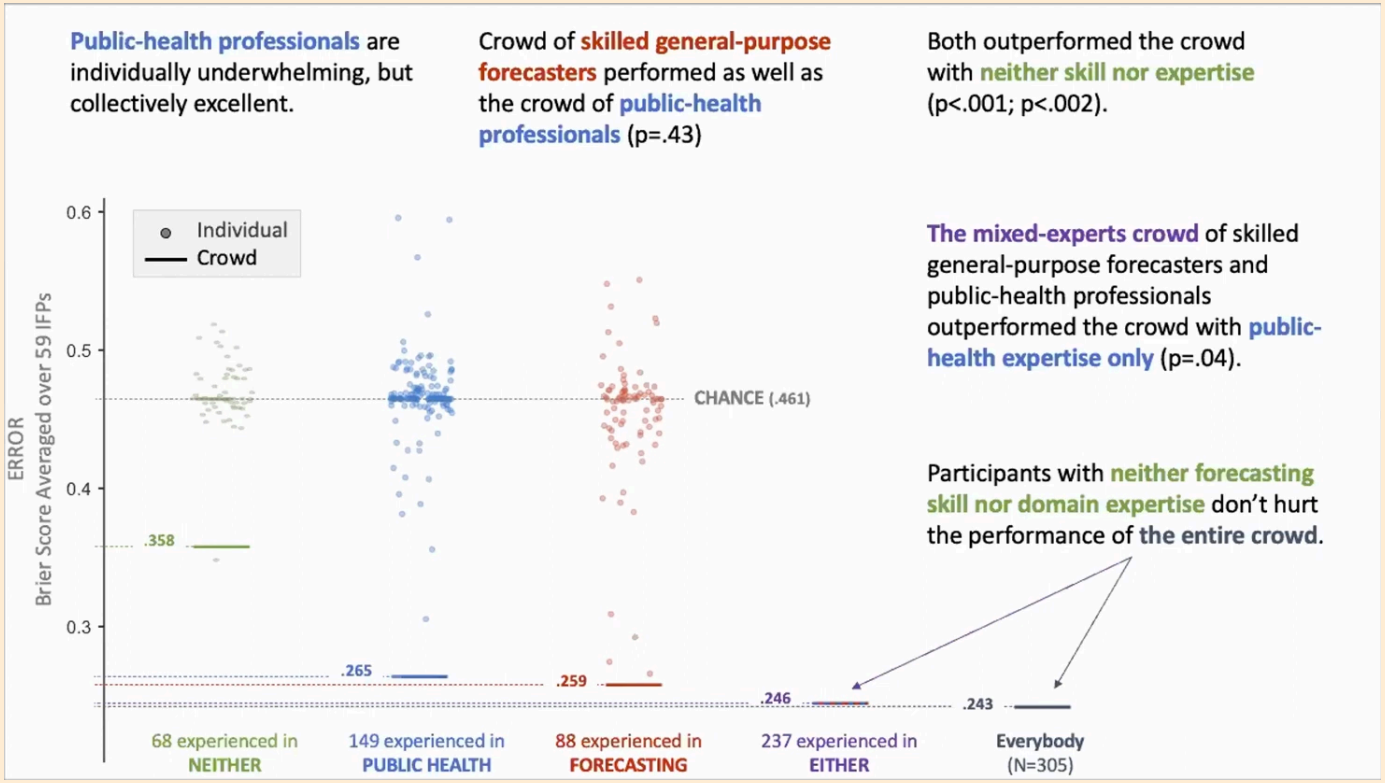
“An important difference from GJP is that FFI's tournament was open to anyone who wanted to participate.”</td>
</tr>
</table>

[13] A perfectly calibrated forecaster expects on average p - p^2 Brier points from their prediction. So this average Brier suggests that a “typical" InTrade prediction was either <4% or >96%. From experience, this feels too confident and suggests that questions were either biased towards low noise or that luck is partly responsible for such good performance.

[14] Per comment: “the 60 FFI supers were selected and evaluated on the same 150 questions (Beadle, 2022, 169-170). Beadle also identified the top 100 forecasters based on the first 25 questions, and evaluated their performance on the basis of the remaining 125 questions to see if their accuracy was stable over time, or due to luck. Similarly to the GJP studies, he found that they were consistent over time (Beadle, 2022, 128-131).”

Tetlock et al. (2023)	For the 25 year timeframe, there were 42 questions about nuclear proliferation, and 40 about boundaries. The study drew on EPJ studies of experts' probabilistic forecasts on slow-motion variables: • Base rate of change lower than 25% over 25 years. • Base rate of at least 5%. Exercises that had at least 25 forecasters and at least 25 forecasts per-participant.	Experts were more accurate than non-experts in nuclear proliferations questions. The Brier of non-experts was 60 % (= 0.08/0.05 - 1) higher for the 25 year timeframe. **Table 2** *Accuracy Scores for Nuclear Proliferation (NP) Forecasts* Brier Score: Timeframe, Baserate, Predict Baserate, Predict No Change, All Forecasters, NP Non-Expert, NP Expert; Log Score: NP Non-Expert, NP Expert 5 .02 .02 .02 .03 .03 .02 .11 .07 10 .10 .09 .10 .05 .06 .03 .18 .11 25 .12 .10 .12 .07 .08 .05 .26 .17 All .08 .07 .08 .05 .06 .03 .18 .12 *Note.* For Log Scores, we set probabilities of 0 at .01 and probabilities of 1 at .99. Experts were as accurate as non-experts in border change/secession questions. Their Briers were the same for each of the timeframes. **Table 4** *Accuracy Scores for Border Change/Secession (BCS) Forecasts* Brier Score: Timeframe, Baserate, Predict Baserate, Predict No Change, All Forecasters, BCS Non-Expert, BCS Expert; Log Score: BCS Non-Expert, BCS Expert 5 .10 .09 .10 .06 .06 .06 .21 .23 10 .13 .11 .13 .07 .07 .07 .25 .27 25 .23 .17 .23 .15 .15 .15 .51 .51 All .15 .13 .15 .09 .09 .10 .33 .34 *Note.* For Log Scores, we set probabilities of 0 at .01 and probabilities of 1 at .99.	**"Cruder operational definition that treated forecasters as experts if they had been educated at the post-graduate level in relevant disciplines and if they saw the topic as central to their professional identity"** "The study has many methodological shortcomings [since it "was never a priority in the larger EPJ project"]: • small sample sizes, • inadequate measures of expertise, • a flawed probability scale[15], and • a rushed schedule that gave forecasters little time to deliberate."
Pandemics			

[15] "Forecasters used the same 11-point, zero-to-one, subjective probability scale as in other EPJ exercises, with equal 0.1 spacing between levels (0, .1, .2, … , .9, 1)."

[Sell et al. (2021)](#) and [Servan-Schreiber (2021)](#)	Hypermind + John Hopkins study. Started a year before the pandemic. *Paper* Health pros (n=388) Hypermind forecasters (n=132 incl. 11 health pros) n=61 settled questions *Talk* Public health pros (n=149) Hypermind forecasters (n=88) (Sample from the talk is the subset of the crowd which was recruited earliest, thus with the most opportunities to forecast questions.)[16]	 *From the paper:* On the face of it, roughly equal. Of the top 10 forecasters: • 4 were public-health professionals, • 3 other health-related professionals • 3 Hypermind forecasters without a public-health background. • 5 vetted Hypermind forecasters.

[16] Personal communication with Servan-Schreiber.

<table>
<tr><td></td><td></td><td>And the 1st place went to one of the very few public-health professionals who was also a skilled Hypermind forecaster.

Key problem: experts got busy with the pandemic, so forecasters updated their forecasts relatively more often.

From the talk:
● Individually forecasters are 3% better (.454 v .467, p=.01).
● Crowds performed similarly.
● Mixed group +7% over experts alone.</td></tr>
<tr><td>McAndrew, Cambeiro, Besiroglu (2021)</td><td>Experienced life science pros (n=10)

Top-1% Metaculus forecasters (n=11)

Consensus: the aggregate of the 2 groups

Only 6 out 23 questions have resolved. They concerned safety, efficacy, and timing of a COVID-19 vaccine.</td><td>Trained forecasters had the highest log scores on average, followed by consensus models, and then subject-matter experts (nonsignificantly, the study is underpowered).

<table>
<tr><th></th><th>25th and 75th percentiles for log score</th><th>Mean scaled rank[17]</th></tr>
<tr><td>individuals all</td><td>[0.42, 2.98]</td><td>—</td></tr>
<tr><td>individual forecasters</td><td>—</td><td>.56 | 80CI: [.18, .94]</td></tr>
<tr><td>individual experts</td><td>—</td><td>.48 | 80CI: [.08, .98]</td></tr>
<tr><td>consensus all</td><td>[0.98, 2.96]</td><td>.58 | 80CI: [.49, .63]</td></tr>
<tr><td>consensus forecasters</td><td>[1.24, 2.90]</td><td>.56 | 80CI: [.43, .72]</td></tr>
<tr><td>consensus experts</td><td>[0.65, 3.07]</td><td>.53 | 80CI: [.35, .73]</td></tr>
</table>
</td></tr>
</table>

[17] Given N log scores, scaled rank assigns a value of 1/N to the smallest log score, a value of 2/N to the second smallest log score, and so on, assigning a value of 1 to the highest log score. (As with log scores, here computed from probability density functions, — the higher rank the better.)

<table>
<tr>
<td>Bosse et al. (2021)</td>
<td>Two semi-mechanistic models

Ensemble of all models submitted to the Forecast Hub

Crowd forecasts based on n=32 forecasters (17 are self-identified experts in forecasting or epidemiology)</td>
<td>Crowd consistently outperformed epidemiological models as well as the Hub ensemble when forecasting cases but not when forecasting deaths.

Weighted Interval Score (WIS, the lower the better) relative to the Hub ensemble
<table>
<tr><td>Two weeks ahead:</td><td>Cases</td><td>Deaths</td></tr>
<tr><td>Hub ensemble</td><td>1</td><td>1</td></tr>
<tr><td>Renewal model</td><td>1.40</td><td>1.79</td></tr>
<tr><td>Convolution model</td><td>—</td><td>1.22</td></tr>
<tr><td>Crowd</td><td>0.89</td><td>1.26</td></tr>
</table>
For cases, our contributions (compared to the Hub ensemble without our contributions) consistently improved performance across all forecasting horizons (e.g., rel. WIS 0.9, two weeks ahead).

For deaths, contributions from the renewal model and crowd forecast together improved performance only for one week ahead predictions and showed an increasingly negative impact on performance for longer horizons (rel. WIS 1.01 two weeks ahead, 1.05 four weeks ahead). Individual contributions from both the renewal model and the crowd forecast were largely negative.</td>
<td>Not clear how good Forecast Hub models were but their affiliations are impressive.

Irrespective of the above, suggests that crowd forecasting might be useful in practice.</td>
</tr>
<tr>
<td>Liptay (2021)</td>
<td>A single superforecaster

CDC-funded panel of experts

n=28 pandemic-related questions from UMass</td>
<td>Forecaster did 10% better than experts as judged by Brier score:
<table>
<tr><td>Superforecaster</td><td>.246</td></tr>
<tr><td>Experts</td><td>.268</td></tr>
</table></td>
<td>As usual, it’s unclear if the panel faced other incentives but forecasting accuracy.</td>
</tr>
</table>

Movies

Pathak et al (2015)

Movie critics: n=40

Betfair, a prediction market: variable n, including “low liquidity markets”

Predicting Oscar winners

Prediction market RMSE was 10%+ better than pundits.

RMSE for 2013 Oscar

	Days before	Categories	Experts	Betfair
Average, n=40	3	24	.20	.18
Nate Silver	3	6	.26	.18
Ben Zauzmer	8-9	21	.25	.20

(Hollywood Stock Exchange seems to be doing 10%..50% worse than Betfair, Intrade, and PredicWise.)

Spann & Skiera (2003)

Hollywood Stock Exchange, a virtual-points prediction market

Two expert predictions: Box Office Mojo, Box Office Report.

HSX is much better than BOR in terms of MAPE (n=24). And recalibrated HSX prediction is nonsignificantly different from BOM (n=140).

MAPE, n=24

HSX	40.62
HSX, recalibrated	36.48
BOM	35.30
BOR	53.40

MAPE, n=140

HSX	31.11
HSX, recalibrated	28.40

		BOM 28.05	
SCOTUS			
Katz et al. (2017)	7,000 participants 600,000 predictions 450 cases	Built an impressively accurate model on top of FantasySCOTUS predictions, and from Ruger et al. (2004) we know that simple models outperform experts.	FantasySCOTUS
Blackman et al. (2012)	The Forecasting Project's decision tree vs FantasySCOTUS vs The Forecasting Project's experts FantasySCOTUS most active users vs other users	75% v 64.7% v 59.1% — the comparison is between different terms.[18] The power predictor average, 7.93 points, was higher than the crowd average, 7.25 points. And "The results do not conclusively prove that the power predictors' forecasts were superior to those of the crowd. Although the power predictors generally do better, the crowd is able to make rather strong predictions to bridge the gap."	Most (seems like at least 75%) active FantasySCOTUS betters sometimes have specialized backgrounds. See a blogpost and ¶35-6 of the paper.
Ruger et al. (2004)	Fairly simple decision tree vs subject matter experts	The model predicted 75% of the cases correctly, which was more accurate than their experts with 59.1%.	The Forecasting Project, SCOTUS
Elections			

[18] It's unclear to me how well they did compare to a prior based on how often SCOTUS reverses the decisions. The historical average is ~70% with ~80% reversals in 2008, the relevant term.

<table>
<tr>
<td>Servan-Schreiber & Atanasov (2015)</td>
<td>Hypermind and 7 statistical models

6 questions on U.S. 2014 midterm elections: majority-control of the Senate and 5 most-undecided states.</td>
<td>Mean Daily Brier Score
<table>
<tr><td>.34</td><td>Hypermind</td></tr>
<tr><td>.41</td><td>Daily Kos</td></tr>
<tr><td>.43</td><td>Huffington Post</td></tr>
<tr><td>.43</td><td>PredictWise</td></tr>
<tr><td>.45</td><td>Models Mean</td></tr>
<tr><td>.46</td><td>Washington Post</td></tr>
<tr><td>.46</td><td>FiveThirtyEight</td></tr>
<tr><td>.48</td><td>New York Times</td></tr>
<tr><td>.68</td><td>Princeton Election Consortium</td></tr>
</table></td>
<td>Low n and errors are somewhat correlated, so not particularly informative.</td>
</tr>
<tr>
<td colspan="4">Miscellaneous</td>
</tr>
<tr>
<td>Cowgill & Zitzewitz (2015)</td>
<td>Corporate setting: demand forecasting, project completion, project quality, external events</td>
<td>MSE prediction market / MSE experts at firms
<table>
<tr><td>Ford</td><td>Google</td><td>—[19]</td><td>—</td></tr>
<tr><td>0.742</td><td>0.727</td><td>0.924</td><td>0.908</td></tr>
</table></td>
<td></td>
</tr>
</table>

[19] Anonymous basic materials conglomerate.

Search criteria

We were given a set of initial studies to branch out from.

- Good Judgement Project
- Tom McAndrew studies
- Hypermind + Johns Hopkins

And some general suggestions for scholarship:

- look for review articles
- look for textbooks and handbooks or companions
- find key terms
- go through researchers' homepages/google scholar

Superforecasting began with IARPA's ACE tournament.[20] We think the evidence in Tetlock's *Expert Political Judgment* doesn't fit: there were no known skilled-amateur forecasters at that point. See Appendix C: Tetlock's Expert Political Judgment.

A Google Scholar search for studies funded by IARPA ACE yielded no studies. We looked at other IARPA projects (ForeST, HCT, and OSI), which sounded remotely relevant to our goals.

We searched Google Scholar for (non-exhaustive list): "good judgment project", "superforecasters", "collective intelligence", "wisdom of crowds", "crowd prediction", "judgemental forecasting", …, and various combinations of these, and "comparison", "experts", …

We got niche prediction markets from the Database of Prediction Markets and searched for studies mentioning them. Hollywood SX and FantasySCOTUS paid off as a result. We also searched for things people commonly predict: sports, elections, Oscars, and macroeconomics.

In the process, we read the papers for additional keywords and references. We also looked for other papers from the authors we encountered.

Regarding AI forecasting

- In more complex domains, like ML, there could be significant returns to knowledge and expertise.

It seems to us that moving from *generalist forecasters* to *competent ML practitioners/researchers* might be better because:

- To predict e.g. scaling laws and emerging capabilities, people need to understand them, which requires some expertise and understanding of ML

[20] For completeness we could mention Galton (1907), the first demonstration of the wisdom of crowds.

- It's unclear whether general forecasters actually outperform experts in a legible domain, even though we believe in the phenomenon of superforecasting, (that some people are much better forecasters than most). We also liked David Manheim's take on Superforecasting.

- We think that this will plausibly reduce ML researchers' aversion to forecasting proposals — and if we were to execute it, we would be selecting good forecasters based on their performance anyway. It seems potentially feasible.

Finally, we note that the above reasoning is heavily limited by a lack of data (lack of it collected and a lack of it made available). We hope that the experimental data gets reanalyzed.

Thanks to Emile Servan-Schreiber, Luke Muehlhauser, and Javier Prieto for comments. These commenters don't necessarily endorse anything in this post, and mistakes are our own. Research funded by Open Philanthropy.

Appendix A: prediction markets vs. opinion pools

Appendix B: Table of less relevant studies

Appendix C: Tetlock's Expert Political Judgment

See also: Database of Prediction Markets

Changelog