Evidence-based Decision Making: Session 6

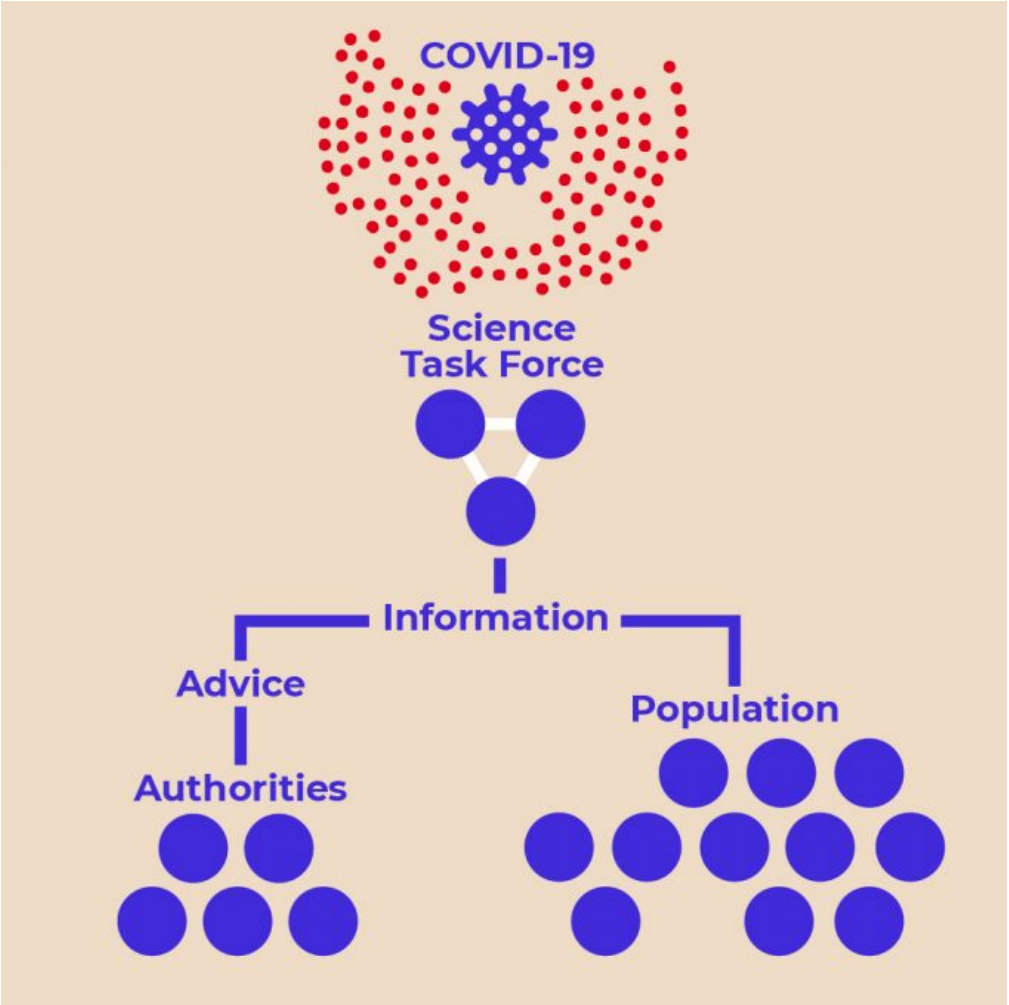
Rui Mata, FS 2021

Goals

Understanding the performance of groups as a process of statistical aggregation and learn about how to predict when crowds vs. experts will do best.

Learn about how psychology is using the tools of aggregation/consensus to change the way economic and political forecasting is conducted.

Exercise



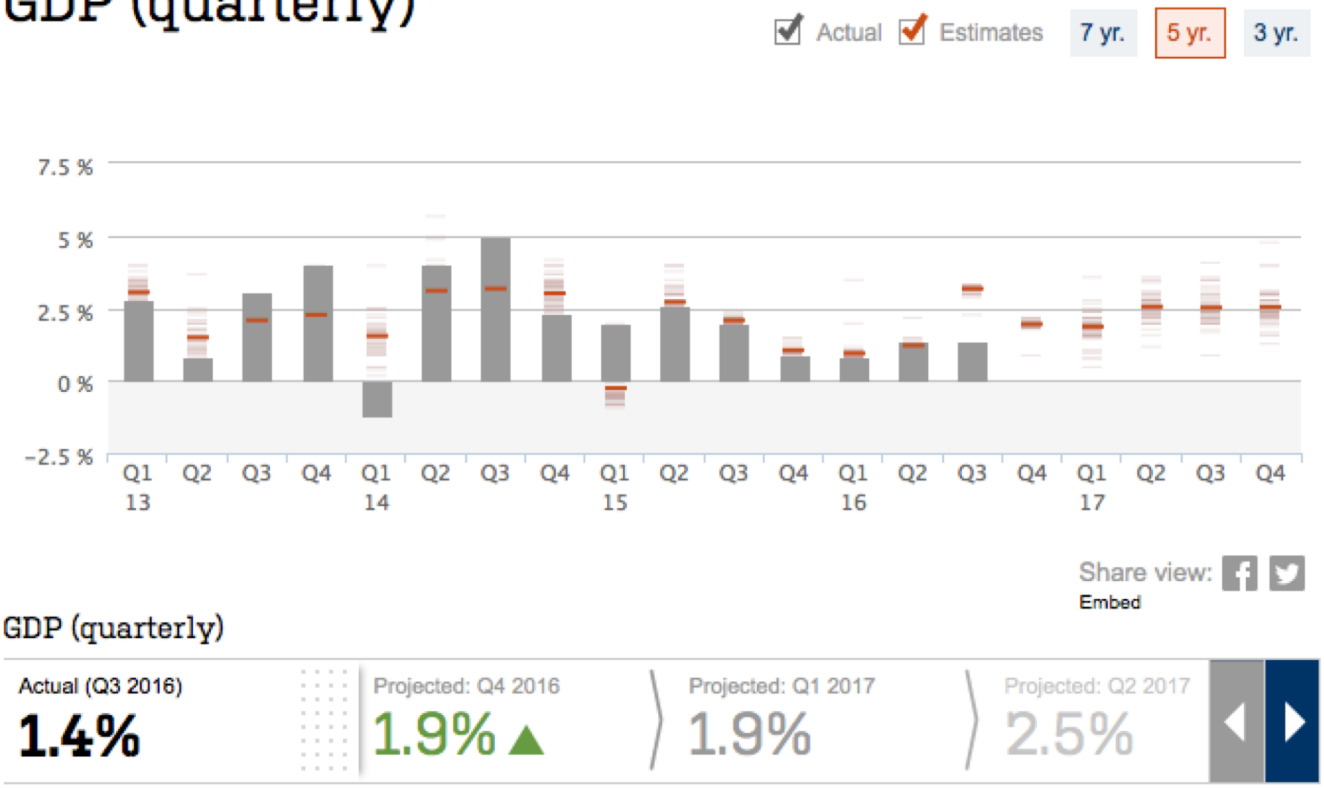
https://sciencetaskforce.ch/en/home/

THE WALL STREET JOURNAL.

Economic Forecasting Survey

The Wall Street Journal surveys a group of more than 60 economists on more than 10 major economic indicators on a monthly basis.

GDP (quarterly)



GDP (quarterly)

Actual (Q3 2016)	Projected: Q4 2016	Projected: Q1 2017	Projected: Q2 2017
1.4%	1.9% ▲	1.9%	2.5%

http://projects.wsj.com/econforecast/#ind=gdp&r=20

Mannes et al. (2014)

1. simulations to show the relative performance of crowds, best judge, or select crowds as a function of enviroment/judge performance

2. show the relative performance of crowds, best judge, or select crowds in real environments

3. surveys/experiments to evaluate people's intuitions about the performance of staticized groups (crowds, select crowds) vs. best judge

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

Aggregation of inferences

Mannes et al. (2014) suggest the success of aggregation relative to a best member (expert) or a team of experts depends on the distribution of knowledge (dispersion) and population bias (bracketing)…

Dispersion in expertise: degree to which members differ in ability to estimate the criterion
Bracketing: frequency with which any two judges fall on opposite sides of the criterion

	Low dispersion in expertise	**High dispersion in expertise**
High bracketing	(A) Whole Crowd	(B) Select Crowd
Low bracketing	(C) Select Crowd	(D) Best Member

Figure 1. Four exemplar judgment environments and the strategies expected to perform the best in each.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

Aggregation of inferences: Simulations

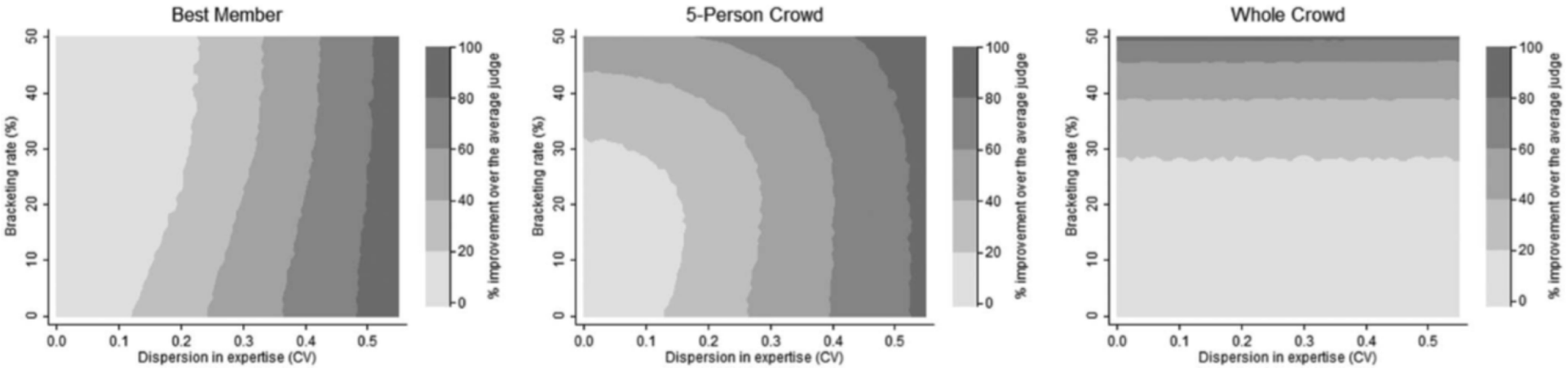


Figure 3. Contour maps of performance across 2,856 simulated judgment environments for three judgment strategies. Five trials of history were used to rank and select judges ($N = 50$). Darker shades of gray indicate greater percent improvement over the average judge. CV = coefficient of variation.

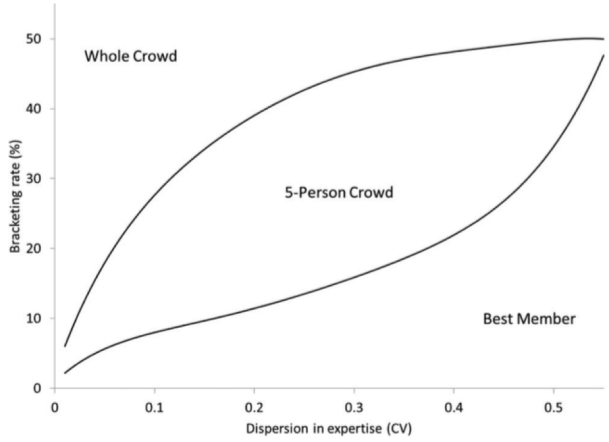


Figure 4. Best-performing strategy for each simulated judgment environment with $N = 50$ judges ranked and selected based on five periods of history. With less (more) history available to select judges, the curves rotate clockwise (counterclockwise). CV = coefficient of variation.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

Aggregation of inferences: Real data

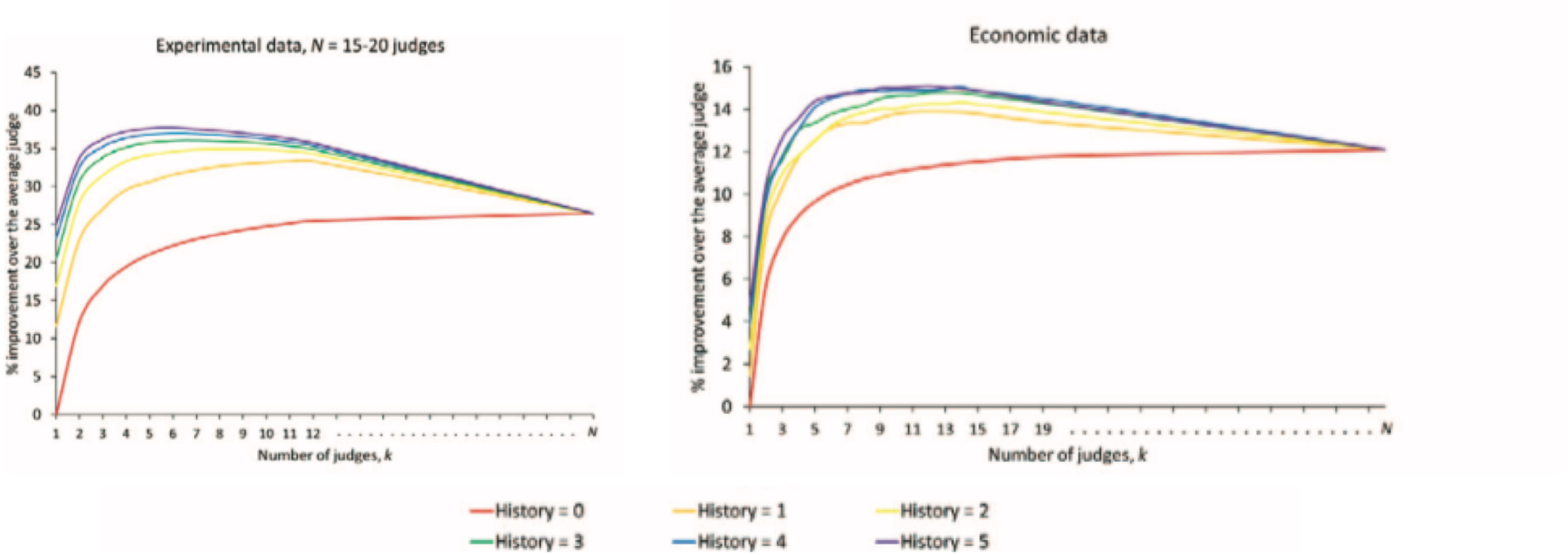


Figure 6. Performance of the best member ($k = 1$), select crowds ($1 < k < N$), and whole crowd ($k = N$) in the experimental and economic data. The experimental data is separated into the 20 sets with $N = 15$–20 judges (top left) and the 20 sets with $N > 20$ judges (top right). Selections based on six levels of history are provided. Performance for omitted values of k (in ellipses) is interpolated.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

Aggregation of inferences: Lay intuitions

Table 2
Ratings of Judgment Strategies in Experiment 1

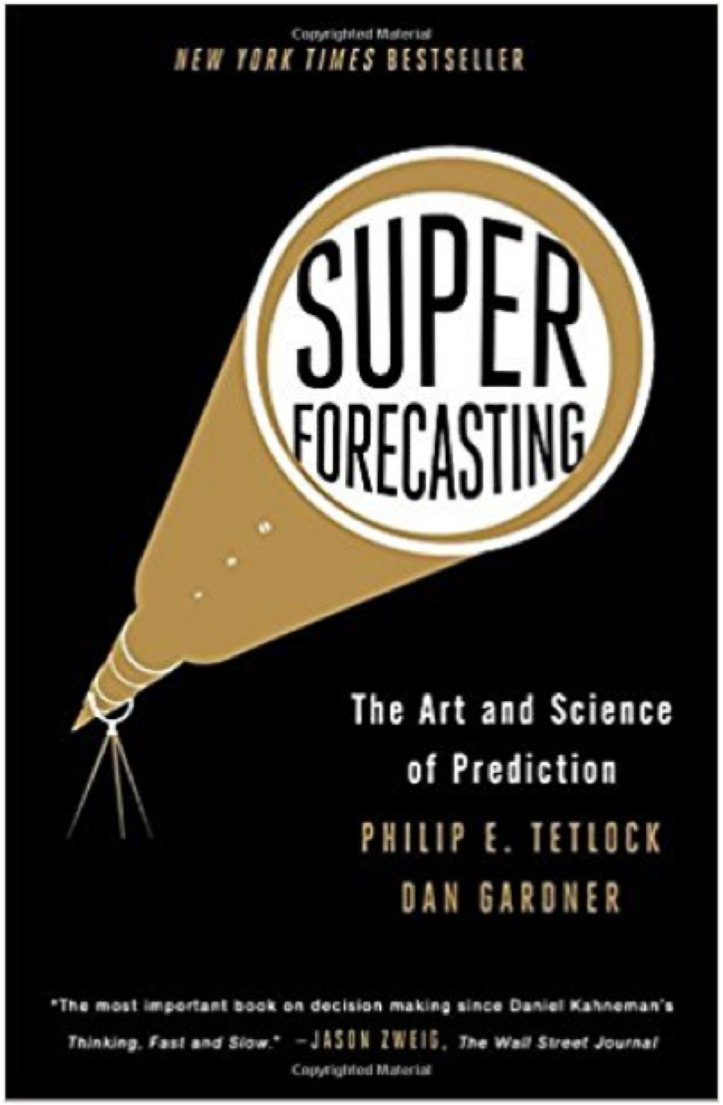
Strategy	*M*	*SD*	Difference in means 1	2	3	4	5
1. Random economist	3.24	1.37	—				
2. Average of all economists	4.71	1.22	1.46***	—			
3. Most accurate economist last year	4.60	1.28	1.35***	−0.11	—		
4. Most accurate economist last 5 years	5.04	1.22	1.79***	0.33***	0.44***	—	
5. Average of 5 most accurate economists last year	5.11	1.20	1.86***	0.40***	0.51***	0.07	—

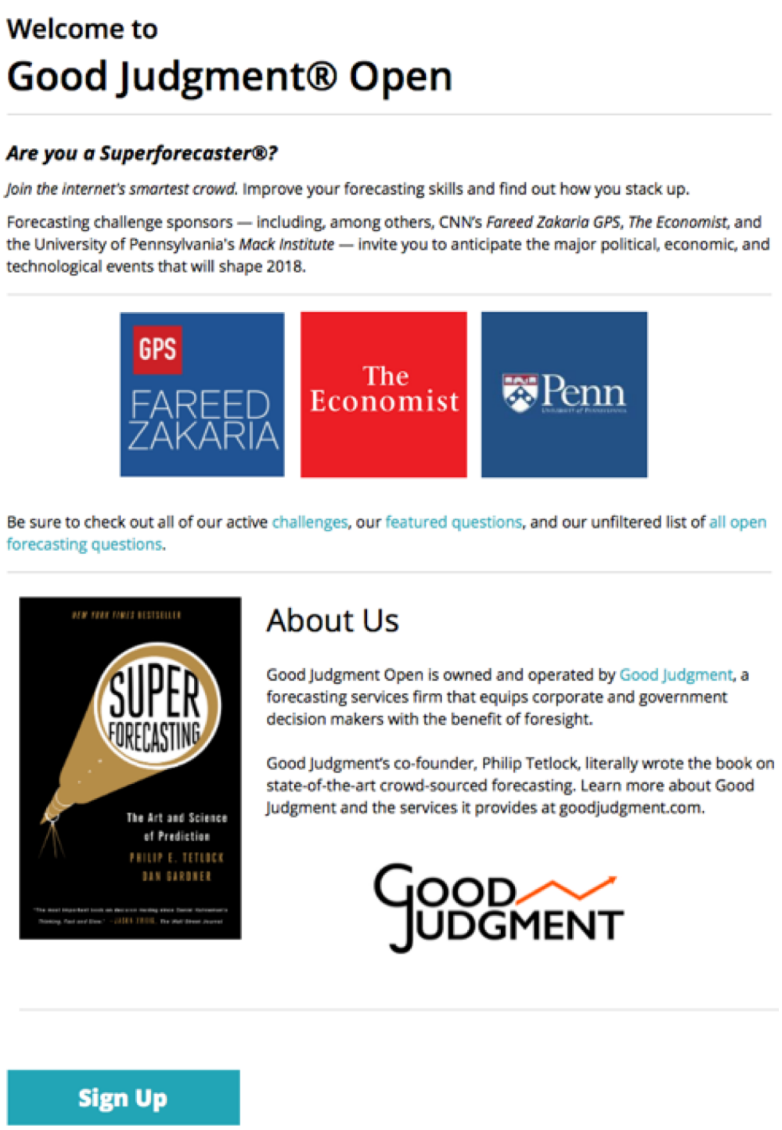
Note. $N = 312$.
*** $p < .005$ (Bonferroni-adjusted, $\alpha_{FW} = .05$).

People seem to have the intuition that the most accurate expert or a team of experts are about the same…
Possible reasons are beliefs about the (lack of) predictability of judges' future performance rather than beliefs about the power of averaging.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

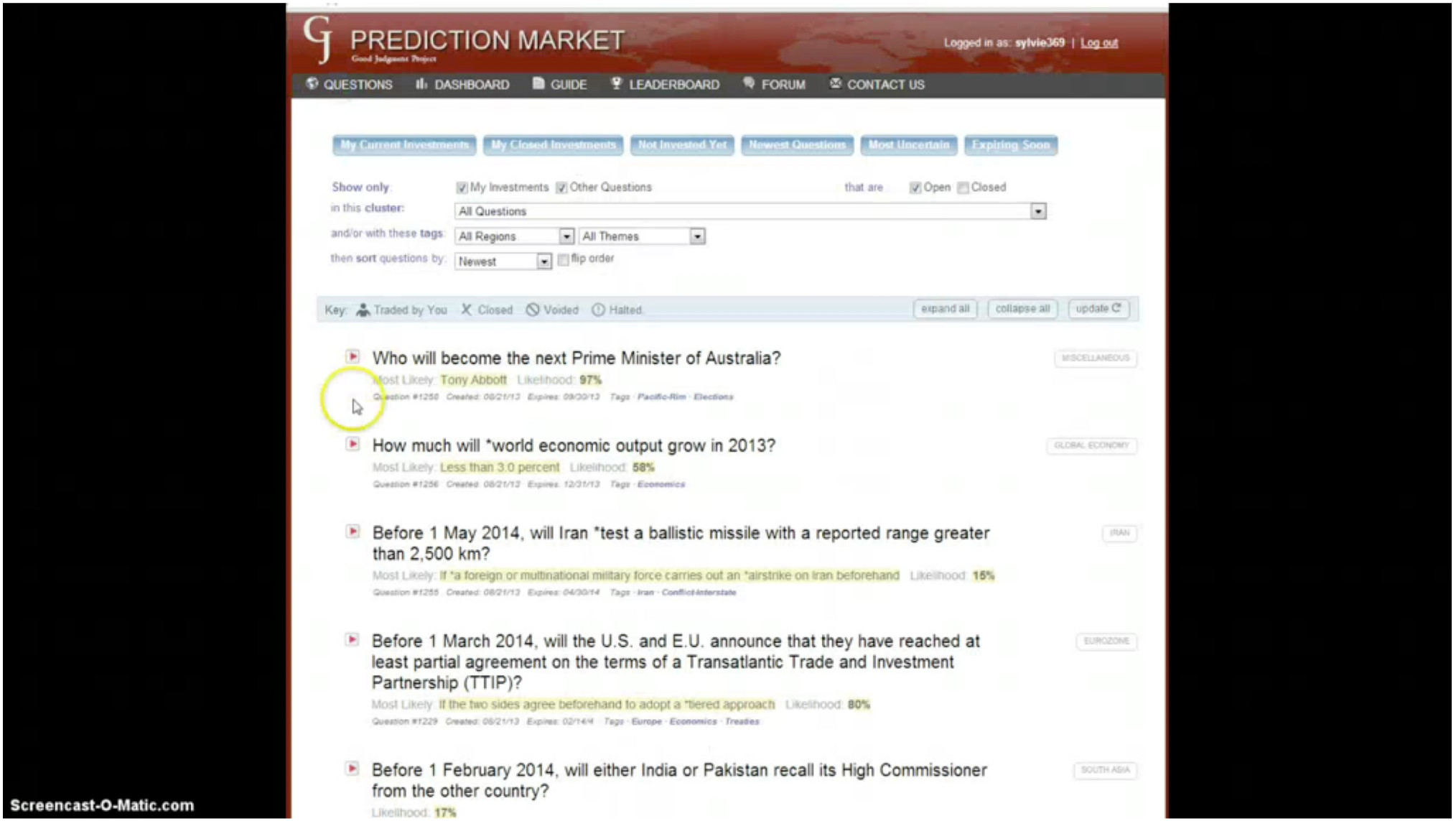
Good Judgment Project



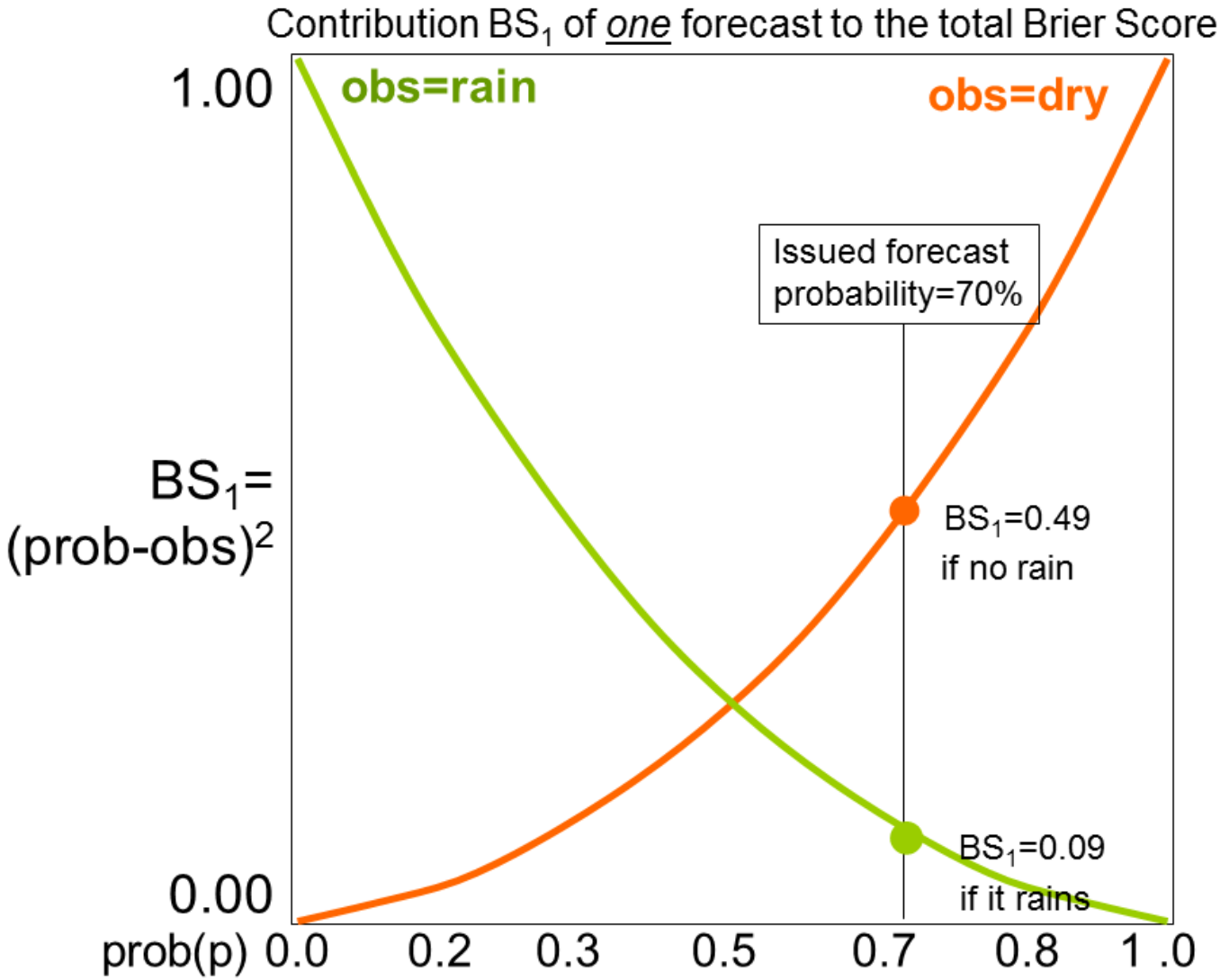


https://goodjudgment.com

Good Judgment Project



Good Judgment Project



https://en.wikipedia.org/wiki/Brier_score

Good Judgment Project

Abstract

Five university-based research groups competed to recruit forecasters, elicit their predictions, and aggregate those predictions to assign the most accurate probabilities to events in a 2-year geopolitical forecasting tournament. Our group tested and found support for three psychological drivers of accuracy: training, teaming, and tracking. Probability training corrected cognitive biases, encouraged forecasters to use reference classes, and provided forecasters with heuristics, such as averaging when multiple estimates were available. Teaming allowed forecasters to share information and discuss the rationales behind their beliefs. Tracking placed the highest performers (top 2% from Year 1) in elite teams that worked together. Results showed that probability training, team collaboration, and tracking improved both calibration and resolution. Forecasting is often viewed as a statistical problem, but forecasts can be improved with behavioral interventions. Training, teaming, and tracking are psychological interventions that dramatically increased the accuracy of forecasts. Statistical algorithms (reported elsewhere) improved the accuracy of the aggregation. Putting both statistics and psychology to work produced the best forecasts 2 years in a row.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., et al. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science*, *25*(5), 1106–1115. http://doi.org/10.1177/0956797614524255

Good Judgment Project

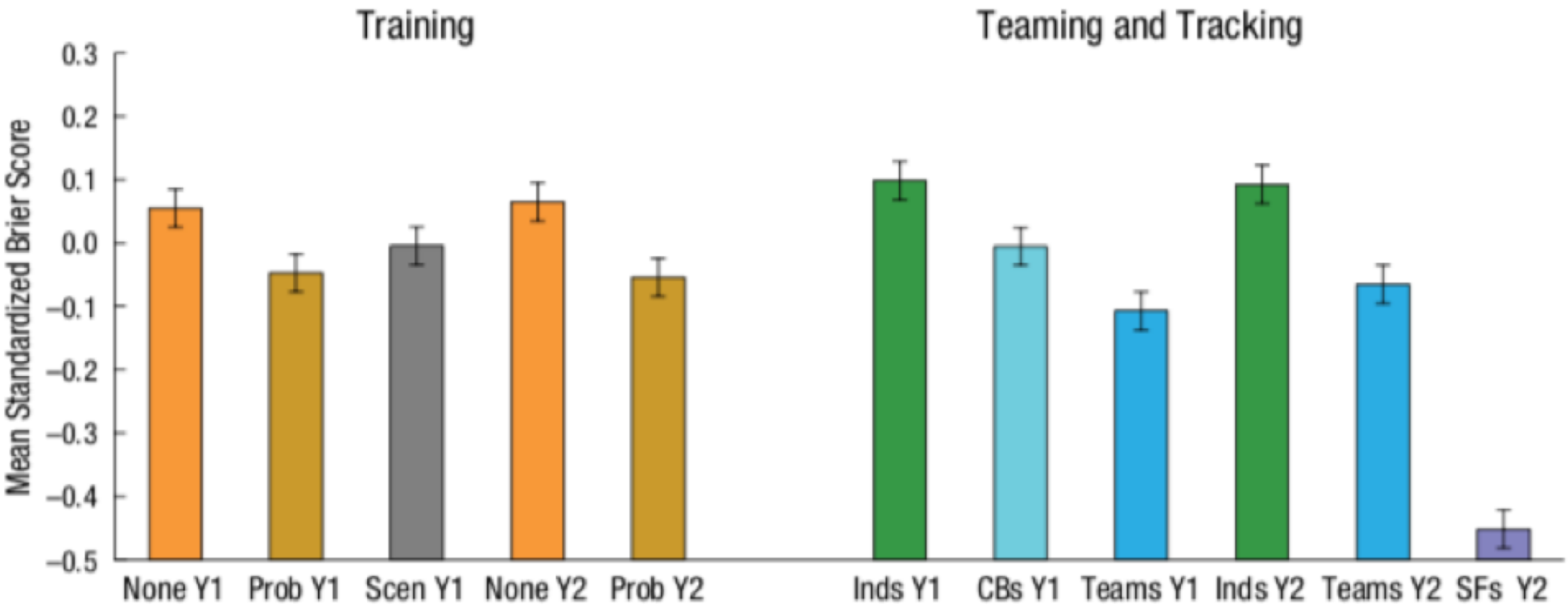


Fig. 1. Effects of training, teaming, and tracking on average Brier scores in Year 1 (Y1) and Year (Y2). The bars at the left show results for the no-training ("None"), probability-training ("Prob"), and scenario-training ("Scen") conditions; the bars at the right show results for independent forecasters ("Inds"), crowd-belief forecasters ("CBs"), team forecasters ("Teams"), and superforecasters ("SFs"). Error bars represent ±2 *SEs*.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., et al. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science*, *25*(5), 1106–1115. http://doi.org/10.1177/0956797614524255

Good Judgment Project

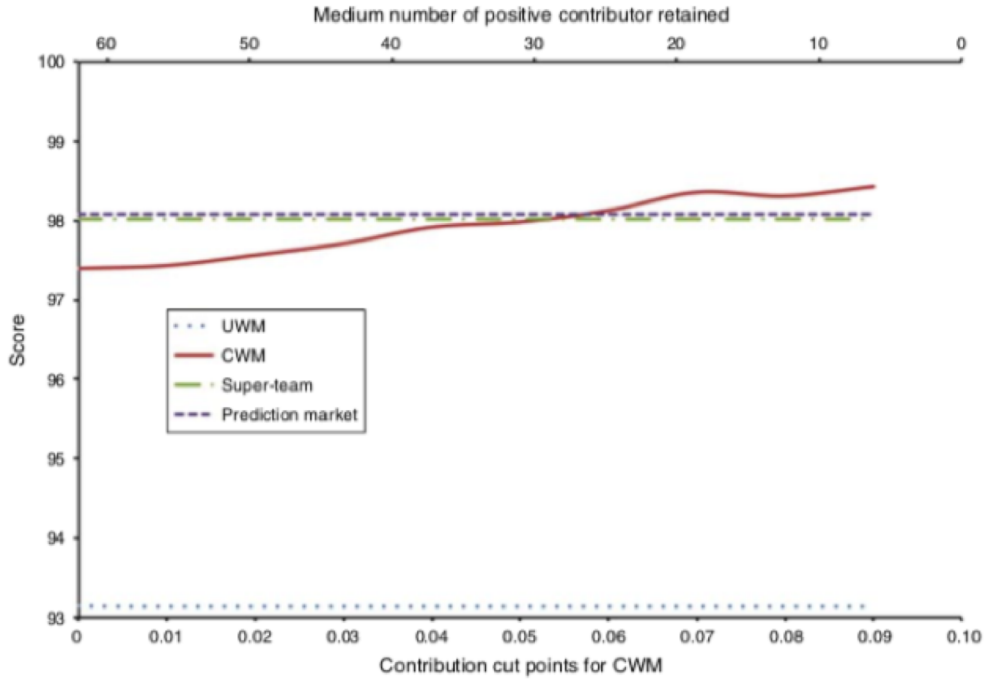
Abstract

Across a wide range of tasks, research has shown that people make poor probabilistic predictions of future events. Recently, the U.S. Intelligence Community sponsored a series of forecasting tournaments designed to explore the best strategies for generating accurate subjective probability estimates of geopolitical events. In this article, we describe the winning strategy: culling off top performers each year and assigning them into elite teams of *superforecasters*. Defying expectations of regression toward the mean 2 years in a row, superforecasters maintained high accuracy across hundreds of questions and a wide array of topics. We find support for four mutually reinforcing explanations of superforecaster performance: (a) cognitive abilities and styles, (b) task-specific skills, (c) motivation and commitment, and (d) enriched environments. These findings suggest that superforecasters are partly discovered and partly created—and that the high-performance incentives of tournaments highlight aspects of human judgment that would not come to light in laboratory paradigms focused on typical performance.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., et al. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. ***Perspectives on Psychological Science, 10***(3), 267–281. http://doi.org/10.1177/1745691615577794

Good Judgment Project

Figure 2 (Color online) CWM Beats Super-Teams and Prediction Market by Increasing the Threshold for Selecting Positive Contributors in Period 1



The Contribution-Weighted Model (CWM) is a version of the select-crowd strategy – it measures the relative contribution of each judge to the group and weights each judge according to their past performance; interestingly, results suggest it is more important to identify the best judges than to weigh each one appropriately.

Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the Contribution-Weighted Model: Robustness and Cost-Benefit Analyses. *Decision Analysis, 13*(2), 128–152. http://doi.org/10.1287/deca.2016.0329

The inner crowd…

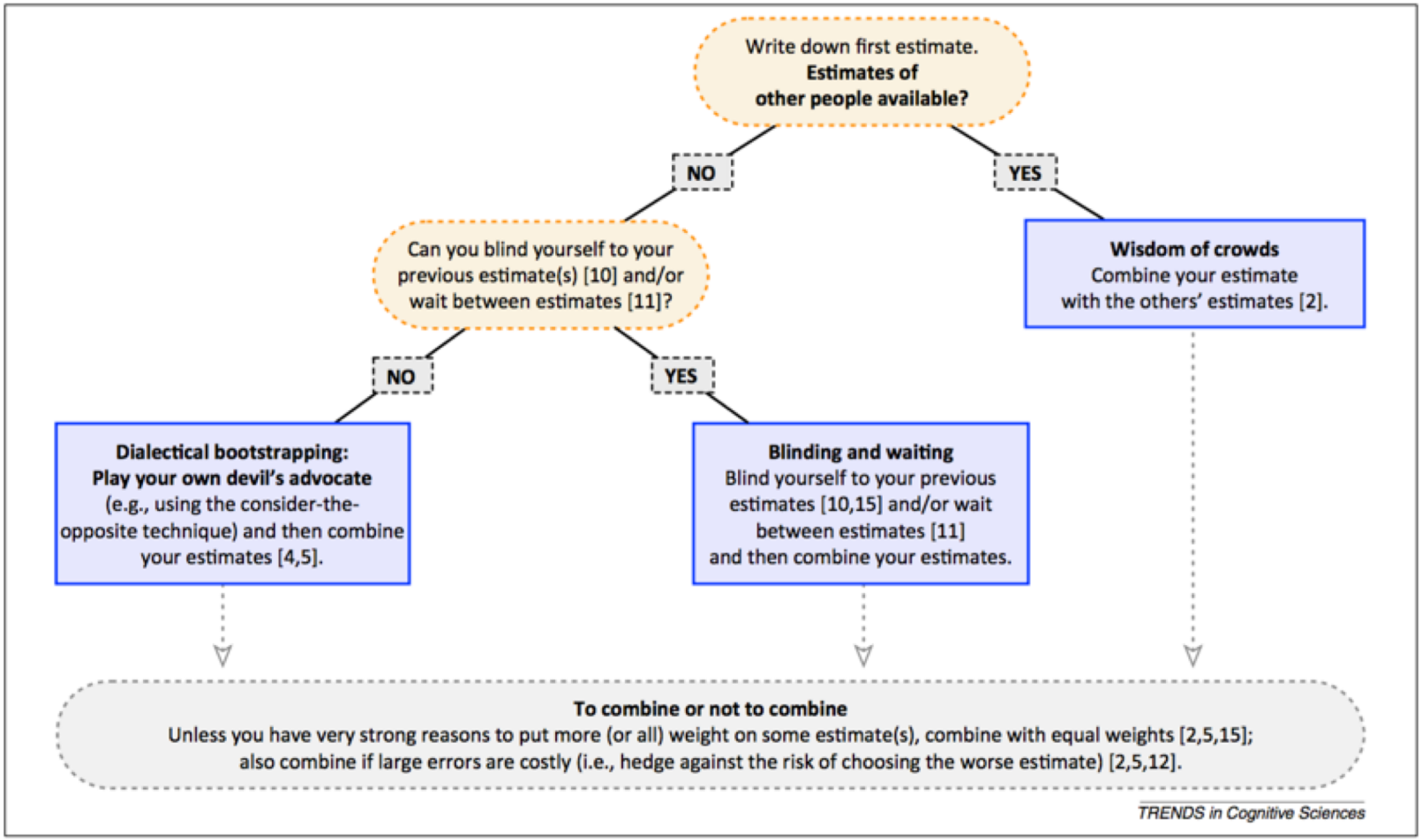


Figure 1. Decision tree for deciding when and how to use the inner crowd.

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences, 18*(10), 504–506.

Summary

Staticized groups can work well. Understanding the performance of groups as a process of statistical aggregation involving different factors - dispersion and bracketing - helps predict when select crowds (or other types of aggregation) will do best.

Aggregating preferences over a whole crowd works best when there is low dispersion of knowledge and high bracketing. Trusting a single expert makes sense if he/she has all the knowledge!

Often, teams of experts seem to provide a good balance by capitalising on dispersion and bracketing. Lay people are not aware of the power of aggregation and are likely not aware of the power of select crowds…