Finding truth even if the crowd is wrong

Drazen Prelec[1,2,3], H. Sebastian Seung[3,4], and John McCoy[3]
[1]Sloan School of Management
Departments of [2]Economics, [3]Brain & Cognitive Sciences, and [4]Physics
Massachusetts Institute of Technology
Cambridge MA 02139
dprelec@mit.edu, seung@mit.edu, jmccoy@mit.edu

Since Galton [1] first reported on the uncanny accuracy of a median estimate (of the weight of an ox, judged by spectators at a country fair), the notion that the 'wisdom of the crowd' is superior to that of any single person has itself become a piece of crowd wisdom, raising expectations that web-based opinion aggregation might replace expert judgment as a source of guidance for individuals and policymakers alike [2, 3]. However, distilling the best answer from diverse opinions is challenging when most people hold an incorrect view [4]. We propose a method based on a new definition of the best answer: it is the one given by those who would be least surprised by the true answer if it were revealed. Because this definition is of interest only when the true answer is unknown, algorithmic implementation is nontrivial. We solve this problem by asking respondents not only to answer the question, but also to predict the distribution of others' answers. Previously, it was shown that this secondary information can be used to create incentives for honesty, if respondents use correct Bayesian reasoning [5]. Here we prove that this information will also reveal which answer is the best answer by our new definition. Unlike multi-item analysis [6, 7] or boosting [8], our method works with a single, sui generis question, thus allowing applications to domains where experts' track records and competencies cannot be established without controversy. Unlike Bayesian models [9, 10, 11, 12, 13] it does not require user-specified prior probabilities. An experiment demonstrates that the method outperforms algorithms based on democratic or confidence-weighted voting [14, 15, 16, 17, 18].

Imagine that you have no knowledge of U.S. geography, and are confronted with the question

> Philadelphia is the capital of Pennsylvania: True or False ?

To find the answer, you pose the question to many people, trusting that the most common answer will be correct. Unfortunately, most people give the incorrect answer ("True"), as shown by the data in Figure 1(a). Why is the majority wrong here? Someone who answers "True" may know only that Philadelphia is an important city in Pennsylvania, and reasonably conclude that Philadelphia is the capital. Someone who answers "False" likely possesses a crucial additional piece of evidence, that the capital is actually Harrisburg.

This elementary example reveals a limitation of the one person, one vote approach. If each respondent's answer is determined by the evidence available to her, the majority verdict will be tilted toward the most widely available evidence, which is an unreliable indicator of truth. The same bias is potentially present in real-world settings, when experts' opinions are averaged to produce probabilistic assessments of risk, forecasts of key economic variables, or numerical ratings of research proposals in peer review. In all such cases, Galton's method of counting opinions equally may produce a result that favors shallow information, accessible to all, over specialized or novel information that is understood only by a minority.

To avoid this problem, one could seek out individuals who believe that they are the most competent to answer the question, or have the best evidence. A popular approach is to ask respondents to report their confidence [13], typically by asking them to estimate their probability of being correct. From their own estimates of the probability they are correct and their True/False answers, one can infer respondents' subjective probability estimates that, e.g., Philadelphia is the capital of Pennsylvania. Averaging these probabilities across respondents produces a confidence-weighted vote. This will improve on an unweighted vote, but only if those who answer correctly are also sufficiently more confident, which is neither the case in our example, nor more generally [4]. As shown by Figure 1(b), the distribution of confidences is roughly the same between the two groups, and cannot override the strong majority in favor of the wrong answer.

One might think that people are simply not good at estimating their confidences, which is true, but the problem is also a theoretical one. Even if the sample contains only individuals whose confidence estimates are derived by correct probabilistic reasoning from available evidence, the distribution of their judgments and confidences is in general insufficient to identify the correct answer. Figure 2 illustrates how identical distributions of votes and confidences can arise from Bayesian reasoning in two different possible worlds models, one where majority opinion is correct (e.g., the Columbia - South Carolina example in Figure 1(d-f)) and one where it is incorrect. Therefore, even if confidence estimates are unbiased and derived by Bayes' rule from evidence, any method that relies only on votes and confidences may not recover the true answer. What is missing is information about the possible worlds model that gives rise to a particular distribution of answers and confidences.

To uncover this model in the context of a single question, our method asks each respondent to predict how others will respond. For a True/False question, the prediction is a number between 0 and 1 indicating the fraction of respondents who will answer "True." As shown in Figure 1(c), those who answer "True" to the Philadelphia question predict that most people will agree and answer "True." On the other hand, those who answer "False" tend to predict that most people will disagree and hence answer "True." This prediction presumably reflects superior knowledge: Respondents who believe that Harrisburg is the capital tend to predict that most people will not know this. The asymmetry between the distributions in Figure 1(c) is marked, suggesting that predictions of others' answers could provide a signal that is strong enough to override majority opinion. To theoretically justify the use of this information, however, we need a precise definition of best evidence and best answer.

Consider a probabilistic possible worlds model in which the world Ω is a random variable taking on values in the set $\{1,\ldots,m\}$ of possible answers to a multiple choice question. The answer X^r given by respondent r is likewise a random variable taking on values in the same set. We assume that the answer is based on a hidden variable, the "signal" or "evidence" available to the respondent, which in turn depends upon the world. Each respondent understands the possible worlds model, reasons correctly from the evidence available to her, and answers honestly. We further assume that respondents with different answers have access to different evidence, and respondents who give the same answer do so based on the same evidence. Therefore a respondent r who answers k, correctly assigns probabilities $\Pr\left[\Omega = i | X^r = k\right]$ to possible answers i, and this

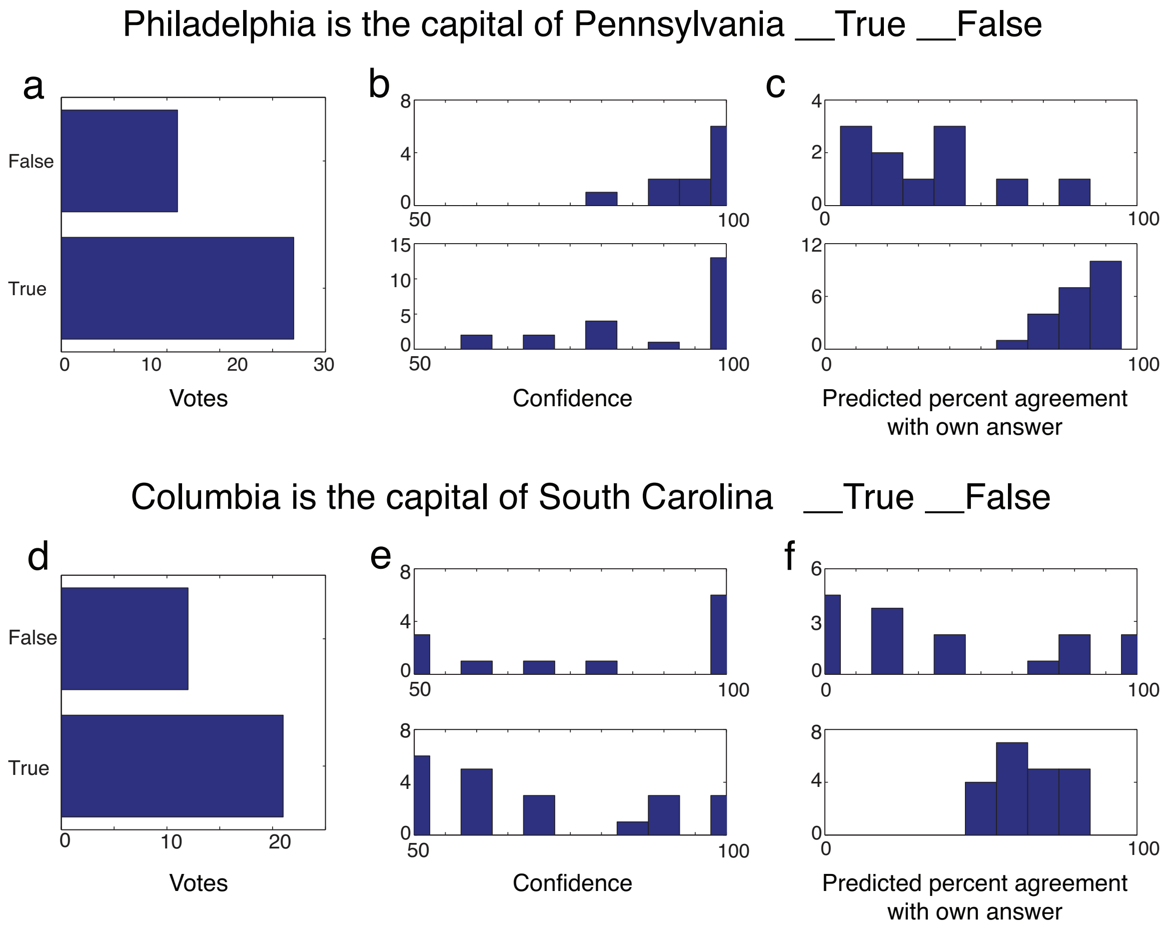

Figure 1: TOP PANEL A question that voting (1) fails to answer correctly, while the "least surprised" principle (2) succeeds (data are from Study 3 described in text). (**a**) The wrong answer wins by a large margin in a democratic vote. (**b**) Respondents are asked to provide confidence estimates expressed as the probability of being correct from 50 (chance) to 100 (certainty). The histograms in the middle panels show that the correct minority (top) and the incorrect majority (bottom) are roughly equally confident, so weighting the votes by confidence does not change the outcome. (**c**) Respondents are asked to predict the frequency of the answer "True," which can be converted into an estimated percent agreement with their own answer. Those who answer "True" believe that most others will agree with them, while those who answer "False" believe that most others will disagree. The LS principle discounts the more predictable votes, leading to a reversal of the majority verdict. BOTTOM PANEL A question that voting (1) and LS (2) both answer correctly. Average confidence is lower than in the Philadelphia example (comparing (**e**) and (**b**)), but majority opinion in (**d**) clearly supports the correct answer "True." Here, the predicted distributions of votes (**f**) are roughly symmetric between the two opinions, and the LS principle does not overturn the majority verdict.

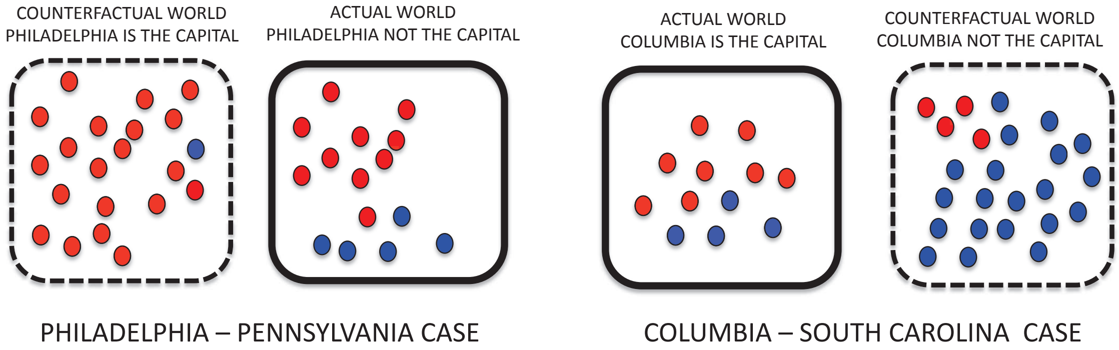
COUNTERFACTUAL WORLD
PHILADELPHIA IS THE CAPITAL
ACTUAL WORLD
PHILADELPHIA NOT THE CAPITAL
ACTUAL WORLD
COLUMBIA IS THE CAPITAL
COUNTERFACTUAL WORLD
COLUMBIA NOT THE CAPITAL
PHILADELPHIA – PENNSYLVANIA CASE
COLUMBIA – SOUTH CAROLINA CASE

Figure 2: Two possible worlds models, giving stylized representations of the Philadelphia-Pennsylvania and Columbia-South Carolina problems, and constructed to show the theoretical limitations of vote counts and confidence distributions. Each model contains an actual world (solid square) and a counterfactual world (dashed square). Opinions are represented by colored chips, Red favoring the Yes answer, Blue favoring the No answer, and are treated as random draws with replacement from the actual world. Although the correct answer is different in the two problems, the models predict the same distribution of opinions, 2:1 in favor of Yes, as there are twice as many Red chips as Blue chips in the actual world in both problems. Moreover, the distribution of confidences will also be the same if respondents derive them by Bayesian reasoning from the correct possible world model. Such a model specifies the distribution of opinions in all possible worlds and their prior probabilities, which are here represented by the total number of Red and Blue chips in each square. Thus, the Yes answer is deemed a priori more likely for the Philadelphia problem, and the No answer for the Columbia problem. A Bayesian respondent draws a chip at random from all chips displayed in given model, and forms a posterior probability that the chip came from one or the other world by consulting the distribution of that color across the two possible worlds. Drawing a Red chip favors the Yes world by the same ratio (2:1) in both models, since there are twice as many Red chips in the left Yes world as in the right No world in both models. Similarly, drawing a Blue chip favors the No world by the same ratio (5:1) in both models. In either model, Bayesian reasoners who draw Red will predict with 67% confidence that the target city is the state capital, while those who draw Blue will predict with 83% confidence that the target city is not the state capital. In the Philadelphia problem, the correct answer will be endorsed by a minority of ideal Bayesian respondents, who will also appear more confident, while in the Columbia problem it will be endorsed by the majority, who will appear relatively less confident. As the observed ‘data’ in the form of votes and confidences will be exactly the same in the two problems, no method based on these two inputs will be able to deduce the correct answer. Some additional information is needed.

function does not depend on r.

The traditional approach to aggregating judgments is the democratic principle which selects the answers most likely to be offered by respondents given the actual world i^*

$$\underset{k}{\operatorname{argmax}} \Pr\left[X^r = k | \Omega = i^*\right] \tag{1}$$

We propose instead that conceptually the answer to select is the one given by respondents who place the highest probability on the actual state of the world. The actual state of the world i^* is unknown. If it were revealed, then, under the assumptions of the above model, the probability $\Pr\left[\Omega = i^* | X^r = k\right]$ would measure (inversely) the surprise of any respondent r who selected answer k. We define the "best answer" as the one given by respondents who would be "least surprised" by the truth (hereafter LS) or

$$\underset{k}{\operatorname{argmax}} \Pr\left[\Omega = i^* | X^r = k\right] \tag{2}$$

By our assumptions, these are also the respondents who possess the best evidence. The best answer to the Philadelphia question should be correct, as those who believe that Philadelphia is the capital would be more surprised by the true answer than those who believe that Philadelphia is not the capital. The best answer is not infallible, because evidence unavailable to any respondent might tip the balance the other way.[1]

To compute (2) without knowing the true value i*, and to clarify the additional information that (2) incorporates, we note that the maximum in principle (2) can be found by comparing $\Pr\left[\Omega = i^* | X^r = k\right]$ and $\Pr\left[\Omega = i^* | X^s = j\right]$ for all answers j and k. Using Bayes' Rule, the ratio between these probabilities can be rewritten as

$$\frac{\Pr\left[\Omega = i^* | X^r = k\right]}{\Pr\left[\Omega = i^* | X^s = j\right]} = \frac{\Pr\left[X^r = k | \Omega = i^*\right]}{\Pr\left[X^r = k\right]} \Big/ \frac{\Pr\left[X^s = j | \Omega = i^*\right]}{\Pr\left[X^s = j\right]} \tag{3}$$

In contrast to the democratic principle which selects answers based only on how likely they are given the actual world, the LST principle incorporates the marginal likelihood of answers and selects answers based on how likely they are in the actual world relative to how likely they are in all worlds, including counterfactual ones. For example, if an answer is salient and so would be often given by respondents independent of the actual world, the LS principle discounts it relative to answers which are only likely in a few worlds.

The marginal likelihoods of different answers cannot easily be elicited from respondents and so we again apply Bayes rule, rewriting the marginal likelihoods in terms of conditional probabilities on different answers. Since $\Pr[X^r = k] = \Pr[X^s = j] \Pr[X^r = k | X^s = j] / \Pr[X^s = j | X^r = k]$ the ratio between the two probabilities

[1] Given our assumptions, the LS principle will always select the correct answer, i^* for dichotomous questions. Since respondents select the answer they believe is most likely to be true, we have $\Pr\left[\Omega = k | X^r = k\right] > 0.5$ for $k = 1, 2$. Hence, $\Pr\left[\Omega = i^* | X^r = i^*\right] > 0.5$, but $\Pr\left[\Omega = i^* | X^r = k\right] = 1 - \Pr\left[\Omega = k | X^r = k\right] < 0.5$, if $k \neq i^*$, and answer i^* will be selected. For $m \geq 3$, LS will select the correct answer whenever $\Pr\left[\Omega = i^* | X^r = i^*\right] > \Pr\left[\Omega = i^* | X^r = k\right]$, $k \neq i^*$, which is to say, whenever respondents who received the signal favoring the true answer believe more strongly in the true answer than those who received other signals. This is a reasonable constraint, but one can construct examples that violate it, such as the following: Suppose six cities are presented as candidates for the capital of Pennsylvania, and respondents who think Harrisburg is the most likely capital are unconfident, assigning it 20% probability, with 16% going to each of the remaining five, while respondents who think that Philadelphia is the capital have Harrisburg as their strong second choice, assigning it 40% probability (with 60% for Philadelphia). The latter group thus assigns more probability to the actual capital (as 40% > 20%), but their incorrect first choice answer — Philadephia — will be selected by the LS principle. This pattern obviously cannot arise if respondents with the correct answer are also the most confident; however, the converse does not hold — LS can select the correct answer even if those who endorse it are the least confident (e.g., Columbia example in Figure 2).

we are comparing is given by

$$\frac{\Pr[\Omega = i^*|X^r = k]}{\Pr[\Omega = i^*|X^s = j]} = \frac{\Pr[X^r = k|\Omega = i^*]}{\Pr[X^s = j|\Omega = i^*]} \frac{\Pr[X^s = j|X^r = k]}{\Pr[X^r = k|X^s = j]} \tag{4}$$

The conditional probability of an answer given the actual world also appears in the democratic principle (1). It can be readily estimated from the relative frequencies of answers even without knowledge of i^*, as answers are by definition sampled from the actual world (e.g., in which Harrisburg is the capital of Pennsylvania). We use respondents' second-order predictions which give their estimates of the frequency of answers in the sample to estimate the conditional probability of one answer on another.[2]

Work in experimental psychology [add cites] has robustly demonstrated that respondents' predictions of the distribution of beliefs in a sample is consistent with them implicitly conditioning on their own answer as an "informative sample of one." Respondents do not know the actual world when giving their second-order predictions, but only the evidence available to them. We thus estimate $\Pr[X^s = j|X^r = k]$ as the average of the second-order predictions of the fraction answering j, given by respondents who answered k, since these second-order predictions reflect an implicit conditioning on answer k.

Consider how this idea works for the Philadelphia question. The answer $k =$"False" is less common than $j =$"True", so the first ratio in (4) is less than one, and the majority answer (1) is incorrect. On the other hand, the second ratio is greater than one, because of the asymmetry in predictions in Figure 1(c), and is actually strong enough to override majority opinion.

For a question with more than two possible answers, one would like to combine all pairwise comparisons into a maximum principle. After taking the logarithm of (4) and rearranging terms,

$$\log\Pr[\Omega = i^*|X^r = k] = \log\Pr[X^r = k|\Omega = i^*] + \log\frac{\Pr[X^s = j|X^r = k]}{\Pr[X^r = k|X^s = j]} + \log\frac{\Pr[\Omega = i^*|X^s = j]}{\Pr[X^s = j|\Omega = i^*]}$$

we may perform a weighted average over all j-s, and drop the rightmost term which does not depend on k. This yields a reformulation of the LS principle as:

$$\underset{k}{\operatorname{argmax}} \Pr\left[\Omega = i^*|X^r = k\right] = \underset{k}{\operatorname{argmax}} \left\{ \log\Pr\left[X^r = k|\Omega = i^*\right] + \sum_j w_j \log\frac{\Pr\left[X^s = j|X^r = k\right]}{\Pr\left[X^r = k|X^s = j\right]} \right\} \tag{5}$$

for any set of weights w_j satisfying $\sum_j w_j = 1$.

To implement the LS principle we estimate the probabilities in (5) from a population of respondents. Let $x_k^r \in \{0,1\}$ indicate whether respondent r endorsed answer k , and y_k^r her prediction of the fraction of respondents endorsing answer k. If we estimate $\Pr\left[X^r = k|\Omega = i^*\right]$ using the arithmetic mean, $\bar{x}_k = n^{-1}\sum_r x_k^r$, then Eq. (5) takes the form

$$\underset{k}{\operatorname{argmax}} \left\{ \log\bar{x}_k + \sum_j w_j \log\frac{\bar{y}_{jk}}{\bar{y}_{kj}} \right\} \tag{6}$$

[2] When will LS endorse majority opinion? From (3) it is evident that this will occur when $\Pr\left[X^r = k\right] = \Pr\left[X^r = j\right]$, leading to an equivalence of (2) and (1). To see what this implies about the estimated conditional probabilities, observe that the distribution over X^r is also the stationary distribution of the Markov matrix $\left[\Pr\left[X^r = k|X^r = j\right]\Pr\left[X^r = j\right]\right]$, as $\Pr\left[X^r = k\right] = \sum_j \Pr\left[X^r = k|X^r = j\right]\Pr\left[X^r = j\right]$, for $k = 1,..,m$. It is known that the stationary distribution of a Markov matrix is uniform if and only if the matrix is doubly stochastic, with all column and row entries summing to one. A symmetric matrix is a special case.

where the estimate $\bar{y}_{kj}$ of $\Pr[X^r = k|X^s = j]$ is based on the predictions y_k^s of respondents s who endorsed answer j. This could be the arithmetic mean $\bar{y}_{kj} = (n\bar{x}_j)^{-1} \sum_s x_j^s y_k^s$ or the geometric mean $\log \bar{y}_{kj} = (n\bar{x}_j)^{-1} \sum_s x_j^s \log y_k^s$. The choice of weights w_j only matters in the case of inconsistencies between the pairwise comparisons. To resolve inconsistencies, one could weight answers equally, $w_j = 1/m$, or, alternatively, weight respondents equally, $w_j = \bar{x}_j$. In the empirical results below, we compute geometric means of predictions and weight respondents equally, and refer to this as the LS algorithm.[3]

As an illustration, we implement the algorithm with data from surveys of knowledge of US state capitals. Each question was like the Philadelphia one above, where the named city was always the most populous in the state. Respondents endorsed “True” or “False” and predicted the distribution of votes by other respondents. Although elementary, these questions have properties that make them well suited for a proof-of-concept test. The problems range in difficulty, and individual states are challenging for different reasons. The prominence of a city is sometimes false (Philadelphia-Pennsylvania), and sometimes a valid cue (Boston-Massachusetts), and many less populous states have no prominent city at all. [4]This should give rise to differences in knowledge that the LS principle should be able to exploit. Given natural variation in backgrounds, it is likely that for most states some subset — the local experts — would know the correct answer. At the same time, their number might be small and their influence diluted by the uninformed majority. The survey tests our main theoretical claim — that LS reduces such dilution of expert votes.

Surveys were administered to three groups of respondents at MIT and Princeton.The True-False votes of the respondents were tallied for each question, and the majority decision was correct on only 29 states in Study 1 ($n = 51$), 36 states in Study 2 ($n = 32$), and 31 states in Study 3 ($n = 33$) (ties counted as 0.5 correct). The LS answers were consistently more accurate, reducing the number of errors from 21 to 12 in Study 1 (matched pair $t_{49} = 2.45$, $p < .01$), from 14 to 9 in Study 2 ($t_{49} = 1.69$, $p < .05$), and from 19 to 4 in Study 3 ($t_{49} = 4.40$, $p < .001$). Our basic empirical finding, that LS outperforms democratic voting, is thus replicated by three separate studies.

In order to compare LS with confidence-weighted voting, Study 3 went beyond the first two studies and, unlike studies 1 and 2, asked respondents to report their confidence with a number from 50% to 100%, as described earlier and in Figure 1. Weighting answers by confidence is more accurate than democratic voting, reducing the number of errors from 19 to 13 ($t_{49} = 2.86$, $p < .01$), but is still less accurate than LS ($t_{49} = 2.64$, $p < .02$). More extreme forms of confidence weighting, such as a policy of only counting the answers of individuals that claim to know the answer for sure (100% confident), or using a logarithmic pool [13], are likewise not as accurate as LS (Table 1).

For complex, substantive questions, we may prefer a probabilistic answer as a quantitative summary of all available evidence. An estimate of the probability that a city is the capital can be imputed to each respondent based on the True/False answers and confidence estimates collected in Study 3 (Figure 1(b)). The LS algorithm operates on the answers to multiple choice questions and so requires that these probability estimates be discretized. The discretization was done by dividing the [0,1] interval into uniform bins or into nonuniform bins using a scalar quantization algorithm (details in Supplementary Materials) with these bins treated as possible answers. In Study 3, each respondent was also asked to predict the average of others' confidence estimates. This prediction, along with their prediction of the distribution of True/False votes,

[3] The results obtained from the alternative choice of weights and means are similar, and are shown in the Supplementary Materials.

[4] There is a mixture of so-called ‘kind’ problems, where majority opinion is correct, and ‘wicked’ problems where it is incorrect [19].

was used to impute to each respondent a prediction of the entire distribution of probability estimates. The LS algorithm used the discretized probability estimates and predictions of other respondents' probabilities to select a bin, and its midpoint served as the best probability according to the LS definition. We compared this with averaging respondents' probabilities.

We found (Table 1) that LS probabilities were more accurate than average probabilities. This was not surprising for questions like Philadelphia-Pennsylvania, for which majority opinion was incorrect. More interestingly, LS outperformed probability averaging even on majority-solvable questions, defined as those for which majority opinion was correct in both Studies 1 and 2. For example, the Jackson-Mississippi question was majority-solvable, but most respondents found it difficult, as judged by their low average confidence. LS not only answered this question correctly, but also more confidently. Other algorithms that put more weight on confidence, such as the logarithmic pool [13] or retaining only the most confident answers, also outperformed probability averaging on majority-solvable problems, but not on majority-unsolvable problems like Philadelphia-Pennsylvania.

In (6), we proposed estimating $\Pr[X^s = k | X^r = j]$ by averaging the predictions of those respondents who answered j. In doing so, we regarded the fluctuations in the predictions between respondents giving the same answer as noise. Alternatively, fluctuations can be regarded as a source of additional information about individual expertise. The LS algorithm (the version of Eq. 6 with $w_j = \bar{x}_j$ and geometric means of predictions) can be rewritten as

$$\underset{k}{\operatorname{argmax}} \left\{ \frac{1}{n\bar{x}_k} \sum_r x_k^r u^r \right\} \tag{7}$$

where we define a score for each respondent r as

$$u^r = \sum_s \sum_{k,j} x_k^r x_j^s \log \frac{\bar{x}_k y_j^r}{\bar{x}_j y_k^s} = \sum_k x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} - \sum_j \bar{x}_j \log \frac{\bar{x}_j}{y_j^r} \tag{8}$$

and $\bar{y}_j$ is the geometric mean of predicted frequencies of answer j, $\log \bar{y}_j = n^{-1} \sum_r \log y_j^r$. The LS algorithm is thus equivalent to selecting the answer which is endorsed by respondents with the highest average score. For respondents who give the same answer, the first term of the score is the same, but the second term (a relative entropy) is higher for those who are better able to predict the actual distribution of answers.[5] If the score is an accurate measure of individual expertise, the best answer might be that of the single respondent with the highest score, rather than the LS algorithm, which selects the answer endorsed by the respondents with the highest average score as in (7). We found that the accuracy of the top scoring person on each question was comparable to the LS algorithm (Table 1). The individual scores may be computed for any single multiple-choice question, unlike approaches which rely on determining the consistency of respondents across multiple questions, although an individual's cummulative performance may also be examined across multiple questions. Figure 3, right panel, shows that the score of an individual respondent, averaged across all 50 states, is highly correlated with his or her objective accuracy (Study 1: r = 0.84; Study 2: r = 0.94; Study 3: $r = 0.82$, $p < .001$ for all three studies). For comparison, we also computed a conventional wisdom (CW) index, defined as the number of states for which a respondent votes with the majority for that state. Because

[5] The score does not simply identify expertise with second-order prediction accuracy as the first term has larger impact on the relative scores. Accuracy at predicting the distribution is not driving the predictive power of the score (details in Supplementary Materials).

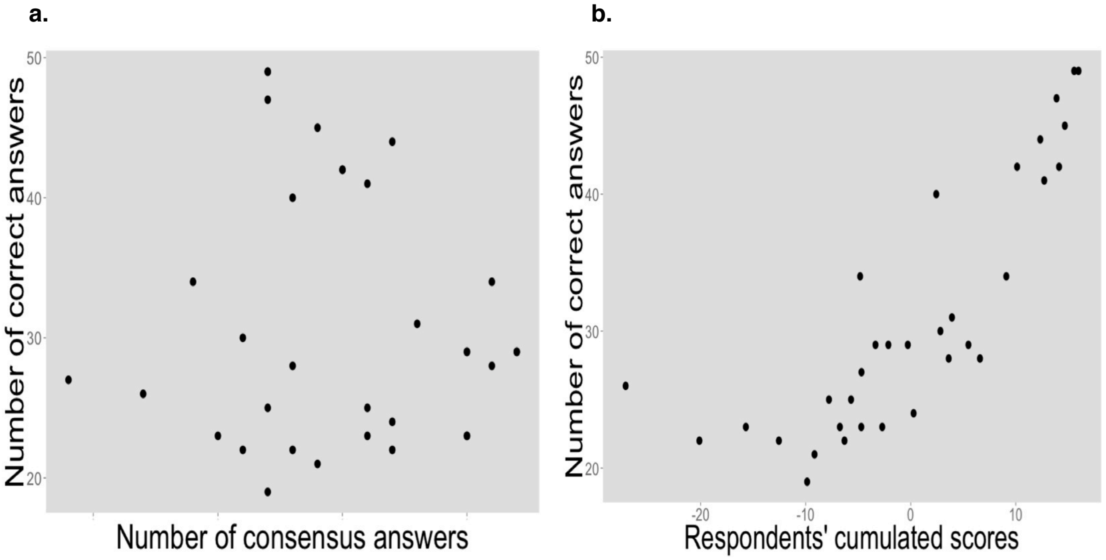

Figure 3: Scoring the expertise of individual respondents (data is from study 3, similar results are shown for studies 1 and 2 in the Supplementary Materials). (**a**) The accuracy of a respondent across all fifty states is uncorrelated with her conformity to conventional wisdom, defined as the number of times she votes with the majority. **(b)** Accuracy is highly correlated with the individual score u^r of (8) cumulated across fifty states.

the majority is correct more than half the time, one might expect that respondents with high CW scores will also be more accurate. However, accuracy and CW are uncorrelated, as shown by the left panel of Figure 3. Respondents' individual scores (8) also outperformed several other approaches as an indicator of expertise, such as principal components analysis (Supplementary Materials).

While these results provide a critical initial test, we are ultimately interested in applying the algorithm to substantive problems, such as assessments of risk, political and economic forecasts, or expert evaluations of competing proposals. Because a verdict in these settings has implications for policy, and truth is difficult or impossible to verify, it is important to guard against manipulation by respondents who may have their own interests at stake. The possibility of manipulation has not been considered in this paper, as we have assumed that respondents gave honest and careful answers. We note, however, that the expertise score of (8) is identical to the payoff of a game that incentivizes respondents to give truthful answers to questions, even if those answers are nonverifiable. In the context of its application to truthfulness, the score (8) was called the Bayesian Truth Serum or BTS score [5, 20].

This scoring system has features in common with prediction markets, which are gaining in popularity as instruments of crowd-sourced forecasting [21]. Like market returns, the scores in (8) sum to zero, thus promoting a meritocratic outcome by an open democratic contest. Furthermore, in both cases, success requires distinguishing one's own information from information that is widely shared. With markets, this challenge is implicit — by purchasing a security, a person is betting that some relevant information is not adequately captured by the current price. Our approach makes the distinction explicit, by requesting a personal opinion and a prediction about the crowd. At the same time, we remove a limitation of prediction markets, which is the required existence of a verifiable event. This, together with the relatively simple input requirements, greatly expands the nature and number of questions that can be answered in a short session. Therefore, in combination with the result on incentives [5], the present work points to an integrated, practical solution to the problems of encouraging honesty and identifying truth.

Acknowledgments

Supported by NSF SES-0519141, Institute for Advanced Study (Prelec), and Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20058. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] Galton, F. Vox populi. *Nature* **75**, 450–451 (1907).

[2] Sunstein, C. *Infotopia: How many minds produce knowledge* (Oxford University Press, USA, 2006).

[3] Surowiecki, J. *The wisdom of crowds* (Anchor, 2005).

[4] Koriat, A. When are two heads better than one and why? *Science* **336**, 360–362 (2012).

[5] Prelec, D. A bayesian truth serum for subjective data. *Science* **306**, 462–6 (2004).

[6] Batchelder, W. & Romney, A. Test theory without an answer key. *Psychometrika* **53**, 71–92 (1988).

[7] Uebersax, J. Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association* **88**, 421–427 (1993).

[8] Freund, Y. & Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).

[9] Chen, K., Fine, L. & Huberman, B. Eliminating public knowledge biases in information-aggregation mechanisms. *Management Science* **50**, 983–994 (2004).

[10] Morris, P. Combining expert judgments: A bayesian approach. *Management Science* **23**, 679–693 (1977).

[11] Winkler, R. The consensus of subjective probability distributions. *Management Science* **15**, B–61 (1968).

[12] Yi, S., Steyvers, M., Lee, M. & Dry, M. The wisdom of the crowd in combinatorial problems. *Cognitive science* (2012).

[13] Cooke, R. *Experts in uncertainty: opinion and subjective probability in science* (Oxford University Press, USA, 1991).

[14] Austen-Smith, D. & Banks, J. Information aggregation, rationality, and the condorcet jury theorem. *American Political Science Review* 34–45 (1996).

[15] DeGroot, M. Reaching a consensus. *Journal of the American Statistical Association* **69**, 118–121 (1974).

[16] Grofman, B., Owen, G. & Feld, S. Thirteen theorems in search of the truth. *Theory and Decision* **15**, 261–278 (1983).

[17] Hastie, R. & Kameda, T. The robust beauty of majority rules in group decisions. *Psychological review* **112**, 494 (2005).

[18] Ladha, K. The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science* 617–634 (1992).

[19] Hertwig, R. Tapping into the wisdom of the crowd-with confidence. *Science* **336**, 303–304 (2012).

[20] John, L., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science* **23**, 524–532 (2012).

[21] Wolfers, j. & Zitzewitz, E. Prediction markets. *Journal of Economic Perspectives* **18**, 107–126 (2004).

	Average probability assigned to the correct answer			Error measure		
Aggregation method	All 50 States	30 majority-solvable states	20 majority-unsolvable states	Number of incorrect answers out of 50	Quadratic error (Brier score)	Log score
Linear pool	61.4	70.1	48.3	13	0.17	-0.52
0/1 Majority vote	N/A	N/A	N/A	19*	N/A	N/A
Logarithmic pool	67.1***	79.8***	48.1	14	0.15*	-0.46**
Counting only 100% confident	70.7***	83.8***	50.9	11.5	0.15	-0.43*
LST algorithm, T/F answers only	N/A	N/A	N/A	4*	N/A	N/A
Top scorer by u^r in each state	81.5***	85.4*	75.6***	4*	0.08**	-0.36*
Average of top 3 scorers by u^r in each state	81.5***	86.7***	73.7***	2.5**	0.06***	-0.24***
Top scorer by u^r across all 50 states	98.8***	100.0***	97.0***	0.5***	0.01***	-0.02***
Probabilistic LST with 2 equal bins	70.8*	79.0	58.6	12	0.15	-0.49
Probabilistic LST with 3 equal bins	85.3***	85.6*	84.9***	3*	0.07**	-0.29*
Probabilistic LST with 5 equal bins	81.8***	78.7	86.6***	7	0.14	-0.59
Probabilistic LST with 2 scalar-quantized bins	81.7***	85.2*	76.6***	4*	0.09*	-0.34*
Probabilistic LST with 3 scalar-quantized bins	84.9***	88.5***	79.5***	5*	0.10	-0.38
Probabilistic LST with 5 scalar-quantized bins	92.7***	95.6***	88.4***	3*	0.06**	-0.31**

Table 1: **Performance of LST compared to baseline aggregation methods**. The table shows the performance of different aggregation methods on data collected in Study 3. The results of methods which do not require elicitations of personal confidence on data from studies 1 and 2 are discussed in the main text and shown in the Supplementary Information. Results are shown for baseline aggregation methods (linear and log pools), different implementations of the LST algorithm, and individual respondent BTS score (9). This includes LST applied to the binary True/False answer, and then averaging the probability of the identified experts either per question or across questions. They also include probabilistic LST with equal-sized or scalar-quantized bins. LST algorithms outperform the baseline methods with respect to Brier scores [13], log scores, and the probability they assign to the correct answer. The performance of the algorithms is shown separately on majority-solvable and unsolvable states where solvable states are defined as those for which the majority decision was correct in both Studies 1 and 2. By this definition there were 30 easy and 20 hard states. Significance assessed against the Linear pool, two-tailed matched-pair (t49), *=<.05, **=<.01, ***=<.001.