MAKING BUSINESS PREDICTIONS BY COMBINING HUMAN AND MACHINE INTELLIGENCE IN PREDICTION MARKETS

Completed Research Paper

Yiftach Nagar **Thomas W. Malone**
MIT Center for Collective Intelligence and Sloan School of Management,
Massachusetts Institute of Technology
Cambridge, MA 02142 USA
{ynagar, malone} @ mit.edu

Abstract

Computers can use vast amounts of data to make predictions that are often more accurate than those by human experts. Yet, humans are more adept at processing unstructured information and at recognizing unusual circumstances and their consequences. Can we combine predictions from humans and machines to get predictions that are better than either could do alone? We used prediction markets to combine predictions from groups of people and artificial intelligence agents. We found that the combined predictions were both more accurate and more robust than those made by groups of only people or only machines. This combined approach may be especially useful in situations where patterns are difficult to discern, where data are difficult to codify, or where sudden changes occur unexpectedly.

Keywords: Track 16. Knowledge Management and Business Intelligence, Track 10. Human-Computer Interaction, Track 18. E-business, Business-Intelligence, Forecasting, Prediction Markets, Collective-Intelligence, Artificial intelligence, Forecasts and implications, Intelligent systems, Decision Support Systems (DSS), Predictive modeling, Enterprise 2.0, Decision making/makers

Introduction

How can we make accurate and reliable predictions[1] about actions or behaviors in complex social systems? Making predictions is a key aspect of intelligence (Hawkins and Blakeslee 2005), crucial to survival in changing environments. This is true for individual human beings, of course, and it is also true for organizations. Since the 1990's, and increasingly in the past decade, companies have been deploying knowledge management and business intelligence systems, and devising complementary work processes to help them organize data, and gain insights from the information they accumulate as they operate. "Informating" (Zuboff 1985) can help companies achieve competitive advantages, and failing to extract value from information can be detrimental in competitive environments. Organizations routinely make predictions about sales, operating costs, customer preferences, business competitor moves, etc. Historically, before the computer age, such forecasts were done by human experts (working individually, or in teams). The advent of information technologies has enabled the development and use of advanced modeling for making such predictions. Many companies nowadays use predictive analytics, running algorithms that churn large amounts of data to detect and predict fraudulent activity, customer intentions to defect, etc. Indeed, given relevant data there are models that can usually make reliable predictions. Yet, when it comes to making predictions about such things as strategic moves of competitors or partners, customer preferences in fashion-driven businesses, and regulatory action, data is often unstructured, which makes the use of models alone problematic (Negash 2004).

The challenge of making accurate, reliable predictions has been the subject of extensive research in many fields, including management, operations research, economics, judgment and decision making, artificial-intelligence and psychology. Many researchers focused either on developing and improving "mechanical" models (based on mathematical, statistical, or algorithmic methods), or on the flaws and techniques of improving the judgment of human forecasters. In addition, some research has focused on techniques for aggregation of predictions from either multiple models, or multiple human forecasters.

In this paper, we argue that it is useful to link knowledge management (KM) and business intelligence (BI) by combining automated predictions based on traditional data mining with live knowledge of real people. Mechanically combining predictions from multiple humans and computer-models may prove a better strategy than relying solely on either humans or models. In what follows, we briefly review previous related work, and then present a laboratory study in which we used prediction markets to combine predictions from humans and artificial-neural-net agents about actions of human groups (football teams). We found the combined human-agent predictions to be more accurate and more robust in comparison to those made by markets of either humans alone or machines alone. We discuss the results of our study and finally, draw some broader conclusions.

Background

Substantial evidence from multiple domains suggests that models usually yield better (and almost never worse) predictions than do individual human experts (e.g. Dawes et al. 1989; Dawes and Kagan 1988; Grove et al. 2000). Whereas models (or machines) are better at information processing and are consistent (Einhorn 1972), humans suffer cognitive and other biases that make them poor judges of probabilities (cf. Kahneman et al. 1982; Kahneman and Tversky 1973; Lichtenstein et al. 1982; Rabin 1996). In addition, *"Such factors as fatigue, recent experience, or seemingly minor changes in the ordering of information or in the conceptualization of the case or task can produce random fluctuations in judgment"* (Dawes et al. 1989), and it is therefore not surprising that models of judges often outperform the judges themselves (Armstrong 2001b; Goldberg 1970; Stewart 2001). When working in groups, humans often exhibit effects such as G*roupthink* (Janis 1972; Janis and Mann 1977) and *group polarization* (see chapter 6 in Brown 1986) that negatively affect their judgment. Nevertheless, humans still have an important role in predicting real-life situations, for at least two good reasons: humans are still better at retrieval and acquisition of many types of information – especially unstructured types of information (Einhorn 1972; Kleinmuntz 1990) and this advantage is not soon to disappear. In addition, humans' common-sense is

[1] In this paper, we use the terms predictions/forecasts (and predictor/forecaster) interchangeably.

required to identify and respond to "broken-leg" situations (Meehl 1954) in which the rules normally characterizing the phenomenon of interest do not hold. Therefore, combining the human and machine/model predictions may help in overcoming and mitigating human and model respective flaws and yield better predictions. The scarcity of both theoretical and empirical work to that end is conspicuous.

Another vast body of theoretical and empirical research suggests that combining forecasts from multiple independent, uncorrelated forecasters that have relevant knowledge and information leads to increased forecast accuracy[2]. This unanimous result holds whether the forecasts are based on human judgment or mathematical models (Armstrong 2001a; Clemen 1989). Further, because it may be difficult or impossible to identify a single forecasting method that is the best (Makridakis and Winkler 1983), "*it is less risky in practice to combine forecasts than to select an individual forecasting method*" (Hibon and Evgeniou 2005). Research of pattern-recognition classifiers in artificial-intelligence offers similar conclusions (cf. Duin and Tax 2000; Ho et al. 1994; Kittler et al. 1998; Lam 2000; Suen and Lam 2000)

Weaving together these threads of inquiry, we hypothesize that it should be possible to combine predictions from multiple humans and multiple models in ways that will emphasize their relative advantages, mitigate their respective flaws, and thus yield better predictions than either humans or models alone. Indeed, it is surprising that this path has hardly been explored. Previous work (Blattberg and Hoch 1990; Bunn and Wright 1991; Einhorn 1972), emphasizes the complementary nature of humans and models in making predictions, but does not stress the potential of improving predictions by combining predictions from multiple humans and models.

But how to combine? Many different ways of combining predictions are explored in the literatures of forecasting and model fusion, including simple and weighted averaging, majority voting, max, min, median, etc., as well as techniques that involve learning, e.g. Bayesian learning. Theoretical and empirical comparisons have shown that no single method or rule of combination are best under all circumstances (see, for example, Armstrong 1989; Clemen 1989; Duin and Tax 2000; Kuncheva 2002; Lam 2000). The simple average is usually considered a good default (Armstrong 2001a), and the most robust against 'classifier peculiarities' (Lam 2000).

Over the past decade, following the success of prediction markets[3] in public settings (Berg et al. 2001), many companies have started using them to efficiently aggregate predictions from employees (Cowgill et al. 2008; Malone 2004; Sunstein 2006; Surowiecki 2004; Wolfers and Zitzewitz 2004). Empirical investigations of prediction markets performance have shown that indeed they yield predictions that are usually at least as accurate and as calibrated as other methods traditionally used for forecasting (Chen and Plott 2002; Cowgill et al. 2008; Hopman 2007; Ortner 1997; Spann and Skiera 2009). Prediction markets also fared well against other methods of combining predictions such as simple average, weighted average and a logarithmic regression (Berg et al. 2008; Chen et al. 2005).

While prediction markets have mostly been used to aggregate predictions from humans, there is no reason why the mechanism cannot be used to aggregate predictions from software agents; yet this option remains mostly unexplored[4]. One exception is a recent study by Perols et al. (2009) who used a prediction market to combine predictions from machine classifiers. In their experiment, depending on the setting, the market mechanism either outperformed or performed on par with 3 benchmark combination mechanisms: simple average, weighted average, and majority. To the best of our knowledge, no one, to this day, has tried to use prediction markets for combining human and model predictions.

It is certainly possible that, in some scenarios, prediction markets will provide only minute improvements in accuracy over other methods, as two recent studies (Goel et al. 2010; Perols et al. 2009) suggest, and

---

[2] (For discussion of the philosophical and the mathematical principles underlying the logic of combining forecasts, see Armstrong 2001a; Larrick and Soll 2006; Makridakis 1989; Sunstein 2005; Winkler 1989).

[3] Also known as information markets, decision markets, electronic markets, virtual markets, idea futures, event futures and idea markets (Tziralis and Tatsiopoulos 2007; Wolfers and Zitzewitz 2004)

[4] Albeit some researchers in machine learning have shown interest in prediction markets, their focus thus far seems to concentrate on properties of market-maker mechanisms (e.g. Chen and Vaughan 2010).

costs of implementation, set-up and training should be considered. However, prediction markets may be appealing in some settings for reasons beyond accuracy improvement. First, as Perols et al. (2009) note, unlike some combination methods that require learning and setup, prediction markets can adjust to changes in base-classifier composition and performance without requiring offline training data or a static ensemble composition configuration. Second, by increasing attentive participation and by tying compensation to performance while giving participants a sense of both fun and challenge, they serve to increase both extrinsic and intrinsic motivation. For instance, human participants in prediction markets have an economic incentive to gather more information that would improve their performance in the markets. Third, the use of markets can also help knowledge discovery and sharing in organizations (Cowgill et al. 2008; Hayek 1945), especially so in large, distributed and/or virtual environments. Fourth, they also induce a sense of participation which supports the legitimacy and acceptance of the predictions made. Finally, Markets can also be open for people to design and run their own 'pet' agents, thus potentially incorporating an open, continuous improvement pattern into the forecasting process. For these reasons, prediction markets are a potentially useful and appealing mechanism for dynamically combining predictions from a varying population of humans and agents in real organizational settings.

We hypothesize, therefore, that combining predictions made by humans and artificial-intelligence agents can outperform both predictions made solely by humans or solely by artificial-intelligence (or statistical) models. We also hypothesize that prediction markets can be a useful mechanism for dynamically combining human and agent predictions.

It is important to realize that we are not claiming that combining human and machine predictions in this way is *always* better, only that it is *sometimes* better. As such, our results can be seen as an existence proof of one situation in which it is better. In the conclusion of the paper below, we speculate about the range of other situations in which this approach may be superior.

Method

To test these hypotheses, we conducted a study whose goal was to compare the quality of predictions made by three different types of predictors: groups of humans, groups of artificial-neural-network agents, and 'hybrid' groups of humans and agents. We used prediction markets to combine the predictions that these three types of groups made. In each case, the groups predicted whether the next play in an American football game would be a "run" or "pass", based on information about the game situation just before the play occurred. This enabled us to emulate a realistic situation where humans and agents had access to different information (e.g., humans had access to unstructured video information about the game situation that would have been difficult or costly to codify for the agents). We chose the domain of making predictions about football games, in part, because it was analogous to more complex real-world predictions such as what actions would a business competitor take next.

We hypothesized that 'hybrid' markets of humans and computers would do better than both markets of computer-agents with no humans, and markets of humans with no computers.

Lab Experiments

We conducted 20 laboratory sessions in which groups of 15 – 19 human subjects participated in prediction markets, both with and without computer agents (median group size was 18; mean 17.55; mode 19; totaling 351 participants overall).

Human participants: Participants were recruited from the general public via web advertising. We encouraged the participation of football fans by stressing the fun part and by clearly stating that knowledge of football could help make higher profits, but specific knowledge about football was not a mandatory requirement. Compensation to participants included a base payment and an additional performance-based bonus that was proportional to the ending balance in each participant's account and could reach up to 75% of the base pay.

AI agents: We used standard 3-layer artificial neural-net agents, developed using the *JOONE* open-source package[5]. For each play, the agents had three pieces of previously coded information: the down number, the number of yards to first down, and whether the previous play was a run or pass. We used a sigmoid transform function for our output layer, which limits the output to within the range of 0 to 1. The agents were trained on a similar dataset of plays from a previous game. During the tuning phase, one agent was designed to make a trade in every market, no matter its confidence. It traded in the direction of the neural network output (i.e. output >= 0.5 means pass and output < 0.5 means run). Then an additional parameter, *BiasForAction* was added to control the agent trading, such that agents traded only when their confidence level was above a threshold. A sensitivity analysis of the agents' performance with different values of bias for action (ranging from 0.01 to 0.3) allowed us to achieve correct classification for 87.5% of the plays in the test set when *BiasForAction* was set at 0.08. However, in that case the agents would trade in only about 25% of the markets. In order to have more agent participation, therefore, we set *BiasForAction* to 0.3 for the experiments reported here. This allowed the agents to participate in all markets, with 75% of the plays in the test set correctly classified.

Of course, there are many other possible kinds of artificial intelligence approaches that could be used here, many of which could presumably make the agents more accurate. As noted above, however, our goal was not to create the best possible artificial intelligence agents, merely to create one example of such agents for experimentation.

Course of experiment: Subject first filled a short questionnaire about their football keenness where they reported their level of interest and their self-assessed level of knowledge, and also answered a 20-question quiz designed to estimate their level of expertise. Then, after initial explanation[6] and training rounds, each experimental session included 20 plays. The same set of 20 plays was shown in all the sessions. For each play, a short video excerpt from the game was shown to all participants. The video was automatically stopped just before the team possessing the ball was about to start a play. At that stage, a new online prediction market was opened and the group of participants (either human participants only, or human participants along with AI agents[7]) started trading contracts of RUN and PASS[8]. We ran the experiments using a custom-tailored version of the *ZOCALO* open-source prediction markets platform[9], and employed its automated market maker to simplify trading and ensure liquidity in the markets. The market was then closed after 3.5 minutes[10], and the video continued, revealing what had actually happened, and stopping before the next play.

The AI agents participated either in the first half (first 10 plays) or the second half (last 10 plays) of the experiment (according to a random draw previously performed). Human participants were told that AI agents would trade in some of the markets but were not told in which, and could not generally tell. Thus in each lab session we collected data from 10 'human only' markets and 10 'hybrid' (humans and agents) markets.

---

[5] Available at http://sourceforge.net/projects/joone/

[6] Participants were given an elaborate verbal explanation on the goal of the experiment, and on trading in the prediction market. In addition, participants were prompted to read a short manual the day before coming to the lab, doing which – as they were truthfully told, would raise their chance to succeed in the markets and make a higher bonus. We regularly checked by vote of hands how many of them actually read the manual and the overwhelming majority did. The manual was also available on participants' screens, though they rarely, if ever, referred to it during the sessions.

[7] We ran 10 neural-net agents. They used the same code and same training dataset, but their logic of trading was also based on the market price, and they were started in a staggered manner so that they encountered different market conditions.

[8] RUN and PASS were two exhaustive options in this case (we eliminated other moves from the video).

[9] Available at http://zocalo.sourceforge.net/

[10] We decided on 3.5 minutes after several pilot sessions in which we monitored participation and noted that in most cases trading had largely stopped by about 3-3:15 minutes. We alerted participants 30 seconds prior to closing each market.

In addition we ran 10 "computer-only" experimental sessions with no human participants, where the agents traded in all 20 markets, predicting the same plays. We thus got a total of 600 observations (10 observations of each of our 3 conditions for each play).

In our analysis, we took the market closing price as representing the collective group estimation of the probability of the football team to either RUN or PASS the ball (see Wolfers and Zitzewitz 2004; Wolfers and Zitzewitz 2006).

Results

Assessing the Outcome: What Makes a Better Predictor

Prediction quality is a multidimensional concept that aims to capture the degree of correspondence between predictions and observations. There are many measures by which predictions can be assessed, but no single measure is sufficient for judging and comparing forecast quality (Jolliffe and Stephenson 2003). Thus, assessment of prediction quality is a matter of analyzing and understanding trade-offs. To compare the three groups of predictors, we therefore look at three criteria common in the forecasting literature: ***Accuracy, Reliability (a.k.a Calibration)*** *and* ***Discrimination*** which, combined, help understand those trade-offs. We augment our analysis with a comparison of accuracy vs. variability, using the Sharpe ratio (Sharpe 1966; Sharpe 1994), commonly used in economics to compare reward-vs.-risk performance, and then also present an analysis based on the ***Receiver-Operating-Characteristic (ROC)*** approach (Swets 1988; Swets and Pickett 1982; Zweig and Campbell 1993) that has been established and widely accepted in many domains as a method of assessing and comparing predictors who make predictions about binary events. The ROC analysis is in itself a trade-off analysis, which sheds more light on our findings.

Accuracy

Accuracy is a measure or function of the average distance/error between forecasts and observations. A common way to assess the accuracy of predictions and to compare the skill of the people or methods that created them is to use a scoring rule.

Table 1 summarizes the evaluations of accuracy for the human-only markets, agent-only markets, and hybrid markets, over the experimental play set, according to three popular scoring rules: the Mean Absolute Error (MAE), the Mean Square Error (MSE) – also known as the Brier Score (Brier 1950), and the Log Scoring Rule[11] (LSR, introduced by Good 1952). The lines to the right of the MAE and the LSR scores indicate where the differences between the scores of the different predictors were found statistically significant[12] ($p<0.05$). Under all of these scoring rules, a score that is closer to zero is better, and under all of them a perfect predictor who assigns a probability estimation of 100% to actual events and a probability of zero to all other potential options will score zero. While we found a weak correlation between the level of knowledge of individuals and their performance, differences in the average levels of football-keenness of the different groups did not have a significant effect on the accuracy of group predictions.

---

[11] To keep the logic of other rules we reversed the original Log scores, such that a lower score is better. $LSR = 2\text{-}Log_{10}(P)$, where P is the prediction (market closing price) of the actual outcome.

[12] To compare the conditions we built a mixed model to account for nesting, and used SAS's PROC MIXED (Littell et al. 2006) with the first-order autoregressive AR(1) error-covariance-matrix structure (ibid.). The squared errors are not normally distributed, which hinders a parametric statistical comparison of the MSE scores. Therefore significance tests for this column are not shown. Distributions of the absolute errors and of the log-predictions are quasi-normal. *Variability of group level aggregates of the level of interest in, or knowledge of football had no significant effect on the results.*

Table 1 – Accuracy of Prediction Markets

	Scoring Rule		
	Mean Absolute Error	Mean Squared Error	LSR
Humans-only Markets	0.42	0.20	0.25
Agents-only Markets	0.35	0.17	0.23
Hybrid Markets	0.35	0.15	0.21

It is up to the decision maker, therefore, to select the rule to be used for evaluation, and this would be done according to the nature of the setting and the corresponding cost functions. For example, in weather forecasting, small errors are tolerable on a daily basis (say, ±1 degree in temperature predictions), but big errors (predicting a very hot day which turns out to be very cold, or failing to predict a tornado) are not. In a production setting, on the other hand, it may be OK to throw away a unit due to a large prediction error on rare occasions, but precision is very important on a regular basis. While there may be some ambiguity in selecting a scoring rule when the cost of errors is unknown, in our case it appears that the number of large errors matters more than the average accuracy (e.g. it is likely that a prediction of 90% and prediction of 95% for a PASS attempt by the offense team would both translate to the same decision by the defense team) and hence, the MSE and the LSR seem more appropriate than the MAE.

Taken together, these results suggest that the hybrid markets were the most accurate. We also note that although the agents were very simple, on average the agent-only markets were more accurate than the human-only markets, as one could expect based on previous evidence. We later turn to use the ROC method to make a comparison that is agnostic to the cost of errors, but first we explore how well our predictors predicted each play and consider a few other criteria.

A Deeper Look at the Play Level

A deeper look at the play level provides better understanding of the behavior of the predictors and reveals several interesting patterns. Figure 1 depicts the mean absolute prediction error (average of 10 observations from 10 markets) of each condition, per play.

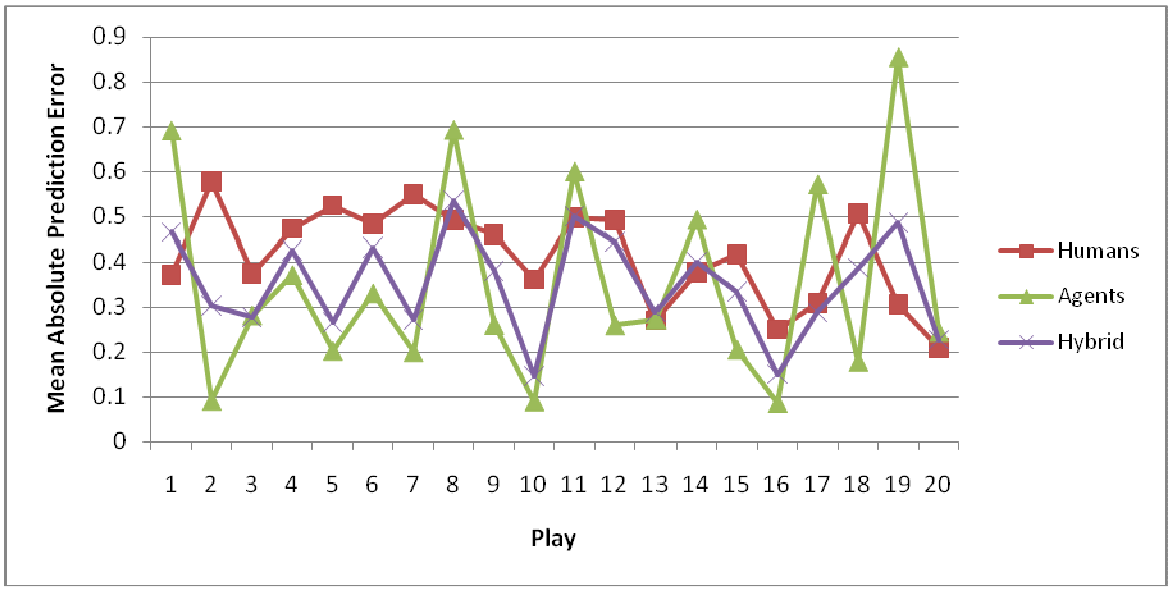

Figure 1 – Mean Prediction Errors of Human, Agents And Hybrid Markets

We note a strong interaction between condition and play. As could be expected, humans and agents predicted differently on different plays. While *on average* agents were more accurate than humans (i.e. had smaller errors), in several cases they made severe errors while humans predicted correctly (notably: plays 4, 12, 16, 29). But why? Informal interviews with participants suggested that they incorporated more information into their decision-making than did the agents. Notably, they gleaned from the video the formation of the offensive and the defensive teams. For example: before both play 4 and play 29, the offense team formed a "Shotgun" formation[13], with a running-back standing next to the quarterback, which to football savvy fans implies a higher probability for a pass attempt. In both those plays, the 'human-only' markets clearly indicated a pass (70% and 77% on average) whereas the 'all-agents' markets indicated a RUN (69% and 85.5% on average, corresponding to 31%, 14.5% predictions for PASS). A few participants also reported that commentary by anchors was helpful, and several others mentioned that the body language of players was revealing.

Beyond Mean Errors: Considering Prediction-Error Variability

Measures of accuracy alone do not provide sufficient information to convey the complexity of the data, as they are essentially comparisons of single numbers representing entire distributions. Two predictors can yield the same mean F (Prediction Error), where F is some scoring rule, and yet offer very different predictions and risk profiles. Therefore, it is important to consider the variability of prediction errors of the different predictors being compared. After assigning economic values to the predictions using scoring rules, the ex post Sharpe ratio (Sharpe 1966; 1994), commonly used in finance to compare reward-vs.-risk performance, enables us to consider accuracy against variability of prediction errors, making the comparison more informative. To keep with the familiar logic of the Sharpe ratio that assumes a higher positive financial return is better, we adjust our scoring rules such that the adjusted MAE score (AMAE) equals 1-MAE and the adjusted MSE score (AMSE) equals 1-MSE. The adjusted Log score is log10(P) where P is the prediction (market closing price) of the actual outcome. We calculated the Sharpe ratio according to equations 3-6 in Sharpe (1994). As a simple and straightforward benchmark, we use an "ignorant" predictor who bets 50% PASS all the time (and whose error variance is therefore zero). The corresponding AMAE, AMSE and ALSR for the benchmark predictor are therefore 0.5, 0.75 and 1.699, correspondingly. The results are summarized in Table 2.

Table 2 - Ex Post Sharpe Ratio for Prediction Markets, Under 3 Scoring Rules

	Scoring Rule		
	AMAE (Benchmark = 0.5)	AMSE (Benchmark = 0.75)	ALSR (Benchmark = 1.699)
Humans-only Markets	0.54	0.40	0.41
Agents-only Markets	0.67	0.39	0.37
Hybrid Markets	0.91	0.73	0.72

Clearly, the hybrid markets yield the highest Sharpe ratio and outperform both the human-only and agent-only markets. This result holds under three different scoring rules. According to the Sharpe ratio index criterion, therefore, the Hybrid markets are more robust, offering a better trade-off between prediction accuracy and variability.

Calibration and Discrimination

Reliability (Murphy and Winkler 1977), (also: Calibration, e.g. Lichtenstein et al. 1982), refers to the degree of correspondence between forecast probabilities and actual (observed) relative event frequencies. For a predictor to be perfectly calibrated, assessed probability should equal percentage-correct where repetitive assessments are being used (ibid.). Figure 2 depicts the calibration diagram for our 3

[13] Mallory, B., & Nehlen, D. (2006, ch. 7-8). *Football offenses & plays*: Human Kinetics Publishers

conditions. Evidently, both the human and hybrid markets were reasonably calibrated, while the agents were not.

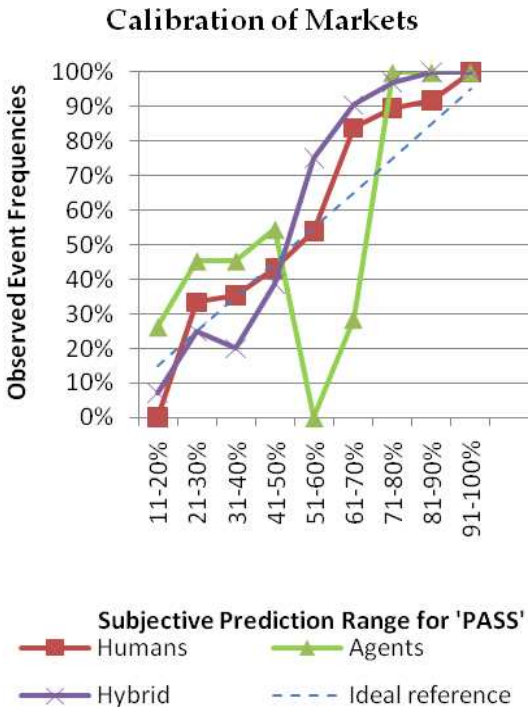

Figure 2

Discrimination (a.k.a Resolution) taps forecasters' ability to do better than a simple predict-the-base-rate strategy. Observers get perfect discrimination scores when they infallibly assign probabilities of 1.0 to things that happen and probabilities of zero to things that do not (Tetlock 2005). It is important to note that calibration skill and discrimination skill are two separate skills. For example, a predictor that always predicts the base-rate of the event will score high on calibration but low on discrimination (for such predictor, the calibration plot will only include a single point, on the diagonal reference line). It has been offered that the MSE can be decomposed as VI+CI-DI where VI is the variability index representing the uncertainty of the phenomena, CI is the calibration index of the forecasts and DI is the discrimination index of the forecasts (Murphy 1973; Murphy and Winkler 1987; see also Tetlock 2005). While the MSE may have drawbacks as a criterion by which to judge the quality of predictions, this decomposition seems nevertheless useful in orienting our understanding of the trade-off between calibration and discrimination of our predictors. Given that the variability of the events in our case is identical for the 3 conditions we want to compare (since they made predictions about the same events) we can draw a plot of (Variability – Discrimination) vs. Calibration for each predictor. For a given variability, we can also draw "efficient front" isopleths of MSE. We present such a plot depicting the performance of our 3 conditions in Figure 3. VI in this study was 0.24. In this plot, the more calibrated a predictor is, the more to the left it would appear (CI closer to zero is better). The more discriminating a predictor is, the lower it would appear. It is evident in this plot that the Hybrid markets were about as discriminating as the agent markets, but more calibrated. It is also clear that compared to human markets, the hybrid markets were slightly less calibrated, but more discriminating. Overall, the hybrid markets are on a more efficient front compared to both agents markets and human markets – as reflected in the MSE scores.

While the agents were more accurate than the humans *on average*, their predictions were less calibrated, and they made more severe errors. For any practical matter, they were *utterly* wrong about 4 out of 20 plays (with errors ranging 60-85%; and wrong to a lesser degree on one other play), potentially rendering them untrustworthy for a decision maker (though, of course that depends on the cost of the errors to the

decision maker). Humans, on the other hand, had only 3 plays where their prediction (average of 10 markets) was in the wrong direction – but in 2 out of those, their average error was less than 0.55 (and in the third, less than 0.58), conveying their uncertainty to the decision maker by a prediction that was very close to the base rate. Then again, they were also very hesitant (non-discriminating) in most other cases, even when predicting the correct outcome, raising doubt about their value as predictors. The 'hybrid' combination of humans and agents proved to be useful in mitigating both those problems. In terms of accuracy or discrimination, it did not fall far from the agents (in fact, according to the MSE and the LSR criteria, the hybrid markets were more accurate than the agents). In addition, it provides better calibration than the agents, and better discrimination than that of the humans. Importantly, the hybrid groups were on average wrong only about a single play (12), yet their prediction for that play (53.5% Run/46.5% Pass) clearly indicates their lack of confidence in this case to the decision maker.

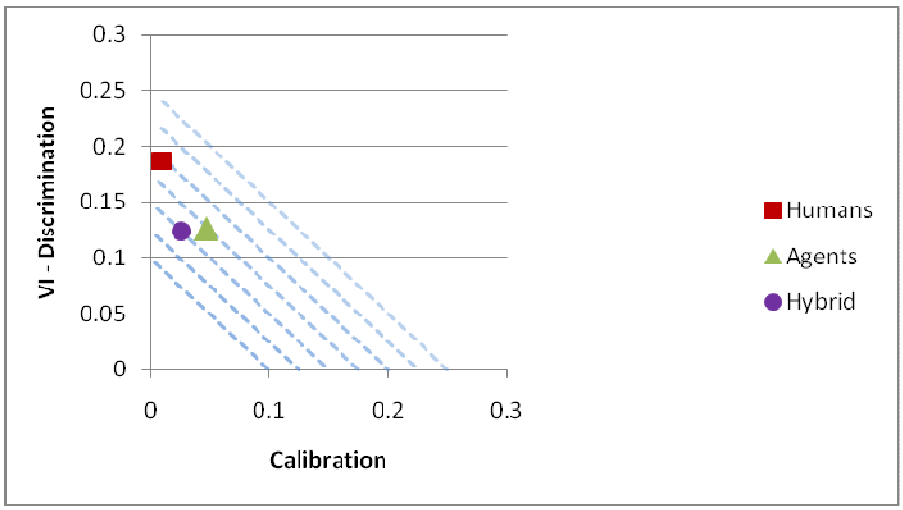

Figure 3 - Variability - Discrimination vs. Calibration (with MSE Isolines. VI=0.24)

ROC Analysis

Our comparisons of accuracy, and of the Sharpe ratio, both rely on attaching values to prediction errors using scoring rules. While we used common rules, they may not represent the actual economic value of predictions (or corresponding errors), and in reality, it is not always possible to determine those values. The Receiver-Operating-Characteristic (ROC) is an established methodology for evaluating and comparing the performance of diagnostic and prediction systems, which does not rely on their unknown economic value, and hence, can provide additional support for our conclusions. ROC has been widely used in many different domains including signal detection, radiology, weather forecasting, psychology, information retrieval etc. (Swets 1973; Swets 1988; Swets and Pickett 1982; Zweig and Campbell 1993). ROC curves are obtained by plotting the hit rate (i.e. correctly identified events) versus the false alarm rate (incorrect event predictions) over a range of different thresholds that are used to convert probabilistic forecasts of binary events into deterministic binary forecasts (Jolliffe and Stephenson 2003). The ROC, plotted for a range of different thresholds, offers a more credible view of the entire spectrum of accuracy of the different predictors (Zweig and Campbell 1993), and serves to highlight the tradeoff between sensitivity and specificity of each predictor. The area under the curve (AUC) serves as a measure of the quality of the predictions, with a perfect predictor scoring 1. The ROC curves of our conditions are presented in Figure 4, and the areas under the curves are depicted in Table 3.

This result suggests that the hybrid prediction markets may provide a better trade-off between sensitivity and specificity when compared to either human-only or agent-only prediction markets. In that, it echoes our previous analyses.

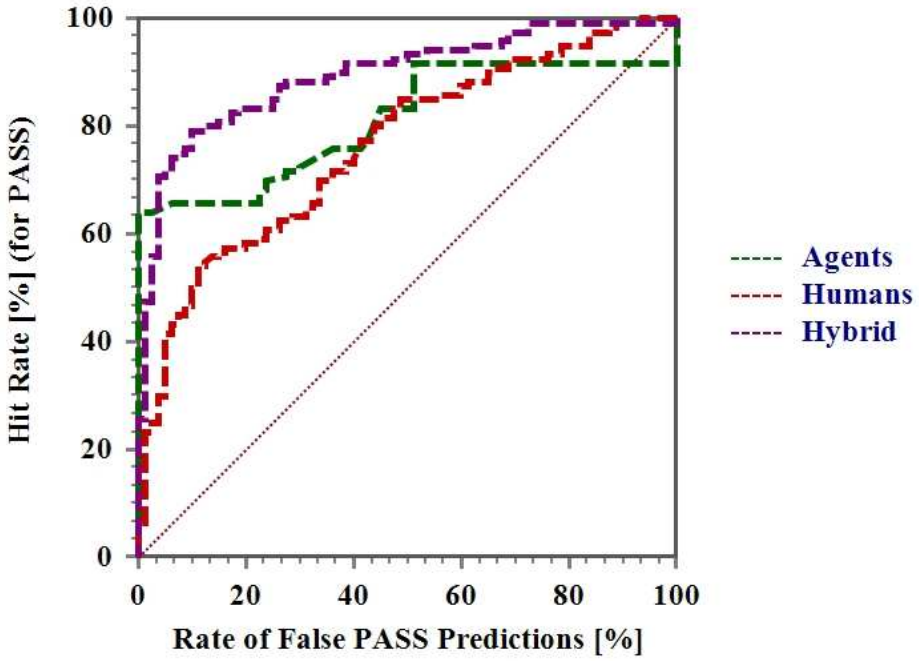

Figure 4 – ROC Curves for predictions of football plays by Human-only, Agent-only and Hybrid prediction markets (20 plays, 10 observations of each play by each condition)

Table 3

	Area under ROC Curve[14]	SE[15]
Humans	0.76	0.033
Agents	0.81	0.031
Hybrid	0.90	0.022

[14] The areas under the curves were calculated in MedCalc software (Schoonjans et al. 1995). MedCalc is available from http://www.medcalc.be/

[15] Standard errors were calculated using the method offered by DeLong, DeLong, & Clarke-Pearson (1988). However they may be inaccurate as we used repeated measurements.

Discussion

We used prediction markets to combine predictions from multiple forecasters, under three conditions: markets of human forecasters, markets of artificial-neural-net agents, and markets where both humans and agents traded together. We used several different measures and criteria to assess and compare the quality of the predictions, including accuracy (measured using 3 common scoring rules), Sharpe ratios, calibration, discrimination and receiver-operating characteristic plots.

The combination of humans and agents proved to be more accurate than either humans or agents according to 2 scoring rules (MSE and LSR). Under the MAE scoring rule, both agent-markets and hybrid markets were more accurate than human markets, though the accuracy of the hybrid prediction markets was indistinguishable from that of the agent markets. The combination of humans and agents provided predictions that were more calibrated than those of the agents, more discriminating than those of the humans and overall providing a better tradeoff of calibration and discrimination compared to the humans or the agents. Predictions made by hybrid humans-and-agents markets also provided the best tradeoff of accuracy and variability of prediction errors, as measured by the Sharpe ratio. An ROC analysis, which does not rely on any assumptions about the cost of errors, also shows that hybrid markets provide a better trade-off between good and bad predictions. Overall, therefore, the combination of human and agent predictions proved more robust, and arguably, superior to either the agents-only predictions or the humans-only predictions in our setting.

What do these results imply about combining predictions of humans and models or agents in general? Naturally, our study has limitations that constrain our ability to generalize its conclusions. We use a limited set of events with binary outcomes, from a single domain. Our implementation of the neural-net agents was simplistic, and, accordingly, their trading in the prediction markets was simplistic too. But, as mentioned above, our goal in this study was not to prove definitively that one method is superior. Rather, it was to provide an existence proof that there are situations where combining predictions from humans and artificial-intelligence agents can outperform those of either group alone. To that end, our results support that claim.

This paper thus contributes to the growing body of knowledge about predictions in the following ways. First, we demonstrate the potential value of mechanically combining predictions from *multiple* models *and multiple* humans. Some previous work (Blattberg and Hoch 1990; Bunn and Wright 1991; Einhorn 1972; Kleinmuntz 1990) suggested that combining human and model predictions may be useful; however, this previous work focused more on the differences between humans and models, devoting attention to analyzing their respective strengths and weaknesses, and did not delve into the details of how to combine them. In addition, none of these authors is explicit about the option of combining *multiple* humans with *multiple* models. To the best of our knowledge, this approach has not been previously suggested or empirically tried, but we hope our results will encourage others to do so in the future.

Second, we show that using artificial intelligence models (e.g. artificial-neural-nets) – as opposed to "traditional" statistical models – in such combinations can be beneficial. Among the advantages of artificial-neural-nets is their the ability to dynamically adapt the model as new data becomes available (Tam and Kiang 1992). Empirical examples elsewhere (ibid.) demonstrate the power of artificial-intelligence in inferring rules from large datasets and supporting the case for use of artificial-intelligence to make better predictions than those of more traditional methods.

Finally, we demonstrate that prediction markets can be a useful way to combine human and machine predictions. As discussed above, many different ways for combining predictions (from either humans or models) are explored in the literature, and no single method is best (or even suitable) for all settings. It is not our goal to argue that prediction markets are always the *best* method for combining multiple predictions, but they are appealing for a number of reasons described in the Background section. Our study further shows that prediction markets can produce valuable predictions in a very short time (minutes), thus potentially serving as a real-time prediction tool. In this sense, our work responds to calls in the literature on predictive analytics and business intelligence for reactive components to help decision makers monitor and respond to time-critical operational processes (e.g. Matteo et al. 2004).

Additional work is required to identify and compare other ways of combining human and machine predictions, and to understand their respective advantages and disadvantages in different contexts. Future work should also examine this approach in more complex domains, and with more sophisticated, domain-specific agents.

Conclusion

As previous research has shown, there are many business and other situations where mechanical predictions based on structured data are superior to predictions made by humans (Grove et al. 2000) and there are many other situations where the factors relevant to predicting are so complex, or where there is so little codifiable data, that no suitable mechanical models even exist. In these cases, the only option is to rely on human judgment.

But we believe there are also many important real-world situations where combining predictions from humans and agents can be valuable. For instance, this approach may be particularly beneficial in complex situations involving the actions of human groups (such as business competitors, customers, partners, or regulators), where it may be difficult to discern or formulate all the rules governing the phenomena of interest but where there is still a fair amount of relevant data that can be analyzed. One prototypical example of such a situation, for instance, would be predicting sales of fashion-driven products. There is a great deal of relevant data that can be analyzed, but some important factors are very difficult to quantify in advance. In such domains, machine learning and other quantitative methods can be useful in building sophisticated and adaptive models based on potentially vast amounts of data (for recent examples, see Bohorquez et al. 2009; Mannes et al. 2008). And humans' tacit knowledge, ability to acquire unstructured information, and intuition can help in both information retrieval, and preventing catastrophic prediction errors.

This approach can thus leverage and combine existing business intelligence and knowledge management approaches in new ways. In general, we believe that there is much additional work to be done to clarify the situations in which combining human and machine intelligence in this way is most useful, and the best ways of making this combination. We hope our initial work reported here will encourage others to further investigate this promising direction.

Acknowledgments

We thank MIT Lincoln Laboratory and the U.S. Army Research Laboratory's Army Research Office (ARO) for funding this project. We are grateful to John Willett for his patience, dedication and help with statistical analyses and to Chris Hibbert for software development, and education about prediction markets. We thank Sandy Pentland, Tomaso Poggio, Drazen Prelec, and Josh Tenenbaum for many discussions out of which this project originated, and benefited greatly. For help with software and experimental design we thank Jason Carver, Wendy Chang, Jeremy Lai and Rebecca Weiss. For their wise comments we thank John Carroll, Gary Condon, Robin Hanson, Haym Hirsh, Josh Introne, Ben Landon, Retsef Levi, David Pennock, Cynthia Rudin, and Paulina Varshavskaya. This paper also benefited from constructive comments of participants and two anonymous reviewers of the NIPS 2010 Crowdsourcing and Computational Social Science Workshop, as well as those of the associate editor and two anonymous reviewers of ICIS 2011. Thanks also go to our research assistants Jonathan Chapman, Catherine Huang, Natasha Nath, Carry Ritter, Kenzan Tanabe and Roger Wong, and to Richard Hill and Robin Pringle for administrative support.

References

Armstrong, J.S. 1989. "Combining Forecasts: The End of the Beginning or the Beginning of the End?," *International Journal of Forecasting* (5), pp. 585-588.

Armstrong, J.S. 2001a. "Combining Forecasts," in *Principles of Forecasting: A Handbook for Researchers and Practitioners,* J.S. Armstrong (ed.). Kluwer Academic Publishers.

Armstrong, J.S. 2001b. "Judgmental Bootstrapping: Inferring Experts' Rules for Forecasting," in *Principles of Forecasting: A Handbook for Researchers and Practitioners,* J.S. Armstrong (ed.). Norwell, MA: Kluwer Academic Publishers.

Berg, J.E., Forsythe, R., Nelson, F., and Rietz, T.A. 2001. "Results from a Dozen Years of Election Futures Markets Research," *Handbook of Experimental Economic Results*), pp. 486–515.

Berg, J.E., Nelson, F.D., and Rietz, T.A. 2008. "Prediction Market Accuracy in the Long Run," *International Journal of Forecasting* (24:2), pp. 283-298.

Blattberg, R.C., and Hoch, S.J. 1990. "Database Models and Managerial Intuition: 50% Model+ 50% Manager," *Management Science* (36:8), pp. 887-899.

Bohorquez, J.C., Gourley, S., Dixon, A.R., Spagat, M., and Johnson, N.F. 2009. "Common Ecology Quantifies Human Insurgency," *Nature* (462:7275), pp. 911-914.

Brier, G.W. 1950. "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* (78:1), pp. 1-3.

Brown, R. 1986. *Social Psychology*, (2nd ed.). New York, NY: Free Press.

Bunn, D., and Wright, G. 1991. "Interaction of Judgemental and Statistical Forecasting Methods: Issues and Analysis," *Management Science* (37:5), May, pp. 501-518.

Chen, K.-Y., and Plott, C.R. 2002. "Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem," *California Institute of Technology, Division of the Humanities and Social Sciences, Working Paper 1131*), March.

Chen, Y., Chu, C.-H., Mullen, T., and Pennock, D., M. . 2005. "Information Markets Vs. Opinion Pools: An Empirical Comparison," in: *Proceedings of the 6th ACM conference on Electronic commerce*. Vancouver, BC, Canada: ACM, pp. 58-67.

Chen, Y., and Vaughan, J.W. 2010. "A New Understanding of Prediction Markets Via No-Regret Learning," *11th ACM conference on Electronic Commerce (EC '10)*, Cambridge, MA: ACM, pp. 189-198.

Clemen, R.T. 1989. "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting* (5:4), pp. 559-583.

Cowgill, B., Wolfers, J., and Zitzewitz, E. 2008. "Using Prediction Markets to Track Information Flows: Evidence from Google," *Dartmouth College*).

Dawes, R.M., Faust, D., and Meehl, P.E. 1989. "Clinical Versus Actuarial Judgment," *Science* (243:4899), pp. 1668-1674.

Dawes, R.M., and Kagan, J. 1988. *Rational Choice in an Uncertain World*. Harcourt Brace Jovanovich San Diego.

DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. 1988. "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics* (44:3), pp. 837-845.

Duin, R., and Tax, D. 2000. "Experiments with Classifier Combining Rules," *First International Workshop on Multiple Classifier Systems (MCS 2000)*, Cagliari, Italy: Springer, pp. 16-29.

Einhorn, H.J. 1972. "Expert Measurement and Mechanical Combination," *Organizational Behavior and Human Performance* (7:1), pp. 86-106.

Gneiting, T., and Raftery, A.E. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association* (102:477), pp. 359-378.

Goel, S., Reeves, D.M., Watts, D.J., and Pennock, D.M. 2010. "Prediction without Markets," *11th ACM conference on Electronic commerce*, Cambridge, MA: ACM, pp. 357-366.

Goldberg, L.R. 1970. "Man Versus Model of Man: A Rationale, Plus Some Evidence, for a Method of Improving on Clinical Inferences," *Psychological Bulletin* (73:6), pp. 422-432.

Good, I.J. 1952. "Rational Decisions," *Journal of the Royal Statistical Society. Series B (Methodological)* (14:1), pp. 107-114.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., and Nelson, C. 2000. "Clinical Versus Mechanical Prediction: A Meta-Analysis," *Psychological Assessment* (12:1), pp. 19-30.

Hawkins, J., and Blakeslee, S. 2005. *On Intelligence*. Owl Books.

Hayek, F.A. 1945. "The Use of Knowledge in Society," *The American Economic Review* (35:4), pp. 519-530.

Hibon, M., and Evgeniou, T. 2005. "To Combine or Not to Combine: Selecting among Forecasts and Their Combinations," *International Journal of Forecasting* (21:1), pp. 15-24.

Ho, T.K., Hull, J.J., and Srihari, S.N. 1994. "Decision Combination in Multiple Classifier Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (16:1), pp. 66-75.

Hopman, J. 2007. "Using Forecasting Markets to Manage Demand Risk," *Intel Technology Journal* (11), pp. 127–136.
Janis, I.L. 1972. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin Boston.
Janis, I.L., and Mann, L. 1977. *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment*. The Free Press New York.
Jolliffe, I.T., and Stephenson, D.B. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley.
Kahneman, D., Slovic, P., and Tversky, A. 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.
Kahneman, D., and Tversky, A. 1973. "On the Psychology of Prediction," *Psychological review* (80:4), pp. 237-251.
Kittler, J., Hatef, M., Duin, R.P.W., and Matas, J. 1998. "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (20:3), pp. 226–239.
Kleinmuntz, B. 1990. "Why We Still Use Our Heads Instead of Formulas: Toward an Integrative Approach," *Psychological Bulletin* (107:3), pp. 296-310.
Kuncheva, L.I. 2002. "A Theoretical Study on Six Classifier Fusion Strategies," *IEEE Transactions on pattern analysis and machine intelligence* (24:2), February, pp. 281-286.
Lam, L. 2000. "Classifier Combinations: Implementations and Theoretical Issues," *First International Workshop on Multiple Classifier Systems (MCS 2000)*, Cagliari, Italy: Springer, pp. 77-86.
Larrick, R.P., and Soll, J.B. 2006. "Intuitions About Combining Opinions: Misappreciation of the Averaging Principle," *Management Science* (52:1), p. 111.
Lichtenstein, S., Baruch, F., and Phillips, L.D. 1982. "Calibration of Probabilities: The State of the Art to 1980," in *Judgment under Uncertainty: Heuristics and Biases,* D. Kahneman, P. Slovic and A. Tversky (eds.). Cambridge University Press.
Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. 2006. *Sas for Mixed Models*, (2nd ed.). SAS Publishing.
Makridakis, S. 1989. "Why Combining Works?," *International Journal of Forecasting* (5:4), pp. 601-603.
Makridakis, S., and Winkler, R.L. 1983. "Averages of Forecasts: Some Empirical Results," *Management Science*), pp. 987-996.
Malone, T.W. 2004. "Bringing the Market Inside," *Harvard Business Review* (82:4), pp. 106-114.
Mannes, A., Michael, M., Pate, A., Sliva, A., Subrahmanian, V.S., and Wilkenfeld, J. 2008. "Stochastic Opponent Modeling Agents: A Case Study with Hezbollah," in *Social Computing, Behavioral Modeling, and Prediction,* H. Liu, J.J. Salerno and M.J. Young (eds.). pp. 37-45.
Matteo, G., Stefano, R., and Iuris, C. 2004. "Beyond Data Warehousing: What's Next in Business Intelligence?," in: *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP*. Washington, DC, USA: ACM.
Meehl, P.E. 1954. *Clinical Vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: University of Minnesota Press.
Murphy, A.H. 1973. "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology* (12:4), pp. 595–600.
Murphy, A.H., and Winkler, R. 1987. "A General Framework for Forecast Verification," *Monthly Weather Review* (115:7), pp. 1330-1338.
Murphy, A.H., and Winkler, R.L. 1977. "Reliability of Subjective Probability Forecasts of Precipitation and Temperature," *Applied Statistics*), pp. 41-47.
Negash, S. 2004. "Business Intelligence," *The Communications of the Association for Information Systems* (13:1), p. 54.
Ortner, G. 1997. "Forecasting Markets—an Industrial Application." University of Technology Vienna.
Perols, J., Chari, K., and Agrawal, M. 2009. "Information Market-Based Decision Fusion," *Management Science* (55:5), pp. 827-842.
Rabin, M. 1996. "Psychology and Economics," *Journal of Economic Literature* (36:1), March, pp. 11-46.
Schoonjans, F., Zalata, A., Depuydt, C.E., and Comhaire, F.H. 1995. "Medcalc: A New Computer Program for Medical Statistics," *Computer Methods and Programs in Biomedicine* (48:3), pp. 257-262.
Sharpe, W.F. 1966. "Mutual Fund Performance," *Journal of business* (39:1), pp. 119-138.
Sharpe, W.F. 1994. "The Sharpe Ratio," *Journal of portfolio management* Fall), pp. 49-58.
Spann, M., and Skiera, B. 2009. "Sports Forecasting: A Comparison of the Forecast Accuracy of Prediction Markets, Betting Odds and Tipsters," *Journal of Forecasting* (28:1), pp. 55-72.

Stewart, T.R. 2001. "Improving Reliability of Judgmental Forecasts," in *Principles of Forecasting: A Handbook for Researchers and Practitioners,* J.S. Armstrong (ed.). Kluwer Academic Publishers, pp. 81-106.

Suen, C.Y., and Lam, L. 2000. "Multiple Classifier Combination Methodologies for Different Output Levels," *First International Workshop on Multiple Classifier Systems (MCS 2000)*, Cagliari, Italy: Springer, pp. 52-66.

Sunstein, C.R. 2005. "Group Judgments: Statistical Means, Deliberation, and Information Markets," *New York University Law Review* (80), p. 962.

Sunstein, C.R. 2006. *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press, USA.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday.

Swets, J.A. 1973. "The Relative Operating Characteristic in Psychology," *Science* (182:4116), pp. 990-1000.

Swets, J.A. 1988. "Measuring the Accuracy of Diagnostic Systems," *Science* (240:4857), pp. 1285-1293.

Swets, J.A., and Pickett, R.M. 1982. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press.

Tam, K.Y., and Kiang, M.Y. 1992. "Managerial Applications of Neural Networks: The Case of Bank Failure Predictions," *Management Science* (38:7), pp. 926-947.

Tetlock, P. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press.

Tziralis, G., and Tatsiopoulos, I. 2007. "Prediction Markets: An Extended Literature Review," *Journal of Prediction Markets* (1:1), pp. 75-91.

Winkler, R.L. 1969. "Scoring Rules and the Evaluation of Probability Assessors," *Journal of the American Statistical Association* (64:327), September, pp. 1073-1078.

Winkler, R.L. 1989. "Combining Forecasts: A Philosophical Basis and Some Current Issues," *International Journal of Forecasting* (5:4), pp. 605-609.

Wolfers, J., and Zitzewitz, E. 2004. "Prediction Markets," *The Journal of Economic Perspectives* (18:2), pp. 107-126.

Wolfers, J., and Zitzewitz, E. 2006. "Interpreting Prediction Market Prices as Probabilities," *CEPR Discussion Paper No. 5676*), May 2006.

Zuboff, S. 1985. "Automate/Informate: The Two Faces of Intelligent Technology," *Organizational Dynamics* (14:2), pp. 5-18.

Zweig, M.H., and Campbell, G. 1993. "Receiver-Operating Characteristic (Roc) Plots: A Fundamental Evaluation Tool in Clinical Medicine," *Clinical Chemistry* (39:4), pp. 561-577.