Re-Interpreting The Condorcet Jury Theorem

(DRAFT)

Lu Hong
Loyola University Chicago

Scott E Page
University of Michigan

October 5, 2015

Perhaps we should not use dice to explain democracy.

The Condorcet Jury Theorem occupies a prominent position within formal political science. Written in 1785 by the French enlightenment philosopher and mathematician, Nicolas de Condorcet, the theorem applies probability theory to provide a logical justification for jury votes, committee decisions, and democratic elections writ large.

The theorem applies to a group of voters who must identify or predict the correct answer from among two alternatives. In the canonical version, each voter independently knows the answer with the same probability. If that probability exceeds one-half, four results follow: the majority identifies correctly with a higher probability than each individual, collective accuracy increases in individual accuracy and in group size, and large groups approach but never achieve perfect accuracy.

In the mid seventeenth century when Condorcet stated the theorem, probability theory existed only in nascent form. Pascal and Fermet had constructed the formal concept of probability in order to derive and evaluate betting strategies in a popular dice game. Soon after, Jacob Bernoulli generalized the theory to a broader class of parlor games and probabilistic phenomena. The axiomatic basis for probability used today did not appear until the mid twentieth century in the work of Kolmogorov (Apostol 1969). Our point being that Condorcet was working along a mathematical frontier.

His conceptualization of a partially informed classification as a random draw or a spin of a roulette wheel, like Nash's representation of iterative thinking as best response functions, represents a seminal contribution to modeling. Nearly every model in formal political science that assumes incomplete information adopts Condorcet's framework of independent signals. The widespread use owes partly to the framework's tractability. Condorcet's model can be tweaked in myriad directions. Common extensions allow for correlated signals (Ladha 1992) common signals (Grofman, et al 1983), and strategic voting (Austen-Smith and Banks (1996), Feddersen and Pesendorfer (1996)).

In this paper, we suggest that Condorcet's construction obscures a more nuanced connection between individual and group accuracy that depends on the nature of the inference task and the diversity of the ways that people make predictions. To make that point, we contrast his framework with a more detailed framework that assumes a state space and a function mapping states of the world into outcomes. The framework generalizes models of collective forecasting, prediction, and classification used in cognitive psychology, machine learning, and econometrics. Following Hong and Page (2009), we refer to Condorcet's approach as the *generated signal* framework and to the alternative as the *interpreted signal* framework.

Stark differences distinguish the two frameworks (Page 2007). In the generated signal framework, signals are draws from a distribution. Often generated signals are assumed to have equal accuracy and be independent. Interpreted signals consist of a representation of the world and a model that maps that representation into outcomes. Those representations and models coarsen a true underlying process that maps a *state of the world* into an outcome.

The state of the world can often be represented within a *feature space* or a set of *attributes*, but in most general case it need not be. The features or attributes of the world may be constructed through an interpretation. The interpretation creates structure by transforming raw experience into categories that contain information.

As in Condorcet's paper, we restrict attention to *yes /no* decisions (formally: *binary classification problems*). Examples of these problems include a jury considering the guilt or innocence of an alleged tax evader or a vote on a policy proposal. In the case of a jury, the state of the world would consist of all relevant facts in the case – the transactions that occurred, the actions of the defendant, and the particulars of the United States Tax Code. An *outcome function* maps that state of the world into an outcome: guilty or innocent. When the jury makes a decision of guilt or innocence, they are correct only if their classification agrees with the outcome.

The dimensionality of the state space and the complexity of the outcome function may preclude a jury member from learning the precise state of the world or from learning the outcome function. In a case with abundant of minutia, a jury member may use an *interpretation*, a partition of the state space. Partitions can take at least three forms. They can ignore some dimensions of the state of the world. They can project multiple dimensions onto a single variable such as monetary value. Or, they can create bins of similar instances based on exemplars.

For each set in a partition, a jury member assigns an outcome, either guilty or innocent, using a *predictive model*. The jury member's predictive model is defined not over the state space but over the partitioning of the state space created by the interpretation. Thus, a jury member does not distinguish between any two states of the world that belong to a common set or category within her partitioning.

While in the generated signal framework, the modeler assumes only an accuracy and correlation among the signals, in the interpretive signal framework, she must make more explicit assumptions about the nature of the classification task. Accuracy and correlation depend on the dimensionality of the state space, the complexity and randomness of the outcome function, the partitions created by the interpretations, and the sophistication of predictive models. Jurors who use finer partitions and better predictive models have more accurate signals. Jurors with similar interpretations and predictive models produce more correlated signals.

The interpreted signal framework does not reveal a flaw in Condorcet's logic so much as it provides a more elaborate model that reveals the conditionality of claims derived within the generated framework. Instead of producing one big theorem: *groups are better than individuals*, it results in a collection of more modest and conditional claims. The framework highlights the importance of diversity. It also obliges us to distinguish complexity from randomness. Complexity can, but need not be, overcome by diverse interpreted signals. Randomness can never be. The framework also introduces new opportunities for strategic and coordinated behavior and allows for deeper models of deliberation (Landemore and Page 2015).

As noted above, we make no claims to the originality of the interpreted signal framework. Hong and Page (2009) introduced the terminology to organize multiple literatures. Most notably, the *ensemble methods* approach from machine learning classifies complex outcome functions using multiple classifiers that rely on distinct features and combination of features (Dietterich 2000). These are a type of interpreted signal. Successful ensemble methods such as *random forest* algorithms generate stark interpretive signals and they then vote (Brieman 2001). Decades of research on random forest methods reveal the crude and misleading nature of Condorcet's theorem. Groups aren't always better. And accuracy, as a rule, stops increasing in any meaningful way as groups increase in size.

Three differences distinguish random forests from most decision making groups. Random forest methods randomly generate decision trees (if-then rules) that partition the states of the world. Only if a tree correctly classifies a sufficient proportion of cases will it be included in the forest. Second, in constructing trees, algorithms seek diversity in features (Brown, et al 2005) as well as negatively correlated predictors (Liua and Yaob 1999). In doing so, random forests and other ensemble methods all but guarantee effective collections of interpretive signals. The group will be more accurate than the individuals. We show that this result need not hold generally.

Third, individuals learn from one another and also respond to incentives. An individual's model may come to resemble the model of those with whom she interacts and reduce group accuracy (Economo et al 2015). Individuals might also have incentives to be different. Making the correct prediction and being in the minority may lead to larger gains in a market setting (Hong, Page, Riolo 2012).

The Condorcet Jury Theorem assumes common preferences or a common goal. They do not apply directly to cases where people differ in their preferences. That said, ideological differences are often defined over instruments and fundamental desires. People may share a common desire for lower rates of unemployment and crime and higher rates of economic growth. They may only differ in how to achieve those ends. Further, though we frame our analysis in the context of a jury, our results apply to any group making a binary classification. This could be an advisory board making on an endorsement, a venture capital firm deciding on whether to fund, a department making a tenure decision, a political body ratifying a treaty, or a board of directors approving a merger.

The remainder of the paper consists of six parts. We first present the generated signal version of The Condorcet Jury Theorem and list the main results. We then present the interpreted signals framework in simplified form using two examples. In the third part, we provide a brief summary of a literature that attempts to disentangle *complexity* from *randomness* and demonstrate its germaneness to The Condorcet Jury Theorem. In the fourth part, we show how within the interpretive framework, overcoming complexity requires diversity and sophistication. We also show that randomness presents an insurmountable obstacle. We write a set of inequalities that bound the accuracy of juries. In the fifth part, we contrast the main results of the generated framework with those of the interpretive framework. We conclude by discussing the potential for the interpretive framework to enable richer models of deliberation and more types of strategic behavior.

The Condorcet Model

The Condorcet model makes three assumptions. There exists an *outcome* that is either *true* or *false*. Each individual receives a signal that is correct, that is it agrees with the outcome, with probability p. And the probability that any individual is correct is independent of the probability that any other individual is correct. These signals are *generated* by some process. The model makes no explicit assumptions about how or why signals correlate with the true outcome. Extensions of the model allow for differences in accuracies and correlations between the predictions (Ladha 1992 Grofman et al 1983).

A1: Binary Outcomes: *The outcome is either* **true (T)** *or* **false (F)**

A2: Generated Signals *Each individual's signal is correct with probability* $p > \frac{1}{2}$

A3: Independence: *The probability an individual receives the correct signal is independent of any other individual's signal.*

The second assumption implies no limit to the number of independent signals. In the interpretive framework, that assumption cannot hold.

Until the last section where we discuss strategic voting and deliberation, we limit attention to *majority rule* and assume sincere voting. We also assume an odd number of jury members to avoid ties. A Condorcet Jury can be defined by two variables: N, an odd integer that represents the size of the jury, and p, the probability that an individual makes a correct classification. Four results follow from the assumptions.[1]

Claim 1. Condorcet Jury Theorem: *For all $N > 1$, the jury classifies correctly (is accurate) strictly more often than any individual.*

Claim 2. Monotonicity in Accuracy: *Jury accuracy monotonically increases in individual accuracy (p) if and only if $p \in (0.5, 1)$.*

Claim 3. Monotonicity in Size: *Jury accuracy monotonically increases in the jury size (N).*

Claim 4. Asymptotic Accuracy : *Jury Accuracy approaches but does not reach one as the jury becomes large ($N \to \infty$).*

These four results are considered non controversial. Yet, as we show later in the paper, none of them necessarily hold in the interpretive framework. They can hold, but they need not. Their veracity hinges on the assumptions made about the complexity of the classification problem, the diversity and sophistication of the individuals, and the level of uncertainty.

[1]We refer the reader to Ladha (1992) and Grofman, et al 1983) for proofs.

Interpreted Signals

In the interpreted signal framework, individuals classify a finite set of **states of the world**, Ω that each occur with equal probability. Individuals use models to approximate an **outcome function** that classifies each state as either true or false. $F : \Omega \rightarrow \{T, F\}$. The outcome function induces the *outcome partition*, a partitioning of the states of the world into *true states* and *false states*. Initially, we assume a deterministic outcome function.

Figure 1 shows an example of a state space and an outcome function. It represents each state of the world as small squares. Grey squares correspond to true states and white squares to false states.

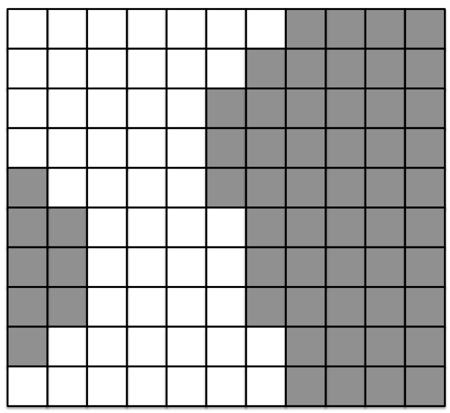

Figure 1: An Outcome Function Defined Over the States of the World

An interpreted signal is a classification of the states of the world. To construct such a signal, an individual first maps the states of the world into an **internal language** using an **interpretation**. Interpretations are *partitions* over the states of the world: $\{X_1, ...X_m\}$ such that any two X_i do not intersect and the union of the X_i equal Ω.

An individual then classifies each set in her partition as either true of false using a *predictive model*. Figure 2 shows an interpretation and a predictive model. The interpretation consists of two large rectangles. The predictive model classifies those states of the world in the rectangle on the right as true and the states of the world in the rectangle on the left as false.

Figure 2: An Interpretation and a Predictive Model

The **interpreted signal** equals the mapping of states to outcomes through the interpretation and the predictive model. The accuracy of an interpretive model equals the proportion of states classified correctly. In our example, the interpreted signal correctly classifies states of the world colored grey and white and incorrectly classifies those states colored black shown in Figure 3.

Interpreted signals make errors for one of two reasons. First, errors can result from a lack of alignment between the interpretation and the outcome partition. An interpreted signal assigns the same classification to the elements of a set. It follows that if a set in the interpretation overlaps both sets of the outcome partition, such as shown in figure 3, then errors must occur. We will refer to an outcome function as *classifiable* given an interpretation if the sets in the interpretation *refine* the outcome partition, that is if there exists no overlap.

Overlap could occur because of the complexity of the function, because the interpretation coarsens improperly, from confusion, or through conflation of states of the world. On complex outcome functions, individuals may rationally choose to stop refining their interpretations.

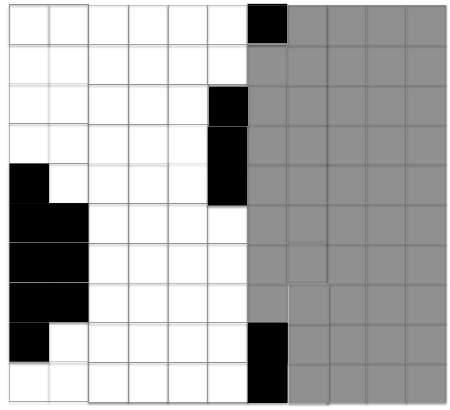

Figure 3: Interpretive Signal Accuracy

To quote Al-Najjar, Casadesus-Massanell, and Ozdenoren (2003) "*The agent believes these variations have no further useful structure; in short, to him they look random.*[2]

To denote interpretive error, for each set X_i in an interpretation, we let $C(X_i)$ equal the *number* of the less likely type of state in set X_i. We refer to these as *interpretive errors.* [3]. If a set has ten true states and three false states, the interpretive error is of size three.

Second, errors can occur if the predictive model incorrectly classifies a set. We refer to these as *classification errors.* If there exist no interpretive errors, this would mean that the predictive model classifies a set consisting only of one type of state as being the other type of state. If interpretive errors exist, classification errors correspond to assigning the less likely type to a set. We will refer to the *best possible* predictive model as one that classifies correctly on every set in the interpretation.[4] Formally, we let $\delta(X_i)$ equal one if the set X_i is classified correctly and zero otherwise. Given these definitions we can state the following claim.

Claim 5. *The accuracy of an interpretive signal equals the sum of the proportion of states*

[2]They refer to this as *complexity.* We take up complexity in the next section.

[3]To avoid notational complication, we assume throughout that within any set in an interpretation a strict majority of states have one of the two outcomes.

[4]Hong and Page (2009) refer to these as *experience generated* interpreted signals.

that are interpreted correctly belonging to sets classified correctly and the proportion of states interpreted incorrectly belonging to states classified incorrectly.[5]

The second part of the claim states that if an individual makes both an incorrect interpretation and an incorrect classification of that set, she makes a correct prediction. This two wrongs make a right property holds because classifications are binary. In the more general case, two wrongs could make a right but they need not.

We have described interpretations in the abstract. In an actual situation, the states of the world may have one or more natural representations in terms of attributes or features. We can use feature spaces to describe three types of interpretations: *threshold functions*, *dimensional reductions*, and *exemplars*.

A *threshold function interpretation* assigns a value to each state of the world by assigning weights to features and classifies those states with values above a threshold as true and all others as false. Recall figure 2. It shows a linear threshold based interpretation. Linear threshold models attach a weight to each feature. Two threshold based interpretations differ if they assign different weights given the same feature space, if they include different features, or if they use distinct functional forms.

A *dimensional reduction interpretation* considers only a subset of features. Consider a left leaning political activist attempting to classify new acquaintances as either friends or not friends. Each new acquaintance has a variety of features – identity characteristics, a past history, a personality type, and ideological leanings. These represent the feature space. The political activist's internal language might consider a single binary feature: whether or not the person is a *Wobbly* or a *Non-Wobbly*. The activist's *predictive model* could then classify all Wobblies as *friends* and all Non-Wobblies as *not friends*. These would be her *interpretive signals*. The accuracy of those signals depends on how closely her predictive model aligns

[5]The formal can be written as follows:

$$p = \frac{1}{\mid \Omega \mid} \sum_{i=1}^{m} \delta(X_i)(\mid X_i - C(X_i) \mid) + (1 - \delta(X_i)) \mid C(X_i) \mid$$

with the outcome function.

To determine the accuracy of a group using majority rule, we compute the probability that a majority makes the correct classification averaged over all states of the world. Example #1 describes the states of the world using a feature space consisting of three binary attributes (*good* or *bad*.) A state of the world will be true if and only if a majority of the attributes are good. Each of three individuals considers a different single attribute to construct an interpreted signal. An individual classifies a state of the world as true if the attribute she considers has a good value and as bad otherwise.

Example #1 Separable Binary Strings Length Three

States of the world: $\Omega = \{(Z_1, Z_2, Z_3) : Z_i \in \{0,1\}\}$ (0=bad, 1=good)

Outcome function: True if a majority of the attributes have value one and False otherwise.

Interpretive Signals: Individual i interprets the state by the value of attribute i. Individual 1's interpretation consists of two sets:
$0\#\# = \{000, 001, 010, 011\}$ and $1\#\# = \{100, 101, 110, 111\}$.
An individual classifies a set as True if and only if her attribute has value one. Each individual has accuracy $p = 0.75$.

Group: Correct with probability one.

In the example, the group is more accurate than each individual as in The Condorcet Jury Theorem. However, the three individuals classify accurately with probability one, so the other three main results – *Monotonicity in Accuracy, Monotonicity in Size* and *Asymptotic Accuracy* do not hold.

Last, we consider *exemplar based interpretations* which reason from past cases (Gilboa and Schmeidler 1995). Interpretations are based on past experiences, so they can be path dependent. Given a state of the world, an individual might choose the most similar past state and assume the two outcomes are the same. Or an individual might only assign the outcome true if the state of the world is near a past true case.

In Example #2, we show a jury whose members use *exemplars.* the states of the world consists of six sandwiches. Sandwiches can be vegetarian or turkey and served on either multigrain, white, or rye bread. The outcome function corresponds to a boss's preferences. The boss likes any vegetarian sandwich but does not like any turkey sandwich. Each of the boss's three employees has seen the boss eating a different type of vegetarian sandwich. One knows that she likes veggie on multi-grain. One knows she likes veggie on rye. And one knows she likes veggie on white.

Each employee has a predictive model that partitions the possible sandwiches into two sets. The first consists of the single type of sandwich he has seen the boss eating. Each employee predicts (correctly) that the boss will like that sandwich. He interprets all other sandwiches as "unknown" and predicts that his boss will not like those. Given these interpretations, each employee classifies correctly in four of six cases. However, if the three employees vote, they classify correctly only one half of the time.

Example #2 Exemplar Based Interpretations

States of the world: $\Omega = \{VM, VR, VW, TM, TR, TW\}$

V= vegetarian, T = turkey, M= multigrain, W=white, and R=rye.

Outcome function: good if vegetarian, bad if turkey.

Interpretive Signals:

Individual 1: VM = good, all others bad. Correct with probability $\frac{4}{6}$.

Individual 2: VW = good, all others bad. Correct with probability $\frac{4}{6}$.

Individual 3: VR = good, all others bad. Correct with probability $\frac{4}{6}$.

Group: Classifies all outcomes as bad. Correct with probability $\frac{1}{2}$.

In Example #2, The Condorcet Jury Theorem does not hold. The group classifies correctly with a lower probability than each individual. This occurs because any state of the world classified differently by two interpretative signals is classified *incorrectly* by the third interpretive signal.

These three classes of interpreted signals are by no means exhaustive. In fact, *any partition* of the set of the states of the world can be used to produce an interpreted signals. That said, the number of interpreted signals that exist in a given situation could be small. Diverse interpreted signals require diverse interpretations, diverse predictive models, or both. An assumption of distinct, low correlation signals requires underlying heterogeneity in models, features spaces, or experiences. Large numbers of such signals may not be a plausible assumption in most settings. Instead, it may be more reasonable to assume a set of possible models as in Economo et al (2015). Generated signals which, like rolls of a die or spins of a wheel, are often treated as though they can be produced in any number desired.

Independently Correct and Independently Interpreted

Our two examples demonstrate that none of the four Condorcet Jury Theorem results need hold for interpreted signals. In Example #1, the interpretive signals were constructed so that each interpretive signal was correct when *pivotal*: when the other two signals disagreed. In Example #2, the signals were constructed so that each was correct only when it was not pivotal. Thus, in neither example were the signals *independently correct*. Had the signals been *independently correct* then the standard results would hold.

Independence of the correctness of two interpretive signals requires strong assumptions. First, the signals must have *correlated interpretations*. Figure 4 shows two independently correct interpreted signals. The two signals partition the states of the world into four sets based on how they are classified. Signal S_1 classifies the top two boxes as true. Signal S_2 classifies the two boxes on the left as true.

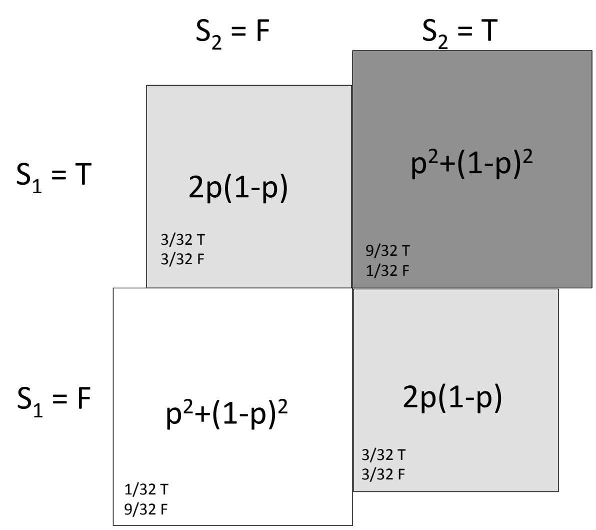

Figure 4: Independently Correct Interpreted Signals

Assuming true and false outcomes are equally likely and that each signal is correct with probability p, it follows that the two signals must both classify as true a proportion $p^2 +$

$(1-p)^2 > \frac{1}{4}$ of the states of the world. In the case of $p = \frac{3}{4}$, the two signals must make the same classification with probability $\frac{10}{16}$ and opposite classifications with probability $\frac{6}{16}$. This shows why independently correct interpreted signals are positively correlated in how they partition states of the world.

A corollary of that result is that interpreted signals that rely on independent partitions must be negatively correlated in their correctness (Hong and Page 2009). The group can then be more accurate than would occur with generated signals (recall Example #1).

Figure 4 shows that it is possible for interpreted signals to be independently correct. However, Hong and Page (2009) show that this can only occur given a unique set of partitionings (up to isomorphism or renaming). These partitions require creating a feature space and creating a set of interpretations that omits a single and distinct feature. Such interpretations could exist, but they would seem an improbable coincidence and are difficult to justify strategically. Furthermore, the partitions depend on p. Different p's require different partitions.

In sum, with interpreted signals, independent correctness appears a dubious assumption. Correlations between signals depend on features of the environment and the individuals. The modeler is therefore obliged to investigate the motivations, expertise, and incentives of the individuals. For example, independent interpretations seem unlikely to arise by chance though they could arise in strategic settings. They would be the equilibrium of a game in which there exists a common internal language and each group member can consider only a single dimension due to cognitive constraints.

Prudence suggests that how we model interpretations should depend on the context. For any number or reasons, we should not expect a single assumption to capture a jury making a decision about the guilt or innocence of a defendant, venture capitalists deciding whether to make an investment, college professors voting on tenure, or a marketing team deciding among two advertising strategies.

Let's compare a jury with a marketing team. The jury members have common information which may limit the set of plausible internal languages. The jury members also may lack expertise and may rely on predictive models with only a few variables. Initially, they are precluded from talking with one another, so we might expect them to derive diverse predictive models. In contrast, the members of a marketing team may have vast experience and use sophisticated predictive models. However, they may also have worked together for a long time. Their models may lack diversity. To first approximation then, we might expect a jury to have diverse, simple models and a marketing team to have sophisticated similar models.

In any given situation the correlation in signals depends on context. An assumption of independent interpretations only makes sense if individuals are restricted to single feature interpretations and have the ability and incentive to coordinate or if each individual has a single piece of private, uncorrelated piece of information. Independently correct interpreted signals (the standard assumption for generated signals) seems unrealistic.

Group Accuracy

In the interpreted signal framework given any group of individuals and any voting rule, we can construct a *collective interpreted signal*. The accuracy of the group equals the accuracy of that signal. Constructing the signal requires two steps. First, we define their *collective interpretation*, the partition of the states of the world created by taking the intersections of the individual interpretations.

Consider three individuals denoted by A, B, and C, whose interpretations each partition the states of the world into two sets Denote these sets as A_T, A_F, B_T, B_F, C_T and C_F to differentiate the sets classified as true and false. The collective interpretation consists of the eight sets created by intersected one set from A's interpretation with one set from B's interpretation and one set from C's interpretation as shown in figure 5. Using the language of set theory, the collective interpretation equals the coarsest common refinement of the individual interpretations.

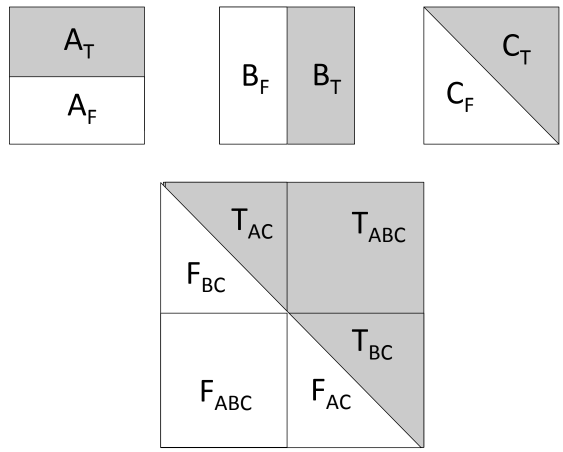

Figure 5: A Collective Interpretative Signal

The second step defines the *majority predictive model* to be the classification produced by majority rule. This is shown at the bottom of figure 5. We can identify sets in collective partition by their classification with subscripts denoting which individuals make the majority classification. Thus, the set F_{AB} denotes the region classified as false by A and B, and the region T_{ABC} denotes the region classified as true by all three individuals.

As can be seen in the figure, whenever A and B classify a state of the world identically so too does C It follows that the majority never disagree with C. Even though the collective interpretation refines the individual interpretations, the majority predictive model does not take advantage of that refinement. It follows that the majority predictive model will not be *best possible* for some outcome functions. Consider first the outcome function shown in figure 6. True outcomes are shaded black and false outcomes are shaded white.

The collective interpretation refines the outcome partition, i.e. the collective interpretations contains a set corresponding to each level set of the outcome function. If the majority rule predictive model did not make any classification errors, the group could be perfectly accurate.

However, as already shown, the majority rule predictive model is the same as that of

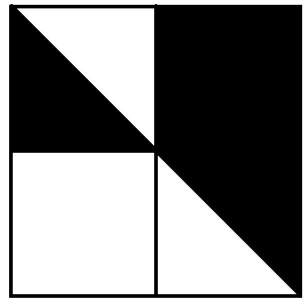

Figure 6: A Classifiable Outcome Function for A, B, and C

individual C, and individual C's predictive model fails to classify the sets F_{BC} and T_{AC} correctly. It follows that the majority rule predictive model is correct with probability three-fourths, which equals the accuracy of each individual's predictive model. Failure stems not from the collective interpretations not being able to classify the outcome function but from the inadequacies of majority rule. Even though the sets F_{BC} or T_{AC} belong to the collective interpretation, each is lumped within a larger set in the individual interpretations.

This same phenomenon occurs in Example #2 where the collective was less accurate than its members. The collective interpretation consists of four sets. Three singleton sets each containing one vegetarian sandwich and one set containing all three turkey sandwiches. The best possible predictive model would classify the first three sets as good and the fourth set as bad and would be perfectly accurate.

Not all outcome functions will be classifiable by the collective interpretation as shown in figure 7. The black sets in the upper lefts and lower right do not belong to the collective interpretation. On this function, the best possible predictive model predicts correctly with probability three fourths, so too does the majority rule predictive model. The failure of The Condorcet Jury Theorem can be ascribed to the collective interpretation and not to the shortcomings of majority rule.

With interpreted signals collective accuracy depends on the group accurately reconstruct the outcome function. The outcome function must be classifiable (no interpretive error) and

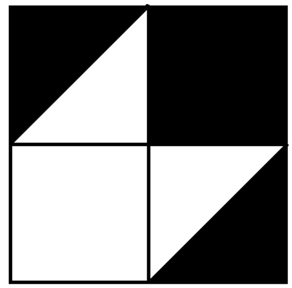

Figure 7: A NonClassifiable Outcome Function for A, B, and C

the majority rule predictive model must not make mistakes (no classification error). With generated signals accuracy arises through random errors canceling through the logic of the Law of Large Numbers. That is a statistical logic. Interpreted signals rely on a logic of function approximation.

Disagreement and Collective Accuracy

Within the interpreted signal framework it is possible to construct bounds on group accuracy. Those bounds depend on the **disagreement**, $D(i, j)$ of interpretations which equals the proportion of states of the world classified differently by individuals i and j. We make the following two assumptions that approximate the standard assumptions for generated signals.

A4: Equal Outcome Probability: *An equal number of states of the world are mapped to each outcome.*

A2': Equal Accuracy: *Each interpreted signal correctly classifies $p > \frac{1}{2}$ of the states of the world that are mapped to each outcome.*

We define the *accuracy gain* to be how much more accurate a group can be than each individual and the *accuracy loss* to denote how much less accurate a group can be. We can state the following claim:

Claim 6. *Given A1,A2', and A4, the minimal disagreement between any two individuals* *bounds the accuracy gain and accuracy loss for a group of size three. Exact bounds increase and then decrease in minimal disagreement with maximal accuracy loss occuring at a lower level of disagreement than maximal accuracy gain.*

pf: see appendix.

Figure 8 shows maximal and minimal accuracy for a group of three individuals who each classify correctly with probability sixty percent.

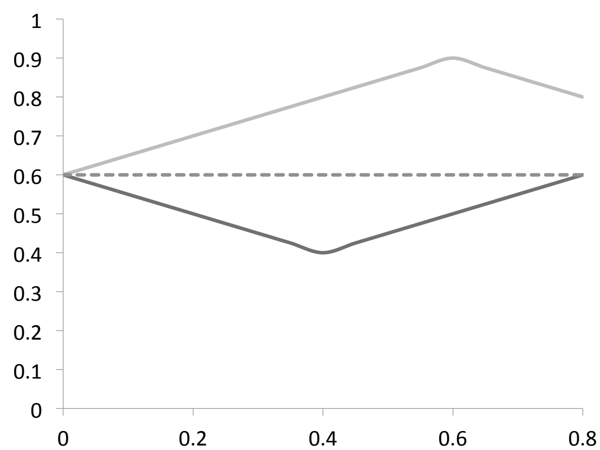

Figure 8: Minimal and Maximal Accuracy As a Function of the Disagreement ($p = 0.6$)

The claim reveals that accuracy losses and gains require differences in interpretations. How much two interpretations differ can vary with the type of interpretations. Figure 9 shows three sets of two interpretations with their disagreement regions shaded black. The two interpretations on the left rely on linear thresholds and have almost no disagreement. The middle interpretations rely on dimensional reductions. We have exaggerated the disagreement regions to capture a situation where relatively more states of the world lie in the disagreement region. The interpretations on the far right rely on exemplars and have moderate disagreement.

Imagine adding a third individual to each of these three groups of size two. Given the previous claim, for the threshold based interpretations on the left, the group could not be

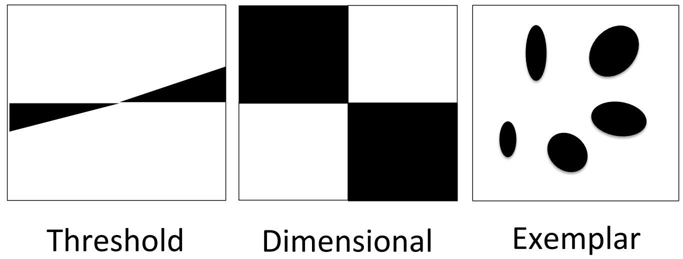

Figure 9: Differences in Disagreement for Three Interpretation Types

much more accurate than either of the two individuals. Even if the third individual classified correctly for the entire disagreement region, the accuracy of the group would exceed the accuracy of each individual by at most a small margin.

The dimensional reduction interpretations shown in the middle of figure 9 include more than half of the states of the world in the disagreement region. This large disagreement provides an opportunity for the third individual to increase collective accuracy. Furthermore, the large disagreement region implies a small maximal accuracy loss. A large disagreement region does not guarantee an accuracy gain but it precludes much accuracy loss.

Moderate levels of disagreement such as those in the right panel create the potential for large gains or losses in accuracy. Figure 10 gives an example of maximal accuracy loss for three individuals who each predict correctly with probability three fifths. In the example, disagreement between each pair of interpretations equals forty percent (the low point in figure 8). Group accuracy equals only two-fifths.

If the majority rule predictive model incorrectly classifies all states of the world in the disagreement region, then no states can be interpreted incorrectly. The collective interpretation must incorrectly classify the outcome function in the disagreement region. Thus, we

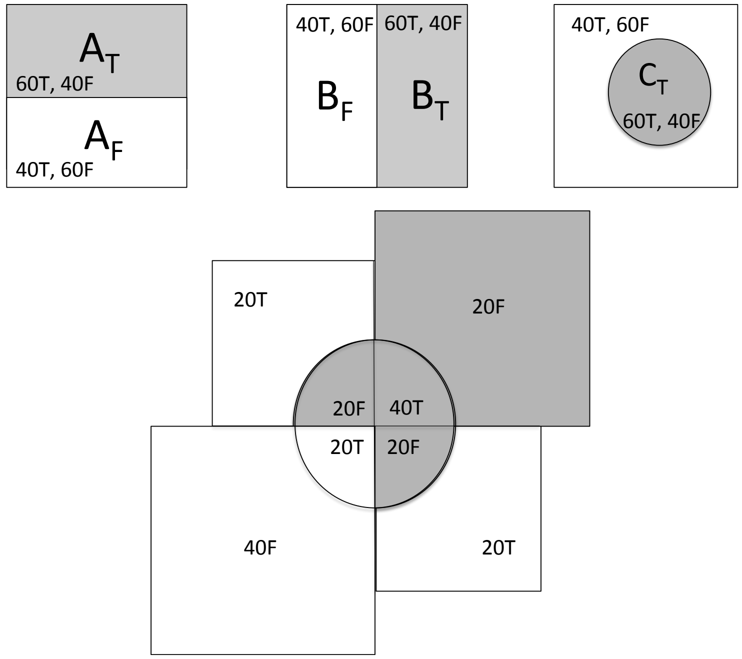

Figure 10: Maximal Jury Failure: p=0.6, d=0.2, Group Accuracy =0.4

can state the following.

Corollary 1. *Maximal accuracy loss occurs when majority rule makes classification errors and the collective interpretation refines the outcome partition.*

Moderate levels of disagreement increase the lack of predictability in small group classifications. Both large accuracy gain and large accuracy loss are possible. Ensemble methods produce accuracy gain by creating a diverse set of classifiers that rely on distinct subsets of features (Brown, et al 2005). They also advantage classifies that make negatively correlated predictions that are also accurate (Liau and Yoab 1999). By definition, a state of the world in the disagreement region is predicted incorrectly by a classifier. Preferring classifications that are negatively correlated in their accuracy leads to greater accuracy in the disagreement region.

Given this logic, the most accurate three person groups have relatively large numbers of two to one votes. The following corollary reinforces that insight.

Corollary 2. *For each p, the maximally accurate group of size three has a maximal number of two to one votes.*

The corollary implies a rather complicated relationship between unanimity and collective accuracy. Groups consisting of individuals with similar interpretations and predictive models often vote the same way. That means the group cannot be very accurate (unless the problem is simple). Conditional on unanimity, one must infer an increased likelihood of similar interpreted signals. That lowers confidence in the group. On the other hand, a two to one vote means that two people have disagreed. That means the third person is pivotal on a state of the world for which the outcome was relatively hard to classify. If it were not, the other two people would not have disagreed.

The inference of unanimity becomes much less complicated if the group makes multiple classifications. Consistent unanimity implies either a lack of diversity or simple problems. In many cases, the outcome of a vote rendering any determination of accuracy is impossible. We don't know if a defendant really was guilty or innocent or whether a policy would have been effective. In those contexts, unanimity should be viewed with suspicion.

Complexity and Randomness

Previously, we have described how interpreted signals can fail to make accurate predictions because of interpretive error or classification error (but not both – owing to the two wrongs make a right property). Either type of error could arise because of the complexity of the outcome function. Complex outcome functions include non linearities and interaction terms between attributes. These result in more distinct regions of true and false outcomes. If the mapping to states of the world to outcomes were not complex, we would expect people to classify accurately. We now introduce the possibility of randomness.

We now relax the assumption that the outcome function must be deterministic. Probabilistic outcome functions could exist for many reasons. One important case involves classification of future events such as the outcome of military maneuvers, elections, and policy choices. To include randomness in the formal model, we redefine the range of the outcome function to be the probability of a true outcome: $F : \Omega \rightarrow [0, 1]$, where $F(\omega)$ equals the probability of a true outcome.

As an example, consider a Polya Process based on an urn containing red and blue balls. In each period, a ball is drawn from the urn and placed back in the urn along with a new ball of the same color. To map this into the interpreted signal framework, let the state of the world equals the initial number of red and blue balls. Let the outcome function correspond to whether the urn contains a majority of red balls or blue balls ten periods into the future. The outcome function will be path dependent (Page 2006). Given a state of the world, the outcome will be a probability distribution. If the current state of the world consists of five red balls and two blue balls, a majority of red balls will be more likely in period ten, but with some probability a majority of blue balls could occur.

Probabilistic outcome functions also arise when multiple initial conditions could have produced the state of the world. In a jury trial, the state of the world could correspond to the available evidence. Multiple past histories could be consistent with that evidence. In some histories, the defendant would be guilty in others the defendant would be innocent. The outcome function would correspond to the likelihood of each type of history. In this case, the uncertainty would never be resolved.

The straightforward distinction between complexity and randomness masks an inferential problem. In constructing an interpretive signal from data an individual must disentangle complexity and randomness. That unpacking has been a central challenge in the development of measures of complexity (Feldman and Crutchfield (1998), Crutchfield and Shalizi (1999)).

Consider a simple outcome function defined over ten real valued features in which any state of the world with an above average first feature is true. That function would be classifiable by an interpretation that only took into account the first feature. Suppose that we introduce randomness in the outcome function and that with a ten percent probability, the outcome switches from true to false or false to true.

Individuals trying to learn the outcome function by creating interpretations and predictive models might infer that the outcome depends in a complicated way on the ten features. Only after seeing a large number of outcomes would the randomness reveal itself. Formal complexity solve this problem by measuring complexity as the information content that remains after all uncertainty has been removed.[6]

The complexity measures literature demonstrates that with sufficient data a model can distinguish randomness from complexity within a time series. The same logic applies to classifications across multiple cases. It follows that complexity and randomness differ in their effects on group accuracy. Complexity produces interpretive and classification errors. However, it need not result in group error if the group has a sufficient and proper diversity of interpreted signals. Randomness though cannot be overcome. No amount of individual sophistication or collective diversity can classify a random event.

In making that observation, we also reveal three "two wrongs make a right" effects. If errors can come from interpretations, classifications, or randomness, then the group will be correct if it makes no errors or exactly two errors. We begin with an example that shows that randomness cannot be overcome.

[6] See Prokopenko et al (2009), and Page (2008) for surveys.

Example #3 Separable Binary Strings With Randomness

$\Omega = \{0,1\}^3$. $x_i = 0$ corresponds to a blue ball being placed in the urn and $x_i = 1$ corresponds to a red ball being placed in the urn. Each are equally likely.

The outcome function is a random draw of a ball from the urn (red = true, blue = false)

Each of three individuals interprets the state by the value of a distinct attribute. Individual 1 maps the state 011 into the set 0## = $\{000, 001, 010, 011\}$. Individual i predicts true if and only if her attribute has value one.

Individual accuracy equals $\frac{2}{3} = \frac{1}{4} + \frac{1}{2}\frac{2}{3} + \frac{1}{4}\frac{1}{3}$.

Group accuracy equals $\frac{3}{4} = \frac{1}{4} + \frac{3}{4}\frac{2}{3}$

In Example #3, the group classifies correctly with probability three-fourths. Increases in the size of the group cannot improve accuracy because of the randomness in the outcomes.

Example #4 NonSeparable Binary Strings Length Three

$\Omega = \{0, 1\}^3$

The outcome function equals true if either two or zero of the attributes have value one and equals false otherwise.

Individual i interprets the state by the value of attribute i. Individual 1 maps the state 011 into the set $0\#\# = \{000, 001, 010, 011\}$.

An individual predicts true if and only if her attribute has value one.

Each individual has accuracy one-half.

Group accuracy equals three-fourths. It incorrectly classifies the states equals 000 and 111.

This example violates the second main result from the generated signal Condorcet Jury Theorem. Each individual voter is correct with probability one-half, yet the jury is correct with probability three fourths. This example could be overcome by introducing two voters who vote no on states 000 and 111 and yes on all other states and two voters who vote no on all states. This shows how with sufficient diversity of interpreted signals, the group can be accurate.

To prove a general result, we define the *state randomness* for $\omega \in \Omega$, r_ω, to equal the probability of the less likely outcome in state ω. The best possible classification will have accuracy $(1 - r_\omega)$ for state ω. We define the *total randomess* to be the average state randomness.

$$R = \frac{1}{\mid \Omega \mid} \sum_{\omega \in \Omega} r_\omega$$

Given this definition, we can state the following claim

Claim 7. *Expected accuracy cannot exceed one minus R,*

The accuracy of an interpreted signal can then be written as a function of *random error* as well as the probability of *interpretive error* and *classification error.* A state of the world is classified correctly if and only either no errors are made or exactly two errors are made. If exactly two errors are made the classification is correct because two wrongs make a right. We will assume equal state randomness for all states in the same set in an interpretation, which we denote by $r(X_i)$. Let $N(X_i)$ equal the number of states of the world in set X_i.We can now state the following claim.

Claim 8. *An interpreted signal classifies a state correctly if and only if no errors occur, i.e. the state is interpreted correctly, classified correctly, and no random error takes place, or if exactly two types of errors occur.* [7]

Let $\bar{C}$ be proportion of states interpreted incorrectly. Assume that the best possible predictive model. The following corollary can be shown.

Corollary 3. *The accuracy of the best possible predictive model equals*

$$\frac{1}{\mid \Omega \mid} \sum_{X_i} N(X_i) \left[(1 - r(X_i))(1 - C(X_i))\right] \; \geq \; (1 - R)(1 - \bar{C})$$

This corollary reveals a subtle benefit from the entanglement of complexity and randomness. If a set in the interpretation has a high degree of randomness, then classification error may be more likely. *The costs of classification error decrease with randomness.*

[7]This can be written formally as follows:

$$\frac{1}{\mid \Omega \mid} \sum_{X_i} N(X_i) \cdot [\delta(X_i)[(1 - r(X_i))(1 - C(X_i)) + r(X_i)C(X_i)] + (1 - \delta(X_i))[r(X_i)(1 - C(X_i)) + (1 - r(X_i))C(X_i)]]$$

Which reduces to the following:

$$\frac{1}{\mid \Omega \mid} \sum_{X_i} N(X_i) \cdot [\delta(X_i)[(1 - 2r(X_i))(1 - 2C(X_i))] + [r(X_i) + C(X_i) - 2r(X_i)C(X_i)]]$$

Complexity and Group Size

The generated signal version of the Condorcet Jury Theorem implies that groups outperform individuals, that more accurate individuals produce more accurate groups, that larger groups outperform smaller groups, and that in the limit of large gropus accuracy approaches one. We have seen how in the interpreted signal framework the first and second results can, but need not hold, and that the fourth result would only hold in rare cases.

We now take up the third result: accuracy in group size. The accuracy of a group depends on how closely its collective interpretation and the majority rule predictive model track the outcome function. To gain intuition on the role of group size, we assume a common feature space for the outcome function and the indivdiuals' interpretations.

The outcome partition separates the feature space into sets as does each individual's interpretation. For complex outcome functions, the outcome partition consists of many small sets as represented in the feature space. If we assume that each individual has limited sophistication, then each individual's interpretation will have fewer sets than the outcome partition.

For convenience, we can think of each individual's interpretation as a *coarsening* of the outcome partition. Formally, this means that the outcome function must be a strict subset of each individual's interpretation. Recall that was true of the outcome function in figure 6 but not of the outcome function in figure 7

As the group increases in size, the collective interpretation becomes finer. If the group has appropriate diversity in its interpretations the collective interpretation will *refine* the outcome partition. The outcome function will be classifiable. At this point, adding more individual to the group could help in terms of making the majority rule predictive model more accurate but it won't help in improving the collective interpretation.

Under the assumption that interpretations coarsen the outcome partition, each outcome function has a complexity expressed in terms of the number of interpretations needed to classify the outcome function. That number (provided that the majority rule predictive model classifies correctly) would be the optimal group size. There would be no advantage to increasing the group further.

If the majority rule predictive model classifies some sets in the outcome partition incorrectly, then individuals could be added to the group who predict correct on those sets until the majority rule predictive model is accurate. At that point, there would be no advantage to increasing group size.

Put more formally, there exists a *size* to each outcome function given as *minimal set of interpreted signals* that satisfies two properties (i) their collective interpretation refines the outcome partition and (ii) their majority rule predictive model correctly classifies the outcome function. At the level of metaphor, this minimal set of interpreted signals plays the same role as Crutchfield and Shalizi's (1999) *minimal causal states*: a set of states that can produce the pattern in a time series.

In sum, *there will exist an optimal group of finite size*. Bigger need not be better.

This optimal group, the minimal set of interpreted signals, need not naturally exist, arise through incentives, or be constructed by an astute manager or central planner. Instead, a group of a given size may be a draw of interpreted signals from a distribution as in Economo et al (2015). Imagine a set of K interpreted signals and let $p(k)$ be the probability that an individual uses interpreted signal k. Whether larger groups do better will depend on that distribution. Ironically, now a law of large number logic comes into play. Large groups will be representative draws from the distribution. Small groups will be biased draws. It follows that small groups will be better, on average, if one or two types of signals are far more likely than others.

Discussion

The Condorcet Jury Theorem and its extensions state that groups are more accurate than their members, group accuracy increases in individual accuracy and group size, and in the limit large groups come close to, but do not achieve perfect accuracy. Those results hold within the model, but they may not hold generally.

In this paper, we have applied an alternative framework based on interpretive signals and found that group accuracy depends on the complexity of the outcome function, the amount of randomness, the dimensionality of the voters' interpretations, the sophistication of the voters' predictive models, and the diversity of the voters – the distinctness of their interpretations and models. Borrowing from the literature on complexity measures, we find that groups can overcome complexity – but only if they have sufficient sophistication and diversity – but that they can never predict random outcomes.

Among our more interesting findings concern the double edge sword of interpretive diversity. If people interpret the world differently (in the model that means different partitions), they create the possibility for gains in group accuracy, but diversity is *necessary but not sufficient*. The predictive models must be such that they're more accurate when others disagree for the group to realize those gains. Groups that exhibit accuracy loss when they classify incorrectly in regions of disagreement.

By including a set of states of the world and constructing partitions, the interpreted framework reveals the shortcomings of majority rule. The intersection of a group members' partitions admits any number of predictive models. Majority rule is just one such rule. This opens up the possibility of much richer models of deliberations as discussed in Landemore and Page (2015).

Deliberation with generated signals consists of Bayesian updating. A juror might reason, "I received a signal innocent but nine other received the signal guilty. That outcome is far more likely if the defendant was guilty than innocent, so I should change my vote. " With interpretive signals, deliberation could cause a person to abandon her interpretation and predictive model, to refine her interpretation or to construct a new predictive model. These forms of deliberation could lead to better results but they need not lead to consensus (Landemore and Page 2015).

The interpretive signals framework predicts consensus when there is a lack of juror diversity or if the classification problem is simple and low dimensional. Jurors found Oliver North guilty on the just three of sixteen charges – accepting a gratuity, obstructing a congressional inquiry, and ordering Fawn Hall, his assistant to destroy documents. These were the least complex of the charges. Thus, they were more likely to produce unanimity.

We conclude with an appeal for greater engagement with the interpreted signal framework. The generated signal framework offers tractability and clean, general results. To the extent it takes into account the complexity of the task, it does so through the probability of a correct vote. The Condorcet Jury Theorem implies that any amount of complexity can be overcome by enlarging the group size. That's not a justifiable claim in light of the past twenty years of research on complexity (Crutchfield and Shalizi 1999).

To determine the accuracy of a group of individual each of whom predicts correctly with probability p, we need to first identify the reason or reasons for their errors. The domain could be high dimensional. The outcome function could be complex. The environment could include randomness. High dimensional domains can be handled by larger groups but only if the groups consider different dimensions in their models. That's a core lesson from the random forest literature. More can only be better if it is different. Complexity can partly be overcome through group size, but the interpretations would have to be finer partitions. They must be capable of capturing interaction. Randomness, of course, can never be overcome.

In each case, the interpreted signals must be diverse. They must rely on different categorizations. These could rely on different structural models, include different variables, or be based on unique experiences. That diversity, though necessary, is not however sufficient. The diversity must be of the right form. Each model must somehow be more likely to be correct when the other models are divided (Fedderson and Pesendorfer 1996, Austen-Smith and Banks 1996). Machine learning methods can produce such interpreted signals. The relevant question for social science is how do societies do the same?

Proof of Claim 2: Let $d = \frac{D(i,j)}{2}$. The probability that a group of size three correctly classifies, $P(p, d)$ lies in the following intervals

$$\begin{array}{ll} [p + d - (1-p), (p-d) + p] & \text{if } p < 2d \\ [p + d - (1-p), p + d] & \text{if } (1-p) \le 2d \le p \\ [p - d, p + d] & \text{if } 2d < (1-p) \end{array}$$

By A2' and A4, each interpretation is equally likely to classify states of the world as true or false. We consider only those states of the world mapped to the outcome true. We refer to these as true states of the world. Given two interpreted signals s_i and s_j, let $\{T_i, T_j\}$ denote the true states of the world classified as true by i and j, let $\{F_i, F_j\}$ denote the true states of the world classified as false by both i and j, and let $\{T_i, F_j\}$ and $\{F_i, T_j\}$ denote the true states in which the two disagree.

By A2', the size of the sets $\{T_i, F_j\}$ and $\{F_i, T_j\}$ must be equal. Let d equal the probability that a true state of the world belongs to one of these two sets. By definition, $D(i, j) = 2d$ and the probability that a true state of the world lies in the set $\{T_i, T_j\}$ equals $p - d$.

Next, let $\{X_{ki}\}$ equal the true states of the world in set $\{T_i, F_j\}$ that the third individual (denoted by k) classifies incorrectly. Define $\{X_{kj}\}$ similarly. We first ignore the constraint that individual k classifies correctly with probability p. The accuracy of the majority equals the probability that a true state of the world belongs to one of three sets $\{T_i, T_j\}$, $(\{T_i, F_j\} \setminus \{X_{ki}\})$ and $(\{F_i, T_j\} \setminus \{X_{kj}\})$. If $\{X_{ki}\} = \{X_{kj}\} = \emptyset$ group accuracy equals $p + d$, and if $\{T_i, F_j\} = \{X_{ki}\}$ and $\{F_i, T_j\} = \{X_{kj}\}$ group accuracy equals $p - d$.

To achieve group accuracy of $p+d$, individual k correctly classifies every true state in the sets $(\{T_i, F_j\}$and $(\{F_i, T_j\}$. The proportion of true states in the union of these sets cannot exceed p. Therefore, if $2d < p$, maximal accuracy equals $p + d$. Otherwise, it equals at most $p - d + p$.

By the same logic, the proportion of true states of the world in $\{X_{ki}\}$ and $\{X_{kj}\}$ cannot exceed $1 - p$. If $2d < (1 - p)$, group accuracy can equal $p - d$ because individual k can incorrectly classify every true outcome in the disagreement region. If $2d > (1-p)$, individual

k must correctly classify at least $2d-(1-p)$ of the true outcomes in the disagreement region so group accuracy cannot fall below $p-d+2d-(1-p)$ which equals $p+d-(1-p)$.

Proof of Corollary 2: Maximal group accuracy occurs at $2d=p$. Given p this is the largest possible disagreement region. From the proof of Claim 2, in the maximally accurate group, the third individual classifies correctly on the entire disagreement set. Therefore, the entire disagreement set consists of two to one votes. Given $p=2d$, the third individual must misclassify all other states of the world. Therefore, *every outcome* which is correct has two to one vote.

References

Al-Najjar N, R. Casadesus-Masanell R, and E. Ozdenoren (2003). "Probabilistic representation of complexity" *Journal of Economic Theory*Vol 111:1 p 49-8.

Apostol, T. M. (1969) *Calculus, Volume II (2nd edition).* John Wiley & Sons, New York, NY.

Austen-Smith, David and Jeffrey S. Banks (1996) "Information Aggregation, Rationality, and The Condorcet Jury Theorem" *The American Political Science Review* 90: 34-45.

Breiman, Leo (2001). "Random Forests". *Machine Learning* 45 (1): 5?32.

Brown, G., J. Wyatt, R. Harris, R. and X. Yao,(2005) "Diversity creation methods: a survey and categorisation." *Information Fusion*, 6(1), pp.5-20.

Crutchfield, J.P. and C. R. Shalizi, (1999) "Thermodynamic Depth of Causal States: Objective Complexity via Minimal Representations", *Physical Review E.* 59:275-283.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In J. Kittler and F. Roli (Ed.) *First International Workshop on Multiple Classifier Systems* pp. 1-15. New York: Springer Verlag

Economo, E, L. Hong and S.E. Page (2015) " Social Structure, Endogenous Diversity, and Collective Accuracy." *Journal of Economic Behavior and Organization.* forthcoming.

Feddersen, T. and W. Pesendorfer (1996) "The Swing Voters? Curse." *American Economic Review* 86: 408?24.

Feldman, D.P. and J. P. Crutchfield, (199) "Measures of Statistical Complexity: Why", *Physics Letters A.* 238: 244-252.

Gilboa, I. and D. Schmeidler. (1995). "Case-based decision theory", *The Quarterly Journal of Economics* , Vol. 110, pp. 605-639.

Grofman, Bernard, Guillermo Owen, and Scott L. Feld. (1983). "Thirteen Theorems in Search of the Truth." *Theory and Decision* Volume 15 (3): 261-278.

Hong, Lu, and Scott E Page. (2009) "Interpreted and Generated Signals" *Journal of Economic Theory* 144: 2174-2196.

Hong, Lu, Maria Riolo, and Scott E. Page (2012) "Incentives, Information, and Emergent Collective Accuracy" *Managerial and Decision Economics* 33:5 pp 323 - 334.

Ladha, Krishna (1992). "The Condorcet Jury Theorem, Free Speech, and Correlated Votes". *American Journal of Political Science*: 36 (3): 617?634.

Landemore, Helene and Scott E. Page (2015) "Deliberation and Disagreement: Problem solving, Prediction, and Positive Dissensus." *Politics, Philosophy, and Economics* forthcoming.

Liua, Y, and X. Yaob, (1999). "Ensemble learning via negative correlation." *Neural Networks*12:10 pp1399?1404

Page, Scott E. (2006), "Essay: Path Dependence," *Quarterly Journal of Political Science*, 1: 87-115, 2006.

Page, Scott E. (2007), *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*, Princeton University Press.

Page, Scott E (2008) *Diversity and Complexity*, (chapters 1,2) Princeton University Press, Princeton, NJ.

Prokopenko, Boshectti, and Ryan "An information-Theoretic Primer on Complexity, Self-Organization, and Emergene," *Complexity* (2009)