General purpose computer-assisted clustering and conceptualization

Justin Grimmer[a] and Gary King[b,1]

[a]Department of Political Science, Stanford University, Encina Hall West, 616 Serra Street, Palo Alto, CA 94305; and [b]Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge MA 02138

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2010.

Contributed by Gary King, December 22, 2010 (sent for review September 23, 2010)

We develop a computer-assisted method for the discovery of insightful conceptualizations, in the form of clusterings (i.e., partitions) of input objects. Each of the numerous fully automated methods of cluster analysis proposed in statistics, computer science, and biology optimize a different objective function. Almost all are well defined, but how to determine before the fact which one, if any, will partition a given set of objects in an "insightful" or "useful" way for a given user is unknown and difficult, if not logically impossible. We develop a metric space of partitions from all existing cluster analysis methods applied to a given dataset (along with millions of other solutions we add based on combinations of existing clusterings) and enable a user to explore and interact with it and quickly reveal or prompt useful or insightful conceptualizations. In addition, although it is uncommon to do so in unsupervised learning problems, we offer and implement evaluation designs that make our computer-assisted approach vulnerable to being proven suboptimal in specific data types. We demonstrate that our approach facilitates more efficient and insightful discovery of useful information than expert human coders or many existing fully automated methods.

Creating categories and classifying objects in the categories "is arguably one of the most central and generic of all our conceptual exercises. It is the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis in general. Without classification, there could be no advanced conceptualization, reasoning, language, data analysis or, for that matter, social science research" (1). An important step in the development of new hypotheses is the adoption of new ways of partitioning objects into categories. In this paper, we develop a method intended to assist in the creation of unique and insightful conceptualizations from a wide array of possible datasets and substantive problems. We focus on creating "clusterings" (i.e., partitions) of a given set of input objects in an "unsupervised" framework (i.e., with no training set).

Illustrations of useful clusterings in particular applications have been found for some of the existing individual cluster analysis methods. However, for a given application, no method exists for choosing before the fact which of these unsupervised approaches will lead to the most useful clusterings or the most insightful discoveries.

Although our approach builds on almost all prior methods, our goal diverges from the existing literature in one crucial respect: Whereas current cluster analysis methods are designed to produce fully automated clustering (FAC), we attempt to create a computer-assisted clustering (CAC) approach. The problem with FAC is that it requires a single, precisely defined objective function that works across applications. This is infeasible given that human beings are typically optimizing a (mathematically ill-defined) goal of "insightful" or "useful" conceptualizations; the definition of "insightful" differs to some degree by user; and codifying human creativity in a mathematical function is either logically impossible or well beyond current technology. (Existing methods, which we describe as FAC, do come with tuning parameters that enable a user to adjust the optimization function, but in our experience most adjustments turn out to have very small empirical effects, typically much smaller than the differences between methods.)

We develop a CAC approach that uses and encompasses all existing automated cluster analysis methods, numerous novel ones we create (based on combinations of existing solutions), and any others a researcher may create by hand or other technique. By using the collective wisdom of the statistical literature on cluster analysis, we generate a single approach applicable across many substantive problems, without having to know ahead of time which method to apply. We are able to do this by requiring interaction between our methodology and a human user.

In part because of the unsupervised learning nature of cluster analysis, the literature offers few satisfactory procedures for evaluating categorization schemes or the methods that produce them. Unlike in supervised learning methods or classical statistical estimation, straightforward concepts like unbiasedness or consistency do not immediately apply. We respond to this challenge by developing a design for evaluation experiments that reveal the quality of the results and the degree of useful information discovered. We implement these experimental designs in a variety of datasets and show that our CAC methods lead to more insightful conceptualizations than either subject matter experts or individual FAC methods can do alone.

In practice, before applying our algorithm and evaluation techniques, researchers may wish to set aside a randomly selected test set of observations. This holdout set could then be used as a way of making the researcher vulnerable to being wrong about the applicability or generality of a new conceptualization. This may also help prevent researchers from choosing clusterings that merely conform to preexisting conceptualizations, although of course researchers may also choose to let these preexisting views help guide their search for new conceptualizations. Below, we demonstrate that the clusterings and conceptualizations we discover in our subset of documents provide a useful way of analyzing the entire collection of documents.

Although our methods apply to categories of any type of object, we apply them here to clustering documents containing unstructured text. The spectacular growth in the production and availability of text makes this application of crucial importance in many fields.

2 Methodology

One way to think about CAC is to imagine presenting an extremely long list of clusterings (ideally, all of them) and letting the researcher choose the best one for his or her substantive pur-

Author contributions: J.G. and G.K. designed research, performed research, contributed new tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: king@harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1018067108/-/DCSupplemental.

poses. However, human beings do not have the patience, attention span, memory, or cognitive capacity to evaluate so many clusterings in haphazard order. Moreover, from the point of view of a human being, many clusterings are essentially the same. (Imagine 10,000 documents sorted into five categories and moving one document from category 3 to 4; these clusterings are essentially the same because few would even be able to perceive the difference.) Thus, we seek to organize these clusterings so researchers can quickly select the one that best satisfies their particular objectives.

Our procedure represents each clustering as a point in a two-dimensional visual space, such that clusterings (points) close together in the space are almost the same (and so can be disregarded except for fine tuning), and those farther apart may warrant a closer look because they differ in some important way. In effect, this visualization translates the uninterpretable chaos of huge numbers of possible clusterings into a simple framework that (we show) human researchers are able to comprehend and use to efficiently select one or a small number of clusterings that conveys the most useful information.

To create our space of clusterings, we follow six steps, outlined here and detailed below. First, we translate textual documents to a numerical dataset (Section 2.1). (This step is necessary only when the items to be clustered are text documents or in general not already numerical; all our methods would apply without this step to objects with preexisting numerical data.) Second, we apply (essentially) all clustering methods proposed in the literature, one at a time, to the numerical dataset (Section 2.2). Each approach represents different substantive assumptions that are difficult to express before their application, but the effects of each set of assumptions are easily seen in the resulting clusters, and it is the resulting clustering that is of most interest to applied researchers. (A new R package we have written makes this relatively fast.) Third, we develop a metric to measure the similarity between any pair of clusterings (Section 2.3). Fourth, we use this metric to create a metric space of clusterings, along with a lower dimensional Euclidean representation useful for visualization (Section 2.4).

Fifth, we introduce a "local cluster ensemble" method (Section 2.5) as a way to summarize any point in the space, including points for which there exist no prior clustering methods—in which case they are formed as local weighted combinations of existing methods, with weights based on how far each existing clustering is from the chosen point. This allows for the fast exploration of the space, ensuring that users of the software are able to quickly identify partitions useful for their particular research question. Sixth and finally, we develop a new type of animated visualization that uses the local cluster ensemble approach to explore the metric space of clusterings by moving around it while one clustering slowly morphs into others (Section 2.6), again to rapidly allow users to easily identify the partition (or partitions) useful for a particular research question. We also introduce an optional addition to our method that creates new clusterings (Section 2.7).

2.1 Standard Preprocessing: Text to Numbers. We begin with a set of text documents of variable length. For each, we adopt common procedures for representing them quantitatively: We transform to lower case, remove punctuation, replace words with their stems, and drop words appearing in fewer than 1% or more than 99% of documents. For English documents, about 3,500 unique word stems usually remain in the entire corpus. We then code each document with a set of (about 3,500) variables, each coding the number of times a word stem is used in that document.

Despite all the information discarded, these procedures are very common (2). The reason is that most human language is highly repetitive, and so this representation is usually more than adequate. For example, we need not read many sentences of a vitriolic blog post about a political candidate before getting the point. Our general procedure also accommodates multiple representations of the same documents. These might include tf-idf or other term weighting representations, part of speech tagging, tokenization rules such as replacing "do" and "not" with "do_not", etc. (3). Likewise, the many variants of kernel methods—procedures to produce a similarity metric between documents without explicitly representing the words in a matrix—could also be included (4).

2.2 The Collective Wisdom of the Statistical Community. Second, we apply a large number of clustering methods, one at a time, to the numerical representation of our documents. To do this, we have written an R package that runs (with a common syntax) every published clustering method we could find that has been applied to text and used in at least one article by an author other than its developer; we have also included many clustering methods that have not been applied to text before. We developed computationally efficient implementations for the methods included in our program (including variational approximations for the Bayesian statistical methods) (5) so that we can run all the methods on a moderate sized dataset relatively fast; new methods can easily be added to the package as well. Although inferences from our method are typically not affected much, and almost never discontinuously, by including any additional individual method, there is no disadvantage in including as many methods as are available.

A complete list of the methods that we include in our application is available in the *SI Appendix*, but the method is extremely flexible. The only requirement is that each "method" form a proper clustering, with each document assigned either to a single cluster or to different clusters with weights that sum to 1.

2.3 Distance Between Clusterings. We next derive a metric for measuring how similar one clustering is to another. We do this stating three axioms that narrow the range of possible choices to only one. First, the distance is a function of the number of pairs of documents not placed together (i.e., in the same cluster) in both clusterings. (We also prove in the *SI Appendix* that focusing on pairwise disagreements between clusterings is sufficient to encompass differences based on all possible larger subsets of documents, such as triples, quadruples, etc.) Second, we require that the distance be invariant to the number of documents, given any fixed number of clusters in each clustering. Third, we set the scale of the measure by fixing the minimum distance to zero and the maximum distance to $\log(k)$. A key point is that none of these axioms requires that one artificially "align" clusterings before judging their distance, as some others have attempted; in fact, we do not even restrict the clusterings to have the same number of clusters.

As we prove in the *SI Appendix*, only one measure of distance satisfies all three axioms, the variation of information. This measure has also been derived for different purposes from a larger number of different first principles by Meila (6).

2.4 The Space of Clusterings. The matrix of distances between each pair in the set of J clusterings can be represented in a J-dimensional metric space. (The clusterings can each have the same number of clusters, if chosen by the user, or differing numbers.) We project this space down to two Euclidean dimensions for visualization. Because projection entails the loss of information, the key is to choose a multidimensional scaling method that retains the most crucial information. For our purposes, we need to preserve small distances most accurately, because they reflect clusterings to be combined (in the next section) into local cluster ensembles. As the distance between two clusterings increases, a higher level of distortion will affect our results less. This leads naturally to the Sammon multidimensional scaling algorithm

(7); in the *SI Appendix*, we define this algorithm and explain how it satisfies our criteria.

An illustration of this space is given in Fig. 1, *Middle*, with individual clusterings labeled (we discuss this figure in more detail below). Nearby points in this space represent similar clusterings, as judged by our distance metric.

2.5 Local Cluster Ensembles. A "cluster ensemble" is a technique used to produce a single clustering by averaging in a specific way across many individual clusterings (8–13). This approach has the advantage of creating a new, potentially better, clustering, but by definition it eliminates the underlying diversity of individual clusterings and so does not work for our purposes. A related technique that is sometimes described by the same term organizes results by performing a "meta-clustering" of the individual clusterings. This alternative procedure has the advantage of preserving some of the diversity of the clustering solutions and letting the user choose, but because no method is offered to summarize the many clusterings within each "meta-cluster," it does not solve the problem. Moreover, for our purposes, the technique suffers from a problem of infinite regress: Because any individual clustering method can be used to cluster the clusterings, a researcher would have to use them all and their combinations to avoid eliminating meaningful diversity in the set of clusterings to be explored. So whether the diversity of clusterings is eliminated by arbitrary choice of meta-clustering method rather than a substantive choice, or we are left with more solutions than we started with, these techniques, although useful for some other purposes, do not solve our particular problem.

Thus, to preserve local diversity and avoid the infinite regress resulting from clustering a set of clusterings, we develop here a method of generating local cluster ensembles, which we define as a new clustering created at a point in the space of clusterings from a weighted average of nearby existing clusterings. The procedure requires three steps. First, we define the weights around a user selected point in the space. Consider point $\boldsymbol{x}^* = (x_1^*, x_2^*)$ in our space of clusterings. The new clustering defined at this point is a weighted average of nearby clusterings with one weight for each existing clustering in the space, so that the closer the existing clustering, the higher the weight. We define the weight for each existing clustering j on a normalized kernel as $w_j = p(\boldsymbol{x}^*, \sigma^2) / \sum_{m=1}^{J} p(\boldsymbol{x}_m, \sigma^2)$, where $p(\boldsymbol{x}^*, \sigma^2)$ is the height of the kernel (such as a normal or Epanechnikov density) with mean $\boldsymbol{x}^*$ and smoothing parameter σ^2. The collection of weights for all J clusterings is then $\boldsymbol{w} = (w_1, \ldots, w_J)$. Note that although we are using a density to define the kernel, the approach requries no statistical or probabilistic reasoning.

Second, given the weights, we create a similarity matrix for the local cluster ensemble, where each clustering casts a weighted vote for whether each pair of documents appears together in a cluster in the new clustering. First, for a corpus with N documents clustered by method j into K_j clusters, we define an $N \times K_j$ matrix $\boldsymbol{c}_j$ that records how each document is allocated into (or among) the clusters (i.e., so that each row sums to 1). We then horizontally concatenate the clusterings created from all J methods into an $N \times K$ weighted "voting matrix" of methods by document pairs, $\boldsymbol{V}(\boldsymbol{w}) = \{w_1\boldsymbol{c}_1, \ldots, w_J\boldsymbol{c}_J\}$ (where $K = \sum_{j=1}^{J} K_j$). The result of the election is a new similarity matrix, which we create as $\boldsymbol{S}(\boldsymbol{w}) = \boldsymbol{V}(\boldsymbol{w})\boldsymbol{V}(\boldsymbol{w})'$. This calculation places priority on those cluster analysis methods closest in the space of clusters.

Finally, we create a new clustering for point $\boldsymbol{x}^*$ in the space by applying any coherent clustering algorithm to this new averaged similarity matrix (with the number of clusters fixed to a weighted average of the number of clusters from nearby clusterings, using the same weights). As we demonstrate in the *SI Appendix*, our definition of the local cluster ensemble approach becomes invariant to the particular choice of clustering method applied to the new averaged similarity matrix as the number of clusterings increase. This invariance eliminates the infinite regress problem by turning a meta-cluster method selection problem into a weight selection problem (with weights that are variable in the method). The *SI Appendix* also shows how our local cluster ensemble approach is closely related to our underlying distance metric defined in Section 2.3. The key point is that the local cluster ensemble approach will approximate more possible clusterings as additional methods are included and of course will never be worse, and usually considerably better, in approximating a new clustering than the closest existing observed point.

2.6 Cluster Space Visualization. Fig. 1 illustrates our visualization of the space of clusterings, when applied to one simple corpora of documents. This simple and small example, which we choose for expository purposes, includes only the biographies of each US president from Roosevelt to Obama (see http://whitehouse.gov).

The two-dimensional projection of the space of clusterings is illustrated in the figure's middle panel, with individual methods

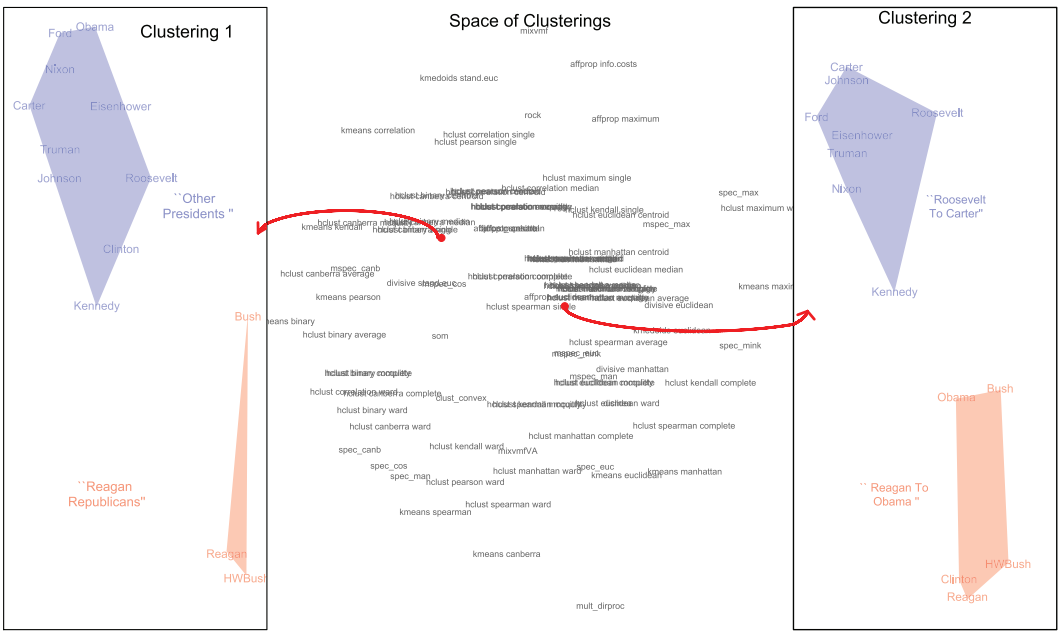

Fig. 1. A clustering visualization. The center panel gives the space of clusterings, with each name printed representing a clustering generated by that method, and all other points in the space defined by our local cluster ensemble approach that averages nearby clusterings. Two specific clusterings (see red dots with connected arrows), each corresponding to one point in the central space, appear to the left and right; labels in the different color-coded clusters are added by hand for clarification, as is the spacing in each.

labeled. Each method corresponds to one point in this space and one set of clusters of the given documents. Points corresponding to a labeled method correspond to results from prior research; other points in this space correspond to new clusterings, each constructed as a local cluster ensemble.

A key point is that once the space is constructed, the labeled points corresponding to previous methods deserve no special priority in choosing a final clustering. For example, a researcher should not necessarily prefer a clustering from a region of the space with many prior methods as compared to one with few or none. In the end, the choice is the researcher's and should be based on what he or she finds to convey useful information. Because the space itself is crucial, but knowledge of where any prior method exists in the space is not, visualization software can easily toggle off these labels so that researchers can focus on clusterings they identify.

The space is formally discrete, because the smallest difference between two clusterings occurs when (for nonfuzzy partitions) exactly one document moves from one cluster to another, but an enormous range of possible clusterings still exists: Even this tiny dataset of only 13 documents can be partitioned in 27,644,437 possible ways, each representing a different point in this space. A subset of these possible clusterings appears in the figure corresponding to all those clusterings the statistics community has come up with, as well as all possible local cluster ensembles that can be created as weighted averages from them. (The arching shapes in the figure occur regularly in dimension reduction when using methods that emphasize local distances between the points in higher dimensional space; see ref. 14.)

Fig.1 also illustrates two points (as red dots) in the middle panel, each representing one clustering and portrayed on one side of the central graph, with individual clusters color coded (and substantive labels added by hand for clarity). Clustering 1, in the left panel, creates clusters of "Reagan Republicans" (Ronald Reagan, George H. W. Bush, and George W. Bush) and all others. Clustering 2, in the right panel, groups the presidents into two clusters organized chronologically.

This figure summarizes snapshots of an animated software program at two points. In general, the software can be set up so a researcher can put a single cursor somewhere in the space of clusterings and see the corresponding set of clusters for that point appear in a separate window. The researcher can then move this point and watch the clusters in the separate window morph smoothly from one clustering to another. Our experience in using this visualization often leads us first to check about 4–6 well-separated points, which seems to characterize the main aspects of the diversity of all the clusterings. Then, we narrow the grid further by examining about the same number of clusterings in the local region. Although the visualization offers an enormous number of clusterings, the fact that they are highly ordered in this simple geography makes it possible to understand with greatly reduced time and effort.

2.7 Optional New Clustering Methods to Add. For most applications, beginning with the collective wisdom of the statistics community, and clusterings constructed from them, helps to narrow down the enormous space of all possible clusterings to a large (indeed larger than has ever before been explored) but yet still managable set of solutions. However, there may well be useful insights to be found outside of the large space that we have already identified. Thus, we offer two methods to explore some of the remaining uncharted space. First, we randomly sample new clusterings from the entire space. Second, we define a Markov chain to move beyond the space of existing clusterings to the area around those clusterings. Details about both algorithms are available in the *SI Appendix*.

3 Evaluation Designs

The most important approach to evaluating a purely unsupervised learning approach to clustering is whether the user, or the user's intended audience, finds the chosen clustering useful or insightful. Thus, a perfectly reasonable approach is to use our method, choose a clustering and gather insight, and be done. However, one may also wish to go further in some circumstances and formally evaluate the clustering solutions.

Common approaches to evaluating the performance of cluster analysis methods, which include comparison to internal or supervised learning standards, have known difficulties. Internal standards of comparison define a quantitative measure indicating high similarity of documents within, and low similarity of documents across, clusters. But if this were the goal, we could define a cluster analysis method with an objective function that optimizes it directly; this may lead to a good answer but not one that is vulnerable to being proven wrong. Indeed, because any one quantitative measure is unlikely to reflect the actual substance a researcher happens to be seeking, "good scores on an internal criterion do not necessarily translate into good effectiveness in an application" (ref. 2, pp. 328–329).

An alternative evaluation approach is based on supervised learning standards, which involve comparing the results of a cluster analysis to some "gold standard" set of clusters, prechosen by human coders. Although human coders may be capable of assigning documents to a small number of given categories, they are incapable of choosing an optimal clustering or one in any sense better than what a CAC method could enable them to create. As such, using a supervised learning "gold standard" to evaluate an unsupervised learning approach is also of questionable value.

Success at facilitating discovery is difficult to formalize mathematically and easy to lead to unfalsifiable approaches. Indeed, some in the statistical literature have even gone so far as to chide those who attempt to use unsupervised learning methods to make systematic discoveries as unscientific (15).

To respond to these problems, we introduce and implement three direct evaluation approaches using insights from survey research and social psychology to compare to elicited human judgment in ways that people are capable of providing. We first evaluate cluster quality, the extent to which intracluster similarities outdistance intercluster similarities (Section 3.1). Cluster quality demonstrates that users of our approach are able to efficiently search through the space of clusterings to identify clusterings that are coherent and useful to others. Second is discovery quality, a direct evaluation by substance matter experts of insights produced by different clusterings in their own data (Section 3.2). This ensures that the clusterings identified are insightful for experts working in a field of study. Third and finally, we offer a substantive application of our method and show how it assists in discovering a specific useful conceptualization and generates new verifiable hypotheses that advance the political science literature (Section 3.3). For this third approach, the judge of the quality of the knowledge learned is the reader of this paper.

3.1 Cluster Quality. We judge cluster quality with respect to a particular corpus by randomly drawing pairs of documents from the same cluster and from different clusters and asking human coders unaware how each document was chosen to rate the similarity of the documents within each pair on a simple three point scale: (*i*) unrelated, (*ii*) loosely related, (*iii*) closely related. (Our extensive pretesting indicated that intercoder reliability suffers with more categories, but coders are able to understand and use effectively this coding scheme. We also found that the average code from 10 graduate students correlated with the average code from the Amazon Mechanical Turk system at 0.99.) The idea is to keep our human judges focused on well-defined tasks they are able to perform well, in this case comparing only two documents at a time. Then the numerical measure of cluster quality is the aver-

age rating of pair similarity within clusters minus the average of pairs in different clusters. (The *SI Appendix* also introduces a way to save on evaluation costs in measuring cluster quality.)

We apply this measure in each of three different corpora by choosing 25 pairs of documents (13 from the same clusters and 12 from different clusters), computing cluster quality, and averaging over the judgments about the similarity of each pair made separately by many different human coders. We then compare the cluster quality generated by our approach to the cluster quality from a preexisting hand-coded clustering. This comparison demonstrates that users of our method are able to identify clusterings that are coherent and are able to efficiently search through the millions of clusterings we present users.

What we describe as "our approach" here is a single clustering from the visualization we chose ourselves without participating in evaluating document similarity. This procedure is biased against our method because if we had let the evaluators use our visualization, our approach would almost by definition have performed much better. Although the number of clusters does not necessarily affect the measure of cluster quality, we constrained our method further by requiring it to choose a clustering with approximately the same number of clusters as the preexisting hand-coded clustering.

Press releases. We begin with 200 press releases we randomly selected from those issued by Senator Frank Lautenberg's Senate office and categorized by him and his staff in 24 categories (http://lautenberg.senate.gov). These include appropriations, economy, gun safety, education, tax, social security, veterans, etc. These represent a difficult test for our approach because the documents, the categorization scheme, and the individual classifications were all created by the same people at great time and expense.

The top line in Fig. 2 gives the results for the difference in our method's cluster quality minus the cluster quality from Lautenberg's hand-coded categories. The point estimate appears as a dot, with a thick line for the 80% confidence interval and a thin line for the 95% interval. The results, appearing to the right of the vertical dashed line that marks zero, indicate that the clustering our method identified had unambiguously higher quality than the author of the documents produced by hand. This provides evidence that the clusterings are organized in a way that allows for the efficient search over many millions of different (but similar) conceptualizations. (We give an example of the substantive importance of our selected clustering in Section 3.3.)

State of the Union messages. Our second example comes from an analysis of all 213 quasi-sentences in President George W. Bush's 2002 State of the Union address, hand coded by the Policy Agendas Project (http://www.policyagendas.org). Each quasi-sentence (defined in the original text by periods or semicolon separators) takes the role of a document in our discussion. The authors use 19 policy topic-related categories, including agriculture, banking & commerce, civil rights/liberties, defense, education, etc. Quasi-sentences are difficult tests because they are very short and may have meaning obscured by the context, which most automated methods ignore.

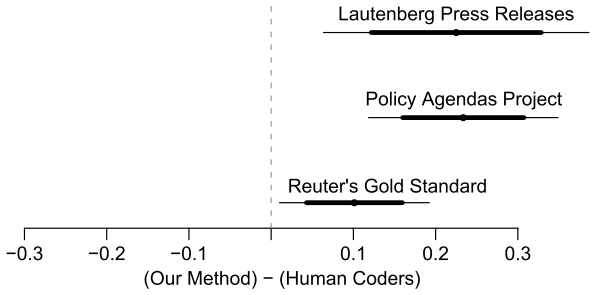

Fig. 2. Cluster quality experiments. Each line gives a point estimate (dot), 80% confidence interval (dark line), and 95% confidence interval (thin line) for a comparison between our automated cluster analysis method and clusters created by hand. Cluster quality is defined as the average similarity of pairs of documents from the same cluster minus the average similarity of pairs of documents from different clusters, as judged by human coders one pair at a time.

The results of our cluster quality evaluation appear as the second line in Fig. 2. Again, using our CAC methods we selected a clustering that turned out to have higher quality than the Policy Agendas Project coding scheme; this can be seen by the whole 95% confidence interval appearing to the right of the vertical dashed line. These results do not imply that anything is wrong with the Policy Agendas Project classification scheme, only that there seems to be more information in than the project's chosen categories may indicate.

Substantively, our CAC approach led us to notice that the largest cluster of statements in Bush's address were those that addressed the 9/11 tragedy, including many devoid of immediate policy implications, and so are lumped into a large "other" category by the project's coding scheme, despite considerable political meaning. For example, "And many have discovered again that even in tragedy, especially in tragedy, God is near." or "We want to be a Nation that serves goals larger than self." This cluster thus conveys how the Bush administration's response to 9/11 was sold rhetorically to resonate with his religious supporters and others, all with considerable policy and political content. For certain research purposes, this discovery may reflect highly valuable additional information.

Reuters news stories. For a final example of cluster quality, we use 250 documents randomly drawn from the Reuters-21578 news story categorization. This corpus has often been used as a gold standard baseline for evaluating clustering (and supervised learning classification) methods in the computer science literature (http://www.daviddlewis.com/resources/testcollections/reuters21578/). In this collection, each Reuters financial news story from 1987 has been classified by the Reuters news organization (with help from a consulting firm) into one of 22 categories, including trade, earnings, copper, gold, coffee, etc. We again apply the same evaluation methodology; the results, which appear as the bottom line in Fig. 2, indicate again that the clustering we identified turned out to have unambiguously higher cluster quality than Reuters's own gold standard classification.

3.2 Discovery Quality. We show here that using our approach leads to more informative discoveries for researchers engaged in real scholarly projects. This is an unusually hard test for a statistical method and one rarely performed; it would be akin to requiring not merely that a standard statistical method possesses certain properties like being unbiased, but also, when given to researchers and used in practice, that they actually use it appropriately and estimate their quantities of interest correctly.

The question we ask is whether the computer assistance we provide helps. To perform this evaluation, we recruited two scholars in the process of evaluating large quantities of text in their own (independent) works in progress, intended for publication (one faculty member, one senior graduate student). In each case, we offered an analysis of their text in exchange for their participation in our experiment. One had a collection of documents about immigration in America in 2006; the other was studying a longer period about how genetic testing was covered in the media. Both had spent many months reading their documents. (To ensure the right of first publication goes to the authors, we do not describe the specific insights we found here and instead only report how they were judged in comparison to those produced by other methods.) Using a large collection of texts from each researcher, we spent about an hour using our method to

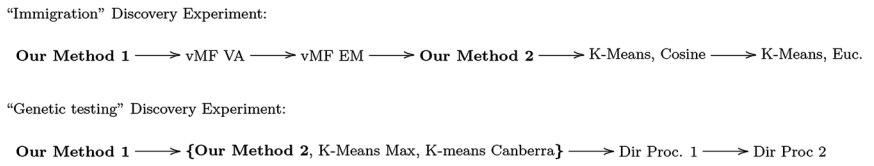

Fig. 3. Results of discovery experiments, where $A \rightarrow B$ means that clustering A is judged to be "more informative" than B in a pairwise comparison, with braces grouping results in the second experiment tied due to an evaluator's cyclic preferences. In both experiments, a clustering from our method is judged to beat all others in pairwise comparisons.

identify two distinct clusterings from our space that we thought provided useful and distinct insights into the data. For comparison, we also applied the popular k-means clustering methodology (with variable distance metrics) and one of two more recently proposed clustering methodologies—the Dirichlet process prior and the mixture of von Mises Fisher distributions, estimated using a variational approximation (16). We used two different clusterings from each of the three cluster analysis methods applied in each case. For our method, we again biased the results against our method and this time chose the two clusterings ourselves instead of letting them use our visualization.

We then created an information packet on each of the six clusterings. This included the proportion of documents in each cluster, an exemplar document, and a brief automated summary of the substance of each cluster, using a technique that we developed. To create the summary, we first identified the 10 most informative words stems for each cluster, in each clustering (i.e., those with the highest "mutual information"). The summary then included the full length word most commonly associated with each chosen word stem. We found through much experimentation that words selected in this way usually provide an excellent summary of the topic of the documents in a cluster.

We then asked the researchers to familiarize themselves with the six clusterings. After about 30 min, we asked each to perform all $\binom{6}{2} = 15$ pairwise comparisons, presented in random order, between the clusterings and in each case to judge which clustering within a pair they thought was "more informative." In the end, we want a cluster analysis methodology that produces at least one method that does well. Because the user ultimately will be able to judge and choose among results, having a method that does poorly is not material; the only issue is how good the best one is.

We are evaluating two clusterings from each cluster analysis method, and so we label them 1 and 2, although the numbers are not intended to convey order. Fig. 3 gives a summary of our results, with arrows indicating dominance in pairwise comparisons. In the first (immigration) example, illustrated at the top of the figure, the 15 pairwise comparisons formed a perfect Guttman scale (17) with "our method 1" being the Condorcet winner (i.e., it beat each of the five other clusterings in separate pairwise comparisons). (This was followed by the two mixtures of von Mises Fisher distribution clusterings, then "our method 2," and then the two k-means clusterings.) In the genetics example, our researcher's evaluation produced one cycle, and so it was close to but not a perfect Guttman scale; yet "our method 1" was again the Condorcet winner. (Ranked according to the number of pairwise wins, after "our method 1" was one of the k-means clusterings, then "our method 2." then other k-means clustering, and then the two Dirichlet process cluster analysis methods. The deviation from a Guttman scale occurred among the last three items.)

3.3 Partisan Taunting: An Illustration of Computer-Assisted Discovery.

We now give a brief report of an example of the whole process of analysis and discovery using our approach applied to a real example. We develop a categorization scheme that advances one in the literature, measure the prevalence of each of its categories in a new out-of-sample set of data to show that the category we discovered is common, develop a new hypothesis that occurred to us because of the new lens provided by our new categorization scheme, and then test it in a way that could be proven wrong. The degree of insight discovered can be judged by the reader.

In a famous and monumentally important passage in the study of American politics, (ref. 18, p. 49ff) Mayhew argues that "congressmen find it electorally useful to engage in…three basic kinds of activities"—credit claiming, advertising, and position taking. This typology has been widely used over the last 35 years, remains a staple in the classroom, and accounts for much of the core of several other subsequently developed categorization schemes (19–21). In the course of preparing our cluster analysis experiments in Section 3.1, we found much evidence for all three of Mayhew's categories in Senator Lautenberg's press releases, but we also made what we view as an interesting discovery.

We illustrate this discovery process in Fig. 4, where the top panel gives the space of clusterings we obtain when applying

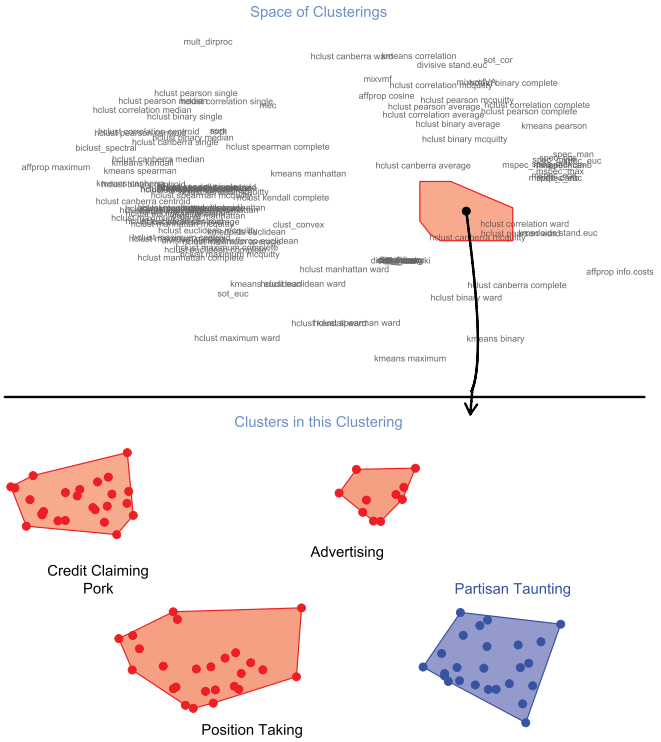

Fig. 4. Discovering partisan taunting. The top portion of this figure presents the space of clustering solutions of Frank Lautenberg's (D-NY) press releases. Partisan taunting could be easily discovered in any of the clustering solutions in the red region in the top plot. The bottom plot presents the clusters from a representative clustering within the red region at the top (represented by the black dot). Three of the clusters (in red) align with Mayhew's categories, but we also found substantial partisan taunting cluster (in blue), with Lautenberg denigrating Republicans in order to claim credit, position-take, and advertise. Other points in the red polygon at the top represent different clusterings, but all clearly reveal the partisan taunting category.

Table 1. Examples of partisan taunting in Senator Lautenberg's press releases

Date	Lautenberg Category	Quote
2/19/2004	civil rights	"The Intolerance and discrimination from the Bush administration against gay and lesbian Americans is astounding."
2/24/2004	government oversight	"Senator Lautenberg Blasts Republicans as 'Chicken Hawks'"
8/12/2004	government oversight	"John Kerry had enough conviction to sign up for the military during wartime, unlike the Vice President [Dick Cheney], who had a deep conviction to avoid military service."
12/7/2004	homeland security	"Every day the House Republicans dragged this out was a day that made our communities less safe."
7/19/2006	health care	"The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then."

our methodology to Lautenberg's press releases (i.e., like Fig. 1). Recall that each name in the space of clusterings in the top panel corresponds to one clustering obtained by applying the named clustering method to the collection of press releases; any point in the space between labeled points defines a new clustering using our local cluster ensemble approach; and nearby points have clusterings that are more similar than those farther apart.

The clusters within the single clustering represented by the black point in the top panel is illustrated in the bottom panel, with individual clusters comprising Mayhew's categories of claiming credit, advertising, and position taking (all in red), as well as an activity that his typology obscures and he does not discuss. We call this new category *partisan taunting* (see blue region in Fig. 4) and describe it below. Each of the other points in the red region in the top panel represent clusterings that also clearly suggest partisan taunting as an important cluster although with somewhat different arrangements of the other clusters. That is, the user would only need to examine one point anywhere within this (red) region to have a good chance at discovering partisan taunting as a potentially interesting category.

Examples of partisan taunting appear in Table 1. Unlike any of Mayhew's categories, each of the colorful examples in the table explicitly reference the opposition party or one of its members, using exaggerated language to put them down or devalue their ideas. Most partisan taunting examples also overlap two or three of Mayhew's existing theoretical category definitions, which is good evidence of the need for this separate, and heretofore unrecognized, category. We did find that the documents were relatively easy to distinguish from Mayhew's existing categories.

Partisan taunting provides a new category of Congressional speech that emphasizes the interactions inherent between members of a legislature. Mayhew's (1974) original theory supposed that members of Congress were atomistic rational actors, concerned only with optimizing their own chance of reelection. Yet legislators interact with each other regularly, criticizing and supporting ideas, statements, and actions. This interaction is captured with partisan taunting but is absent from the original typology. In the *SI Appendix*, we detail how analyzing partisan taunting provides additional insights in addition to Mayhew's (1974) original typology.

Our technique has thus produced a new and potentially useful conceptualization for understanding Senator Lautenberg's 200 press releases. Although asking whether the categorization is "true" makes no sense, this modification to Mayhew's categorization scheme would seem to pass the tests for usefulness given in Section 3.1. We now show that it is also useful for out-of-sample descriptive purposes and separately for generating and rigorously testing other hypotheses suggested by this categorization.

We begin with a large out-of-sample test of the descriptive merit of the new category, for which we analyze all 64,033 press releases from all 301 senator-years during 2005–2007. To do this, we developed a coding scheme that includes partisan taunting, other types of taunting (to make sure our first category is well defined), and other types of press releases, including Mayhew's three categories. We then randomly selected 500 press releases and had three research assistants assign each press release to a category (we had approximately 83% agreement and resolved disagreements by reading the press releases ourselves). Finally, we applied the supervised learning approach to text analysis given by ref. 22 to the entire set of 64,033 press releases to estimate the percent of press releases that were partisan taunts for each senator in each year. (By setting aside a portion of this training set, we verified that the Hopkins–King methodology produced highly accurate estimates in these data.)

Overall, we find that 27% of press releases among these 301 senator-years were partisan taunts, thus confirming that this category was not merely an idiosyncrasy of Senator Lautenberg. Instead partisan taunting seems to play a central role in the behavior of many senators. Indeed, it may even define part of what it means to be a member of the party in government. The histogram in the left panel of Fig. 5 gives the distribution of taunting behavior in our data; it conveys the large amount of taunting across numerous press releases, as well as a fairly large dispersion across senators and years in taunting behavior.*

Finally, analyzing Senator Lautenberg's press releases led us to consider the role of taunting behavior in theories of democratic representation. Almost by definition, partisan taunting is antithetical to open deliberation and compromise for the public good (23). Thus, an important question is who taunts and when—which led us to the hypothesis that taunting would be less likely to occur in competitive senate seats. The idea is that taunting is most effective when a senator has the luxury of preaching to the choir and warning his or her partisans of the opposition (which has few votes); if instead a politician's electoral constituency is composed of large numbers of opposition party members, we would expect partisan taunting to be less effective and thus less used. If true, this result poses a crucial tension in democratic representation. Deliberation is seen as a normative good, but the degree to which a representative is a reflection of his or her constituency is also often seen to be an important component of democracy (24, 25). However, if our hypothesis is empirically correct, then democracies may have a zero sum choice between deliberation, which occurs more often in the absence of partisan taunting and thus in the most competitive states, and reflection, which by definition occurs in the least competitive states.

By using our large dataset of press releases, we construct an out-of-sample test of our hypothesis. The right panel of Fig. 5 gives the results. Each dot in this figure represents one senator-year, with red for Republicans and blue for Democrats. The horizontal axis is the proportion of the 2004 two-party vote for George W. Bush—a measure of the size of the underlying Republican coalition in each state, separate from all the idiosyncratic features of individual senatorial campaigns. We also portray the dominant patterns with a smoothed (LOESS) line for the Republicans (in red) and Democrats (in blue). The results overall clearly support the hypothesis: As states become more Republi-

*The top 10 senator-year taunters include Baucus (D-MT), 2005; Byrd (D-WV), 2007; Thune (R-SD), 2006; Ensign (R-NV), 2005; McConnell (R-KY), 2006; Biden (D-DE), 2005; Reid (D-NV), 2005; Coburn (R-OK), 2007; Sarbanes (D-MD), 2006; Kennedy (D-MA), 2007.

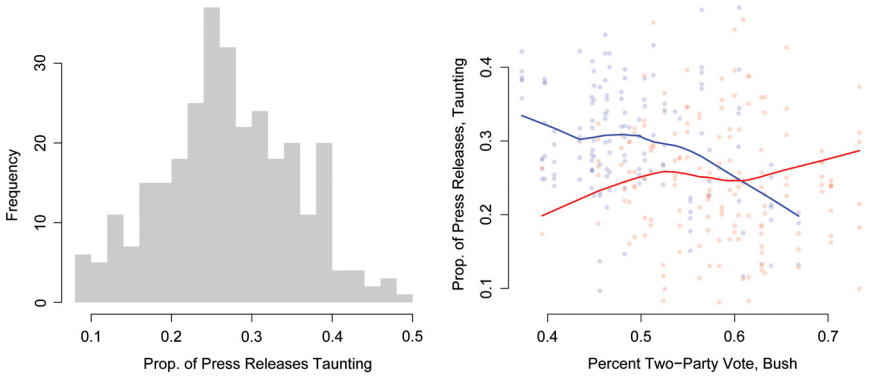

Fig. 5. Partisan taunting hypothesis verification. The left panel shows the distribution in partisan taunting in senators' press releases, and the right panel demonstrates that taunting is more likely when senators are in less competitive states. Each of the 301 points in the right panel represents the results of an analysis of one year's worth of a single senator's press releases, with blue for Democrats and red for Republicans.

can (moving from left to right), partisan taunting by Republicans increases, whereas partisan taunting by Democrats declines.

Of course, much more can be done with this particular empirical example, which is in fact the point: Our clustering methodology helped us choose a new categorization scheme to understand an aspect of the world in a new way, a new concept represented as a new category, a new hypothesis capable of being proven wrong, and a rigorous out-of-sample validation test for both describing and explaining the variation in the prevalence of this category among all senators.

4 Concluding Remarks

We introduce in this paper a computer-assisted approach to unsupervised learning through cluster analysis. We also develop empirically based procedures for evaluating this and other cluster analytic methods and their resulting clusterings that use human judgment in a manner consistent with their cognitive strengths. Through a variety of examples, we demonstrate how this approach can relatively easily unearth new discoveries of useful information from large quantities of unstructured text.

Given the ongoing spectacular increase in the production and availability of unstructured text about subjects of interest to social scientists, and the impossibility of assimilating, summarizing, or even characterizing much of it by reading or hand coding, the most important consequence of this research may be its potential for scholars to help efficiently unlock the secrets this information holds.

For methodologists and statisticians working on developing new methods of cluster analysis, this research also offers techniques for evaluating their efforts. Research that follows up on our strategy by creating new ways of encompassing existing methods might be designed to make the process easier, visualized in other ways, or computationally faster. Most of the research currently being done is focused on developing individual (i.e., nonencompassing) methods; we know that, by definition, any one individual method cannot outperform the approach proposed here, but new individual methods may be able to improve our approach if included in the cluster methods we encompass. For that purpose, we note that the most useful new individual methods would be those that fill empty areas in the space of clusterings, especially those outside the convex hull of existing methods in this space. Methods that produce clusterings for many datasets close to others would not be as valuable.

ACKNOWLEDGMENTS. For helpful advice, coding, comments, or data we thank John Ahlquist, Jennifer Bachner, Jon Bischof, Matt Blackwell, Heidi Brockman, Jack Buckley, Jacqueline Chattopdhyay, Patrick Egan, Adam Glynn, Emily Hickey, Chase Harrison, Dan Hopkins, Grace Kim, Elena Llaudet, Katie Levine, Elena Llaudet, Scott Moser, Jim Pitman, Matthew Platt, Ellie Powell, Maya Sen, Arthur Spirling, Brandon Stewart, and Miya Woolfalk.

1. Bailey KD (1994) *Typologies and Taxonomies: An Introduction to Classification Techniques* (Sage, Beverly Hills, CA).
2. Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval* (Cambridge Univ Press, New York).
3. Monroe Burt, Colaresi M, Quinn K (2008) Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Polit Anal* 16:372–403.
4. Shawe-Taylor J, Cristianini N (2004) *Kernel Methods for Pattern Analysis* (Cambridge Univ Press, Cambridge).
5. Jordan Michael, Ghahramani Z, Jaakkola T, Saul L (1999) An introduction to variational methods for graphical models. *J Mach Learn Res* 37:183–233.
6. Meila M (2007) Comparing clusterings: An information based distance. *J Multivariate Anal* 98:873–895.
7. Sammon J (1969) A nonlinear mapping for data structure analysis. *IEEE T Comput* C-18:401–409.
8. Strehl Alexander, Grosh J (2003) Cluster ensembles: A knowledge reuse framework for combining multiple partitions. *J Mach Learn Res* 3:583–617.
9. Fern X, Brodley C (2003) Random project for high dimensional data clustering: A cluster ensemble approach. *Proceedings of the Twentieth International Conference on Machine Learning* (International Machine Learning Society, Washington).
10. Law M, Topchy A, Jain A (2004) Multi-objective data clustering. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, Washington).
11. Caruana R, Elhawary M, Nguyen N, Smith C (2006) Meta clustering. *ICDM'06. Sixth International Conference on Data Mining* (SIAM, Bethesda, MD), pp 107–118.
12. Gionis A, Mannila H, Tsaparas P (2005) Clustering aggregation. *Proceedings of the 21st International Conference on Data Engineering* (IEEE Computer Society, Tokyo).
13. Topchy A, Jain AK, Punch W (2003) Combining multiple weak clusterings. *Proceedings IEEE International Conference on Data Mining* (IEEE Computer Society, Melbourne).
14. Diaconis P, Goel S, Holmes S (2008) Horseshoes in multidimensional scaling and local kernel methods. *Ann Appl Stat* 2:777–807.
15. Armstrong JS (1967) Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. *Am Stat* 21:17–21.
16. Blei D, Jordan M (2006) Variational inference for dirichlet process mixtures. *Bayesian Analysis* 1:121–144.
17. Guttman L (1950) The problem of attitude and opinion measurement. *Measurement and Prediction* 4:46–59.
18. Mayhew D (1974) *The Electoral Connection* (Yale Univ Press, New Haven, CT).
19. Fiorina M (1989) *Congress, Keystone of the Washington Establishment* (Yale Univ Press, New Haven, CT).
20. Eulau H, Karps P (1977) The puzzle of representation: Specifying components of responsiveness. *Legis Stud Quart* 2:233–254.
21. Yiannakis DE (1982) House members communication styles: Newsletters and press releases. *J Polit* 44:1049–1071.
22. Hopkins D, King G (2010) A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54 pp:229–247 http://gking.harvard.edu/files/abs/words-abs.shtml.
23. Gutmann A, Thompson D (1996) *Democracy and Disagreement* (Harvard Univ Press, Cambridge, MA).
24. Miller WE, Stokes DE (1963) Constituency influence in Congress. *Am Polit Sci Rev* 57:45–56.
25. Pitkin HF (1972) *The Concept of Representation* (Univ of California Press, Berkeley, CA).