The Leadership Quarterly xxx (2021) 101535

Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

Full length article

Predicting leadership perception with large-scale natural language data

Sudeep Bhatia [a,*], Christopher Y. Olivola [b,c], Nazlı Bhatia [a], Amnah Ameen [a]

[a] University of Pennsylvania, United States
[b] Tepper School of Business, Carnegie Mellon University, United States
[c] Department of Social and Decision Sciences, Carnegie Mellon University, United States

ARTICLE INFO

Keywords:
Implicit leadership theories
Leadership perception
Word embeddings
Machine learning
Computational linguistics

ABSTRACT

We present a computational method for predicting, and identifying the correlates of, leadership perceptions for prominent individuals. Our approach proxies knowledge representations for these individuals using high-dimensional semantic vectors derived from large-scale news media datasets. It then applies machine learning techniques to build a model that maps these vectors onto participant ratings of leadership effectiveness. This method greatly outperforms other approaches and achieves accuracy rates comparable to human participants in predicting leadership effectiveness judgments. Crucially, it relies on attributes and associations identified by established theories of leadership perception—notably implicit leadership theories—as guiding lay leadership perception. Thus, our model appears to have learnt the same implicit leadership cues as our human participants. In addition, we show that our approach can be used to not only predict leadership effectiveness judgments, but also to identify dimensions that people associate with effective leadership, as well as quantify the extent of this association for each dimension. We illustrate the broad applicability of our method by using it to predict leadership perceptions for over 6000 individuals in the public sphere, and to algorithmically uncover the particular traits, concepts, and attributes that people most strongly associate with effective leaders.

Introduction

Leadership researchers have long recognized the importance of understanding how leaders are perceived by their followers (Avolio & Bass, 2004; Craig & Gustafson, 1998; Giessner & van Knippenberg, 2008; Hais et al., 1997; Phillips & Lord, 1981; Schyns, 2006; Tyler, 1986). Although academics may disagree on the properties of effective leaders, people often possess their own implicit leadership theories (ILTs) that specify the structure and content of cognitive categories that distinguish leaders from non-leaders and good leaders from bad leaders (Lord et al., 1984, 1982; Schyns, 2006). Effective leaders may not always be perceived as such, but the perception of leadership ability is key to an individual's success as a leader (Antonakis & Eubanks, 2017; Carnes et al., 2015; Olivola & Todorov, 2010).

Researchers have used a wide range of qualitative and quantitative techniques to study various aspects of leadership (for reviews, see Dinh et al., 2014; Gardner et al., 2010; Lovelace et al., 2019; Stentz et al., 2012), including leadership perception and people's ILTs (e.g., Offermann & Coats, 2018; Olivola et al., 2014). When it comes to studying how specific leaders or leadership types are viewed, one of the most popular approaches has been to survey people directly and elicit their perceptions of actual or hypothetical leaders (Braun et al., 2019; Fraser & Lord, 1988; Lord et al., 1984; Nye & Forsyth, 1991). Another popular method has been to analyze the contents of biographies and historical documents, in an effort to identify consistent themes and descriptive patterns (i.e., historiometrics: Bedell et al., 2006; DeChurch et al., 2011; Eubanks et al., 2010, 2016; Hunter et al., 2011; Ligon et al., 2012; Mumford, 2006; Mumford et al., 2007; O'Connor et al., 1995; Parry et al., 2014; Simonton, 1986; Spangler et al., 2012; Strange & Mumford, 2002; Vessey et al., 2014; Yammarino et al., 2013).[1]

These "classic" qualitative and quantitative methodological approaches to studying leadership, while clearly useful, have important limitations (e.g., Ligon et al., 2012; Spangler et al., 2012). Surveys of leadership perception are both costly (in terms of money and time spent) and slow, especially if one wants to obtain participant judgments for a large set of leaders. In addition, the data obtained in any given study can only tell us about the particular leaders and evaluation

* Corresponding author at: Department of Psychology, University of Pennsylvania, Philadelphia, PA, United States.
E-mail address: bhatiasu@sas.upenn.edu (S. Bhatia).

[1] In contrast to surveys, this historiometric method focuses on leadership as a social influence process that can best be understood in context (DeChurch et al., 2011). By examining leadership in context, it allows researchers to identify rich sets of factors, including those external to the leader being studied, that potentially influence leadership perceptions, as well as leadership styles and performance contributions (Ligon et al., 2012; Parry et al., 2014; Spangler et al., 2012).

https://doi.org/10.1016/j.leaqua.2021.101535
Received 28 November 2019; Revised 4 March 2021; Accepted 30 March 2021
Available online xxxx
1048-9843/

dimensions that were explicitly measured in that study; investigating perceptions for a new set of leaders or evaluation dimensions that researchers previously ignored requires surveying a whole new sample of respondents. Current historiometric techniques are also limited in that they rely on small numbers of texts and small samples of human judges to hand-code those texts. The use of human judges to evaluate large quantities of text data is not only slow (due to the time it takes human judges to carefully read and code texts), but also problematic as it can lead to fatigue in judges as well as common method bias (Parry et al., 2014).

These limitations restrict the ability of leadership perception theories to be used in large-scale practical applications. Ideally, existing methodological and theoretical insights concerning leadership perception should be extended to the thousands of well-known individuals who play important roles in business, politics, religion, and society. Applications such as these could be used to assess employee feelings regarding their organization's leadership, thereby measuring job performance and informing leadership and personnel decisions. These applications could also inform politics and policy by automatically measuring public opinion regarding leaders. Perhaps most importantly, such applications can also guide theoretical research: By studying the attributes and associations that best predict leadership perception on a very large scale, we would better understand the content and structure of people's ILTs.

What is needed, then, is a new methodological approach that can analyze and predict leadership perceptions in an automated manner on a very large scale. This approach should build off existing theories of implicit leadership perception, and apply their insights to successfully predict how people judge the leadership ability of the many influential individuals who exist in the public sphere, and are therefore frequently mentioned in news and social media. It should also be able to automatically uncover the traits, concepts, and attributes that people most strongly associate with effective leaders, thereby providing novel theoretical insights regarding the psychological underpinnings of leadership perception. In this paper, we introduce and test such an approach.

Implicit leadership theories

Implicit leadership theories (ILTs; Eden & Leviatan, 1975; Lord et al., 1984)—i.e., "the implicit and naïve conceptualizations people hold of their leaders" (Offermann & Coats, 2018, p. 513)—represent the cognitive structures and schemas associated with people's expectations of leader traits and behaviors compared to those of non-leaders. These schemas are formed as people attempt to organize the world around them into hierarchical categories composed of superordinate, basic, and subordinate levels (Rosch, 1978). Members of a given category are hypothesized to share certain attributes that differentiate them from category non-members.

Within the broad domain of leadership (Lord et al., 1984), the superordinate level is the most inclusive, and comprises the simple differentiation between a leader and a non-leader. According to this theory, although there are presumably few attributes that are common to all leaders, there are also few, if any, attributes that are shared between leaders and non-leaders. The basic level is composed of eleven different types of leaders depending on context (e.g., political leaders, sports leaders, etc.), and the subordinate level contains further differentiating attributes within each basic level of categorization. Lord et al. (1984) asked participants to describe attributes that apply to leaders and corresponding non-leaders within particular categories at the basic level (e.g., "business leader" vs. "businessman"), and showed that the resulting attributes are internally consistent within each category and meaningfully differentiate leaders from non-leaders. In this way, ILTs also document what people view as prototypical of a leader (vs. non-leader). This implicit schema of the prototypical leader subsequently gets activated when people need to make judgments concerning leadership, such as when they evaluate someone's leadership potential or effectiveness, and it influences these judgments. That is, the prototype of a leader becomes an important benchmark that people use to evaluate actual leaders. For example, surveys that ask participants to describe an actual leader they know about and those asking participants to describe a leader without prompting a specific person both yield very similar response factor structures (Eden & Leviatan, 1975; Gioia & Sims, 1985).

Moreover, leaders who overlap (in people's minds) with the prototype of an effective leader are rated as effective (Nye & Forsyth, 1991), whereas those who don't are evaluated negatively (Heilman, 1983; Junker et al., 2011), regardless of their actual effectiveness. A domain where this overlap has possibly produced particularly detrimental effects is gender and leadership perceptions. The influential "think manager, think male" paradigm (Schein et al., 1996; Schein, 1973, 1975) applied the above methods to investigate whether people's mental prototype of a leader is gendered and found that both men and women associate leadership more with masculine than feminine traits. More recent work, including a large-scale meta-analysis, also found support for this close relationship between implicit leadership perceptions and masculinity (Braun et al., 2017; Heilman, 1983; Koenig et al., 2011). Although this body of work does not causally test the relationship between gendered ILTs and perceptions of leader effectiveness, related theories built on the idea of a fit between gender-stereotypic expectations and leadership, do show that women are rated as less effective leaders even when, objectively, they are just as qualified as men (Eagly & Karau, 2002; Heilman, 1983).

The impact of ILTs is not limited to ratings of effectiveness, as they also influence important outcomes, such as job satisfaction, which increases when employees see their leaders as conforming to their expectation of a prototypical leader (Epitropaki & Martin, 2005). Taken together, the literature on ILTs suggests that people have an intuitive idea of what a leader is (and isn't), which guides their judgments of actual leaders, and which differs from their mental representation of a non-leader.

Importantly, ILTs (and other aspects of leadership perception) represent subjective beliefs, and thus need to be distinguished from objective leadership qualities. Just because an individual is widely perceived to be a particular kind of leader or to possess certain traits, does not mean this is actually the case. Indeed, there is plenty of evidence that leader perceptions can be systematically wrong, and even bias high-stakes leadership selection decisions. For example, research has shown that the selection of political leaders (Olivola & Todorov, 2010; Olivola et al., 2012, 2018) and business leaders (Graham et al., 2017; Stoker et al., 2016) may be biased by appearance-based impressions that do not correspond to reality (e.g., Republicans voting for a Democratic candidate who has a more stereotypically Republican-looking face than the *actual* Republican candidate – Olivola et al., 2012, 2018). Consequently, ILTs—and the current paper—concern people's beliefs and perceptions of leaders, not those leaders' actual characteristics and performance.

The ILT literature has also uncovered a number of attributes that people tend to associate with leadership. Perhaps the most prominent leadership attribute classification is that of Offermann et al. (1994), who asked participants to list traits associated with a leader, a supervisor, and an effective leader, and did so without providing explicit definitions of these categories or cuing a specific person from these categories. That study revealed eight dimensions of leadership: Sensitivity, Dedication, Tyranny, Charisma, Attractiveness, Masculinity, Intelligence, and Strength. Importantly, the factor structure of these eight dimensions did not depend on whether participants evaluated a leader, an effective leader, or a supervisor. However, as might be expected, a separate sample of participants rated leaders and effective leaders higher on Sensitivity, Dedication, Charisma, and Intelligence, compared to supervisors. Leaders and supervisors were rated higher on Tyranny compared to effective leaders, whereas effective leaders were rated highest in Strength, followed by leaders, and lastly

supervisors. On the other hand, leaders, effective leaders, and supervisors, were not rated differently in terms of their Masculinity and Attractiveness. Importantly, these eight dimensions of leadership seem to have withstood the test of time: a study published ten years later, involving a diverse sample of employees tested over time, supported the original eight-dimensional factor structure (Epitropaki & Martin, 2004); and a recent follow-up study, carried out more than twenty years after the original, mostly found support for this same factor structure (Offermann & Coats, 2018), with the addition of one more dimension: Creativity.

Computational methods in leadership research

Building on the idea that people have a shared understanding of the traits associated with effective leadership, and therefore intuitively form judgments concerning a given individual's potential to be an effective leader, this paper proposes a novel, automated method for uncovering which individuals, and groups of individuals, are more (vs. less) likely to be perceived as effective leaders. A central contribution of this method is its ability to identify the particular evaluative dimensions (e.g., traits, concepts, and attributes) that people more frequently associate with effective leadership. Critically, this automated method relies on novel machine learning techniques and new types of 'big data' not previously used in leadership perception research.

The explosion of big data over the past decade has yielded many novel research opportunities. Large-scale datasets of human activity obtained from the Internet, in particular, have been used to study various aspects of human psychology and behavior (e.g., Griffiths, 2015; Harlow & Oswald, 2016; Hawn, 2009; Jones, 2017; Moat et al., 2014, 2016). Yet, despite their potential, big data and machine learning techniques remain highly uncommon in organizational and management studies (George et al., 2016; Tonidandel et al., 2018; Wenzel & Van Quaquebeke, 2018), and rarely used in leadership research in particular, with only a few exceptions (e.g., Batistič et al., 2017; Choudhury et al., 2019; Doldor et al., 2019; Harrison et al., 2019; La Bella et al., 2018; Sieweke & Santoni, 2020; Spisak et al., 2019; Wang et al., 2017; Zhu et al., 2019). For example, Doldor et al. (2019), Sieweke and Santoni (2020), and Wang et al. (2017) apply computational, data-driven topic modeling to text data in order to identify the major recurring topics (or "themes") in leadership research literatures (Sieweke & Santoni, 2020; Wang et al., 2017) and the feedback comments given to political leaders in the UK (Doldor et al., 2019). Batistič et al. (2017) and Zhu et al. (2019) carry out co-citation and co-occurrence analyses of the leadership research literature, in order to visualize this literature and track how it has evolved. Spisak et al. (2019) apply machine learning techniques to analyze how contextual variables and self-reported personality-trait ratings are related to (numerical) 360-degree performance evaluations for managers. Choudhury et al. (2019) combine multiple data types (text and video) and computational techniques (data-driven topic modeling, dictionary-driven sentiment analysis, and convolutional neural networks) to identify distinct CEO communication styles. As a final example, Harrison et al. (2019) apply an approach similar to ours (word embedding models that are fine-tuned on training data) to transcripts of CEO speech during earnings calls in order to develop an indirect linguistic measure of CEOs' Big Five personality traits.

The method we propose in this paper combines several techniques that are related to those used in those prior papers. Like Choudhury et al. (2019), Doldor et al. (2019), Sieweke and Santoni (2020), and Wang et al. (2017), we analyze text data using bottom-up (i.e., data-driven) computational methods. In addition, we "sharpen" our model predictions by applying machine learning algorithms to participant ratings data, like Spisak et al. (2019). As mentioned above, Harrison et al. (2019) come closest to our approach, as they also utilize word embedding models (by applying Word2Vec to transcripts of CEO speech) that are fine-tuned (through gradient boosting) to training data produced by a small sample ($N = 3$) of trained human coders. However, unlike all of this prior work, our goal is to understand and predict lay leadership perceptions. For this reason, we use very large-scale natural language datasets of news articles, as well as novel computational techniques to derive semantic representations from these datasets. These representations mimic what people know and associate with various individuals, and thus can be used as inputs into an algorithm for predicting leadership perceptions. For example, in the current paper, we use this method to predict the extent to which prominent individuals are widely perceived to be effective leaders.

Beyond predicting high-level leadership-related judgments, we also demonstrate how this method can provide insights regarding the particular traits and concepts that people tend to associate with (their notion of) effective leadership. For example, we show that this approach is able to "re-discover" the original ILT leadership traits (Offermann et al., 1994), as well as uncover other traits that lay people associate with effective leadership. Moving beyond traits, we combine our method with semantic dictionaries (Pennebaker et al., 2001; Warriner et al., 2013) to identify the psychological concepts that correlate with perceptions of effective leadership.

Word embeddings

The crucial insight underlying our approach is that people's knowledge of the world around them, including of actual and potential leaders, is reflected in the word co-occurrence statistics of the natural language and text data they produce and are exposed to (Firth, 1957; Harris, 1954). In the context of leadership perception, well-known individuals who are similar to each other are frequently mentioned together in speech and text. Likewise, traits and characteristics most associated with these individuals are usually mentioned alongside these individuals, including traits used to guide leadership perceptions. Thus, by analyzing frequencies of word co-occurrences between individuals, as well as between individuals and specific traits, it is possible to uncover *semantic representations* for individuals, in order to predict how they are perceived in terms of leadership ability.

It is important to note that people's knowledge about specific individuals is situated in their broader knowledge about the world, so that a potential or actual leader is not only associated with a small set of closely-related individuals and characteristics, but also with a vast array of places, objects, and concepts. To fully characterize, and to most accurately predict, leader perceptions, semantic representations obtained from natural language word co-occurrence statistics also need to proxy the broader knowledge structures that people possess. Fortunately, recent advances in computational linguistics have made the recovery of such broad representations feasible. The most promising methods for this type of analysis involve *word embeddings* (Landauer & Dumais, 1997; Mikolov et al., 2013; see also reviews in Bhatia et al., 2019; Jones et al., 2015; Lenci, 2018; Mandera et al., 2017; Turney & Pantel, 2010; Young et al., 2018). Word embedding models represent words and phrases (such as the names of individuals and personality traits, as well as various places, objects, and concepts) as vectors in a multi-dimensional space. These vectors are typically obtained through a form of dimensionality reduction on large-scale natural language word co-occurrence data. The use of dimensionality reduction implies that closely-associated words—which tend to be discussed and referred to in similar contexts in human language—have similar representations, and are thus closer to each other in the resulting space. Similar to factor analysis, such an approach also identifies relationships between words and phrases that may be related to each other without directly co-occurring. For example, two individuals who are seldom mentioned in the same context, but nonetheless co-occur systematically with a set of common traits and attributes, would be assigned similar representations.

Since word embedding models are built on very large language datasets, they are able to uncover and quantify rich representations

for nearly all words and phrases used in language. Additionally, although embedding models perform a type of dimensionality reduction, the overall dimensionality of the vectors is usually very large (e.g., 300 dimensions), implying that they possess rich and nuanced knowledge about the world. For this reason, representations generated by word embedding models have had great success at predicting human similarity judgments, categorization, cued recall, free association, and other memory-based phenomena studied by psychologists (see Günther et al., 2019; Jones et al., 2015; Lenci, 2018; Mandera et al., 2017, for reviews). Unsurprisingly, these representations are also useful for modeling language use in humans (e.g., Garten et al., 2018), and, in turn, for facilitating natural language processing in machines (see Turney & Pantel, 2010; Young et al., 2018, for reviews). Although much of the existing work has applied these techniques to core topics in the study of language, memory, and cognition, some applications of this approach have branched out. Bhatia (2017a) and Bhatia and Walasek (2019) extended this approach to study high-level judgments involving real-world objects and events, including probability judgments and forecasting. A number of recent studies applied a variant of this technique to model the stereotypes and prejudices that bias social judgments (Bhatia & Bhatia, 2021; Bhatia, 2017b; Caliskan et al., 2017; Caliskan & Lewis, 2020; DeFranza et al., 2020; Garg et al., 2018; Lewis & Lupyan, 2020), or to study ideological biases in media representations (Bhatia et al., 2018; Holtzman et al., 2011; Hopkins, 2018). Bhatia (2019a) used this technique to study how associations influence which objects come to mind in memory-based decisions, and Bhatia and Stewart (2018) used it to study how people weigh attributes in naturalistic multiattribute choices. Bhatia (2019b) has shown that word embedding models are able to predict the perceived riskiness of various hazards with a high degree of accuracy. Most recently, Bhatia and Olivola (2021) used this technique to predict how consumers perceive a wide range of product brands, and Richie et al. (2019) used it to predict people's responses to a wide array of judgment problems across different domains in psychology. In all of these applications, word embedding models are used to proxy people's knowledge representations regarding everyday objects and concepts, and to predict judgments about these objects and concepts (see Bhatia et al., 2019 for a review).

Overview of our approach

Here, we apply word embeddings to leadership perception. Specifically, we use a word embedding model trained on a very large dataset of news articles, to derive high-dimensional vector semantic representations for prominent individuals. We then use these representations to predict how people evaluate these individuals in terms of leadership ability. This involves training a machine learning algorithm on participant ratings for some target individuals, in order to extrapolate ratings for all individuals present in our embeddings vocabulary (i.e., all individuals for whom we have word embedding representations). Fig. 1 provides an illustration of our approach.

To test the validity of this approach, we examine how well it predicts the leadership effectiveness ratings given by survey participants (i.e., actual human judgments). Specifically, our first set of analyses compare our model predictions for hundreds of leaders with participant data for those same leaders. Our proposed method can also be applied on a very large scale with minimal effort, and we demonstrate its practical value and power by extending the approach to predict leadership effectiveness perceptions for thousands of individuals in the public sphere (for whom we did not collect participant data).

Beyond merely predicting leadership effectiveness judgments, another important feature of this method is that it also allows us to uncover the psychological cues that best predict these judgments, thereby providing insights into the specific correlates of leadership perception. These cues take the form of implicit associations with various traits, concepts, and attributes, and can be contrasted with established theories of leadership perception, such as ILTs. If our algorithm truly captures core aspects of the human psychology of leadership perception, we would expect it to possess the same set of associations identified by these theories. We therefore test for these associations, in order to see whether our approach can "re-discover" the ILT and personality traits associated with leadership effectiveness judgments (Judge & Bono, 2000; Offermann et al., 1994).

Importantly, the associations uncovered by our model can be used, not just to re-examine prior findings, but also to motivate novel research on leadership perception. Using our approach, it is possible to test for associations between perceived leadership effectiveness and a nearly unlimited set of words (including words pertaining to established constructs in psychology and leadership research). Consequently, we can algorithmically measure the sets of traits, concepts, and attributes that are most associated with perceived leadership effectiveness in natural language and, by doing so, develop hypotheses that can be tested further in more controlled settings, such as laboratory or field experiments. These associations can also help us intuitively better understand the "black box" of word embedding features that best predict leadership perception. Here, we apply this approach to two established semantic dictionaries: (i) the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2001) and (ii) a dictionary of valence-arousal norms (Warriner et al., 2013). The former is a comprehensive repository of psychological constructs, including those related to social and personal concerns, as well as affective and cognitive processes. The latter codes the valence and arousal of nearly 14,000 English words. In this way, this method allows for a comprehensive and flexible examination of factors that lay people associate with effective leadership, and which are manifested in language.

Finally, as previously noted, we stress that this approach is designed to uncover how leaders are *perceived* by people, and this is altogether different from the task of determining the actual (i.e., objective) qualities and behaviors of leaders—which our proposed method is not designed for. Thus, whereas our approach can predict, for example, the extent to which a given leader is widely perceived to be, or to have been, effective, it has little to say about that individual's actual leadership ability and effectiveness. That said, some influential theories of leadership assign a critical role to the perception of leadership. For example, Lord and Maher define leadership as "the process of being perceived by others as a leader" (Lord & Maher, 1991, p. 11), and they argue that the mere perception of leadership, whether based on behaviors or traits, is sufficient to influence others. Consequently, being able to predict the extent to which individuals are widely perceived to be leaders may, to a large extent, be equivalent to predicting whether those individuals are, in fact, leaders. Therefore, our method can also be considered a means to predict actual leadership (i.e., influence), and not just the perception (and thus potential illusion) of leadership.

Methods

Datasets

Word embeddings

Although there are many well-known word embedding models based on natural language data, we chose to use the Google News embedding space for our analysis. This is a publicly available set of 300-dimensional vector representations for over 3 million words and concepts. Importantly, it contains representations of multi-word phrases, including names of well-known individuals (e.g., "*Barack Obama*"). This is in contrast to other popular pretrained models (e.g., those trained on the Common Crawl dataset – Bojanowski et al., 2017; Pennington et al., 2014), which mostly contain representations for single words (e.g., "*Barack*" or "*Obama*"), and thus cannot be used to identify the representations of individuals typically referred to by their first and last names.

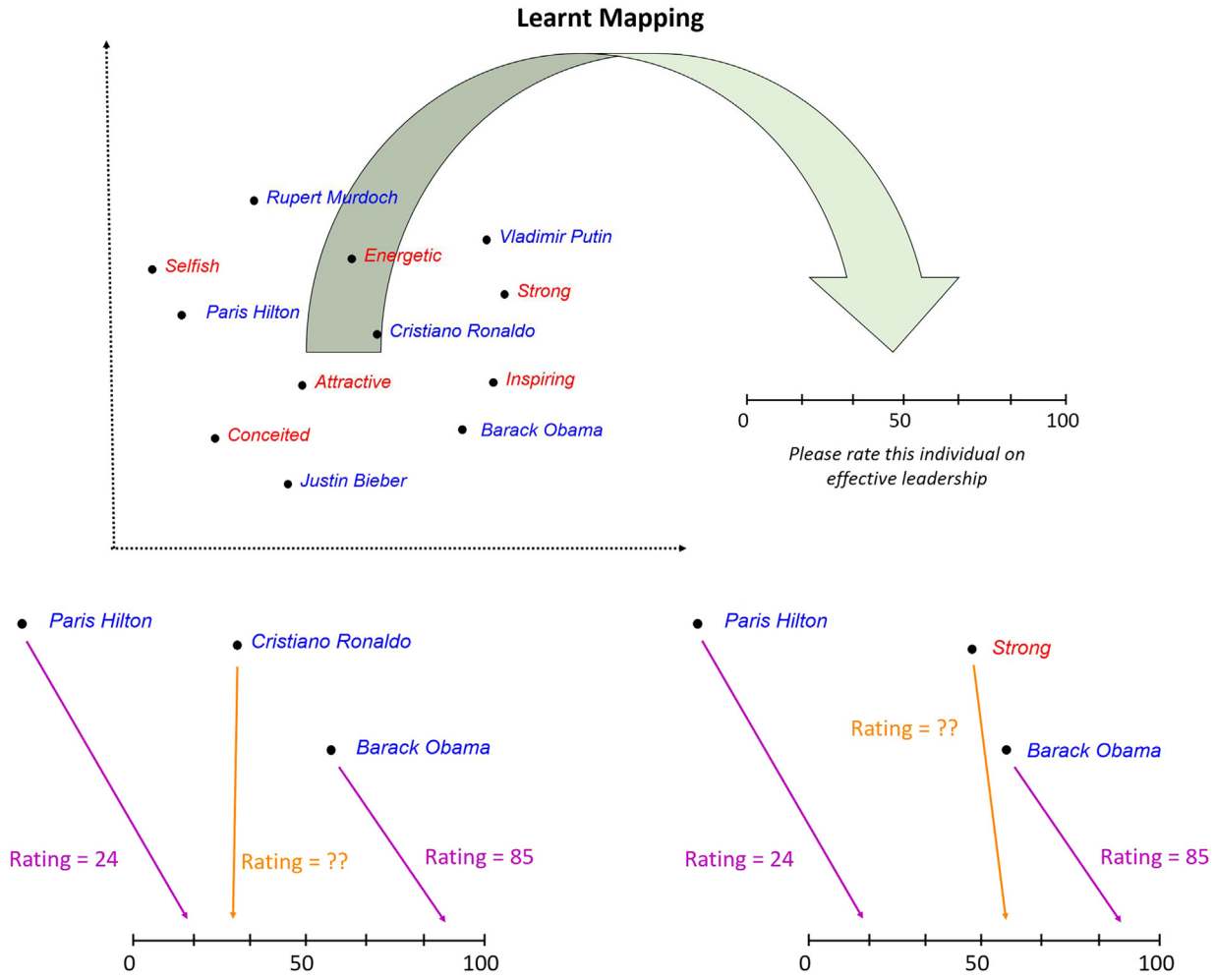

Fig. 1. Illustration of our modeling approach to predicting leadership ratings. The top panel displays a hypothetical two-dimensional space with representations for various well-known individuals (blue) and for various human traits (red). Our approach involves learning a mapping from this space to a rating scale for leadership effectiveness. The bottom-left panel illustrates how participant ratings obtained for some individuals (purple arrows) can be used to predict the ratings for other individuals (orange arrow), while the bottom-right panel illustrates how this approach can be used to predict ratings for other traits (orange arrow). The space used in our model has 300-dimensional representations for over 3 million words and phrases, which include the names of more than 6000 well-known individuals.

Additionally, the Google News embedding space is derived from a very large (100-billion-word) corpus of news articles collected by Google News until 2013. This corpus consists of articles from numerous news outlets from around the world, making it particularly suitable for studying lay representations of popular individuals, including existing leaders, which are likely informed by exposure to news media. Another advantage of this corpus is that it does not include many *obscure* historical figures (see our discussion in the subsequent sections), and thus mirrors the fact that these historical figures are unlikely to be known to our human participants. In this way, there is a close match between the information that our (contemporary) human participants are exposed to and the information contained in the Google News corpus, making embeddings derived from this corpus suitable for our analysis.

The Google News embeddings were trained using the Word2Vec method (Mikolov et al., 2013), which applies a combination of the continuous bag-of-words and skip-gram algorithms. The former algorithm attempts to generate vector representations for words in order to best predict a word based on its linguistic context. The latter algorithm attempts to do the inverse of this: generate vectors to predict neighboring words of a given word. Together, these two algorithms generate embedding spaces in which words that appear in similar contexts and share related meanings are located close to each other. In prior work, embedding spaces derived from the Word2Vec method, such as the Google News space, have been shown to accurately predict human similarity judgments and associative judgments. In particular, people judge words located close to each other in the Google News space as being more similar to, and more strongly associated with, each other (Bhatia, 2017a; Mandera et al., 2017; Pereira et al., 2016). These findings indicate that the Word2Vec method can be used to uncover the kinds of representations people have of the world around them.[2]

[2] Note that research in natural language processing has begun to utilize context-aware embedding training methods that rely on transformer architectures, such as Google BERT (Devlin et al., 2018). Although these are undoubtedly powerful methods, they are most useful for vectorising sentences rather than individual words or short phrases. In order to obtain vector representations for prominent individuals using these methods, it would be necessary to place the name of the individuals in a sentence (i.e., context), yet it is not clear how best to do this without introducing biases into our model. Additionally, as with the two Common Crawl models cited above, context-aware models do not typically contain specialized representations of multi-word phrases. Thus, even though these models can give embedding outputs for phrases like *"Barack Obama"* the outputs are based on the individual representations of the words rather than a holistic representation of the full name (i.e., representations for "*Barack*" and "*Obama*" rather than for "*Barack Obama*"). Although the individual word representations do interact, they are unlikely to capture the complete semantics of the full name. This is in contrast to the Google News Word2Vec model, which contains a specialized representation for many full names.

Pantheon individuals

We utilized the Pantheon 1.0. dataset in order to obtain the set of individuals over which our model could be applied (Yu et al., 2016). This dataset of 11,341 globally famous individuals was compiled using biographies from *Wikipedia* pages written in more than 25 different languages, and was edited using additional online resources. In addition to the names of these individuals, the Pantheon dataset also contains their date of birth, country of origin, gender, domain, industry, occupation, and a metric indicating each individual's popularity. Manual verification by Yu et al. (2016) ensured that the information in this dataset is highly accurate.

We searched the vocabulary of our Google News embeddings to find embedding representations for these 11,341 Pantheon individuals. As the vast majority of these individuals have both first and last names, most of their representations in the Google News vocabulary took the form of multi-word phrases separated by an underscore (e.g., "*Barack_Obama*"). Names represented by a single word in the Pantheon dataset, however, were also associated with single words in the Google News vocabulary (e.g., "*Plato*"). For each individual, we searched the vocabulary for the underscore-separated multi-word or single-word phrase corresponding to their name in the Pantheon dataset, with the first letter of each word uppercased (e.g., "*Barack_Obama*" or "*Plato*"). If we found a match for this phrase, we used the corresponding vector in our analysis. If not, we searched for the full-lowercased phrase corresponding to the name (e.g., "*barack_obama*" or "*plato*") and used the corresponding vector. If this too did not reveal a match, we searched for the full-uppercased phrase (e.g., "*BARACK_OBAMA*" or "*PLATO*") and used that vector. Finally, if all three of these letter-case formats failed to turn up a match, we removed the name from our list. Overall, 6,627 of the 11,341 Pantheon individuals were present in the embedding vocabulary, and thus used in our analysis. For each of these individuals i, we obtained a 300-dimensional vector representation $\boldsymbol{x_i}$, which was normalized to unit magnitude (i.e., $||\boldsymbol{x_i}|| = 1$).

These 6,627 individuals had a mean birth year of 1784 (SD = 629.67; *range* = 3500 BCE to 2008), and a median birth year of 1946. Many of these individuals were from the USA (n = 1,595) and Great Britain (n = 604). There were 1,106 women, 5,484 men and 37 individuals without a known gender. They spanned all eight major domains of the Pantheon dataset, with the most commonly represented domain being the arts (actors, musicians, film directors, designers, artists, etc.), followed by sports (soccer players, chess masters, coaches, etc.), institutions (politicians, diplomats, military personnel, religious figures, etc.), science and technology (biologists, mathematicians, inventors, etc.), humanities (writers, philosophers, etc.), public figures (social activists, celebrities, criminals, etc.), business and law (business people and lawyers), and exploration (explorers and astronauts). These individuals also spanned 27 different industries (with the most common being film and theater, followed by government) and 87 occupations (with the most common being politician, followed by actor). Below, we carry out our analysis for subsets of these individuals, based on their gender and domain. Variables such as birth year, country of origin, industry, and occupation, have many sub-categories with few data points, making it difficult to analyze these variables in a statistically reliable way (though we discuss the implications of our results for these variables whenever possible).

The subset of 6,627 individuals analyzed in this paper has a similar distribution to the full Pantheon dataset. Specifically, the top-three most common countries, birth years, industries, and domains, are the same for both datasets. The largest difference between the two datasets involves the popularities of the individuals. The reason for this is that the Word2Vec vocabulary for multi-word phrases is limited to highly frequent phrases, and individuals whose names are mentioned infrequently in natural language are unlikely to have embedding representations. Subsequently, the average popularity of individuals, measured using the Pantheon historical popularity index (HPI), in the full Pantheon dataset (HPI = 2,138,981) is substantially lower than that of individuals in the subset analyzed in this paper (HPI = 3,064,051). One corollary of this is that the subset of individuals analyzed in the current paper has slightly fewer historical figures. Overall, the average birth year in the full Pantheon dataset is 1655, whereas the average birth year in the subset analyzed in this paper is 1784. Correspondingly, the subset of individuals analyzed in the current paper has slightly more women (as women are underrepresented in earlier time periods). The proportion of women in the full Pantheon dataset is 13%, whereas the proportion of women in the dataset analyzed in the current paper is 17%. Other differences between the datasets pertain to the domain: There are more individuals from the arts and sports domains, and fewer individuals from the institutions and science-technology domains, in our analyzed dataset. Although these are systematic differences, we do not believe that they hurt the quality of our analyses. In the supplemental materials Table S1, we summarize the birth centuries, domains, and genders of individuals included in, and excluded from, the analyses in the current paper.

Leaders

Finally, we understand that not all individuals in our dataset would be considered leaders according to most definitions of leadership. For example, we have actors, popular musicians, and entertainers in the dataset who are famous but are not necessarily leaders. In order to address this issue, we divided our data based on the occupational categories; specifically, in terms of those who would be thought of as leaders vs. non-leaders. When making this distinction, we relied on extant literature that lays out necessary components of leadership. Specifically, we considered individuals in occupations where they can set directions, command a following, and influence the outcomes of others (Middlehurst et al., 1993) as leaders. We should note that this definition also encompasses informal leaders (Hackman, 1992; Neubert & Taggar, 2004; Neubert, 1999). Informal leaders are those perceived as leaders in the eyes of other organizational members, and they command the influence that comes with such a perception, despite not necessarily being formally designated as leaders or holding a hierarchical position that would be associated with leadership. Nevertheless, either due to characteristics that ascribe them status (for example, being male) or to processes that allow them to achieve such a status (such as being seen as competent), informal leaders occupy a position comparable to formal leaders in the eyes of followers and other organizational members (Arnoff & Wilson, 1985). In this way, the idea of informal leadership is closely related to implicit leadership and, as such, our classification includes both formal and informal leaders.[3] Specifically, within our dataset, the occupations categorized as "leaders" include politician, military leader, religious leader, businessperson, scientist, and other occupations, such as social activist, that fit the above definition (see Appendix C for the full lists of occupations, in our dataset, that fall under the "leaders" and "non-leaders" categories, respectively). Therefore, we refer to individuals in these occupations as "leaders", though we recognize that whether or not an individual is considered a leader depends on a number of important factors, and also that not all members of this set are leaders in the conventional sense, and that not all leaders are necessarily in (one of) these occupations. In our analysis, we contrast our results for this "leader" set of individuals with our results for individuals who fall outside one of these "leader" occupations. For simplicity, we refer to members of this latter group as "non-leaders".

Leadership ratings survey

The Google News embeddings for the 6,627 individuals in our dataset allow us to describe these individuals using high-dimensional

[3] The inclusion of informal leaders in our "leader" classification finds additional support in influential leadership theories that largely equate leadership status with being perceived as a leader (Lord & Maher, 1991). According to this compelling conceptualization of leadership, even in the absence of a formal leadership position, merely being perceived as a leader grants influence and leadership status.

vector representations, $\boldsymbol{x_i}$. These representations can, in turn, be used to predict lay participants' ratings of these individuals on leadership efficacy. However, in order to a build model capable of predicting participant ratings using our quantitative embedding representations, we first needed training data. We obtained these training data by asking 210 online participants (99 female, 105 male, 6 other; mean age = 32), recruited from Prolific Academic, to rate a subset of 299 target individuals[4] (drawn from our pool of 6,627 individuals) in terms of leadership effectiveness. Specifically, each participant was presented with the names of 50 target individuals, and asked to rate how effective of a leader they thought each target individual is, was, or would be, on a 0–100 scale. Our focus on effective leadership was based on the results of Offermann et al. (1994), which we use in our analysis below. As with Offermann et al., our participants were explicitly instructed to provide their ratings based on their personal belief of what makes a good leader. They were allowed to select an "I don't know [this person]" response option whenever they did not recognize a target individual's name. The target names were presented on separate pages, and in random order. Our participants were all US citizens with a Prolific Academic approval rating of 90% or higher. They were overwhelmingly Caucasian (160 Caucasian, 50 non-Caucasian) and nearly half identified as Democrat (100 Democrat, 28 Republican, 82 other).

The 50 target individuals presented to each participant were randomly selected (without replacement) from the larger set of 299 target individuals used in our study. These target individuals were selected based on two factors. First, they were the most popular individuals (as determined by the number of views their Wikipedia page received between 2008 and 2013) in our dataset of 6,627 Pantheon individuals with representations in the Google News vocabulary. This was done to ensure that our participants were familiar with the individuals they were rating. Second, we constrained the set of targets to have a maximum of 75 individuals from the arts domain. That is, we initially started with a larger set ($N > 299$) of the most popular individuals, and removed all but the 75 most popular artists until we were left with 299 targets in total (i.e., the 75 most popular artists, along with the 224 most popular individuals who were not classified as artists[5]). This was necessary, as the most popular individuals were predominantly musicians and actors/actresses, and we wanted to ensure that we also obtained participant ratings for political leaders, scientists, athletes, and other occupations, to cover a broad range of leaders and leadership domains (e.g., Eubanks et al., 2016; Olivola et al., 2014). The resulting 299 target individuals had a mean birth year of 1809 (SD = 510.51), and a median birth year of 1954. The oldest individual in this set was Moses (estimated birth year: 1391 BCE), and the youngest individual was Justin Bieber (birth year: 1994). As with the original dataset (of 6,627 individuals) from which this set (of 299 rated targets) was derived, the most common country of origin was the USA, the most common domain was the arts, the most common industry was government (followed closely by film and theatre), and the most common occupation was politician. These rated targets were comprised of 67 females, 222 males, and four individuals of unknown gender.

Although the above procedures should yield an average of about 35 ratings per target individual, there were six target individuals for whom we obtained fewer than four ratings each: Sachin Tendulkar, Carlos Slim Helú, Didier Drogba, Salman Khan, Sunny Leone, and Thierry Henry. These are all non-Americans, who were unfamiliar to our US participants, and received mostly "I don't know [this person]" responses instead of leadership ratings. We excluded these six targets from the set used to train our predictive model. Thus, our predictive model was built using data obtained for the remaining 293 target individuals.

Predictive model

The Google News embeddings and the survey data give us a set of individuals who are described by 300-dimensional vectors and rated in terms of leadership effectiveness. We can use these data to build a computational model that maps points in the Google News embedding space to human ratings, and can thus be used to extrapolate from the subset of individuals rated in our survey (N = 293) to make leadership rating predictions for the much larger set of individuals in our embedding space (N = 6,627). Due to the high dimensionality of our embeddings, we used a ridge regression for our model. This is a type of regularized regression that assumes a linear relationship between the predicted variable (in our case, mean participant ratings, y_i) and observable features (in our case, embeddings for target individuals, $\boldsymbol{x_i}$), but penalizes the model fits based on sum-of-squares of the weight vector. This penalty avoids multicollinearity problems that commonly arise with high-dimensional data. It also allows us to have more embedding dimensions (300) than we do observations for training the model (N = 293). More specifically, our ridge regression attempted to learn a weight vector $\boldsymbol{w}$ that weighs and aggregates embeddings $\boldsymbol{x_i}$ to predict mean participant ratings $y_i \sim \boldsymbol{w}\cdot\boldsymbol{x_i}$. The weight vector $\boldsymbol{w}$ was learnt by minimizing the loss function $||y_i - \boldsymbol{w}\cdot\boldsymbol{x_i}||^2 + \alpha\cdot||\boldsymbol{w}||^2$. Here, $||\cdot||$ specifies the Euclidean norm and α is a hyperparameter that determines the weight on the ridge penalty.[6] In our analysis, we fit the ridge regression on our data using Python's scikit learn machine learning library (Pedregosa et al., 2011). We adopted all of scikit learn's default assumptions regarding the implementation of the ridge regression (including setting $\alpha = 1$ and allowing for a flexible additive intercept for predicting y_i).

The ridge regression applied to the 300-dimensional embeddings permits a large degree of flexibility in our model fits. This can lead to fits that appear to predict our observed participant data well, yet do not generalize well to out-of-sample data. To avoid this issue, we measured predictive accuracy using leave-one-out cross validation. In a setting with N total observations (e.g., N = 293 target individuals for whom we have both participant ratings and embedding representations), this approach divides the dataset into two portions: training data (N – 1 observations) and test data (one observation). It then uses the training data to fit the ridge regression. The best-fit weights from this regression are then combined with the embedding representation for the test observation, to make a prediction regarding the participant rating for the test observation. This procedure is repeated N times, so that each of the N observations are part of the test data once. The predicted ratings, for each target individual, obtained using this method are then contrasted with the observed ratings given by participants, and standard measures like correlation and mean-squared-error are used to calculate the accuracy of the model's predictions. The benefit of this method is that its measured prediction accuracy is a good indicator of how well the model would predict ratings for truly out-of-sample individuals (such as the 6328 individuals for whom we did not obtain participant ratings data).

After establishing the predictive accuracy of our model through cross validation, we used the entire survey dataset of N = 293 target individuals (for whom we have both participant ratings and embedding representations) to train a general predictive model of leadership perception—i.e., a model that predicts individuals' leadership ratings from their embedding representations. This general predictive model was then used to infer leadership ratings for the remaining 6328 indi-

[4] We had initially decided on a sample of 300 target individuals. The reason for this is that our algorithms require a fair amount of training data, and we had found, in prior work (e.g., Bhatia, 2019b; Richie et al., 2019), that 300 observations are enough to give reasonable fits. Sampling more individuals would have been unnecessary whereas sampling fewer individuals would have likely resulted in poorer model predictions. Unfortunately, after running the study, we realized that our final dataset only had 299 individuals, due to a minor technical error. Since 299 is very close to 300 (our initially intended sample size), we decided not to re-run our study.

[5] Note that all other domains had fewer than 75 individuals in our list.

[6] Our analysis also allowed for an intercept in the regression model.

viduals, for whom we did not obtain participant ratings data. This was done by multiplying the embedding representations of these 6328 unrated individuals with the best-fitting weight vector derived from our model fit on the $N = 293$ rated individuals. We also used our general model to measure the traits most associated with effective leadership by similarly multiplying the best-fit weight vector with embedding representations for trait words and other psychological constructs. We discuss the details of these analyses in the results section below. As robustness checks, we also fit additional regularized regression techniques, as well as additional hyperparameter values for the Ridge regression, on our data. We discuss these in more detail below, though we retain our focus on the default Ridge model to avoid overfitting and post hoc analysis.

Summary

The above sections have outlined a number of techniques for using machine learning and large-scale natural language data to study leadership perception. Here, we have used methods such as word embeddings (specifically, the Google News Word2Vec embeddings) to uncover semantic representations from natural language. These embeddings represent words and phrases as high-dimensional vectors and, by doing so, proxy the structure of knowledge and associations in language, and subsequently in the minds of our human participants. For this reason, they can be used as inputs into machine learning models that map words and phrases (in our case, the names of prominent individuals) onto variables such as perceived leadership effectiveness. In order to perform this mapping, we have used regularized regression techniques, such as ridge regression. These techniques are particularly suitable for big data analysis as they allow for a very large number of variables to enter as predictors. Finally, in order to test our models, we apply cross validation (specifically, leave-one-out cross validation), which separates the entire dataset into two portions: training and test. The former is used to fit the regression model, whereas the latter is used to evaluate its predictive accuracy. Cross validation ensures that the flexibility inherent in big datasets and machine learning models does not lead to overfitting and that the final model is capable of generalizing to novel settings.

Table 1 provides a summary of these techniques. Researchers interested in learning more about them can consult Jurafsky and Martin (2019) for an overview of word embeddings and their use in natural language processing, and Müller and Guido (2016) for a hands-on tutorial applying regularized regression and other machine learning techniques in Python. In addition, popular pretrained embedding models can be easily downloaded and used with the magnitude Python module (https://github.com/plasticityai/magnitude). Finally, we make all our code, data, and analyses available for interested researchers at https://osf.io/52w7r/.

Results

Summary of participant ratings

Participants failed to recognize the target (i.e., selected the "I don't know [this person]" response option) a total of 2,609 times out of the 10,500 responses collected across all participants and trials. Consequently, we obtained a total of 7,891 'leadership effectiveness' ratings from our 210 participants, resulting in an average of 28.5 ratings ($SD = 12.02$) for each of our 299 rated target individuals. As discussed above, there were six target individuals for whom we obtained fewer than four ratings. These were excluded from the set used to train our model, which was thus built using data for the remaining 293 target individuals.

The mean rating given by our participants was 48.56 ($SD = 29.85$) and the median rating was 52. These ratings spanned the entire range of our scale, with a minimum of 0 and a maximum of 100. The highest rated target individual in our dataset was Barack Obama ($M = 85.91$, $SD = 21.34$), followed by Abraham Lincoln, Winston Churchill, and George Washington. The lowest rated target individual in our dataset was the Zodiac Killer ($M = 9.54$, $SD = 14.96$), followed by Ed Gein, Jeffery Dahmer, and Paris Hilton. As Table 2 shows, leadership ratings differed as a function of gender and domain, with women and individuals in the arts receiving the lowest average ratings. Table 2 also shows that individuals belonging to one of our 38 leadership occupations obtained higher leadership ratings, on average, than those belonging to the other ("non-leader") occupations. Summary statistics for specific occupations, industries, birth years, and countries of origin are not shown due to the large number of unique categories for these variables.

Predictive accuracy of our model

Overall accuracy

Our computational model predicted out-of-sample participant ratings extremely well. This is shown in Table 2, which lists the model correlation and root-mean-squared-error values, and illustrated in Fig. 2, which plots average participant ratings for each target individual against our out-of-sample prediction for that target individual. The predictions across all target individuals ($N = 293$) have a large out-of-sample correlation ($r = 0.78$, $p < 0.001$) and small root-mean-squared-error ($RMSE = 10.25$). These correlations are comparable to those reported in previous papers that attempted to predict human judgment with word embeddings. For example, Bhatia (2019b) achieved out-of-sample correlations between 0.70 and 0.84 when predicting perceptions of risk for various hazards, while Richie et al. (2019) achieved out-of-sample correlations between 0.59 and 0.88 when predicting judgments across different psychological domains. Moreover, and as we discuss below, our correlations are very close to the split-half correlation in our dataset, which is the theoretical upper-bound in prediction.

As robustness checks, we also replicated our tests with additional regularization techniques, such as the lasso regression, as well as support vector regressions with radial basis function, polynomial and sigmoid kernels. For these additional techniques, as well as for our basic ridge regression, we considered a wide range of hyperparameter values (rather than the default scikit learn values). We found that accuracy rates could be slightly improved using this type of flexibility, with a maximum out-of-sample correlation of 0.79 for the support vector regression with a radial basis function kernel. These results suggest that our ridge model may not necessarily be the best predictive model, but also that accuracy gains with other techniques are fairly small. Further improvements may be possible with additional techniques, such as the square-root Lasso regression (Belloni et al., 2012), not considered in this paper. For expositional simplicity, we will focus on the ridge regression throughout the rest of this paper. The results we obtain with alternative regularization techniques are presented in the supplemental materials Table S2.

Table 2 and Fig. 2 also provide the results for subgroupings of target individuals separated by gender and domain, as well as for individuals categorized as leaders and individuals categorized as non-leaders (based on their occupation, as previously explained). They show that our main (ridge regression) model maintains high predictive accuracy rates when the analysis is restricted to various subgroupings, suggesting that it can also capture differences in participant ratings within each gender, within particular domains, and for both leaders and non-leaders. In fact, the model-participant correlations exceed $r = 0.6$ for all subgroupings except individuals in the sports domain and those in the business and law domain. Moreover, these correlations are all significantly positive ($p < 0.05$) except for individuals in the business and law domain (only directionally positive). The weak performance of our model in this latter domain is driven partially by

Table 1
Summary of machine learning methods used in the current paper.

Technique	Use	Implementation	Additional Resources
Pretrained Embeddings			
High-dimensional vector representations of words and phrases obtained from word distribution statistics in natural language data.	Simple but powerful method for representing the semantic content of words and phrases, for use as inputs in predictive models.	Google News Embeddings: 300-dimensional vectors for over 3 million words and phrases, trained on a large corpus of news articles using Word2Vec method.	Mikolov et al. (2013) for technical details; Jurafsky and Martin (2019; chapter 6) for a tutorial; Bhatia et al. (2019) for an application to judgment.
Regularized Regressions			
Statistical technique for mapping large numbers of predictor variables onto a continuous output.	Allows for accurate predictions, avoiding multicollinearity problems common in large datasets.	Ridge regression: A linear regression technique which minimizes the loss function: $\|\|y_i - \boldsymbol{w}\cdot\boldsymbol{x}_i\|\|^2 + \alpha\cdot\|\|\boldsymbol{w}\|\|^2$.	Müller and Guido (2016; chapter 2) for a tutorial and implementation in Python.
Cross Validation			
Model validation technique that trains and tests models on separate portions of the full dataset.	Useful for measuring out-of-sample predictive accuracy and generalizability beyond training data. Avoids overfitting problems common in large datasets.	Leave-one-out cross validation: A version of cross validation in which each observation serves as the test data once.	Müller and Guido (2016; chapter 5) for a tutorial and implementation in Python.

its overestimation of the predicted rating for Donald Trump. Both the Google News embeddings and the Pantheon dataset were published in 2013. Thus, the embeddings largely associate him with his business and media activities prior to his entry into politics, and the Pantheon dataset likewise categorizes him as part of the business and law domain (rather than the institutions domain, which is typically reserved for politicians and presidents). By contrast, our participants (who were surveyed in 2019, more than two years into his presidency) likely rated his effectiveness as a political leader, and most of them probably disapproved of his political leadership (recall that a minority of our participants identified as Republican). For these reasons, our trained model gave him a higher than average leadership rating of 52.11, whereas our survey participants gave him an average rating of only 23.13. Excluding Donald Trump from the analysis increases the measured correlation between predicted and observed ratings from 0.18 to 0.36, among individuals in the business and law domain. It is worth noting that the observed correlation of 0.78 for the full sample of 293 individuals does not change (i.e. remains 0.78) if we exclude Donald Trump or even all past US presidents.

Table 2 also shows that our model is able to correctly predict not only the variations in ratings for individuals within subgroupings but also the differences in ratings across subgroupings. Specifically, and in line with our survey participants, the model attaches lower ratings to non-leaders (vs. leaders), women (vs. men), and individuals in the arts domain (vs. other domains). In fact, there is a perfect rank correlation between the model's predicted average ratings across the eight domains and participants' average ratings across those same domains, suggesting that the model can capture differences in aggregate ratings across different subgroupings of individuals with a very high degree of accuracy.

Comparison with alternate metrics and models

As another way to evaluate the predictive power of our model, we contrasted its measured accuracy rate against three alternate metrics. The first is the average pairwise participant correlation. This metric is obtained by selecting pairs of participants and calculating the correlations between their ratings (across multiple pairs) to determine how well, on average, one participant's leadership ratings predict another participant's leadership ratings. In our analysis, we estimated the average pairwise participant correlation by randomly sampling 1000 pairs of participants from our data and measuring the correlation for the target individuals rated by both participants in each pair (typically, far fewer than 50 targets were rated by the same two participants). This resulted in an average correlation of $r = 0.38$ ($p < 0.001$), which is less than half that achieved by our computational model. In other words, our model is better able to predict a given participant's ratings than can the ratings given by another participant from the same population who evaluated the same target individuals.

Our second metric is the expected participant-to-aggregate correlation. This metric quantifies the degree to which the average individual participant's leadership ratings are correlated with the aggregate ratings of all remaining ($M - 1$) participants. We performed this analysis on our 210 survey participants, which yielded a correlation of $r = 0.59$ ($p < 0.001$). This is still smaller than the correlation achieved by our computational model. In other words, our model is better able to predict aggregate participant ratings than can the ratings given by a single participant from the same population who evaluated the same target individuals. This also means that our model is better able to predict individual participant ratings than can the aggregate —and therefore far less noisy—ratings of many other participants drawn from the same population.

Table 2
Statistics of observed and predicted ratings for individuals rated in our survey. 'Count' refers to the total number of rated individuals in a category, whereas 'Obs. Rating' and 'Pred. Rating' indicate, respectively the average observed and average out-of-sample predicted ratings for these individuals. 'Correlation' and 'RMSE' indicate the correlation and root-mean-squared-error between the observed and out-of-sample predicted ratings for the individuals in a category.

	Count	Obs. Rating	Pred. Rating	Correlation	RMSE
All	293	47.15	47.21	.78	10.25
Leaders	134	55.74	55.16	.76	11.13
Non-leaders	159	39.91	40.49	.62	9.45
Male	222	49.76	49.52	.76	10.51
Female	67	39.19	39.97	.77	9.25
Arts	74	37.32	37.54	.65	8.84
Business & Law	12	52.37	48.97	.18	14.34
Exploration	1	63.56	60.40	N/A	3.16
Humanities	29	54.05	53.78	.71	8.04
Institutions	62	59.22	57.36	.74	11.25
Public Figure	43	40.67	43.38	.82	11.88
Science & Technology	16	59.60	58.44	.64	7.59
Sports	56	43.22	44.45	.33	10.21

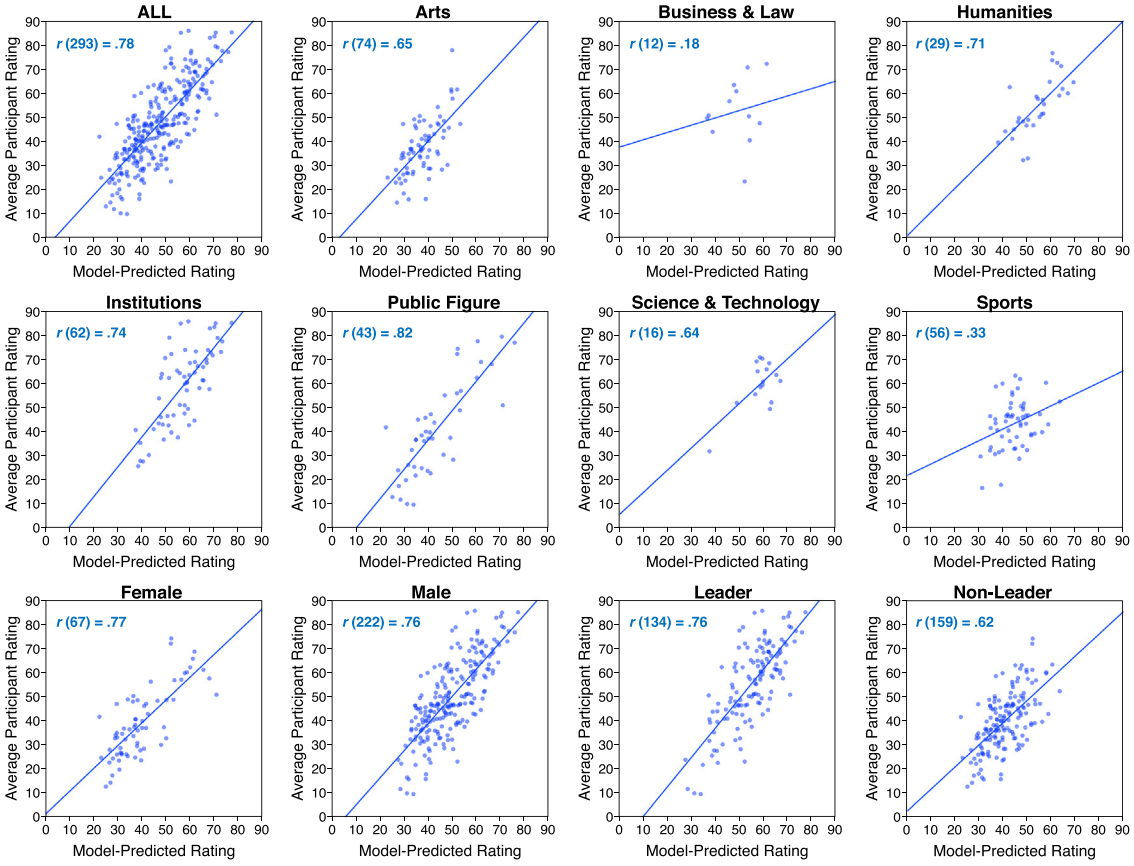

Fig. 2. Participant ratings plotted against our model-predicted ratings, for the full sample of target individuals (top-left graph), and for various subgroups of target individuals (all other graphs). Each data point represents a single target individual, and solid lines represent best linear fits. Standard parametric correlations between the two variables (r) and their associated sample sizes (in parentheses) are also presented.

Our third metric is the split-half reliability in our participant data, which, like the pairwise participant and participant-to-aggregate correlations, is a function of the degree of variability or inconsistency (i.e., non-consensus) across participants. This metric quantifies the degree to which the aggregated ratings of half the participants are correlated with the other half of participants. To obtain the split-half reliability, we randomly divided the participants into two groups, calculated each group's average ratings for the target individuals, and finally calculated the correlation between these two sets of average ratings. The expected correlation generated by this procedure was the split-half reliability. Since it determines how well half the participant sample predicts the ratings of the other half, split-half reliability is a good measure of the maximum predictive power that can be achieved with the existing data (in a very large sample of participants, this would be the theoretical upper bound on the predictive accuracy attainable by any single model applied to the data). We estimated split-half reliability by performing the above procedure (randomly splitting the sample into two equal-sized groups) 1000 times. This yielded an average correlation of $r = 0.79$ ($p < 0.001$), which is only slightly larger than the out-of-sample correlation achieved by our computational model. Thus, our model's ability to predict the aggregate leadership ratings of human participants is especially impressive, and very close to the highest predictive power possible for these data.

We also considered an alternate "meta-data" statistical model for predicting participant ratings. This model uses the individual-level meta-data provided in the Pantheon dataset as independent variables for predicting the aggregate participant ratings data. Specifically, instead of predicting the aggregate rating y_i by applying a ridge regression using best-fit weights $\boldsymbol{w}$ on the 300-dimensional embedding vector $\boldsymbol{x_i}$, this alternate meta-data model applies a ridge regression using best-fit weights $\boldsymbol{v}$ on the 104-dimensional vector $\boldsymbol{z_i}$ generated from dummy variables for gender, occupation, country of origins, and birth century (20th, 19th, 18th, 17th, 16th or earlier) to predict y_i. Here, we used occupation instead of industry or domain, since the former provides a higher level of detail regarding the target individual (the meta-data model's predictions are slightly worse if occupation is replaced with industry or domain). We also used multiple birth century dummies rather than a single, continuous birth year predictor, as doing so improves the meta-data model's predictive accuracy (probably because birth year has a complex, nonlinear relationship with participant ratings). As with our embeddings-based computational model, we implemented the ridge regression using the default parameters in the scikit learn package, and evaluated model predictions using leave-one-out cross validation (we also carried out a standard linear regression for this analysis, but found that its predictions were significantly worse).

Overall, the alternate meta-data model (i.e., ridge regression applied to the dummy variable vector $\boldsymbol{z_i}$) achieved an out-of-sample correlation of $r = 0.64$ ($p < 0.001$). Although this value is higher than the pairwise participant correlation and participant-to-aggregate correlation, it is still lower than the out-of-sample prediction attained using our proposed model. Thus, the Google News embeddings predict human ratings better than manually annotated meta-data for the target individuals.

The top-left graph in Fig. 3 compares our model's correlation (with human ratings) against the correlations achieved by the alternate

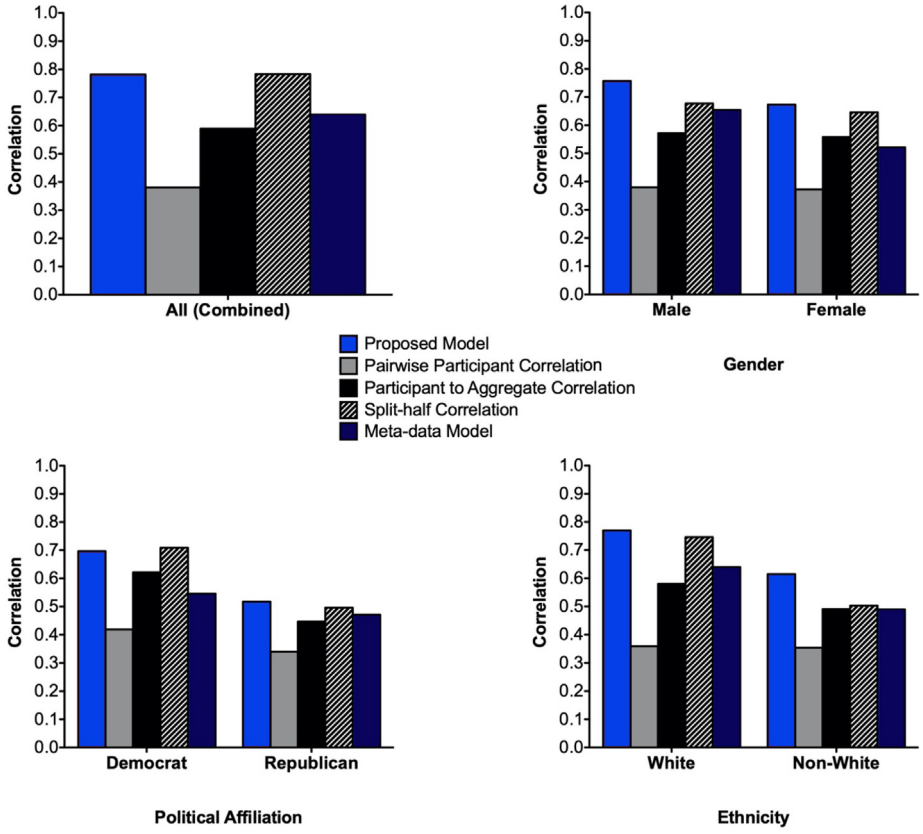

Fig. 3. The predictive power achieved by our model (light blue bars) and other benchmarks, across all study participants (top-left graph), and for subgroups of study participants, divided by gender (top-right), political affiliation (bottom-left), and ethnicity (bottom-right).

meta-data model, as well as the pairwise participant, participant-to-aggregate, and split-half reliability correlations, for the entire sample of survey participants (N = 293).

Analysis of participant subsamples

In the preceding section, we have shown that our model is able to accurately predict aggregate participant ratings for our entire sample of survey participants. However, as previously explained, this sample is skewed in favor of Caucasian Democrats. To test whether (and to what extent) our model's predictive power extends beyond this demographic, we calculated its out-of-sample correlations separately for Caucasian and non-Caucasian participants, as well as for Democrat and Republican participants. We also calculated these correlations separately for male and female participants, to see whether our model's performance depends on the gender of the rater. As with our previous analysis, we compared the correlations achieved by our model with those achieved by the alternate meta-data model, as well as the participant, participant-to-aggregate, and split-half reliability correlations, for each participant subsample.

Fig. 3 presents these correlations, and shows that our model achieves out-of-sample correlations that exceed $r = 0.5$ (all $ps < 0.001$) for every participant subsample. It also shows that our model is poorer at predicting the ratings of female participants, non-Caucasian participants, and Republican participants. This could occur if the Google News embeddings, which we use to represent our target individuals, are more reflective of the beliefs, associations, and knowledge structures held by men, Caucasians, and Democrats. It could also occur if female participants, non-Caucasian participants, and Republican participants have more variable ratings. In fact, there was greater variance in the ratings of our Republican participants than those of our Democrat participants ($M_{var.}$ = 648.66 vs. 571.71); however, women's ratings varied only slightly more than men's ($M_{var.}$ = 631.59 vs. 620.01), while the ratings of non-Caucasians actually varied less than those of Caucasians ($M_{var.}$ = 616.54 vs. 621.54), so systematic differences in judgment variability could only explain why our model is worse at predicting the ratings of the Republicans (vs. Democrats) in our survey sample. Finally, these correlation results could be explained by the simple fact that we have fewer women than men (average number of raters per target individual: M_n = 12.57 vs. 13.58), non-Caucasians than Caucasians (M_n = 6.25 vs. 20.64), and Republicans than Democrats (M_n = 3.49 vs. 12.82) in our survey sample, thereby making the former subgroups' aggregate ratings noisier and thus harder to predict. Consistent with this last explanation, we find that the three alternate metrics (pairwise participant, participant-to-aggregate, and split-half reliability correlations) and the alternate meta-data model's predictive power, are all also lower for the female (vs. male), non-Caucasian (vs. Caucasian), and Republican (vs. Democrat) subsamples (see Fig. 3). Thus, the variations in our model's performance can largely be explained by corresponding variations in participant numbers, which suggests that our model could predict ratings equally well across subgroups if we had equal numbers of raters in each. Note that, although not shown in Fig. 3, we also replicated our tests for the subjects in our study who did not identify as either Republican or Democrat. Our model achieved an out-of-sample correlation of 0.68 for these subjects, indicating that it is able to accurately capture the leadership perceptions of independents.

Fig. 3 also shows that our model's predictive correlations always exceed those of the alternate meta-data model, as well as the pairwise participant and participant-to-aggregate correlations. They also equal, and sometimes exceed, the split-half reliability correlations (we previously mentioned that split-half reliability is an upper bound on predictive performance, but note that this is only the case with a large

number of participant observations; we have only a small number of participants in the Republican and non-white categories, which reduces their split-half correlations). Overall, the favorable performance of our model relative to the alternate model and metrics, for all subsamples of our survey participants, suggests that its predictive power is highly robust, and that it closely mimics the leadership judgments of many different participant demographics.

Extrapolating leadership ratings

One benefit of our computational approach is that we can use the best-fitting model, parameterized with weight vector $\boldsymbol{w}$, obtained from the ridge regression trained on our participant data, to predict leadership ratings for all 6,627 individuals in our embedding space. As with the target individuals used in our survey, each of these 6,627 individuals is represented by a vector $\boldsymbol{x_i}$, which means that $\boldsymbol{w}\cdot\boldsymbol{x_i}$ (plus the linear regression intercept) is the predicted average participant rating for that individual. Obtaining leadership ratings in this manner allows us to study leadership perception on a much larger scale than is possible through human participant data alone.

Figures S1–S3 in the supplemental materials provide graphs of the distributions and summary statistics of the predicted leadership ratings obtained from the best-fitting model trained on our participant ratings, for all 6,627 individuals and for subsets of these individuals (note that some individuals have predicted ratings outside the 0–100 scale used in our study; this happens because the ridge regression model used in our algorithm involves a linear function and thus does not explicitly restrict its predictions to the 0–100 range). The individuals our model predicts will be perceived as the most effective leaders are Jomo Kenyatta, a Kenyan anti-colonial activist and the first president of Kenya (prediction = 163.57), Abraham Lincoln (prediction = 163.39), and Francis Drake, an English sea captain and explorer (prediction = 158.67), while the individuals predicted to be perceived as the least effective leaders are Elisha Cuthbert, a Canadian actress and model (prediction = −71.16), Travis Barker, the drummer of punk rock band Blink-182 (prediction = −57.48), and Deryck Whibley, the lead vocalist and guitarist of punk rock band Sum-41 (prediction = −55.77).

Overall, explorers (individuals in the domain of exploration) obtain the highest average predicted ratings (M = 87.39, SD = 25.10, N = 44), while artists obtain the lowest average predicted ratings (M = 36.56, SD = 28.10, N = 2,282). Out of the 87 occupations in the Pantheon dataset with more than 5 rated individuals, astronauts (M = 95.17, SD = 29.16, N = 16) obtain the highest average predicted ratings, whereas pornographic actors obtain the lowest average predicted ratings (M = −11.40, SD = 21.27, N = 7). Mirroring the responses of our survey respondents, individuals in one of the 38 leadership occupations obtain higher predicted ratings (M = 68.36, SD = 25.69, N = 2,672) than those in the other occupations (M = 37.57, SD = 25.84, N = 3,955).

We also observe systematic time period effects, with individuals born in the 16th century or earlier obtaining the highest average predicted ratings (M = 71.66, SD = 20.95, N = 572), followed by individuals born in the 17th century (M = 70.57, SD = 17.93, N = 64), 18th century (M = 70.37, SD = 23.16, N = 157), 19th century (M = 69.47, SD = 23.80, N = 826), and 20th century (M = 43.39, SD = 28.98, N = 5,008). There are also country-level differences in predicted ratings, though much of this cross-country variance is likely the product of small sample sizes for smaller countries (e.g., Barbados, which obtains the lowest predicted average rating, only has one individual represented in our dataset: Rihanna, whose predicted leadership rating is −17.61).

We also observe substantial gender differences in predicted ratings. Overall, male target individuals obtain an average predicted rating of 53.06 (SD = 28.27, N = 5,484), whereas female target individuals obtain an average predicted rating of 34.72 (SD = 32.60, N = 1,106). This difference of almost 20 points is statistically significant ($t(6{,}588) = 19.14$, $p < 0.001$) and corresponds to a Cohen's d = 0.60; a medium to large effect. Overall, the female individuals in our dataset who obtain the highest predicted leadership ratings are Kalpana Chawla, an Indian-American astronaut who died in the Space Shuttle Columbia disaster (predicted rating = 156.03), Eleanor Roosevelt (predicted rating = 152.48), and Sojourner Truth (predicted rating = 147.12).

One reason why we may be observing gender differences in predicted ratings could be due to domain-level differences in gender. For example, women make up only 11% of the explorers (who obtain the highest predicted leadership ratings), but 29% of the artists (who obtain the lowest predicted leadership ratings), in our dataset. To rigorously control for these confounds, we regressed our predicted ratings on gender, with dummy variables for each of our eight domains. We still found a significant negative coefficient associated with being female: $\beta = -11.84$ ($t = -13.51$, $p < 0.001$, 95% CI = [-13.55, −10.12]), indicating that women are, on average, predicted to receive leadership ratings that are nearly 12 points lower than men, even after controlling for domain. We repeated this regression with additional controls, such as birth century, the Pantheon historical popularity index, and industry, and we obtained similar results ($\beta = -8.06$, $t = -9.58$, $p < 0.001$, 95% CI = [−9.72, −6.41]). Thus, our model predicts a systematic gender bias in leadership perceptions, which is robust to additional controls.

Trait and concept associations

Opening the black box of machine learning

Our computational model was built to learn a mapping from the high-dimensional Google News embedding space to a (one-dimensional) scale representing people's leadership effectiveness ratings. This mapping, which is based on the "knowledge" contained in the Google News text corpus and the training dataset collected through our survey (participants' ratings of 293 target individuals), reflects a set of latent beliefs or associations concerning the nature of effective leadership. Consequently, we can explore this conceptual map, which our model has produced, to address a variety of interesting questions regarding the general beliefs and associations that underlie shared perceptions of effective leaders: What are these beliefs and associations? What are the features of the semantic representations that our model learns to prioritize when judging effective leadership? Answering these questions will also help us open up the black box of our machine learning approach, and better understand the cues that our model relies on to make predictions.

By themselves, the 300 dimensions of our embedding space that our model weighs and aggregates do not have an intuitive interpretation: These dimensions can be seen as reflecting the factor structure of linguistic co-occurrences between words (i.e., concepts), and words that load highly onto one or more of these dimensions are not necessarily relevant to leadership perception. However, the weight vector, $\boldsymbol{w}$, reflects our model's attempt to predict evaluations of leadership effectiveness (i.e., to connect the 300-dimensional embedding space to participants' leadership effectiveness ratings), and thus captures important details about the concepts that our model does associate with effective leadership. In particular, we can obtain the embeddings representation for any given word and concept, and multiply this with the weight vector to obtain a predicted "pseudo-rating" for that concept. Even though the word or concept itself may not be an individual, the predicted pseudo-rating would nonetheless capture the degree to which our model associates that word or concept with leadership. Therefore, words or concepts given high pseudo-ratings would be ones that are especially associated with individuals judged to be high in effective leadership. These words and concepts may include specific occupations and broader career domains, countries and time periods,

as well as personality traits, person descriptors and adjectives, and various other socially relevant cues.

Implicit leadership traits

To evaluate these socially relevant cues in detail, we first obtained the set of human traits identified by prior research on leadership perception, and multiplied the embeddings representations of these traits with our best-fit weight vector (derived from our survey data) to obtain predicted pseudo-ratings for these traits. This allowed us to identify the traits that our model strongly associates with effective leaders. In particular, we used the same set of traits as Offermann et al. (1994). Those authors explored the content and factor structure of ILTs for effective leaders. Their extensive analysis, described earlier in this paper, identified 41 traits that combined to form eight factors: sensitivity, dedication, tyranny, charisma, attractiveness, masculinity, intelligence, and strength (see Appendix A for the full list of traits that make-up each factor). They found that factors such as dedication, intelligence, strength, and charisma were especially associated with effective leaders (in the minds of participants), whereas masculinity, attractiveness, and tyranny were not. Critically, 36 out of these 41 traits are represented in our model's Google News embedding vocabulary. The remaining five traits involve multiple words (e.g., power-hungry, well-dressed) and are not represented in the model's vocabulary (although our model contains representations for common multi-word phrases, such as names, these five multi-word traits are rare enough to be excluded from its vocabulary). For each of the 36 traits in our model's vocabulary, we obtained a vector representation $\boldsymbol{x_i}$ and used our best-fitting ridge regression weights, $\boldsymbol{w}$ (trained on our aggregate participant data) to calculate the pseudo-rating $\boldsymbol{w}\cdot\boldsymbol{x_i}$ (plus the intercept). As we explained above, this pseudo-rating can be seen as a measure of the strength of association between a given trait and the concept of effective leadership.

Fig. 4A plots the predicted pseudo-rating for each of the individual traits in out model's vocabulary. This figure shows that traits such as inspiring, bold, and dedicated are most associated with effective leadership, whereas others such as manipulative, obnoxious, and selfish are least associated with effective leadership. Fig. 4B provides a scatterplot that relates our model's average predicted pseudo-ratings for the traits that make up the eight factors to the mean factor ratings in Offermann et al.'s study (Table 5 in their paper). This figure reveals a strong positive relationship ($r = 0.74$ with $p < 0.05$, across the eight data points in the scatterplot), indicating that our model has learnt many of the same associations between particular traits and effective leadership held by participants in Offermann et al.'s study (despite our model not being explicitly trained on these associations). In other words, our computational model not only predicts peoples' leadership judgments, it does so by implicitly relying on many of the same psychological cues (i.e., associations with traits) that people seem to use when they judge leadership effectiveness. Similar results with alternative regression models are presented in the supplemental materials Table S2.

Personality (Big Five) traits

Prior work has also examined the relationship between personality traits and leadership perception. In particular, Judge and Bono (2000) found—in a large-scale study of leaders spanning over 200 organizations—that some Big Five personality trait (self-) ratings were correlated with ratings of leader effectiveness. Specifically, their results revealed that openness to experience and extraversion were stronger predictors of leader effectiveness than the other Big Five factors. We therefore attempted our analysis with traits that make up the five dimensions of the Big Five personality inventory (Goldberg, 1992; McCrae & Costa, 1987): Openness to experience, conscientiousness, extraversion, agreeableness, and emotional stability. Specifically, and mirroring our analysis with ILTs above, we used the 45 traits comprising the five factors of personality. In particular, we used the traits from Goldberg's Big Five classification (Goldberg, 1992) since this scale uses single-word adjectives (e.g. extraverted), whereas other commonly used Big Five scales, such as the NEO-PI (McCrae & Costa, 1987), are at the sentence level (e.g. enjoy talking with other people) (recall that our model contains only representations for individual words). Goldberg's traits include adjectives such as intellectual and creative (for the openness dimension), organized and systematic (for the conscientiousness dimension), extraverted and talkative (for the extraversion dimension), kind and cooperative (for the agreeableness dimension), and unemotional and relaxed (for the emotional stability dimension). Appendix B provides a full list of the traits associated with each factor.

For each of these traits, we obtained a vector representation $\boldsymbol{x_i}$ and used our best-fitting ridge regression weights, $\boldsymbol{w}$, to calculate the pseudo-rating $\boldsymbol{w}\cdot\boldsymbol{x_i}$ (plus intercept), which is a measure of the strength of association between a given trait and the concept of effective leadership. As shown in Fig. 5A, we found that the traits with the highest predicted pseudo-ratings were bold and innovative. In contrast, the traits with the lowest predicted pseudo-ratings were systematic and unemotional. Fig. 5B provides a scatterplot that relates our model's average predicted pseudo-rating on each of the Big Five personality factors to the correlations between Big Five ratings and leadership effectiveness ratings that Judge and Bono (2000) obtained in their study (Table 2 in their paper). This figure reveals a fairly strong positive relationship ($r = 0.61$), indicating that our model has learnt many of the same associations (or lack thereof) between Big Five personality traits and effective leadership ratings (despite not being explicitly trained on these associations), as those found among the leaders in Judge and Bono's original study. This figure also shows that traits that make up the openness (to experience) and extraversion personality dimensions are more associated with effective leadership ratings than those making up the other three factors, thus replicating the findings of Judge and Bono (2000). Traits that make up the emotional stability dimension, it seems, are least associated (in people's minds) with effective leadership. In other words, individuals who are typically mentioned in the context of openness (to experience) and extraversion traits are more likely to be seen as effective leaders than those mentioned in the context of emotional stability traits. Similar results with alternative regression models are presented in the supplemental materials Table S2.

Psychological (LIWC) constructs

In the previous two sections, we have attempted to measure pseudo-ratings for trait words, and by doing so, understand the characteristics and attributes of individuals who are most associated with perceived leadership effectiveness. In this section, we expand our analysis beyond traits to consider a wider set of psychological constructs that could be relevant to the study of leadership. Importantly, these constructs have received very little, if any, attention in the leadership perception literature. Consequently, although the outcomes of this analysis cannot necessarily be correlated with existing human data, they can nonetheless be used to motivate hypotheses that can be tested in future behavioral and experimental work. In particular, we applied our analysis to the constructs used in the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2001). These constructs span numerous domains, including social concerns (e.g., constructs such as family and friends), personal concerns (e.g., constructs such as money and religion), affective processes (e.g., constructs such as anxiety and sadness), cognitive processes (e.g., constructs such as insight and causation), perceptual processes (e.g., constructs such as see and hear), and biological processes (e.g., constructs such as health and sexuality). Each of these constructs takes the form of a set of words that characterize that construct. For example, the 'family' construct in LIWC has words such as "aunt" and "babies", whereas the 'money' construct has words such as "account" and "audit". The words linked to each construct are often used in automated text analysis to measure the

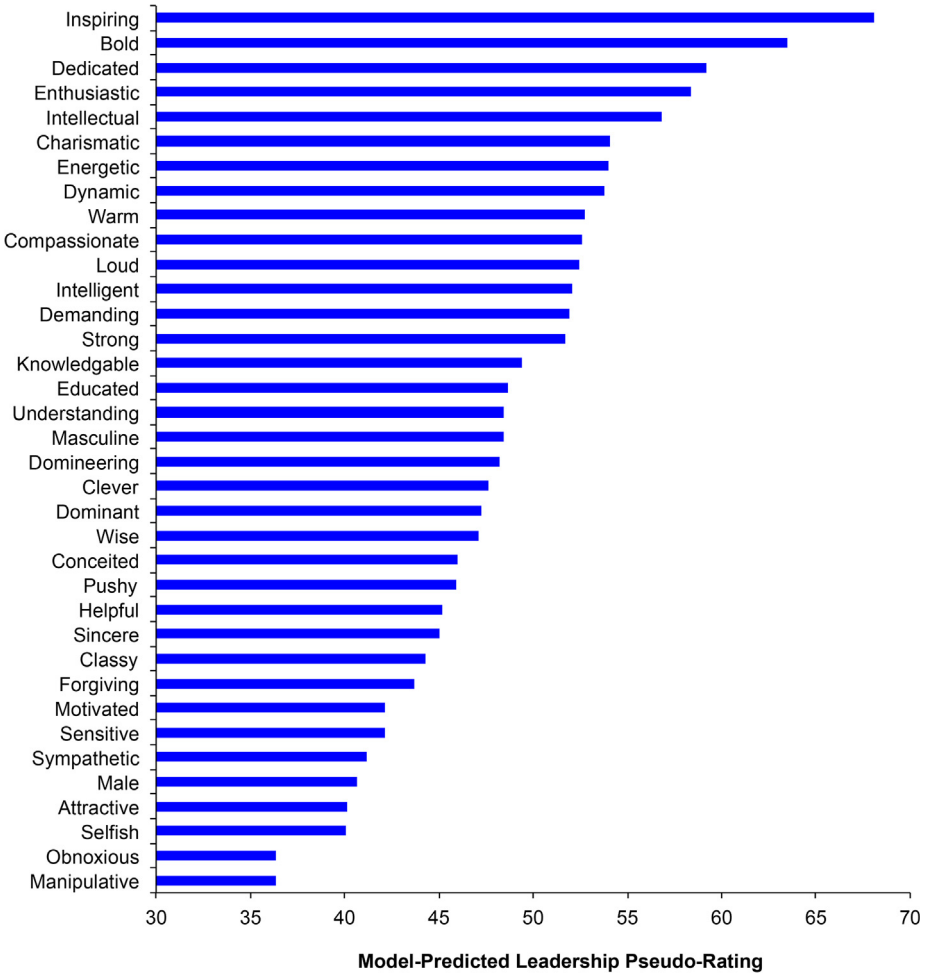

Fig. 4A. Model-predicted pseudo-ratings for the 36 traits examined by Offermann et al. (1994). Traits are ordered in terms of their model-predicted pseudo-ratings, from highest (top) to lowest (bottom).

degree to which the constructs are reflected in the text. In prior work, this approach has been shown to predict the personalities, traits, and emotional states of the individuals who generated the text (for reviews, see Pennebaker et al., 2003).

We replicated our above analysis on all LIWC constructs except those associated with linguistic and psycholinguistic variables (e.g., negation, present tense, personal pronouns). Specifically, we calculated the predicted pseudo-ratings of our model for words that make up 33 distinct constructs corresponding to social concerns, personal concerns, affective processes, cognitive processes, perceptual processes, and biological processes. We then averaged the pseudo-ratings for the words in each construct to determine the aggregate pseudo-rating of the construct. These aggregate pseudo-ratings are shown in Fig. 6. From this figure, we can see that constructs such as religion, achievement, and work have the highest pseudo-ratings and are thus most associated with perceived leadership effectiveness. In contrast, constructs such as anger, friend, and sexual are least associated with perceived leadership effectiveness. Thus, it seems that individuals who are frequently mentioned in the context of religion, achievement, and work are typically seen as being effective leaders, whereas those more frequently mentioned in the context of anger, friendship, and sexuality, are typically seen as being ineffective leaders.

Valence and arousal

Our final analysis applies the above approach to a much larger and more comprehensive set of words. Specifically, we calculated pseudo-ratings for a dataset of 13,915 words compiled by Warriner et al. (2013). These are all high-frequency words in the English language, and, importantly for our analysis, received ratings of valence and arousal by human participants. These ratings range from 1 (low valence or low arousal) to 9 (high valence or high arousal) and describe what human participants felt while reading each word. Words rated high in valence engender feelings of happiness, whereas those rated low in valence engender feelings of unhappiness. Words rated high in arousal engender feelings of excitement, whereas those rated low in arousal engender feelings of calmness.

Figs. 7A and 7B present word clouds for the 100 words with the highest and lowest predicted pseudo-ratings in our model, respectively. The color and lightness of these words are based on their valence and arousal ratings in Warriner et al. (2013), and their size corresponds to their pseudo-rating. We can see that the words most associated with effective leadership have mostly positive valence, and include religious words (e.g., apostle), words associated with commemoration (e.g., wreath), and leadership (e.g., pioneer). Conversely, words least associated with effective leadership have mostly negative valence, and typically involve sexual deviance (e.g., swinger), crime

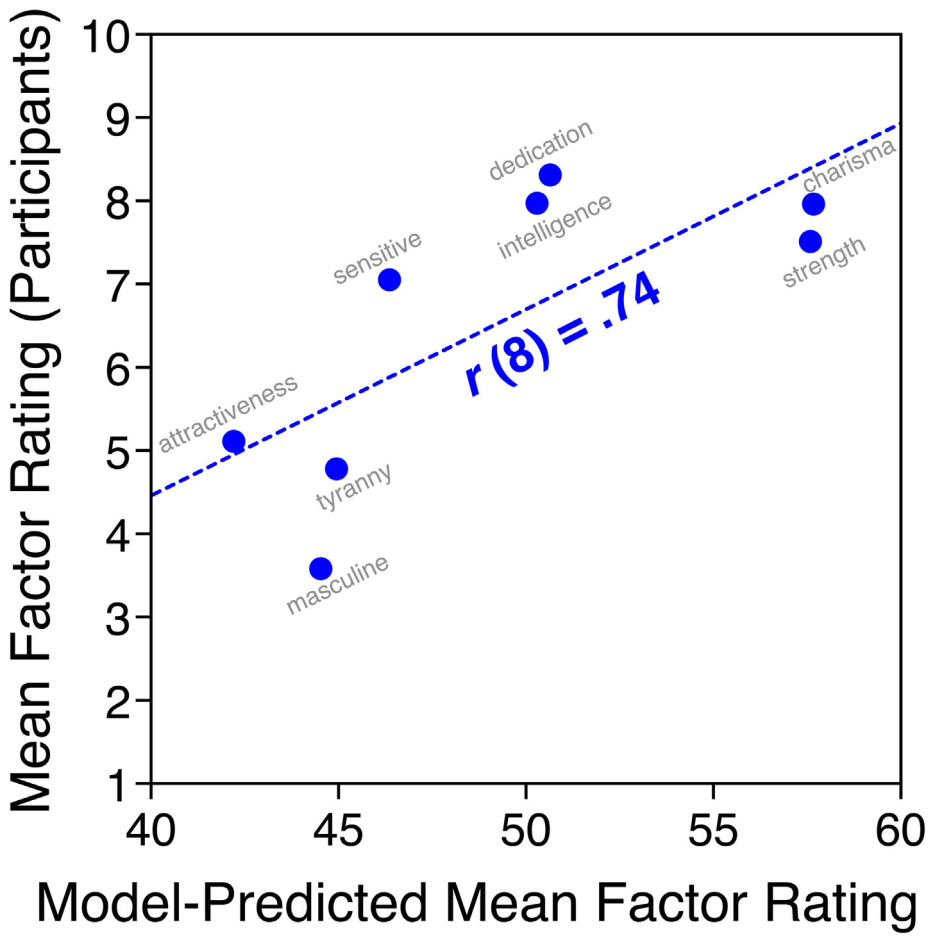

Fig. 4B. Factor ratings obtained by Offermann et al. (1994) plotted against our model-predicted average pseudo-ratings for the traits that make up the factors. The dashed line represents the best linear fit. The standard parametric correlation between the two variables (r) and its associated sample size (in parentheses) are also presented.

(e.g., murderer), and mental illness (e.g., psychopath), and often all three. Overall, we find a correlation of 0.23 between predicted pseudo-rating and valence ($p < 0.001$), indicating that positive valence words are typically associated with individuals rated high in effective leadership. We also find a correlation of −0.15 between predicted pseudo-rating and arousal ($p < 0.001$), indicating that high arousal words are typically associated with individuals rated low in effective leadership.

Taken together, these findings further highlight the power of our method, which does more than just predict people's leadership effectiveness judgments: the conceptual map it has produced can also help us uncover the traits, constructs, and concepts that underlie those judgments. Some of the traits identified by our model have already been documented in past work (Judge & Bono, 2000; Offermann et al., 1994), and our approach replicates those prior findings using a vast amount of naturally-occurring data. Other factors, such as those examined in our analyses of psychological constructs, valence, and arousal, provide novel insights regarding a large range of everyday concepts that appear to be relevant to leadership perception. While it may not be surprising that positively valenced and low arousal words (i.e., those suggesting a level of control) are associated with effective leadership, the fact that our model can predict these associations in large scale text data without having to be trained on them opens up vast possibilities for leadership research, as we discuss below, in our General Discussion.

General discussion

In this paper, we presented a novel, big-data-driven computational approach to predicting leadership effectiveness perceptions. In particular, we utilized a word embedding model trained on a large natural language corpus of news articles, to obtain high dimensional vector representations for well-known individuals. These vector representations proxy what people know about, and associate with, these individuals, and we used these vectors to predict the leadership effectiveness ratings given by survey participants (i.e., actual human judgments). Our approach achieved high out-of-sample accuracy rates in predicting participant ratings. These accuracy rates exceeded various benchmarks, and even approached participant split-half reliability (the theoretical upper-bound on such predictions). In addition, our approach maintained a high accuracy rate even when we separated participants by gender, ethnicity, and political affiliation. Thus, our approach is able to predict the leadership perceptions of the aggregate survey sample, as well as various demographic and political subgroups. Moreover, we were able to predict leadership perceptions for different types of target individuals, and our accuracy rates remained high when we limited our analyses to individuals typically considered to be leaders (i.e., individuals in leadership occupations), to individuals in a specific domain (e.g., arts or sports), or to individuals of a particular gender.

Importantly, we showed that our approach can go beyond merely predicting leadership effectiveness judgments, as it allows us to

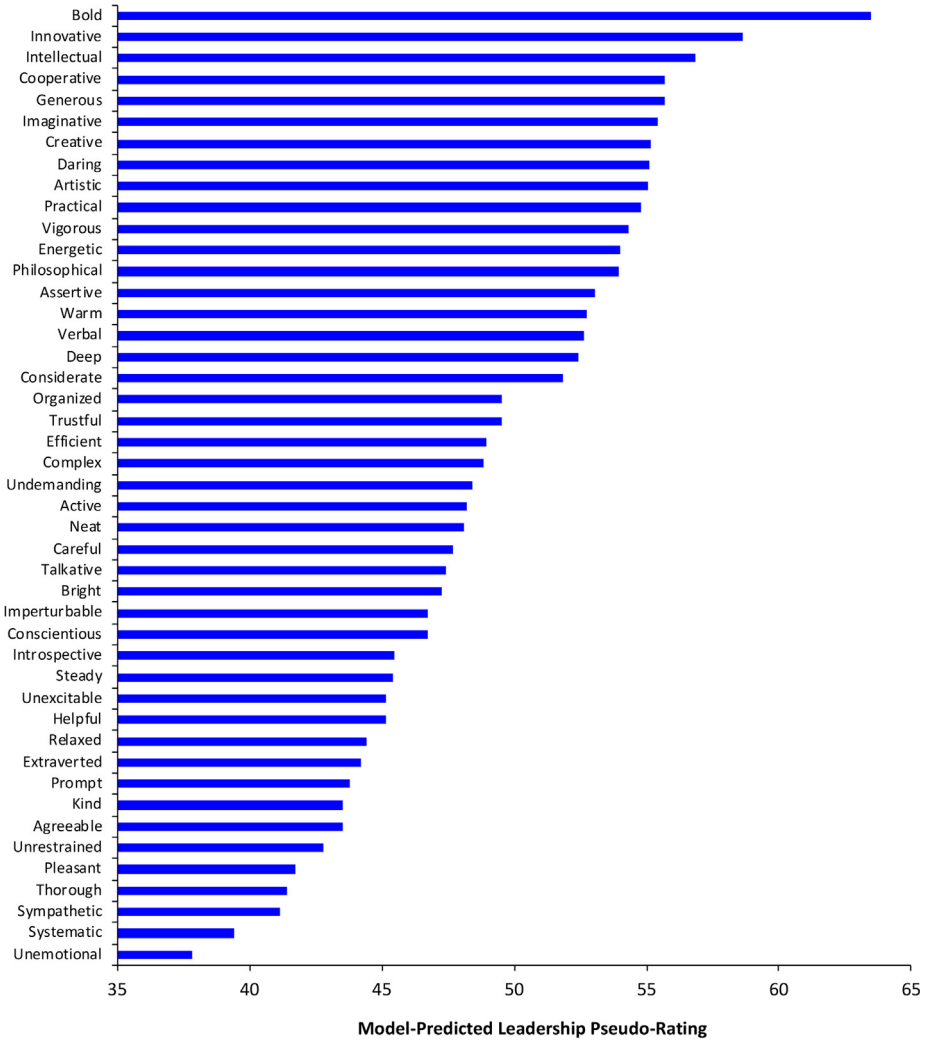

Fig. 5A. Model-predicted pseudo-ratings for the 45 Big Five traits from Goldberg (1992). Traits are ordered in terms of their model-predicted pseudo-ratings, from highest (top) to lowest (bottom).

uncover the psychological cues that best predict these judgments, thereby providing insights into the specific correlates of leadership perception. In particular, we first demonstrated that our approach successfully "re-discovers" the ILT and personality traits that—prior research had found—are associated with leadership effectiveness judgments (Judge & Bono, 2000; Offermann et al., 1994). We also showed how the associations uncovered by our model can be used to examine previously unexplored correlates of leadership perception, and thus motivate novel research. Specifically, we mapped the associations between perceptions of effective leadership and a large variety of words that (mentally) invoke a host of psychological concepts (Pennebaker et al., 2001), as well as words differing in their perceived valence and arousal (Warriner et al., 2013). In addition to identifying many previously unexamined potential drivers of leadership perception, these analyses illustrate the generalizability of our approach, which makes it possible to test for associations between perceived leadership effectiveness and a nearly unlimited set of words. These associations, in turn, help us better understand the "black box" of word embedding features that best predict leadership perception.

Our findings regarding the evaluative dimensions associated with effective leadership lay the groundwork for future empirical work. This work can test whether constructs related to religion (which we found is positively associated with effective leadership) or sexuality (which we found is negatively associated with effective leadership) do in fact bias people's leadership judgments, using standard experimental techniques to complement our data-driven machine learning methods. Future work could also attempt to understand the psychological underpinnings of these associations, so as to extend existing ILTs and better predict and influence perceptions of leadership. Finally, since many of the associations documented in this paper stem from linguistic co-occurrences in news media, future work could attempt to use our results to characterize how the media influences perceptions of effective leadership.

An interesting—albeit, disappointing—finding in our study was that the women in our dataset consistently received lower leadership ratings, and this female "evaluative penalty" persisted even when we accounted for gender imbalances in the different domains we examined. When considering extant research that shows gendered evaluations of both leadership perceptions (Schein, 1973, 1975) and leadership effectiveness (Forsyth et al., 1997; Heilman, 1983) as well as the still-prevalent underrepresentation of women in managerial positions (Warner et al., 2018), this finding is perhaps not surprising. However, the fact that we were able to observe this trend in large-scale language data using objective, machine learning methods speaks to the prevalence of the gender bias in the domain of leadership. Given the importance of gender equality in leadership, it is critical that future research continues to examine the manifestation of gender bias in this context using different large-scale language data sources.

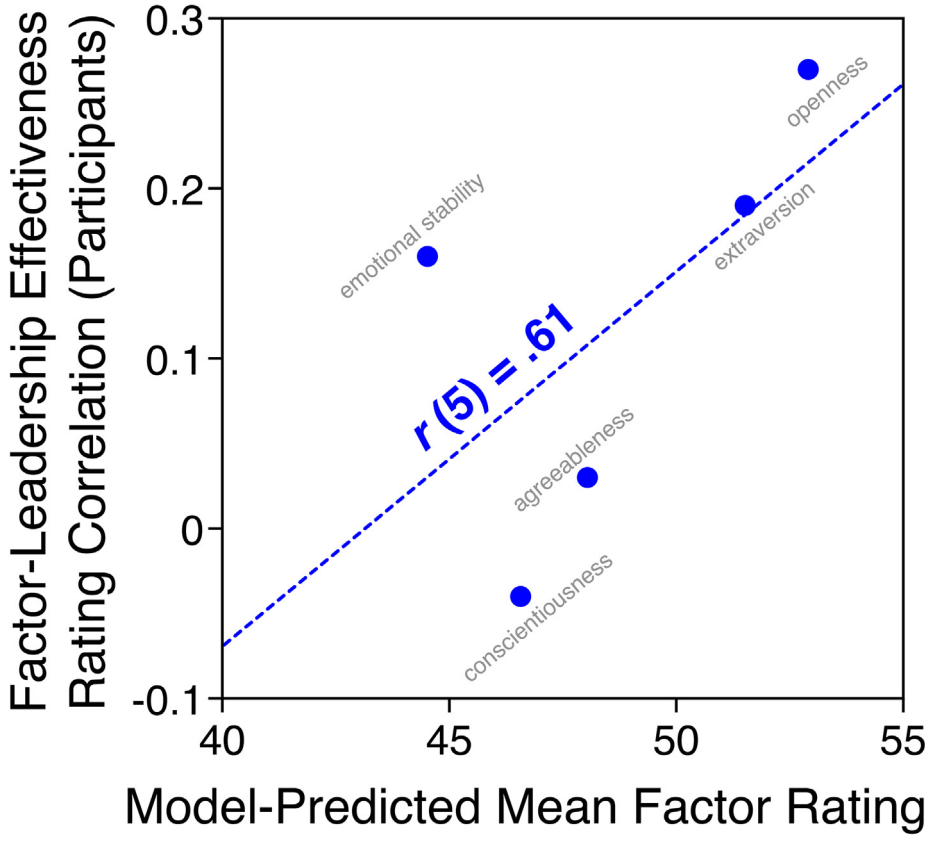

Fig. 5B. Correlations (between Big Five factor ratings and leadership effectiveness ratings) obtained by Judge and Bono (2000) plotted against our model-predicted average pseudo-ratings for the traits that make up the Big Five factors. The dashed line represents the best linear fit. The standard parametric correlation between the two variables (*r*) and its associated sample size (in parentheses) are also presented.

One unique benefit of the approach presented in this paper is that it can be applied with relative ease to predict leadership perceptions for nearly any individual, as long as the individual's name appears in the model's vocabulary (i.e., is discussed in the texts used to train the embeddings model). Exploiting this feature of our approach, we were able to predict leadership evaluations for over 6,000 well-known individuals for whom we did not collect participant data. We make these predictions freely available for other researchers at https://osf.io/52w7r/.

The word embedding model that we used is not just a practical computational tool for predicting leader perceptions; its neural network architecture and statistical assumptions are based on psychological theories of memory, and word embeddings are often seen as mimicking the structure of knowledge and associations in peoples' minds (see discussions in Bhatia, 2017a; Jones et al., 2015; Mandera et al., 2017). In this way, our approach can connect directly to theories of leader knowledge, leader concept representation, and leader associations, specifically captured in implicit theories of leadership (Lord et al., 1984), and especially to neural network models of leadership perception (Lord & Dinh, 2014; Lord & Emrich, 2001; Lord et al., 2001a, 2001b). According to ILTs, people have a prototypical idea of a leader, stored in memory, which gets activated in situations where this idea needs to be accessed. That is, leadership perceptions about a specific target are formed in relation to how the target fares in comparison to the prototypical leader. In this way, our embeddings model —which hinges on the closeness of similar words in the vector space, corresponding to how the conceptual associations of these words are represented in people's mind—also reflects an implicit understanding of leadership. For this reason, our approach is capable of predicting leadership perceptions not only for individuals, but also with regard to traits: Since traits occupy the same vector space as individuals, representations for traits can be added as inputs into the trained algorithm. Traits with higher predicted pseudo-ratings would be ones that our model associates with leadership. Consistent with prior work on leadership perception (especially Offermann et al., 1994), we found that our model associates traits such as inspiring, bold, and dedicated with effective leadership, and traits such as manipulative, obnoxious, and selfish with ineffective leadership. Thus, our model implicitly relies on the same set of trait associations used by human judges, to predict leadership effectiveness perceptions.

To the best of our knowledge, the current paper is the first to use word embeddings to model and predict leadership effectiveness perceptions and evaluations. Moreover, the approach we presented is novel in other ways. First, the data we used to predict leader perceptions were derived from an extremely large, and highly diverse, corpus of human language (over 100 billion words drawn from Google News articles). The rich vocabulary contained in this dataset means we can obtain vector representations for a wide range of individuals (and also a rich set of traits)—i.e., our approach is highly generalizable. At the same time, it is worth noting that, unlike other methodological approaches in leadership research that examine text data, such as historiometrics, this dataset was neither extracted from a leadership context (e.g., we did not analyze published biographies of leaders) nor designed specifically for the purpose of studying leader perception. Nonetheless, we showed it is still possible to predict leadership effectiveness judgments, even from text data that are not (primarily) about those leaders. We believe this aspect of our approach is noteworthy for both methodological and theoretical reasons. On the methodological side, our method retains the advantages of more qualitative methods, such as historiometrics, by studying leadership perception in context,

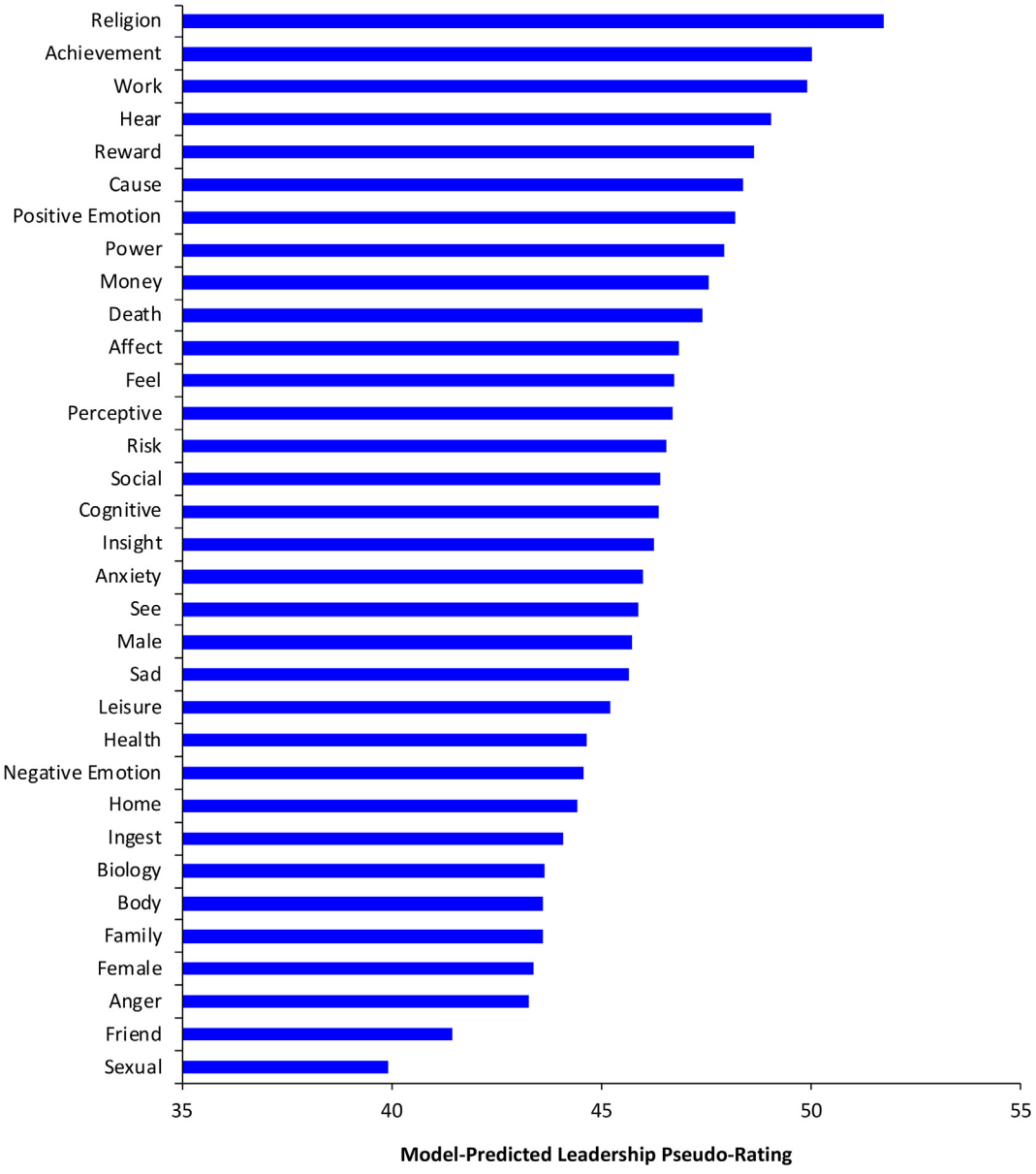

Fig. 6. Model-predicted pseudo-ratings for the 33 LIWC constructs. Constructs are ordered in terms of their model-predicted pseudo-ratings, from highest (top) to lowest (bottom).

but with the added ability to quickly and efficiently examine a large variety of contexts (i.e., a large and diverse set of texts). On the theoretical side, we believe our method captures a realistic representation of the ways people think of and evaluate leaders. It is not always the case that people (and followers) evaluate leaders specifically in their capacity as leaders. For example, we might often read about the president of the United States meeting with White House staff, but we might also sometimes read about the U.S. president vacationing with family over Thanksgiving. It is likely that both of these events inform our perceptions of the specific individual who is currently the U.S. president and of the prototypical American president more generally. Our dataset, which consists of a large sample of news articles, would include both kinds of events (among others involving the U.S. president), and is thus able to capture the diverse set of contexts that people may be exposed to when forming leadership perceptions.

Second, the continued growth in text corpora being made available (and processed for word embedding-based analyses), ensures that future researchers will be able replicate, and extend, our approach to predict leadership effectiveness associations for ever larger, and richer, lists of individuals (as well as traits). For example, researchers could extend this approach to non-English text corpora, in order to map leadership effectiveness associations in other countries and cultures (e.g., Chong & Thomas, 1997; Ensari & Murphy, 2003; Gerstner & Day, 1994; Sy et al., 2010; Vogel et al., 2015). Although we were able to predict (certain) leadership effectiveness associations among American participants using an English language corpus, these perceptions may not be entirely shared by non-Western (e.g., Korean or Turkish) people. Given the rapid growth of text data across languages, we look forward to seeing this approach extended to other linguistic, and cultural, populations (e.g., DeFranza et al., 2020; Khadilkar & KhudaBukhsh, 2020; Lewis et al., 2020; Lewis & Lupyan, 2020). Applying our approach to other linguistic groups would allow researchers to map out cultural variations and "universals" in leadership perception—e.g., which traits are most (vs. least) associated with effective leadership across countries. Moreover, the recent availability of historical natural language data also opens up new avenues for

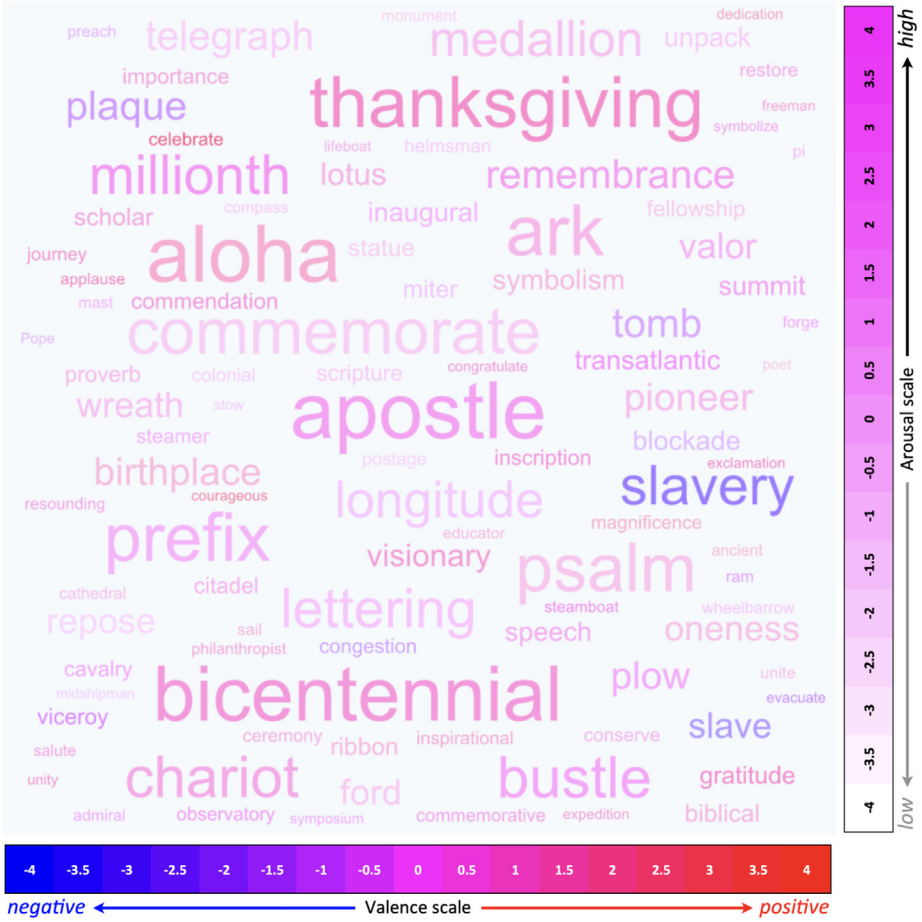

Fig. 7A. The 100 Warriner et al. (2013) words with the highest predicted pseudo-ratings in our model. Color and lightness indicate valence and arousal, respectively (see color and brightness keys, on the bottom and right). Font size indicates word association strength (i.e., pseudo-ratings), with larger fonts indicating stronger associations.

studying historical leader perceptions. For example, we can train word embedding models to reflect associations going back as far as the start of the 20th century and quantify changes in leadership perceptions over the past 100 years (see Bhatia & Bhatia, 2021; Garg et al., 2018; Khadilkar & KhudaBukhsh, 2020, for similar applications of word embeddings to the study of gender and ethnic stereotypes).

Third, as already explained, the approach we propose can be used to quickly, and efficiently, predict leadership effectiveness perceptions for most leaders (as well as for other evaluation dimensions, beyond 'leadership effectiveness'). In many cases (for which some prior ratings of relevant leaders have been collected), this requires no (new) participant data. Indeed, we have already exploited the data from our survey study to generalize predictions to a large number of additional leaders. Even in the "worst case" scenarios (where no relevant prior ratings are available), our approach significantly reduces the amount of data that need to be collected: rather than eliciting ratings for all leaders (and dimensions) of interest, researchers only need a subset of these to train the predictive model.

The method we presented offers a number of useful applications. For example, leadership researchers and advisors could use this approach to quickly predict how a given leader (or set of leaders) will be perceived on a given trait dimension (or set of trait dimensions); in fact, they can just as easily do so for a vast range of leaders (and trait dimensions). Note that our method can serve two purposes. It can serve as a substitute for participant data, thereby eliminating or drastically reducing the need to continuously administer surveys. In doing so, it would help free leadership researchers and advisors from the constraints associated with existing data-collection and coding approaches, and thus greatly expand the range, size, speed, and adaptive quality of the available data at their disposal. Alternatively, it can serve as a complement to, and even work in tandem with, participant data, in order to boost its predictive power (as we showed in this paper).

Of course, our approach is not without limitations. First, this method is fundamentally constrained by the size and richness of the text corpora used to derive its vectors. Specifically, any given text corpus (at least today) is unlikely to contain every single individual's name, and may even lack the names of some well-known leaders. The resulting word embedding model will be unable to make predictions concerning individuals who are missing from the corpus it was built on. We should note that this limitation (the absence of some names in a given text corpus) will often be a *desirable* feature of our approach, since lesser-known individuals are precisely the ones least likely to appear in a text corpus (i.e., to be discussed in those texts), which reflects the fact that they are also the ones least likely to be represented in people's minds. In other words, the non-random absence of certain individual names in a text corpus actually mirrors people's (lack of) knowledge. Therefore, we can expect that individuals for whom our method cannot make predictions (due to their absence in the model vocabulary) will often be those whom many people do not recognize anyway, and thus cannot evaluate. Still, we cannot rule out the possibility that some of the leaders who failed to appear in our embedding model vocabulary may be important to study and/or representative of the population of leaders. This omission may thus limit

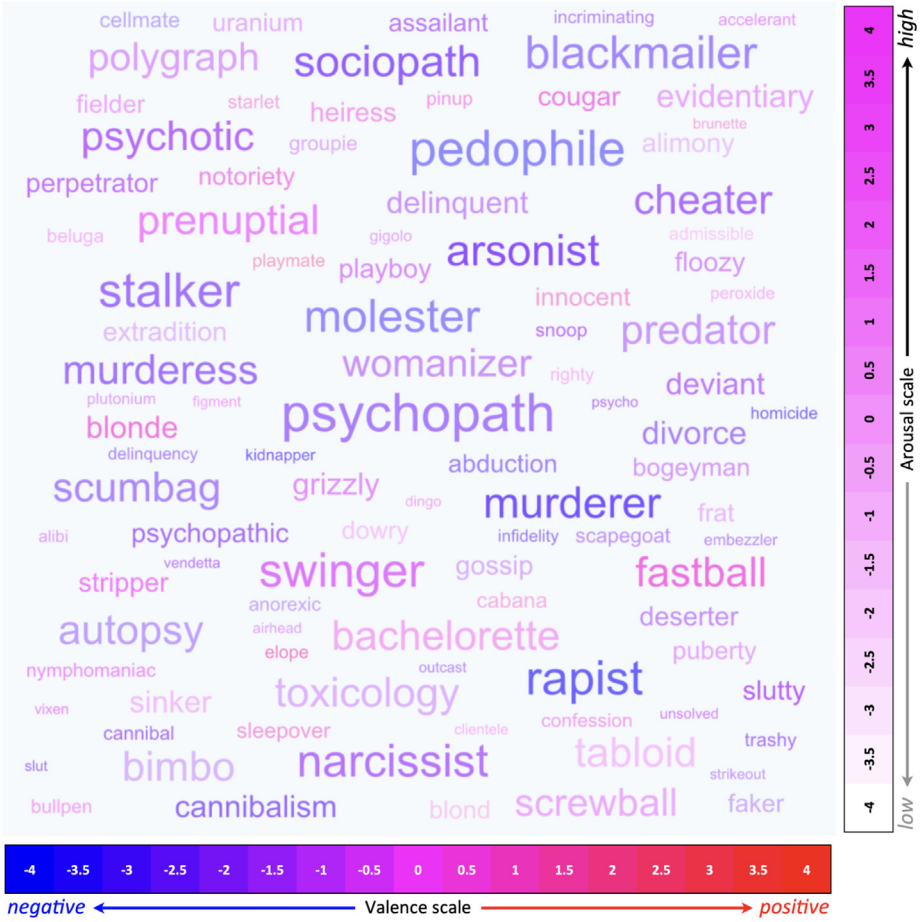

Fig. 7B. The 100 Warriner et al. (2013) words with the lowest predicted pseudo-ratings in our model. Color and lightness indicate valence and arousal, respectively (see color and brightness keys, on the bottom and right). Font size indicates word association strength (i.e., pseudo-ratings), with larger fonts indicating stronger associations.

the extent to which our findings generalize beyond the set of leaders in our dataset, especially if the omission is non-random. Another issue related to the size and richness of the text corpora being used is whether, and to what extent, it reflects people's views. For example, the Google News corpus is unlikely to perfectly represent the views of the average American. Yet, the less it represents people's views, the more disconnected its resulting embedding model will be from their perceptions of leadership.

Fortunately, researchers will be able to address the above issues by collecting larger and richer text corpora, as well as by combining various existing text corpora. Indeed, while we used an existing set of vectors built on the Google News corpus, other researchers could instead build vectors using text data extracted from other sources, such as Wikipedia, social media platforms (e.g., Facebook, Twitter, etc.), and electronic books (e.g., biographies, historical texts, etc.). In fact, we strongly encourage leadership researchers to collectively extend the diversity, size, and conceptual richness of the text data we can use to build word embedding-based vector spaces. Doing so will provide a collective resource that can be used to develop increasingly better models of leadership perception. Researchers should also test different techniques for building vector representations on these data (such as those outlined in Bhatia et al., 2019; Jones et al., 2015; Lenci, 2018; Mandera et al., 2017; Turney & Pantel, 2010; Young et al., 2018). Although the pre-trained Word2Vec embeddings used in this paper already yield highly accurate predictions (and can be easily extended to study a very large set of leaders), even greater predictive power may be possible with other techniques.

A second, related issue concerns the potential short-term dynamics of leader perception. The sizes of text corpora needed for our approach make it currently unsuitable for modeling rapid changes in leader perception (e.g., following a political advertisement campaign, public relations event, or scandal; see Smith & French, 2011) or perceptions of emerging leaders. Indeed, although we were able to accurately predict leadership perceptions for most individuals, our model made the greatest prediction error for Donald Trump (as previously explained, our model was trained on text data prior to Trump becoming president, whereas our participants were asked to give leadership ratings in 2019, during Trump's presidency). Fortunately, with the growth of digital data and the wide availability of online social data and news media data, it is becoming possible to train embedding models almost in real time. Thus, we expect this will not be a major issue for leadership researchers in the near future.

Third, the participants in our survey study were recruited from Prolific Academic and, therefore, do not perfectly represent the US population. Future researchers should, if possible, try to administer their surveys to representative samples of the population of interest (e.g., by hiring a survey company). That said, online samples, such as Prolific Academic, are far more representative of the US population than the student samples that many researchers used to recruit.

Finally, we stress that our approach, much like ILTs, is useful for understanding *subjective* leadership impressions—the beliefs and perceptions that people hold of leaders—, and not for probing *actual* leadership qualities. That's not to say these kinds of methods have nothing to offer in the latter regard; indeed, one could rely on linguistic

co-occurrences to identify (with varying levels of accuracy) certain objective leader characteristics (e.g., identifying gender by relying on the fact that female [vs. male] leaders' names are more likely to co-occur with words such as "she" [vs. "he"] and "her" [vs. "his"]). Nonetheless, the current paper focused on the perceptions of leaders, which do not always reflect—and therefore need to be distinguished from—the "true" qualities and actual behaviors of those leaders. That said, as we previously noted, some leadership researchers have convincingly argued that the mere perception of leadership grants influence, and thus reality to a person's leadership status (Lord & Maher, 1991). The implication is that being perceived as a leader is a core component of ("actual") leadership (Lord & Maher, 1991). Consequently, the method presented in this paper can also be used to identify whether, and to what extent, an individual is (actually) a leader.

In summary, we present a novel, general, and highly accurate method for the measurement and study of leadership perception. Our approach relies on the same set of trait associations used by lay people to judge leadership effectiveness, and can thus closely mimic their leadership effectiveness ratings. We hope leadership scholars will use this approach to advance (and accelerate) their research, and that practitioners (e.g., leadership advisors) will use it to better monitor the evolution of leader perceptions.

Funding

Funding was received from the National Science Foundation grant SES-1847794 and the Alfred P. Sloan Foundation.

Appendix A. Implicit leadership theory factors and their associated traits

Factor	Traits
Sensitivity	Sympathetic, Sensitive, Compassionate, Understanding, Sincere, Warm, Forgiving, Helpful
Dedication	Dedicated, Motivated, Hardworking, Goal-oriented
Tyranny	Domineering, Pushy, Dominant, Manipulative, Power-hungry, Conceited, Selfish, Obnoxious, Demanding, Loud
Charisma	Energetic, Charismatic, Inspiring, Enthusiastic, Dynamic
Attractiveness	Well-groomed, Attractive, Well-dressed, Classy
Masculinity	Male, Masculine
Intelligence	Intellectual, Educated, Intelligent, Knowledgeable, Clever
Strength	Strong, Bold

Appendix B. The Big-5 personality inventory factors and their associated traits

Factor	Traits
Surgency (Extraversion)	Extraverted, Talkative, Assertive, Verbal, Energetic, Bold, Active, Daring, Vigorous, Unrestrained
Agreeableness	Kind, Cooperative, Sympathetic, Warm, Trustful, Considerate, Pleasant, Helpful, Generous
Conscientiousness	Organized, Systematic, Thorough, Practical, Neat, Efficient, Careful, Steady, Conscientious, Prompt
Emotional Stability	Relaxed, Imperturbable, Unenvious, Unemotional, Unexcitable, Undemanding
Intellect (Openness to Experience)	Intellectual, Creative, Complex, Imaginative, Bright, Philosophical, Artistic, Deep, Innovative, Introspective

Appendix C. The leaders and non-leaders occupation lists

Affiliation	Occupation
Leaders	Anthropologist, Archaeologist, Architect, Astronomer, Biologist, Business person, Chemist, Computer scientist, Critic, Cyclist, Designer, Diplomat, Economist, Explorer, Extremist, Fashion designer, Geographer, Geologist, Historian, Inventor, Journalist, Judge, Linguist, Military personnel, Mountaineer, Nobleman, Philosopher, Physician, Physicist, Political scientist, Politician, Psychologist, Public worker, Religious figure, Social activist, Sociologist, Statistician, Writer
Non-Leaders	Actor, Artist, Astronaut, Athlete, Baseball player, Basketball player, Boxer, Celebrity, Chef, Chess master, Coach, Comedian, Comic artist, Companion, Composer, Conductor, Cricketer, Dancer, Film director, Game designer, Golfer, Gymnast, Hockey player, Lawyer, Mafioso, Magician, Martial artist, Model, Musician, Painter, Photographer, Pilot, Pirate, Pornographic actor, Presenter, Producer, Race car driver, Referee, Sculptor, Singer, Skater, Skier, Snooker player, Soccer player, Swimmer, Tennis player, Wrestler

Appendix D. Supplementary material

Supplementary data for this article can be found online at https://doi.org/10.1016/j.leaqua.2021.101535.

References

Antonakis, J., & Eubanks, D. L. (2017). Looking leadership in the face. *Current Directions in Psychological Science, 26*(3), 270–275.

Arnoff, J., & Wilson, J. P. (1985). *Personality in the Social Process*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Avolio, B. J., & Bass, B. M. (2004). Multifactor leadership questionnaire (MLQ). *Mind Garden, 29*.

Batistič, S., Černe, M., & Vogel, B. (2017). Just how multi-level is leadership research? A document co-citation analysis 1980–2013 on leadership constructs and outcomes. *The Leadership Quarterly, 28*(1), 86–103.

Bedell, K. E., Hunter, S. T., Angie, A. D., & Vert, A. (2006). A historiometric examination of Machiavellianism and a new taxonomy of leadership. *Journal of Leadership and Organizational Studies, 12*(4), 50–72.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica, 80*(6), 2369–2429.

Bhatia, N., & Bhatia, S. (2021). Changes in gender stereotypes over time: A computational analysis. *Psychology of Women Quarterly, 45*(1), 106–125.
Bhatia, S. (2017a). Associative judgment and vector space semantics. *Psychological Review, 124*(1), 1–20.
Bhatia, S. (2017b). The semantic representation of prejudice and stereotypes. *Cognition, 164*, 46–60.
Bhatia, S. (2019a). Semantic processes in preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(4), 627–640.
Bhatia, S. (2019b). Predicting risk perception: New insights from data science. *Management Science, 65*, 3800–3823.
Bhatia, S., Goodwin, G. P., & Walasek, L. (2018). Trait associations for Hillary Clinton and Donald Trump in news media: A computational analysis. *Social Psychological and Personality Science, 9*(2), 123–130.
Bhatia, S., & Olivola, C. Y. (2021). Computational brand perception: Fine-tuned word embedding techniques for predicting consumer brand-trait associations. Working Paper.
Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences, 29*, 31–36.
Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition, 179*, 71–88.
Bhatia, S., & Walasek, L. (2019). Association and response accuracy in the wild. *Memory & Cognition, 47*(2), 292–298.
Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.
Braun, S., Hernandez Bark, A., Kirchner, A., Stegmann, S., & van Dick, R. (2019). Emails from the Boss—Curse or blessing? Relations between communication channels, leader evaluation, and employees' attitudes. *International Journal of Business Communication, 56*(1), 50–81.
Braun, S., Stegmann, S., Hernandez Bark, A. S., Junker, N. M., & van Dick, R. (2017). Think manager—think male, think follower—think female: Gender bias in implicit followership theories. *Journal of Applied Social Psychology, 47*(7), 377–388.
Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.
Caliskan, A., & Lewis, M. (2020). Social biases in word embeddings and their relation to human cognition. Working Paper.
Carnes, A., Houghton, J. D., & Ellison, C. N. (2015). What matters most in leader selection? The role of personality and implicit leadership theories. *Leadership & Organization Development Journal, 36*(4), 360–379.
Chong, L. M. A., & Thomas, D. C. (1997). Leadership perceptions in cross-cultural context: Pakeha and pacific islanders in New Zealand. *The Leadership Quarterly, 8*(3), 275–293.
Choudhury, P., Wang, D., Carlson, N. A., & Khanna, T. (2019). Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal, 40*(11), 1705–1732.
Craig, S. B., & Gustafson, S. B. (1998). Perceived leader integrity scale: An instrument for assessing employee perceptions of leader integrity. *The Leadership Quarterly, 9*(2), 127–145.
DeChurch, L. A., Burke, C. S., Shuffler, M. L., Lyons, R., Doty, D., & Salas, E. (2011). A historiometric analysis of leadership in mission critical multiteam environments. *The Leadership Quarterly, 22*(1), 152–169.
DeFranza, D., Mishra, H., & Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology, 119*(1), 7–22.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
Dinh, J. E., Lord, R. G., Gardner, W. L., Meuser, J. D., Liden, R. C., & Hu, J. (2014). Leadership theory and research in the new millennium: Current theoretical trends and changing perspectives. *The Leadership Quarterly, 25*(1), 36–62.
Doldor, E., Wyatt, M., & Silvester, J. (2019). Statesmen or cheerleaders? Using topic modeling to examine gendered messages in narrative developmental feedback for leaders. *The Lesadership Quarterly, 30*(5) 101308.
Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573.
Eden, D., & Leviatan, U. (1975). Implicit leadership theory as a determinant of the factor structure underlying supervisory behavior scales. *Journal of Applied Psychology, 60*(6), 736–741.
Ensari, N., & Murphy, S. E. (2003). Cross-cultural variations in leadership perceptions and attribution of charisma to the leader. *Organizational Behavior and Human Decision Processes, 92*(1), 52–66.
Epitropaki, O., & Martin, R. (2004). Implicit leadership theories in applied settings: Factor structure, generalizability, and stability over time. *Journal of Applied Psychology, 89*(2), 293–310.
Epitropaki, O., & Martin, R. (2005). From ideal to real: A longitudinal study of the role of implicit leadership theories on leader-member exchanges and employee outcomes. *Journal of Applied Psychology, 90*(4), 659–676.
Eubanks, D. L., Antes, A. L., Friedrich, T. L., Caughron, J. J., Blackwell, L. V., Bedell-Avers, K. E., & Mumford, M. D. (2010). Criticism and outstanding leadership: An evaluation of leader reactions and critical outcomes. *The Leadership Quarterly, 21*(3), 365–388.
Eubanks, D. L., Palanski, M. E., Swart, J., Hammond, M. M., & Oguntebi, J. (2016). Creativity in early and established career: Insights into multi-level drivers from Nobel Prize winners. *The Journal of Creative Behavior, 50*(4), 229–251.
Firth, J. R. (1957). *Papers in Linguistics*. London, UK: Oxford University Press.
Forsyth, D. R., Heiney, M. M., & Wright, S. S. (1997). Biases in appraisals of women leaders. *Group Dynamics: Theory, Research, and Practice, 1*(1), 98–103.
Fraser, S. L., & Lord, R. G. (1988). Stimulus prototypicality and general leadership impressions: Their role in leadership and behavioral ratings. *Journal of Psychology, 122*(3), 291–303.
Gardner, W. L., Lowe, K. B., Moss, T. W., Mahoney, K. T., & Cogliser, C. C. (2010). Scholarly leadership of the study of leadership: A review of The Leadership Quarterly's second decade, 2000–2009. *The Leadership Quarterly, 21*(6), 922–958.
Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635–E3644.
Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods, 50*(1), 344–361.
George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big Data and data science methods for management research. *Academy of Management Journal, 59*(5), 1493–1507.
Gerstner, C. R., & Day, D. V. (1994). Cross-cultural comparison of leadership prototypes. *The Leadership Quarterly, 5*(2), 121–134.
Giessner, S. R., & van Knippenberg, D. (2008). "License to fail": Goal definition, leader group prototypicality, and perceptions of leadership effectiveness after leader failure. *Organizational Behavior and Human Decision Processes, 105*(1), 14–35.
Gioia, D. A., & Sims, H. P. (1985). On avoiding the influence of implicit leadership theories in leader behavior descriptions. *Educational and Psychological Measurement, 45*(2), 217–232.
Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26–42.
Graham, J. R., Harvey, C. R., & Puri, M. (2017). A corporate beauty contest. *Management Science, 63*(9), 3044–3056.
Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition, 135*, 21–23.
Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*(6), 1006–1033.
Hackman, J. R. (1992). Group influences on individuals in organizations. In M. D. Dunnette & L. H. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., pp. 199–267). Palo Alto, CA: Consulting Psychologists Press.
Hais, S. C., Hogg, M. A., & Duck, J. M. (1997). Self-categorization and leadership: Effects of group prototypicality and leader stereotypicality. *Personality and Social Psychology Bulletin, 23*(10), 1087–1099.
Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods, 21*(4), 447–457.
Harris, Z. S. (1954). Distributional structure. *Word, 10*(2–3), 146–162.
Harrison, J. S., Thurgood, G. R., Boivie, S., & Pfarrer, M. D. (2019). Measuring CEO personality: Developing, validating, and testing a linguistic tool. *Strategic Management Journal, 40*(8), 1316–1330.
Hawn, C. (2009). Take two aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care. *Health Affairs, 28*(2), 361–368.
Heilman, M. E. (1983). Sex bias in work settings: The Lack of Fit model. *Research in Organizational Behavior, 5*, 269–298.
Holtzman, N. S., Schott, J. P., Jones, M. N., Balota, D. A., & Yarkoni, T. (2011). Exploring media bias with semantic analysis tools: Validation of the contrast analysis of semantic similarity (CASS). *Behavior Research Methods, 43*(1), 193–200.
Hopkins, D. J. (2018). The exaggerated life of death panels? The limited but real influence of elite rhetoric in the 2009–2010 health care debate. *Political Behavior, 40*(3), 681–709.
Hunter, S. T., Cushenbery, L., Thoroughgood, C., Johnson, J. E., & Ligon, G. S. (2011). First and ten leadership: A historiometric investigation of the CIP leadership model. *The Leadership Quarterly, 22*(1), 70–91.
Jones, M. N. (2017). *Big data in cognitive science*. New York, NY: Taylor and Francis.
Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254). New York: Oxford University Press.
Judge, T. A., & Bono, J. E. (2000). Five-factor model of personality and transformational leadership. *Journal of Applied Psychology, 85*(5), 751–765.
Junker, N. M., Schyns, B., van Dick, R., & Scheurer, S. (2011). The importance of leader categorization for commitment, job satisfaction, and well-being with particular consideration of gender role theory. *Zeitschrift Fur Arbeits-Und Organisationspsychologie, 55*(4), 171–179.
Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd Ed. draft). Retrieved from: URL https://web. stanford. edu/~ jurafsky/slp3.
Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin, 137*(4), 616–642.
Khadilkar, K., & KhudaBukhsh, A. R. (2020). An unfair affinity toward fairness: Characterizing 70 years of social biases in B^{H}ollywood. *35th AAAI Conference on Artificial Intelligence (AAAI)*.
La Bella, A., Fronzetti Colladon, A., Battistoni, E., Castellan, S., & Francucci, M. (2018). Assessing perceived organizational leadership styles through twitter text mining. *Journal of the Association for Information Science and Technology, 69*(1), 21–31.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics, 4*, 151–171.

Lewis, M., Cahill, A., Madnani, N., & Evans, J. (2020). Local similarity and global variability characterize the semantic space of human languages. Working Paper.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour, 4*(10), 1021–1028.

Ligon, G. S., Harris, D. J., & Hunter, S. T. (2012). Quantifying leader lives: What historiometric approaches can tell us. *The Leadership Quarterly, 23*(6), 1104–1133.

Lord, R. G., Brown, D. J., & Harvey, J. L. (2001). System constraints on leadership perceptions, behavior and influence: An example of connectionist level processes. In M. A. Hogg & R. S. Tindale (Eds.), *Blackwell handbook of social psychology: Group processes* (pp. 283–310). Oxford, England: Blackwell.

Lord, R. G., Brown, D. J., Harvey, J. L., & Hall, R. J. (2001). Contextual constraints on prototype generation and their multilevel consequences for leadership perceptions. *The Leadership Quarterly, 12*(3), 311–338.

Lord, R. G., & Dinh, J. E. (2014). What have we learned that is critical in understanding leadership perceptions and leader-performance relations? *Industrial and Organizational Psychology, 7*(2), 158–177.

Lord, R. G., & Emrich, C. G. (2001). Thinking outside the box by looking inside the box: Extending the cognitive revolution in leadership research. *The Leadership Quarterly, 11*(4), 551–579.

Lord, R. G., Foti, R. J., & De Vader, C. L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance, 34*(3), 343–378.

Lord, R. G., Foti, R. J., & Phillips, J. S. (1982). A theory of leadership categorization. In J. G. Hunt, U. Sekaran, & C. Schriesheim (Eds.), Leadership: Beyond Establishment Views (pp. 104–121). Carbondale, IL: Southern Illinois Univ. Press.

Lord, R. G., & Maher, K. J. (1991). *Leadership and Information Processing: Linking Perceptions and Performance*. Boston, MA: Routledge.

Lovelace, J. B., Neely, B. H., Allen, J. B., & Hunter, S. T. (2019). Charismatic, ideological, & pragmatic (CIP) model of leadership: A critical review and agenda for future research. *The Leadership Quarterly, 30*(1), 96–110.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81–90.

Middlehurst, R., & Society for Research into Higher Education, Ltd., London (England). (1993). Leading Academics. Open University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Moat, H. S., Olivola, C. Y., Preis, T., & Chater, N. (2016). Searching choices: Quantifying decision making processes using search engine data. *Topics in Cognitive Science, 8*(3), 685–696.

Moat, H. S., Preis, T., Olivola, C. Y., Liu, C., & Chater, N. (2014). Using big data to predict collective behavior in the real world. *Behavioral and Brain Sciences, 37*, 92–93.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists*. O'Reilly Media Inc..

Mumford, M. D. (2006). *Pathways to Outstanding Leadership: A Comparative Analysis of Charismatic, Ideological, and Pragmatic Leaders*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Mumford, M. D., Espejo, J., Hunter, S. T., Bedell-Avers, K. E., Eubanks, D. L., & Connelly, S. (2007). The sources of leader violence: A comparison of ideological and non-ideological leaders. *The Leadership Quarterly, 18*(3), 217–235.

Neubert, M. J. (1999). Too much of a good thing or the more the merrier? Exploring the dispersion and gender composition of informal leadership in intact manufacturing teams. *Small Group Research, 30*, 635–646.

Neubert, M. J., & Taggar, S. (2004). Pathways to informal leadership: The moderating role of gender on the relationship of individual differences and team member network centrality to informal leadership emergence. *The Leadership Quarterly, 15*(2), 175–194.

Nye, J. L., & Forsyth, D. R. (1991). The effects of prototype-based biases on leadership appraisals: A test of leadership categorization theory. *Small Group Research, 22*(3), 360–379.

O'Connor, J., Mumford, M. D., Clifton, T. C., Gessner, T. L., & Connelly, M. S. (1995). Charismatic leaders and destructiveness: An historiometric study. *The Leadership Quarterly, 6*(4), 529–555.

Offermann, L. R., & Coats, M. R. (2018). Implicit theories of leadership: Stability and change over two decades. *The Leadership Quarterly, 29*(4), 513–522.

Offermann, L. R., Kennedy, J. K., & Wirtz, P. W. (1994). Implicit leadership theories: Content, structure, and generalizability. *The Leadership Quarterly, 5*(1), 43–58.

Olivola, C. Y., Eubanks, D. L., & Lovelace, J. B. (2014). The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership Quarterly, 25*(5), 817–834.

Olivola, C. Y., Sussman, A. B., Tsetsos, K., Kang, O. E., & Todorov, A. (2012). Republicans prefer Republican-looking leaders: Political facial stereotypes predict candidate electoral success among right-leaning voters. *Social Psychological and Personality Science, 3*, 605–613.

Olivola, C. Y., Tingley, D., & Todorov, A. (2018). Republican voters prefer candidates who have conservative-looking faces: New evidence from exit polls. *Political Psychology, 39*(5), 1157–1171.

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior, 34*(2), 83–110.

Parry, K., Mumford, M. D., Bower, I., & Watts, L. L. (2014). Qualitative and historiometric methods in leadership research: A review of the first 25 years of The Leadership Quarterly. *The Leadership Quarterly, 25*(1), 132–151.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825–2830.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Mahway: Lawrence Erlbaum Associates.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*(1), 547–577.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *EMNLP*, 1532–1543.

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology, 33*(3–4), 175–190.

Phillips, J. S., & Lord, R. G. (1981). Causal attributions and perceptions of leadership. *Organizational Behavior and Human Performance, 28*(2), 143–163.

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra. Psychology, 5*(1).

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27–48). Erlbaum.

Schein, V. E. (1973). The relationship between sex role stereotypes and requisite management characteristics. *Journal of Applied Psychology, 57*, 95–100.

Schein, V. E. (1975). The relationship between sex role stereotypes and requisite management characteristics among female managers. *Journal of Applied Psychology, 60*, 340–344.

Schein, V. E., Mueller, R., Lituchy, T., & Liu, J. (1996). Think manager–think male: A global phenomenon? *Journal of Organizational Behavior, 17*, 33–41.

Schyns, B. (2006). The role of implicit leadership theories in the performance appraisals and promotion recommendations of leaders. *Equal Opportunities International, 25*(3), 188–199.

Sieweke, J., & Santoni, S. (2020). Natural experiments in leadership research: An introduction, review, and guidelines. *The Leadership Quarterly, 31*(1) 101338.

Simonton, D. K. (1986). Dispositional attributions of (presidential) leadership: An experimental simulation of historiometric results. *Journal of Experimental Social Psychology, 22*(5), 389–418.

Smith, G., & French, A. (2011). Measuring the changes to leader brand associations during the 2010 election campaign. *Journal of Marketing Management, 27*(7–8), 721–738.

Spangler, W. D., Gupta, A., Kim, D. H., & Nazarian, S. (2012). Developing and validating historiometric measures of leader individual differences by computerized content analysis of documents. *The Leadership Quarterly, 23*(6), 1152–1172.

Spisak, B. R., van der Laken, P. A., & Doornenbal, B. M. (2019). Finding the right fuel for the analytical engine: Expanding the leader trait paradigm through machine learning? *The Leadership Quarterly, 30*(4), 417–426.

Stentz, J. E., Clark, V. L. P., & Matkin, G. S. (2012). Applying mixed methods to leadership research: A review of current practices. *The Leadership Quarterly, 23*(6), 1173–1183.

Stoker, J. I., Garretsen, H., & Spreeuwers, L. J. (2016). The facial appearance of CEOs: Faces signal selection but not performance. *PLoS ONE, 11* e0159950.

Strange, J. M., & Mumford, M. D. (2002). The origins of vision: Charismatic versus ideological leadership. *The Leadership Quarterly, 13*(4), 343–377.

Sy, T., Shore, L. M., Strauss, J., Shore, T. H., Tram, S., Whiteley, P., & Ikeda-Muromachi, K. (2010). Leadership perceptions as a function of race-occupation fit: The case of Asian Americans. *Journal of Applied Psychology, 95*(5), 902–919.

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods, 21*(3), 525–547.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141–188.

Tyler, T. R. (1986). The psychology of leadership evaluation. In H. W. Bierhoff, R. L. Cohen, & J. Greenberg (Eds.), *Justice in social relations* (pp. 299–316). Boston, MA: Springer.

Vessey, W. B., Barrett, J. D., Mumford, M. D., Johnson, G., & Litwiller, B. (2014). Leadership of highly creative people in highly creative fields: A historiometric study of scientific leaders. *The Leadership Quarterly, 25*(4), 672–691.

Vogel, R. M., Mitchell, M. S., Tepper, B. J., Restubog, S. L. D., Hu, C., Hua, W., & Huang, J. (2015). A cross-cultural examination of subordinates' perceptions of and reactions to abusive supervision. *Journal of Organizational Behavior, 36*(5), 720–745.

Wang, Y., Bowers, A. J., & Fikis, D. J. (2017). Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of EAQ articles from 1965 to 2014. *Educational Administration Quarterly, 53*(2), 289–323.

Warner, J., Ellmann, N., & Boesch, D. (2018). Women's leadership gap. Retrieved from: https://www.americanprogress.org/issues/women/reports/2018/11/20/461273/womens-leadership-gap-2/.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207.

Wenzel, R., & Van Quaquebeke, N. (2018). The double-edged sword of big data in organizational and management research: A review of opportunities and risks. *Organizational Research Methods, 21*(3), 548–591.

Yammarino, F. J., Mumford, M. D., Serban, A., & Shirreffs, K. (2013). Assassination and leadership: Traditional approaches and historiometric methods. *The Leadership Quarterly, 24*(6), 822–841.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine, 13*(3), 55–75.

Yu, A. Z., Ronen, S., Hu, K., Lu, T., & Hidalgo, C. A. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data, 3*(1) 150075.

Zhu, J., Song, L. J., Zhu, L., & Johnson, R. E. (2019). Visualizing the landscape and evolution of leadership research. *The Leadership Quarterly, 30*(2), 215–232.