Supplementary Information for

Social Media-Predicted Personality Traits and Values Can Help Match People to their Ideal Jobs

Margaret L. Kern, Paul X. McCarthy, Deepanjan Chakrabarty and Marian-Andrei Rizoiu

Address correspondence to:
E-mail: Peggy.Kern@unimelb.edu.au

This PDF file includes:

Supplementary text
Figs. S1 to S2
Tables S1 to S2
References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1917942116

Supporting Information Text

Initial exploration of personality digital fingerprints. Several existing studies have found distinctive Big 5 profiles for specific occupations, as well as correlations between personality and career satisfaction. For instance, Williamson and colleagues (1) found that engineers tended to be less conscientious, extraverted and emotionally stable than a general sample of over 75,000 non-engineer professionals. Studies by Lounsbury and colleagues have found that human resource managers tended to be more open, agreeable, extraverted, emotionally stable, and less conscientious than other professionals (2); scientists who were more extraverted, open, agreeable, and emotionally stable than other scientists reported greater career satisfaction (3); and information technology professionals who were higher in extraversion and openness were more likely to report being satisfied with their job and career (4). Other studies identified evidence for different clusters of traits based on librarian sub-specialities (5), and distinctive Big 5 patterns for Finnish physicians, depending on gender and specialty (6).

Each of these studies focused on single occupations, comparing Big 5 trait scores of participants from a specific profession against a broader population, or identifying personality-based patterns within a specific profession. Rather than focusing on one occupation at a time, our analyses explored whether distinctive personality profiles appeared across a range of occupations. In addition, whereas these studies assessed personality using various self-reported Big 5 measures, our analyses automatically assessed Big 5 from language expressed on Twitter. We also added 5 of Schwartz's Basic Values (7). We did not expect to directly replicate the findings described above, but rather our analyses examined whether we could conceptually replicate the idea that different occupations have distinctive personality fingerprints.

As an initial exploration, we hand-curated a small dataset of 1,035 users across 9 occupations. Our selection was based on the patterns described above, lists that we encountered on Twitter, and our own curiosities. Extending the studies described above, we included software developers, elementary school librarians, health care professionals, and scientists. To provide a contrast to these cognitively-intensive professions, we added one type of athlete (tennis professionals). In addition, as we began curating a list of users, we came across lists of female futurists and top chief information officers, which we chose to include out of curiosity. For transparency, we include all occupations that we explored.

For each of the 9 occupation, we used publicly available lists of people, which we manually mapped to their Twitter user id. Most of these lists featured users selected for their extrinsic career success, as indicated by independent rankings or peer validation in mainstream media. We assumed that the others were successful in their careers, due to receiving high ranks on Twitter. Supplemental Table S1 summarizes the sources of and number of users selected from each occupation.

The initial hand-curated occupations were selected mainly opportunistically, where we could readily find public samples of Twitter users with clearly defined common occupations. For example, most of the world's top-ranked tennis players have active Twitter profiles. Similarly, many of the world's most productive open source software developers on Github also have active Twitter profiles. Leading Chief Information Officers, Science Stars, and Female Futurists were readily identifiable through news stories and had a clear public profile.

For the larger scale replication, occupations for prediction were selected on the basis of those with the largest sample sizes from the sample data we had drawn from Twitter. As such, these were not selected randomly and are likely reflective of more popular self-reported occupations of Twitter users.

Our selection process means that the sample is limited in a number of ways. The sample is based on English-speaking Twitter users. There needs to be a certain amount of linguistic data available to adequately determine digital fingerprints. Further, as popular social media networks continually shift, the extent to which Twitter will provide a useful source of data in the future is questionable. As such, care should be taken in generalizing the pattern of results to non-Twitter users, non-English speakers, and users with limited textual data. However, the methods could be applied to other linguistic data - whether that comes from Twitter users or other social media platforms.

Developing the Vocation Map: Measuring distances between similar and dissimilar professions. Before building the full vocation map, we first considered whether the similarity between professions can be measured using the Euclidean distances between the personality digital fingerprints of the occupations – the 10 dimensional vector of the Big 5 traits and 5 Basic Value scores. We selected two job titles that logically have different responsibilities and require different skill sets - *Congressperson* and *Game Designer*. Supplemental Table S2 indicates the 10 occupations that were farthest and closest from these occupations, based on the Euclidean distance between the personality digital fingerprints of these occupations. *Congressperson* and *Game Designer* appeared to be quite opposite professions; the job titles that were most distant from *Congressperson* related to game programming (e.g., *Video Game Engineer*, *Game Designer* or *Game Programmer*), and the most distant job titles from *Game Designer* reflected US politics and management positions. In contrast, the closest occupations were very similar to the target profession. For instance, *US Senator*, *Senator* and *Congressional Representative* are types of *congresspeople*, while *Game Engineers*, *Game Programmers*, and *Senior Game Designers* require similar skills sets as *Game Designers*.

Choosing the number of clusters. The Partitioning Around Medoids (PAM) (8) algorithm has one hyper-parameter k – the number of clusters – which is chosen ahead of execution. We chose to trade off between cluster granularity (where a higher number of clusters is better) and ease of interpretation (where a lower number of clusters is better). Supplemental Figure S1 verifies the adequacy of choosing 20 clusters by using the silhouette plot. The plot measures the quality of a clustering to determine how well each object lies within its cluster. The red vertical dashed line shows the average silhouette score for all the samples in this clustering, based on 20 clusters. The height of each cluster corresponds to the number of professions in the cluster and the x-axis value represents the silhouette score for each profession. A high average silhouette width indicates good clustering. A given partition is said to be consistent if all clusters have a significant number of samples with silhouette scores larger than the mean of the silhouette scores of all the samples combined, as occurs here (indicated by all 20 silhouettes extending beyond the dashed red line). The plot thus supports the use of 20 clusters.

Additional measures of prediction performance. In the main text, we reported the accuracy of predicting occupations starting from personality fingerprints. The accuracy has the advantage of being an intuitive measure (i.e., the probability that a prediction will be correct), but suffers from being overly optimistic with unbalanced datasets.

Here, we compare the predictions with the observed profession using three standard Machine Learning performance measures: *precision*, *recall*, and *f1*. Precision measures how many of the predictions were correct. Recall measures the completeness of the prediction – how many of the true answers were correctly uncovered. The f1 is the harmonic mean of precision and recall – a classifier needs to achieve both a high precision and a high recall in order to obtain a high f1. These standard performance measures can be compared to a random classifier, which would pick a label at random and obtain precision, recall, and f1 scores of 0.1. Figures S2a to S2c show the prediction precision, recall, and f1 score respectively. We observed that all five classifiers achieved up to five times better performance than the random classifier, with XGBoost demonstrating the best classification. In addition, although the absolute precision, recall, and f1 scores were lower than the accuracy score shown in the main text (as they are different measures), they show the same relations between the performances of the five classifiers.

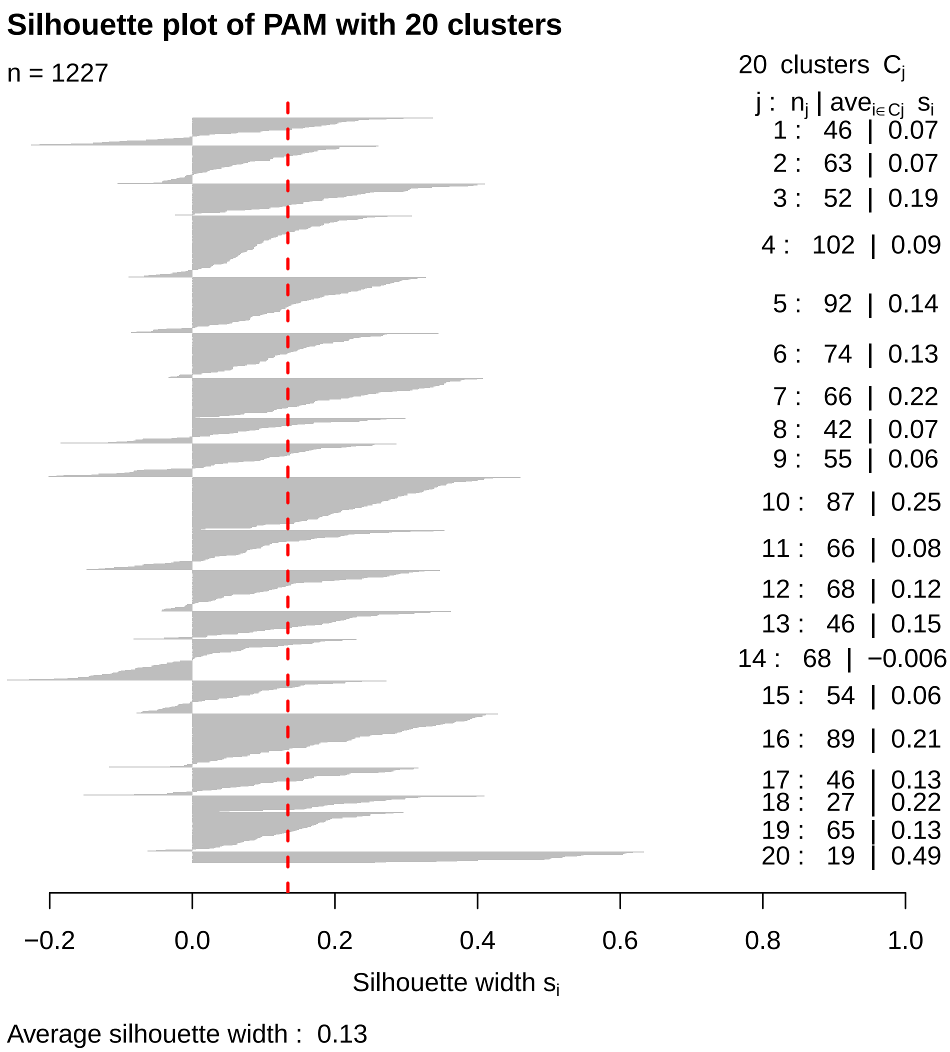
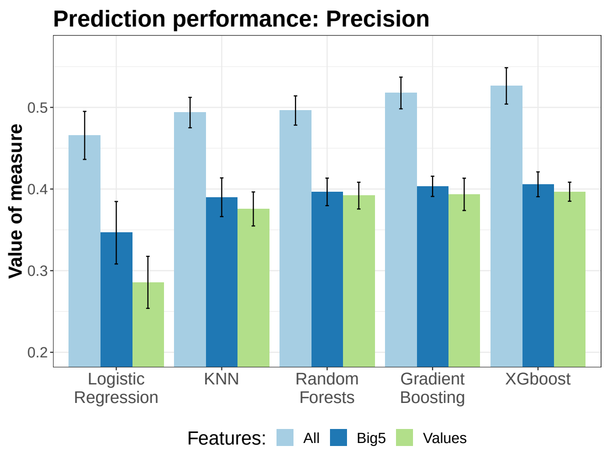
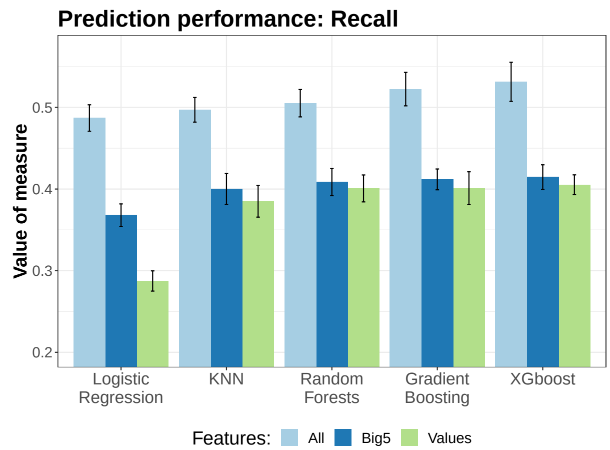


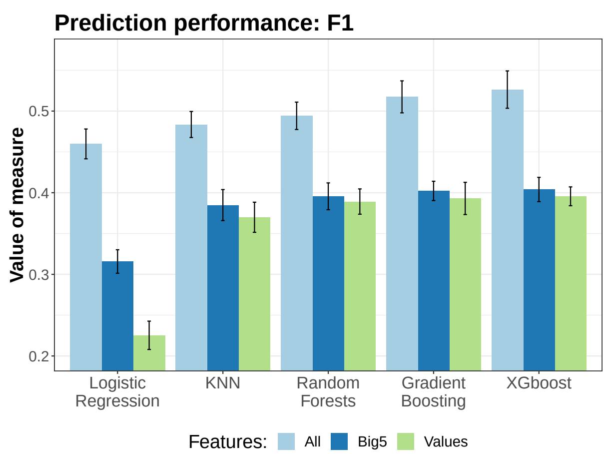
Fig. S1. Silhouette plot of PAM with 20 clusters.



(a)



(b)



(c)

Fig. S2. Performances of predicting professions using the Big 5 features, the 5 Basic Values features and both features sets. Performances are measured using *precision* **(a)**, *recall* **(b)**, and the *f1 score* **(c)**. Across five machine learning classifiers, we predicted the professions of individuals with over five times better precision than random selection using predicted personality traits or values data alone.

Table S1. Proof of concept occupations, sources, and number included

Occupation	Source	Count
Software Developers	Contribution and Influence of Developers on Github https://gist.github.com/paulmillr/2657075	236
Architects	Twitter Advanced Search-Ranked Architects *(manually filtered to remove companies)*	23
Elementary School Librarians	Twitter Advanced Search-Ranked Elementary School Librarians *(manually filtered to remove companies)*	31
Doctors and Health Care Professionals	Twitter Advanced Search-Ranked Doctors *(manually filtered to remove companies)*	43
Female Futurists	List of the world's top female futurists https://rossdawson.com/blog/list-of-the-worlds-top-female-futurists/	119
Science Stars	The Top 50 science stars of Twitter (AAAS) https://www.sciencemag.org/news/2014/09/top-50-science-stars-twitter	50
Top 100 Chief Information Officers	Top 100 Most Social CIOs on Twitter 2016 https://www.huffingtonpost.com/vala-afshar/2016-top-100-most-social-_b_9765538.html	93
Tennis Professionals (Male and Female)	Top Ranked Global Tennis Players on Twitter twitter.com/WTA/lists/players/members (WTA) twitter.com/ATPWorldTour/lists/players/members (ATP)	170
Chemists	100 Chemists on Twitter https://stuartcantrill.com/2014/09/22/100-chemists-on-twitter/ *(Aimed to address perceived gap in Science Stars List)*	95
Total		**1035**

Table S2. Closest (top) and farthest (bottom) occupations (and their cluster) to *Congressperson* and *Game Designer*.

#	Occupation	Cluster name	#	Occupation	Cluster name
	Top 10 Most Similar Roles to Congressperson			**Top 10 Most Similar Roles to Game Designer**	
1	US Senator	US Representative	1	Game Engineer	Software Programmer
2	Senator	US Representative	2	Game Programmer	Software Programmer
3	Congressional Representative	US Representative	3	Video Game Programmer	Software Programmer
4	US Representative	US Representative	4	Senior Game Designer	Software Programmer
5	Member of Congress	US Representative	5	Game Developer	Software Programmer
6	Policy Officer	US Representative	6	Software Engineer	Software Programmer
7	Governor	US Representative	7	Programmer	Software Programmer
8	Inspector	US Representative	8	Video Game Designer	Software Programmer
9	County Treasurer	US Representative	9	Lead Designer	Software Programmer
10	Lieutenant Governor	US Representative	10	Compiler	Software Programmer
	Top 10 Most Different Roles to Congressperson			**Top 10 Most Different Roles to Game Designer**	
1	Concert Promoter	Concert Manager	1	Director of Athletics	Student Services Director
2	Video Game Engineer	Software Programmer	2	Policy Officer	US Representative
3	Game Designer	Software Programmer	3	US Senator	US Representative
4	Game Engineer	Software Programmer	4	Lieutenant Governor	US Representative
5	Talent Buyer	Concert Manager	5	Recruiting Coordinator	District Manager
6	Gaming Manager	Network Manager	6	Congressman	US Representative
7	Game Programmer	Software Programmer	7	Salon Manager	Fashion Marketer
8	Video Game Programmer	Software Programmer	8	Student Affairs Vice President	Student Services Director
9	Senior Game Designer	Software Programmer	9	Congressional Representative	US Representative
10	Game Developer	Software Programmer	10	Senator	US Representative

References

1. Williamson JM, Lounsbury JW, Han LD (2013) Key personality traits of engineers for innovation and technology development. *Journal of Engineering and Technology Management* 30(2):157–168.
2. Lounsbury JW, Steel RP, Gibson LW, Drost AW (2008) Personality traits and career satisfaction of human resource professionals. *Human Resource Development International* 11(4):351–366.
3. Lounsbury JW, et al. (2012) An investigation of the personality traits of scientists versus nonscientists and their relationship with career satisfaction: Relationship of personality traits and career satisfaction of scientists and nonscientists. *R&D Management* 42(1):47–59.
4. Lounsbury JW, Moffitt L, Gibson LW, Drost AW, Stevens M (2007) An Investigation of Personality Traits in Relation to Job and Career Satisfaction of Information Technology Professionals. *Journal of Information Technology* 22(2):174–183.
5. Williamson J, Pemberton A, Lounsbury J (2008) Personality traits of individuals in different specialties of librarianship. *Journal of Documentation* 64(2):273–286.
6. Mullola S, et al. (2018) Personality traits and career choices among physicians in Finland: employment sector, clinical patient contact, specialty and change of specialty. *BMC Medical Education* 18(1).
7. Schwartz SH (2012) An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture* 2(1).
8. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ (2006) Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms* 5(4):475–504.