A DATA ANALYTICS APPROACH TO PERSONA DEVELOPMENT FOR THE FUTURE MOBILE OFFICE

Amudha V. Kamaraj, Atefeh Katrahmani, Mengyao Li, John D. Lee
Department of Industrial & Systems Engineering
University of Wisconsin – Madison, Madison, WI

The concept of using automated vehicles as mobile workspaces is now emerging. Consequently, the in-vehicle environment of automated vehicles must be redesigned to support user interactions in performing work-related tasks. During the design phase, interaction designers often use *personas* to understand target user groups. Personas are representations of prototypical users and are constructed from user surveys and interview data. Although data-driven, large samples of user data are typically assessed qualitatively and may result in personas that are not representative of target user groups. To create representative personas, this paper demonstrates a data analytics approach to persona development for future mobile workspaces using data from the occupational information network (O*NET). O*NET consists of data on 968 occupations, each defined by 277 features. The data were reduced using dimensionality reduction and 7 personas were identified using cluster analysis. Finally, the important features of each persona were identified using logistic regression.

INTRODUCTION

Highly automated vehicles equipped with SAE Level 4 and Level 5 automation may soon be possible (Hancock, Nourbakhsh, & Stewart, 2019). With these levels of automation, vehicle users can give navigational responsibilities to vehicle automation. By transferring navigational responsibilities, time that was previously spent driving can now be used for non-driving related tasks. The possibility of users engaging in non-driving related tasks has spurred conversations on the future of work in automated vehicles by converting the vehicle into a mobile workspace (Janssen et al., 2019; Kun, Shaer, Riener, Brewster, & Schartmüller, 2019; Pollmann, Stefani, Bengsch, Peissner, & Vukelić, 2019). In this paper, this mobile workspace is referred to as the *auto-mobile office* and is defined as an automated vehicle that can support users in performing work-related tasks during commutes (Li, Katrahmani, Kamaraj, & Lee, 2020).

Data on the time spent traveling by US workers shows that workers spend an average of 50 minutes traveling for work daily (Boyle, Lee, & Sadun, 2019). Efforts to convert travel time to work time are already being pioneered by automakers like Audi, Volvo, and BMW (Alessandrini, Campagna, Delle Site, Filippi, & Persia, 2015). As part of Audi's 25th-hour project, the automaker has identified three modes of time use in the vehicle – quality time, productive time, and regeneration time (Savov, 2017). For an auto-mobile office to be useful, designers should consider modifying the in-vehicle environment to support the interactions that arise from users engaging in work-related tasks.

Persona-based Approach for the Auto-Mobile Office

Discussions around the topic of the auto-mobile office have focused on the nature of work-related tasks and technologies like voice, augmented reality and tangible interfaces that can support these tasks (Kun et al., 2019). While these issues are important, before determining the nature of the task, designers can benefit from defining the target user group. A commonly used concept for defining target user groups is that of the *persona.* A persona refers to a fictional representation that is created based on user data to represent different target user groups concretely (Cooper & others, 2004; Pruitt & Grudin, 2003).

The utility of the persona approach is evidenced by its implementation in several complex and novel designs (D'Souza & Lincoln, 2004; Dharwada, Greenstein, Gramopadhye, & Davis, 2007). For systems like the auto-mobile office, a task-related approach to design may lead to the binary classification of work as either knowledge work or manual work. Here, knowledge work refers to work done in an office environment and manual work includes work that might be done in a factory setting (Drucker, 1999). Knowledge work is uniquely suited for mobile workspaces as communication, data, and mobile computing are some of the bare necessities needed for knowledge work (Davis, 2002). Advances in embedded computing make it possible to integrate these necessities into an automated vehicle. Although suited for the auto-mobile office, using the generalized term 'knowledge work' risks overlooking the specific needs of certain user groups leading to poor system design. Complex designs like the auto-mobile office are faced with the need to avoid binary distinctions and design inclusively. A persona-based approach can support inclusive design by identifying different groups of target users within knowledge workers and their needs.

Existing Drawbacks of Persona-Based Approaches

Although the persona-based approach offers advantages in identifying target user groups, several issues have been identified in implementing the approach (Chapman, Love, Milham, ElRif, & Alford, 2008; Chapman & Milham, 2006). The first concerns the curse of dimensionality associated with user data. Personas are often based on user interviews or survey data. Although this is a data-driven process, these methods are limited in terms of sample size and so the data might not fully represent the user population. A large survey might produce a more representative sample. However, with large datasets, the high dimensionality of the data hinders data analysis, and effective quantitative methods for data analysis are rarely used. Second, data are often analyzed subjectively, and the resulting personas are prone to errors and biases. To address these issues, we propose a data analytic approach that can assist designers in constructing personas with large datasets.

Quantitative Methods for Persona Development

Existing quantitative methods for developing personas include latent semantic analysis (LSA), exploratory factor analysis (EFA), and cluster analysis. Miaskiewicz et al. applied LSA to help library staff understand faculty and graduate student needs before designing an institutional repository (Miaskiewicz, Sumner, & Kozar, 2008). Although useful, textual data alone is insufficient in developing personas. McGinn and Kotamraju (2008) and Zhang et al. (2016) implemented EFA to analyze large-sample survey data, and hierarchical cluster analysis to analyze user clickstream data, respectively. While these methods are useful to quantitatively assess user data, they do not effectively integrate dimensionality reduction to deal with large sample sizes and visualization techniques to aid in data interpretation (McGinn & Kotamraju, 2008; Zhang, Brown, & Shankar, 2016). To address this, we present a method for analyzing high-dimensional data and uses data visualization techniques that facilitate the interpretation of results.

The system of interest here is the auto-mobile office and the target users are generally encompassed in the very broad category of knowledge workers. Generating a single persona to represent knowledge workers is insufficient to understand users and their needs. Thus, we sought data that offers a more granular view of the different types of knowledge workers. The Occupational Information Network (O*NET) database provides a granular description of work associated with several occupations and its workers' characteristics. Analysis of this dataset can construct the personas of knowledge workers that may benefit from the concept of the auto-mobile office.

METHODS

The O*NET database is a publicly available database that contains a list of 968 occupations and 277 features that describe the work associated with occupations and the characteristics of its workers (O*NET OnLine, 2018). The features listed in the database are associated with three measurement scales: importance, level, or extent of the activity. Levels are rated on a scale of 0–7. The level rating indicates the degree to which a feature is needed to perform tasks within an occupation, and this was determined to be the only scale relevant to the analysis presented here.

For the proposed analysis, four categories of features that use the level scale are selected: worker abilities, worker knowledge, worker skills, and work activities. These four categories together contain 161 variables that are rated on a scale of 0–7. Worker abilities are the attributes of the individual that influence performance (e.g., cognitive abilities, physical abilities). Worker knowledge is an organized set of principles and facts applied to general domains (e.g., biology, engineering, and technology). Worker skills indicate the capacities that facilitate learning or the more rapid acquisition of knowledge (e.g., social skills, technical skills). Work activities include data on the general types of activities that are required in an occupation (e.g., drafting, operating vehicles, interacting with computers).

This data can be used to find: (1) the grouping of occupations and (2) the most important features in each occupation group. Following the grouping and feature importance of each group, the O*NET database can be referenced to find the tasks of specific occupations, tools and skills needed to complete tasks, and the context of work. Extracting this information can help designers develop a set of personas that can guide design decisions for developing the auto-mobile office. To deal with the high dimensionality of the data and interpretation, a method that combines dimensionality reduction, clustering, and logistic regression is proposed (see Figure 1). First, dimensionality reduction extracts a low dimensional representation of high dimensional data. Following this, groupings with the data are identified using cluster analysis. Finally, each group or cluster is interpreted by extracting its most important features using logistic regression.

Dimensionality reduction

Dimensionality reduction techniques reduce high-dimensional data into a low-dimensional space while preserving the properties of the high dimensional data. While principal component analysis (PCA) is a commonly used method for dimensionality reduction, it is limited in its use when dealing with complex and non-linear data (Jolliffe & Cadima, 2016). A recently developed technique for data reduction is Uniform Manifold Approximation and Projection (UMAP) and it deals with complex data and has good runtimes, reproducibility, and the ability to accommodate non-linear relationships in the data (McInnes, Healy, & Melville, 2018). For these reasons, UMAP is used for the dimensionality reduction of the data.

Cluster Analysis

Cluster analysis is used to identify similar groups within the data. The low-dimensional representation obtained from dimensionality reduction is used to cluster the data into groups and visualize them (Kaski & Peltonen, 2011). In cluster visualization, the data points within a group share similar features compared to the data points in other groups. In the context of the O*NET database, occupations that share similar features are grouped.

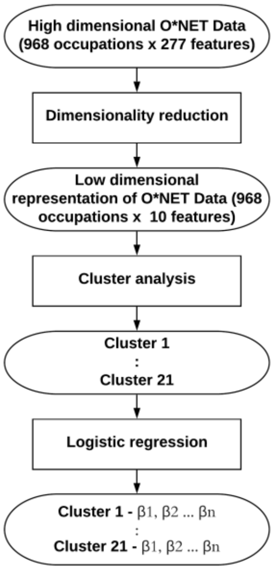

High dimensional O*NET Data (968 occupations x 277 features)
Dimensionality reduction
Low dimensional representation of O*NET Data (968 occupations x 10 features)
Cluster analysis
Cluster 1
:
Cluster 21
Logistic regression
Cluster 1 - β1, β2 ... βn
:
Cluster 21 - β1, β2 ... βn

Figure 1. Flowchart of steps for persona generation using dimensionality reduction, cluster analysis, and logistic regression.

Logistic regression

The cluster analysis produces *n* clusters where each cluster contains similar occupations. Following this grouping of occupations, the next objective is to identify representative features that differentiate the clusters. These clusters along with the representative features can be used to construct a persona for each cluster.

Here, logistic regression is used to extract the representative features of each cluster (Kleinbaum, Dietz, Gail, Klein, & Klein, 2002). These models are popular for classification problems along with estimating the importance of each feature in the classification. Instead of fitting a straight line like in a linear regression model, logistic regression models use a logistic function to determine the output (Molnar, 2019). The coefficients of each feature (β) in a logistic regression function can then be used as an indicator of feature importance. Here, the classification problem is the cluster membership and the coefficients determine the most important features of cluster *n.* A binary outcome of 1 is assigned to cluster *n* and 0 for all other clusters. The resulting coefficients of each feature for cluster *n* define the feature importance.

RESULTS

Implementation of UMAP and cluster analysis was achieved using R.3.6.2 with the *umap* package (Konopka, 2018) and the *clvalid* package (Brock, Pihur, Datta, & Datta, 2008). UMAP was used to reduce the dimensions of the data from 968 occupations with 161 features to 968 occupations with 10 features. These reduced dimensions were then used to cluster the occupations. The *clValid* package was used to determine the optimal number of clusters (n = 21) and the algorithm (divisive analysis clustering). The cluster category that each occupation belongs to is extracted and overlaid on the reduced representation obtained using UMAP. The 21 groups of occupations are shown in Figure 2. The clusters containing the knowledge worker occupations are determined by assessing the occupations in each cluster. The ones determined to contain occupations that will benefit from the auto-mobile office are labeled and highlighted. Following clustering analysis, cluster interpretation is achieved via logistic regression using the *glmnet* package in R (Hastie & Qian, 2014). The resulting logistic regression includes a set of β coefficients for each feature and these coefficients are used to identify the features that distinguish one cluster from another. In total, 7 clusters of knowledge workers were determined to benefit from using the auto-mobile office based on these features.

DISCUSSION

Following dimensionality reduction and cluster analysis, a total of 7 clusters were determined to be composed of knowledge worker occupations that can benefit from the auto-mobile office. Based on assessing the nature of occupations using the features extracted from logistic regression, the 7 clusters are categorized as (1) office and administrative support occupations, (2) computer and mathematical operations, (3) social science and engineering occupations, (4) business, financial, and sales operations, (5) media occupations, (6) management occupations, and (7) education instruction occupations.

Each cluster is used to construct a persona by extracting data from the O*NET database on tasks, tools, skills, and context of work associated with the occupations. Figure 3 shows the personas of three knowledge workers extracted from three different clusters. The first persona is a chief executive officer from the management occupations cluster. The second is a software developer from the computer and mathematical occupations cluster. The third is a human resources assistant from the office and administrative support.

An examination of these three personas reveal shared and unique needs between them. Shared needs can be seen from the context of work that deals primarily with telephones, e-mails, and face-to-face discussions. These shared needs can be supported by embedding technologies in automated vehicles that support these contexts. Unique needs between the personas are evidenced through the tools needed for each persona. For example, the primary skill of software developers is programming, and the tools needed to support them include access to computer servers. This points to the need for high-performance computing that is not needed for the other personas. This is also further supported by the features of importance that differentiate each persona using the coefficients derived from logistic regression. From Figure 3 (right), programming skills along with knowledge of computers and electronics have been identified as important features for software developers. Thus, designers can refer to these personas to identify (1) shared needs between different user groups, and (2) unique needs for a specific user group.

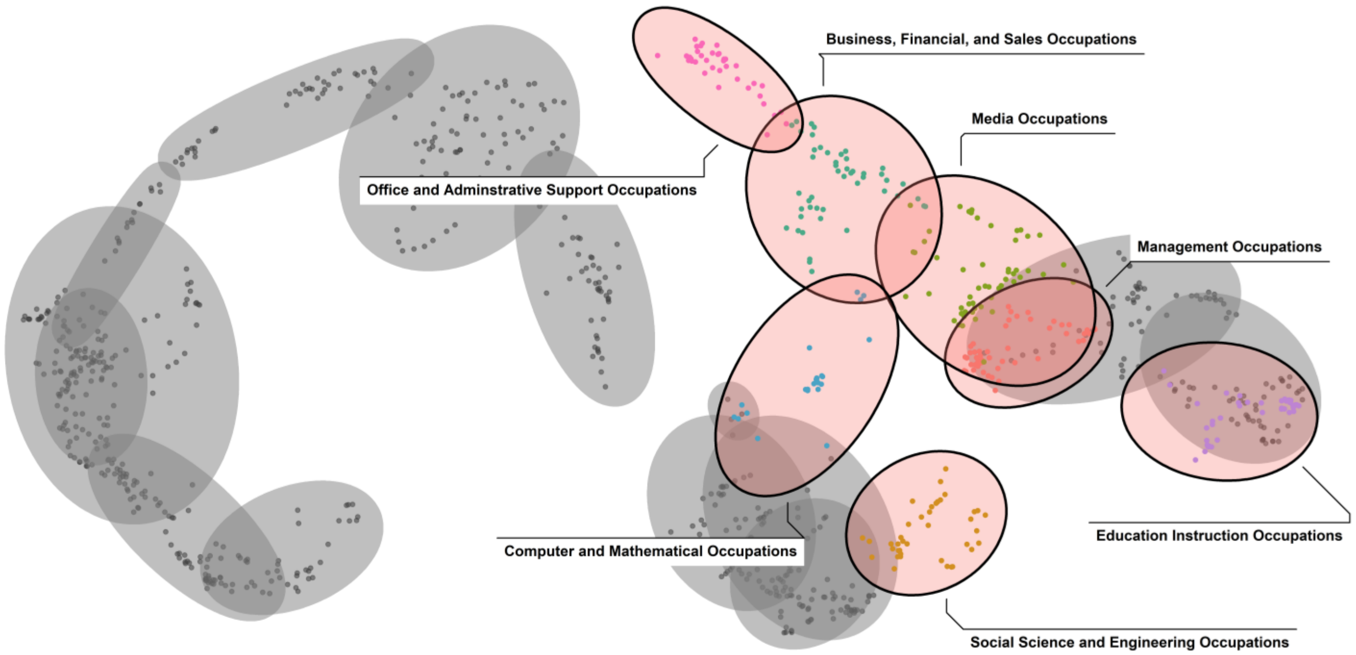


Figure 2. Low dimensional representation of O*NET data using UMAP. Clusters highlighted in pink contain knowledge worker occupations that can benefit from the auto-mobile office.

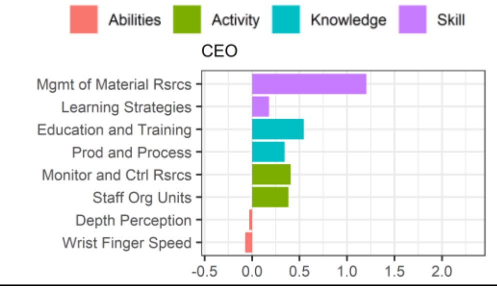
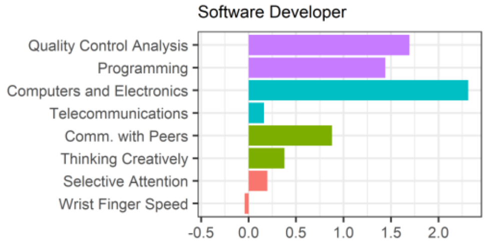
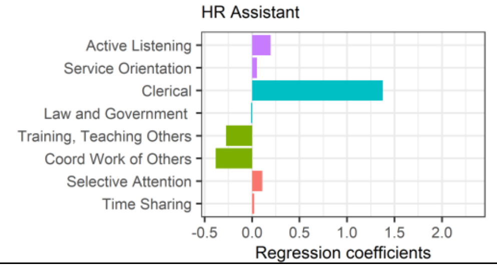
Ginni Griffith CEO Task Direct or coordinate an organization	Tools Mobile phones, computers, PDAs Skills Decision making, problem-solving, critical thinking Context E-mail, face-to-face discussions, telephone	
Andi Smith Software Developer Task Modify existing software to correct errors	Tools Computer servers, computers, PDAs Skills Programming, decision making, problem-solving Context E-mail, face-to-face discussions, teamwork	
Bailey Alcott Human Resources Asst Task Process, verify and maintain personnel documentation	Tools Computers, calculators, fax machines Skills Listening, speaking, writing, critical thinking Context E-mail, telephone, teamwork	

Figure 3. The persona of three groups of knowledge workers extracted from the cluster analysis (left), Features of importance (right).

An additional benefit of identifying design needs through personas is that they can also be used as metrics to evaluate the design. By satisfying the evaluation criteria outlined by the personas, designers may be able to ensure inclusive design.

Limitations and future work

Using the O*NET database to arrive at the different groupings of occupations is a useful first step to create personas that are grounded in user data. The O*NET data provides a macro-level view of the different user groups and this limits persona development. A more granular view of users that incorporates demographic data can improve the quality of the personas. One way to achieve this is to create a crosswalk between different sources of data. For example, designers can gather survey data from target users to supplement data obtained from cluster analysis. Future work pertinent to this project includes creating a crosswalk between the O*NET database and the American Time Use Survey (ATUS) that provides diary data for people who work in the O*NET occupations.

CONCLUSIONS

Dimensionality reduction combined with hierarchical cluster analysis and logistic regression identified groups of occupations from the O*NET database. This process aided in developing seven personas for the auto-mobile office design. Thus, the data analytic technique presented here allows designers to analyze large datasets effectively. This offered a representative sample of a broad range of occupations that direct observation might neglect. For the auto-mobile office, these insights may help analysts understand the work of today through a lens that might help us envision the future of work.

ACKNOWLEDGMENTS

We thank the members of the University of Wisconsin-Madison Cognitive Systems Laboratory for their insightful discussions and comments. This project is supported by NSF Grant CMMI-1839484.

REFERENCES

Alessandrini, A., Campagna, A., Delle Site, P., Filippi, F., & Persia, L. (2015). Automated vehicles and the rethinking of mobility and cities. *Transportation Research Procedia*, *5*(2015), 145–160.

Boyle, L. N., Lee, J. D., & Sadun, R. (2019). *Towards Work in Automated Vehicles*.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid, an R package for cluster validation. *Journal of Statistical Software*, *25*(4), 1–22.

Chapman, C. N., Love, E., Milham, R. P., ElRif, P., & Alford, J. L. (2008). Quantitative evaluation of personas as information. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(16), 1107–1111.

Chapman, C. N., & Milham, R. P. (2006). The personas' new clothes: methodological and practical arguments against a popular method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(5), 634–636.

Cooper, A., & others. (2004). *The inmates are running the asylum:[Why high-tech products drive us crazy and how to restore the sanity]* (Vol. 2). Sams Indianapolis.

D'Souza, M. E., & Lincoln, N. H. (2004). A persona-centric approach to developing complex computer systems: Lessons from the field. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(5), 917–921.

Davis, G. B. (2002). Anytime/anyplace computing and the future of knowledge work. *Communications of the ACM*, *45*(12), 67–73.

Dharwada, P., Greenstein, J. S., Gramopadhye, A. K., & Davis, S. J. (2007). A case study on use of personas in design and development of an audit management system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(5), 469–473.

Drucker, P. F. (1999). Knowledge-worker productivity: The biggest challenge. *California Management Review*, *41*(2), 79–94.

Hancock, P. A., Nourbakhsh, I., & Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. *Proceedings of the National Academy of Sciences*, *116*(16), 7684–7691.

Hastie, T., & Qian, J. (2014). Glmnet vignette. *Retrieve from Http://Www. Web. Stanford. Edu/~ Hastie/Papers/Glmnet_Vignette. Pdf. Accessed September*, *20*, 2016.

Janssen, C. P., Kun, A. L., Brewster, S., Boyle, L. N., Brumby, D. P., & Chuang, L. L. (2019). Exploring the concept of the (future) mobile office. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, 465–467.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Kaski, S., & Peltonen, J. (2011). Dimensionality reduction for data visualization [applications corner]. *IEEE Signal Processing Magazine*, *28*(2), 100–104.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*. Springer.

Konopka, T. (2018). R-package: umap. *Uniform Manifold Approximation and Projection*.

Kun, A. L., Shaer, O., Riener, A., Brewster, S., & Schartmüller, C. (2019). AutoWork 2019: workshop on the future of work and well-being in automated vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, 56–62.

Li, M., Katrahmani, A., Kamaraj, A. V., & Lee, J. D. (2020). Defining A Design Space of the Auto-Mobile Office: A Computational Abstraction Hierarchy Analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

McGinn, J., & Kotamraju, N. (2008). Data-driven persona development. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1521–1524.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Preprint ArXiv:1802.03426*.

Miaskiewicz, T., Sumner, T., & Kozar, K. A. (2008). A latent semantic analysis methodology for the identification and creation of personas. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1501–1510.

Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.(2019). *URL Https://Christophm. Github. Io/Interpretable-Ml-Book*.

O*NET OnLine. (2018). National Center for O*NET Development. O*NET OnLine Help: Data Collection Information. Retrieved November 29, 2019, from https://www.onetonline.org/help/online/data

Pollmann, K., Stefani, O., Bengsch, A., Peissner, M., & Vukelić, M. (2019). How to Work in the Car of the Future? A Neuroergonomical Study Assessing Concentration, Performance and Workload Based on Subjective, Behavioral and Neurophysiological Insights. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Pruitt, J., & Grudin, J. (2003). Personas: practice and theory. *Proceedings of the 2003 Conference on Designing for User Experiences*, 1–15.

Savov, V. (2017, July). Audi's 25th Hour project makes time the ultimate driving luxury. *The Verge*. Retrieved from https://www.theverge.com/2017/7/10/15947784/audi-25th-hour-autonomous-car-driving-work-time

Zhang, X., Brown, H.-F., & Shankar, A. (2016). Data-driven personas: Constructing archetypal users with clickstreams and user telemetry. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5350–5359.