International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hihc20

A Survey of 15 Years of Data-Driven Persona Development

Joni Salminen, Kathleen Guan, Soon-Gyo Jung & Bernard J. Jansen

To cite this article: Joni Salminen, Kathleen Guan, Soon-Gyo Jung & Bernard J. Jansen (2021): A Survey of 15 Years of Data-Driven Persona Development, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2021.1908670

To link to this article: https://doi.org/10.1080/10447318.2021.1908670

Published online: 12 Apr 2021.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2021.1908670

A Survey of 15 Years of Data-Driven Persona Development

Joni Salminen [a,b], Kathleen Guan[c], Soon-Gyo Jung[a], and Bernard J. Jansen[a]

[a]Qatar Computing Institute, Hamad Bin Khalifa University, Doha, Qatar; [b]Turku School of Economics, University of Turku, Turku, Finland; [c]Faculty of Brain Sciences, University College London, London, UK

ABSTRACT

Data-driven persona development unifies methodologies for creating robust personas from the behaviors and demographics of user segments. Data-driven personas have gained popularity in human-computer interaction due to digital trends such as personified big data, online analytics, and the evolution of data science algorithms. Even with its increasing popularity, there is a lack of a systematic understanding of the research on the topic. To address this gap, we review 77 data-driven persona research articles from 2005–2020. The results indicate three periods: (1) *Quantification* (2005–2008), which consists of the first experiments with data-driven methods, (2) *Diversification* (2009–2014), which involves more pluralistic use of data and algorithms, and (3) *Digitalization* (2015–present), marked by the abundance of online user data and the rapid development of data science algorithms and software. Despite consistent work on data-driven personas, there remain many research gaps concerning (a) shared resources, (b) evaluation methods, (c) standardization, (d) consideration for inclusivity, and (e) risk of losing in-depth user insights. We encourage organizations to realistically assess their data-driven persona development readiness to gain value from data-driven personas.

1. Introduction

Personas, as imaginary people describing real user segments (An, Kwak, Salminen et al., 2018), are considered a powerful technique for user understanding and user-centric design in human-computer interaction (HCI). Personas are relevant and potentially useful for researchers and practitioners facing user-centric decision-making tasks in a variety of industries and application domains, including software development (Aoyama, 2007; Hang Guo & Razikin, 2015), design (Goodman-Deane et al., 2018; Miaskiewicz & Luxmoore, 2017), e-health (Holden et al., 2017; Wöckl et al., 2012), marketing/advertising (An, Kwak, Jung et al., 2018), cybersecurity (Dupree et al., 2016; Kim et al., 2019), video games (Ishii et al., 2018; Tychsen & Canossa, 2008), online news (An, Kwak, Jung et al., 2018; Watanabe et al., 2017), recommender systems (Hou et al., 2020; Konstantakis et al., 2020), and so on. In the era of personified big data (Spiliotopoulos et al., 2020), personas are particularly useful for segmenting diverse online user populations (Salminen, Jansen et al., 2018). Moreover, personas are necessary for going "beyond segmentation" (Jenkinson, 1994, p. 72) in order to "give faces to data" (Salminen, Jansen et al., 2019, p. 148) to facilitate the adoption of shared mental models about users and enhance stakeholders' empathetic understanding of who their users are (Nielsen, 2019). Personas can yield a positive return on investment for organizations that deploy them (Drego et al., 2010). Therefore, personas are a formidable field of study.

HCI literature advocates multiple approaches to persona development (Pruitt & Grudin, 2003). Brickey et al. (Brickey et al., 2012) found that 81% of efforts to develop personas evaluated in current academic literature applied qualitative methods, such as interviews, field studies, usability tests, and ethnography (among others). However, manual persona development (MPD) has been criticized for developing personas that are not based on rigorous empirical data (Chapman & Milham, 2006) because MPD often uses small samples, one-time data collection, and non-algorithmic methods.

In turn, efforts toward data-driven persona development (DDPD) have gained increasing interest from scholars and practitioners for their use of large amounts of data to permit algorithmic analysis. In this study, we define DDPD as *the use of algorithmic methods to create accurate, representative, and refreshable personas from numerical data*. Manual and data-driven methods can be used in conjunction, but a lack of resources often incentivizes researchers and practitioners to choose one or the other (Thoma & Williams, 2009). In the face of this choice, DDPD provides some distinct relative benefits (B):

"Personified big data" is becoming more common (Choi et al., 2020; Stevenson & Mattson, 2019), which requires

CONTACT Joni Salminen joolsa@utu.fi Qatar Computing Institute, Hamad Bin Khalifa University, RC1, Doha, Qatar

- **B01: Enhanced objectivity**. MPD is associated with a high degree of subjectivity, which hinders the validity of the personas that are developed (An, Kwak, Jung et al., 2018). Researchers see DDPD as a "way to over come subjectivity [of MPD] both in interpretation and segmentation of available data" (Jansen et al., 2017, p. 2128). Furthermore, DDPD approaches tend to be replicable, and they use large sample sizes to increase the user representativeness of the personas (Chapman & Milham, 2006; Siegel, 2010). This statistical robustness boosts both the validity and credibility of the developed personas.
- **B02: Decreased cost**. MPD typically requires several months to complete from start to finish, involving financial costs in tens of thousands of dol lars when conducted by consultants (Drego et al., 2010). The high-cost fac tor makes high-quality personas inaccessible for organizations with limited financial resources (e.g., start-up companies and nonprofit organizations). DDPD mitigates this cost by relying on automation in the critical processes of persona creation, including data collection and analysis, thus offering ways to "democratize" persona development for organizations of all kinds.
- **B03: Updatability**. Shifts in user demographics and behaviors are typical in many fast-moving industries, such as e-commerce (Salminen, Jung, Jan sen et al., 2019a), Web search engines (Jenkinson, 1994), and social media platforms (Li et al., 2016). When the underlying user data changes, person as become outdated unless refreshed using the new user data. For MPD, updating personas requires excessive amounts of costly and non-scalable manual labor, resulting in the personas often not being updated at all. DD PD can capture the change of user behavior over time (Jansen et al., 2019; Jung et al., 2019), as it relies on automated processes for periodic data collection and easy re-analysis using standard algorithms.
- **B04: Scalability and "data readiness."** Manual analysis of data is costly and requires specific expertise, which makes many MPD efforts incompati ble with large datasets that are increasingly common with the rise of soci al media and Web analytics (An, Kwak, Jung et al., 2018). The more numb er of distinct user segments in the baseline data, the more difficult it is to discover them using MPD. In turn, large datasets are not a concern for DD PD methods, as data science and machine learning algorithms have been developed to process large amounts of data.

persona development to keep up with the changing times (Jansen, Jung et al., 2020). Bigger datasets make scaling of MPD difficult because the manual analysis of 100,000 online user comments to find the users' pain points, for example, is not feasible. While MPD works for small data (e.g., a dozen or so user interviews), it is not feasible for navigating the big data environment of today's user data. Instead, DDPD efforts and research are needed to keep personas relevant in the era of online analytics (Jansen, Salminen et al., 2020; Salminen, Jansen et al., 2018; Salminen, Jung, Jansen et al., 2019a; Wu & Yu, 2020). For this, the promise of data science algorithms is exciting: consider the use of machine learning, where persona developers can manually analyze a subset of data and train algorithms to analyze the rest. It is important to note that manual analysis still has a key role in persona development through enriching data-driven personas with in-depth insights; see, e.g., (Salminen, Şengün et al., 2018).

When personas were first introduced to HCI in the late 1990s by Alan Cooper, the Internet was a nascent technology, with limited tools to collect and process large amounts of user data. Since that era of the "invention of personas," the methodologies and platforms for collecting and automatically analyzing user data have advanced by orders of magnitude. Thus, scholars and practitioners in HCI and other disciplines are enticed by DDPD for two main reasons:

(1) the benefits of DDPD (the main ones listed above), and
(2) the emerging trends taking place in the field of user analytics, of which there are three main examples:

   a. personified big data about social media and online users that provides the "raw material" for persona development,
   b. application programming interfaces (APIs) that enable the real-time collection of this user data, and
   c. the rapid development of data science algorithms and open-source systems for scalable and repeatable analysis of the user populations toward identifying core segments that become the bases of the personas.

Thus, the increasing availability of online user data from Web analytics and social media platforms provides pivotal opportunities for persona development (An, Kwak, Jung et al., 2018). This development has dramatically increased the feasibility of DDPD as a means by which to use online sources where "big data" about users or customers is available (Del Vecchio et al., 2018). Such personified big data is increasingly available in social media platforms and online analytics tools (e.g., Google Analytics, YouTube Analytics, Twitter Analytics). Simultaneously, programming languages (e.g., R, Python) and frameworks (e.g., scikit-learn) for data science applications have also evolved a lot, making a variety of statistical techniques and computational approaches accessible for persona development.

Overall, due to these benefits and emerging trends, DDPD is receiving increasing interest from HCI scholars and practitioners alike (Mcginn & Kotamraju, 2008; Brickey et al., 2010; Laporte et al., 2012; Miaskiewicz et al., 2008), prompting remarks such as "there is a shift from using qualitative data towards using quantitative data for persona development" (Mijač et al., 2018, p. 1427).

Nevertheless, the scholarly literature presently lacks a systematic overview and evaluation of the multitude of methods and approaches currently being used for DDPD, along with how they respectively contribute to the strengths and weaknesses of DDPD. This research gap increases the challenge to position work and to identify pivotal opportunities in this emerging field of study. To address this critical gap, we evaluate the current research on DDPD.

We (1) systematically collect, analyze, and synthesize relevant literature within this domain, (2) provide an overview of the main DDPD methods and their strengths and weaknesses, (3) offer an understanding of the current status of the field, as well as (4) derive implications for future research and practice, including themes and strategies. To this end, we formulate the following research questions (RQs):

RQ01: *How have the DDPD research interests and methodologies developed over time?*

RQ02: *What are the critical DDPD challenges and research gaps?*

RQ03: *What are the critical DDPD trends and future outlooks?*

Following the approach of prior literature reviews in computer science (Dillahunt et al., 2017; Hussain et al., 2009), we collect and analyze 77 research articles that developed personas using quantitative methods and were published between January 2005 and December 2020. This manuscript presents an expanded analysis of previously published literature analysis (Salminen, Guan et al., 2020) with renewed data collection and multiple additional analyses regarding temporal coverage, methodological diversity (GIN index), application domains, and venues, as well as further conceptual development of the emerging research periods and limitations of DDPD for research and practice. We now cover articles through to the end of 2020. These additional analyses and conceptualizations further substantiate the outlook of the current state-of-the-art and research gaps for future research.

The reader should note that our aim is not to claim that DDPD is the *only* or necessarily the *best* way of creating personas. Rather, DDPD, in general, as well as its specific methodologies, have limitations (see Section 7 for discussion). While MPD and DDPD methods have been subject to criticism, they share some general shortcomings. First, personas are one form of user-centric design, and alternative methods can, in some use cases, be superior (Salminen, Jung et al., 2020). Second, personas may simplify people down to archetypes, which makes them useful only to a certain degree (Marsden & Haag, 2016; Turner & Turner, 2011). However, we believe that DDPD has great potential for both research and design practice, as its basic premise is to enable the representation of digital user data in a user-friendly manner for various design tasks. These strengths substantiate the need for a systematic review.

2. Related research

2.1. Methods of persona development

Mulder and Yaar (2006) refer to three primary ways of creating personas: (1) qualitative personas, (2) qualitative personas with quantitative validation, and (3) quantitative personas, which we refer to as DDPD. Other researchers refer to hybrid personas that use mixed methods (Pruitt & Grudin, 2003; Salminen, Şengün et al., 2018). Fundamentally, all methods are based on three main steps: (a) user data collection, (b) segmentation and clustering, and (c) synthesis of the (qualitative or quantitative) data to present user segments and their attributes as persona profiles (Wöckl et al., 2012; Zhu et al., 2019).

2.2. A short history of DDPD

The earliest literary reference to the concept of "data-driven persona" to our knowledge was by Williams (2006). The phrase was further popularized by Mcginn and Kotamraju (2008) with their article "Data-Driven Persona Development." Nonetheless, the underlying concept likely goes even further back. One could argue that personas have always been intended to be based on real user/customer data, regardless of whether the data are in qualitative or quantitative formats. Perhaps, it is only the availability and abundance of that data – the recent emergence of "personified big data" – that has changed over time. For example, Gaiser et al. (2006, p. 521) note that "*In order to fulfill standards of a scientific method, personas can't be created arbitrarily. Personas have to be grounded in data, at best, both qualitative and quantitative data of surveys with the target audience*." In a similar vein, Pruitt and Grudin (2003, p. 1) note that "*[personas] provide a conduit for conveying a broad range of qualitative and quantitative data, and focus attention on aspects of design and use that other method do not*."

2.3. Previous literature reviews of DDPD

The literature has widely acknowledged the methodological diversity within DDPD. Zhu et al. (2019) cite several methods, including affinity diagrams, decision trees, exploratory factor analysis (EFA), hierarchical clustering, k-means clustering, latent semantic analysis (LSA), multidimensional scaling analysis (MSA), and weighted graphs. Angela Minichiello et al. (2017) provide a similar record of semi-automated methods: cluster analysis (including both hierarchical and k-means), factor analysis, principal component analysis (PCA), and LSA. These overviews, however, are superficial, as they typically only list the methods without any further evaluation. In the few literature reviews that provide a more extensive overview of DDPD (Brickey et al., 2012; Jon Brickey et al., 2010; Tu, Dong et al., 2010), the focus is solely on clustering methods. There are conceptual articles that discuss the role of personas in the era of online analytics (Salminen, Jansen et al., 2018), compare and contrast methodological arguments against qualitative personas (Chapman & Milham, 2006) or quantitative ones (Siegel, 2010), or provide guidelines for successful persona development (Pruitt & Grudin, 2003). However, these articles do not place focus on or utilize systematic methodologies to review DDPD methods.

2.4. Research gap

We were unable to locate any previous systematic literature reviews on DDPD apart from scoping literature reviews that focused on clustering or superficially listing other quantitative methods (Brickey et al., 2012; Brickey et al., 2010; Tu, Dong et al., 2010). The scope is limited in these incidents. A plethora of algorithms have been applied for DDPD, but there are no assessments of their strengths and weaknesses. As such, it is necessary to systematically survey these attempts and generate useful insights for both persona researchers and practitioners. As noted by Dillahunt et al. (2017, p. 1), "*literature reviews have proved useful and influential by identifying trends and gaps in the literature of interest and by providing key directions for short- and long-term future work*." In the following section, we present our methodology for meeting this goal.

3. Methodology

We consulted two academic databases: Google Scholar (GS) and ACM Digital Library (DL). We chose these two databases due to their comprehensiveness (GS) and relevance to the

topic of DDPD (DL). We carried out identical literature searches in GS and DL during April and December 2020. We also followed the recommended search strategy for systematic reviews by carrying out snowball sampling to find additional articles (Radjenović et al., 2013). The search phrases were devised based on the authors' previous knowledge of the field. They included references to DDPD (quantitative personas, data-driven personas, procedural personas) and specific methodologies (automatic persona generation, personas AND cluster analysis OR clustering OR principal component analysis OR factor analysis OR conjoint analysis OR latent semantic analysis OR matrix factorization). We used both the plural and singular of the word "persona." To limit our search to only articles written in English, we included negative search words in Spanish ("y," "con," "de") as "persona" means "person" in Spanish. The initial search yielded 190 unique articles, of which 119 came from ACM DL and 71 from Google Scholar.

The following specific search phrases used for searching:

- "automatic persona generation"
- "data-driven personas"
- "procedural personas"
- "quantitative personas"
- +personas + "cluster analysis"
- +personas + "clustering"
- +personas + "conjoint analysis"
- +personas + "factor analysis"
- +personas + "latent semantic analysis"
- +personas + "matrix factorization"
- +personas + "principal component analysis"
- +persona + "cluster analysis"
- +persona + "clustering"
- +persona + "conjoint analysis"
- +persona + "factor analysis"
- +persona + "latent semantic analysis"
- +persona + "matrix factorization"
- +persona + "principal component analysis"

Following the searches, we manually screened the articles by reading the abstracts. The articles that passed the screening went through a subsequent full-text review (Figure 1).

At this stage, we applied snowball sampling, allowing us to identify other potential research articles. We subsequently assessed all the articles retrieved via snowball sampling (N = 44) for full-text review. The inclusion criteria at each stage were:

- **full research article** (no short articles, books, or theses) [*screening stage*]
- **published in a peer-reviewed journal or a conference proceedings volume** [*screening*]
- **written in English** [*screening*]
- **empirical paper that develops personas using quantitative data** [*screening/assessment*]

Note that, based on these selection criteria, the number of articles was 190 search-retrieved and 44 snowball-retrieved, resulting in a total of 234 records. Out of these, 11 (4.7%) were duplicates between the two databases, leaving 223 unique articles. Out of these, we discarded non-English articles (N = 11, 4.7%), non-peer-reviewed articles (N = 40, 17.2%), non-full

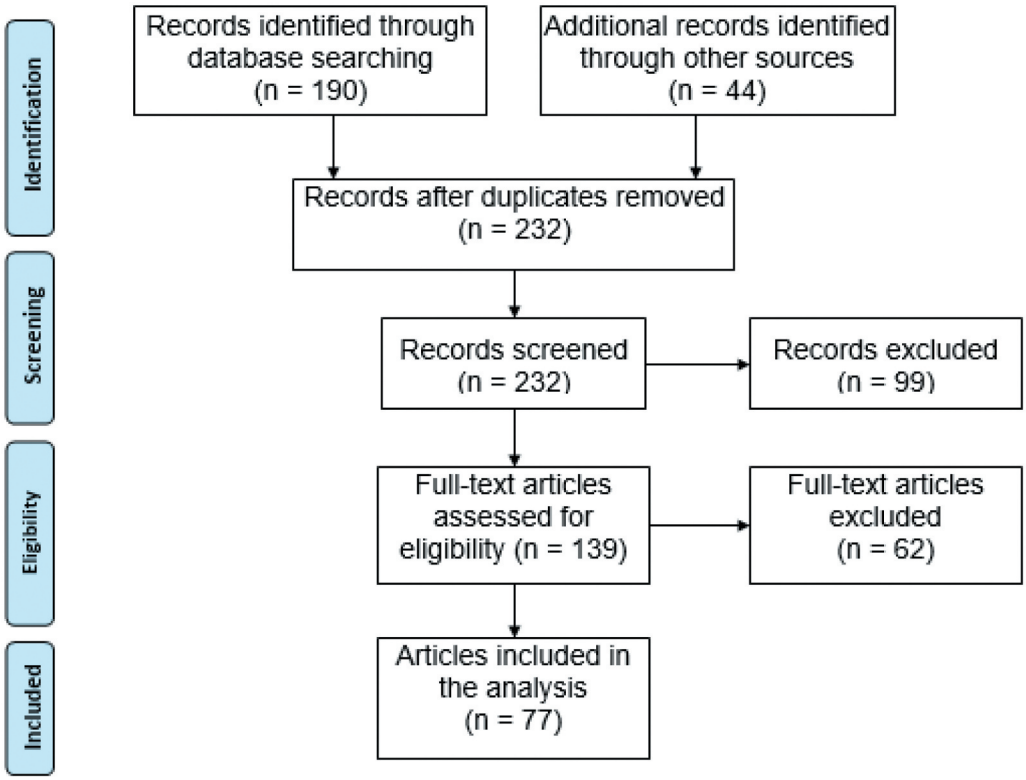

Figure 1. PRISMA flow chart (Van Laar et al., 2017) of the literature collection.

articles (N = 24, 10.3%), and articles not developing data-driven personas using algorithms and quantitative data (N = 93, 41.0%). In total, 146 articles (62.9%) were excluded (note that summing up the class percentages does not match this number because a paper can have many exclusion criteria).

The final collection includes 77 articles, of which 51 (66.2%) were retrieved via searches and 26 (33.8%) via snowball sampling. We kept 26.8% of the search-retrieved articles and 59.1% of the snowball-retrieved articles. The 77 articles are shown in Table A1 (Appendix A).[1] We extracted the data using a standardized data extraction form (Torgerson, 2003) that contained the following information from each selected paper:

- **Basic information**: article title, publication year, keywords
- **Publication information**: name of publication venue and its type (conference/journal)
- **Author information**: authors' institution locations (countries of affiliations)
- **Methodology**: persona development methodology used (e.g., clustering)
- **Mixed methods**: if mixed methods were used (yes/no)
- **Data source**: the source of the data used for persona development (e.g., survey, social media)
- **Data size**: numerical information about the data (number of analysis units, participants)
- **Validation**: validation metrics and methods applied in the research article
- **Future work**: the article authors' suggestions for future work

The resulting dataset was analyzed using descriptive statistics to address the research questions. The following sections present our findings.

4. Research interest in DDPD

4.1. Research articles over time

The earliest paper applying DDPD was written in 2005 by Aoyama (Aoyama, 2005). The researcher applied conjoint analysis to create personas for software embedded in digital consumer products. The first DDPD journal article was published in *IEEE Transactions on Software Engineering* in 2012 (Brickey et al., 2012). Figure 2 shows a stagnating number of DDPD articles per year at first, followed by an increase since 2014. In 2020, the publication count reached its peak at 17 articles. Conference papers were more frequent (N = 57, 74.0%) than journal articles (N = 20, 26.0%) (see Figure 2), perhaps indicative of the significant influence of conference venues in computer science research traditionally. In the third period, there is also an increase in publication numbers relative to earlier years. The first and second periods were characterized by a low number of research articles (as noted in Figure 2). In contrast, the third period saw an average of 9.7 publications per year, a 314.95% increase over the second period (M = 2.33) and a 643.6% increase from the first period (M = 1.3).

4.2. Prominent work

We retrieved citation counts from Google Scholar in December 2020. Table 1 shows the most cited articles. Most typically, the articles have 0 citations (Mode = 0, Mean = 26, Max = 704). There is a weak positive correlation between the years-of-age of the articles and the number of citations ($r = 0.26$). The most cited article (Li et al., 2016) uses dialogs from tweets and movie scripts to develop persona-based conversation models, demarking interest among natural language processing researchers to adopting the concept of persona for dialogue systems.

4.3. Application domains

The application domains were identified based on the contexts (e.g., source of data, industry) and ultimate goals for the development of applicable personas.

- **Healthcare** was one of the top two most common application domains (N = 9, 11.7%). These papers applied personas in order to assist doctors in understanding different patient groups (Goodman-Deane et al., 2018; Holden et al., 2017; Vosbergen et al., 2015). For example, Vosbergen et al. (2015) developed personas of coronary heart disease patients to ultimately establish persona-specific education interventions for different subsets of the patient population (therefore improving patient compliance).

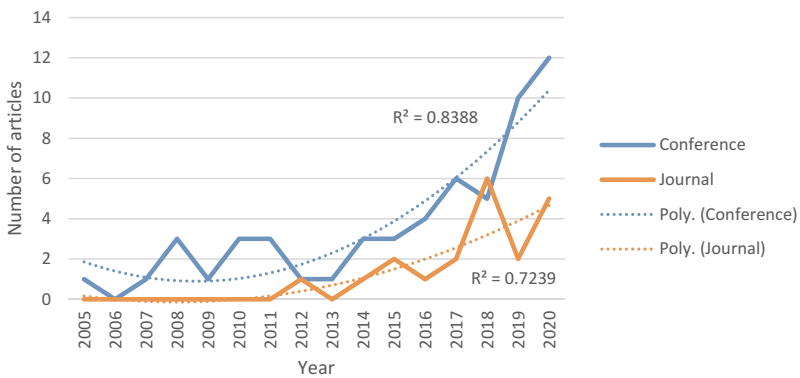

Figure 2. Conference and journal publications over time.

Table 1. Top 10 most cited articles. Citation counts were retrieved from Google Scholar.

Title (Year)	Authors	Citations
A Persona-Based Neural Conversation Model (2016)	Li et al. (2016)	704
Data-Driven Persona Development (2008)	Mcginn and Kotamraju (2008)	155
Learning Latent Personas of Film Characters (2013)	Bamman et al. (2013)	181
Defining Personas in Games Using Metrics (2008)	Tychsen and Canossa (2008)	126
Persona-and-Scenario Based Requirements Engineering for Software Embedded in Digital Consumer Products (2005)	Aoyama (2005)	100
Persona-Scenario-Goal Methodology for User-Centered Requirements Engineering (2007)	Aoyama (2005)	72
A Latent Semantic Analysis Methodology for the Identification and Creation of Personas (2008)	Miaskiewicz et al. (2008)	75
Evolving personas for player decision modeling (2014)	Holmgard et al.(2014)	63
Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry (2016)	Zhang et al. (2016)	65
Invoking the User from Data to Design (2014)	Tempelman-Kluit and Pearce (2014)	38

- **Knowledge management** in the context of education was another top domain (also N = 8, 10.4%). This domain refers to the process of sharing and managing the use of information dissemination in an institutional setting, such as a university's library databases. For example, one study created personas based on library chat support transcripts to comprehend the needs of students using the university library (Tempelman-Kluit & Pearce, 2014).
- Several studies also took place in the context of **social media** (N = 10, 13.0%), including (An, Kwak, Jung et al., 2018; Salminen, Jansen et al., 2019), which analyzed big data from online communities to understand user characteristics and behaviors.
- Researchers also created personas to improve **software** companies' understanding of their customers (N = 5, 6.5%), such as an exploration of the feasibility of a personal safety mobile application for women in India (Hang Guo & Razikin, 2015).
- Finally, researchers applied personas to **games** testing and design (N = 5, 6.5%) to map out the characteristics and narratives of protagonists in the gaming worlds (Tychsen & Canossa, 2008).

5. Methods for DDPD

5.1. Digital data sources

In total, 47% (N = 36) of the articles reported the use of surveys, making it also the most common source for data collection. The second most popular data source was the web and social media data (N = 23, 29.9% of total). This category includes social media platforms (e.g., YouTube An, Kwak, Salminen et al., 2018), discussion forums (Huh et al., 2016), as well as user click logs (Thoma & Williams, 2009), and telemetry (Zhang et al., 2016). Two articles also notably used device-collected data, including GPS signals (Guo & Jianhua, 2018) and physical comfort levels (Dos Santos et al., 2014). Even though this use of device-collected data was marginal, it reveals how "personal big data" can provide interesting information about users, for example, in fields such as health and wellness. Also, behavioral data describing actual user interactions is becoming increasingly common (Minichiello et al., 2018). Ten articles (13.0%) used more than one data source. The most common data source combination was surveys and interviews (N = 8, 80.0% of the multiple data sources). The authors regarded this as a way of enhancing both the breadth (through quantitative data) and depth (through qualitative data) of the generated personas.

5.2. Popularity of methods

We manually tallied the individual methods mentioned in each paper to extract their frequencies (see Table 2). K-means clustering was by far the most popular method (N = 15, 19.5% of total articles), followed by non-negative matrix factorization (N = 13, 16.9%). In total, clustering methods were used in more than a third of the articles (N = 26, 33.8%). In total, 24 articles (31.2%) combined both quantitative and qualitative methods. Furthermore, 37 articles (48.1%) combined multiple quantitative methods, such as k-means clustering with principal component analysis.

5.3. Evaluation approaches

Validation of DDPD tends to be informal and limited. Only a couple of authors had a systematic method for how they

Table 2. Most popular DDPD methods.

Method	Description	Frequency
K-means clustering (KMC)	This is a machine learning algorithm that classifies a dataset using a predetermined prime number (k) of clusters.	N = 15 (19.5%)
Non-negative matrix factorization (NMF)	This is a matrix factorization method, in which matrices are constrained as non-negative. A matrix is decomposed into two matrices to extract sparse and meaningful features.	N = 13 (16.9%)
Hierarchical clustering (HC)	This is a machine learning algorithm that computes the distances between different elements to produce clusters in a hierarchical order based on similarity.	N = 7 (9.1%)
Latent semantic analysis (LSA)	This is a machine learning algorithm that uses singular value decomposition to detect hidden semantic relationships between words.	N = 5 (6.5%)
Principal component analysis (PCA)	This is a linear dimension-reduction algorithm used to extract information by removing non-essential elements with relatively fewer variations.	N = 5 (6.5%)

selected whom to consult during validation. For example, Miaskiewicz & Luxmoore (2017) systematically identified specific surveyed users to represent the personas and further interviews based on k-means distance measures; afterward, they quantitatively compared these individuals' characteristics with the generated personas. Salminen, Şengün et al. (2018) consulted qualitative data from social media users in the geographical region in the forms of Instagram profiles and semi-structured interviews. The researchers used these to further enrich and improve the automatically generated personas. Furthermore, while some studies engaged subject-matter experts (Dupree et al., 2016; Mcginn & Kotamraju, 2008), these evaluations varied, ranging from brief discussions to quantitative coding of interrater agreement levels to the extent that user observations and subject expert evaluations led to substantial and significant modifications in the finalized personas; these were unspecified in all the articles.

5.3.1. Quantitative evaluation of DDPD

The validation of the personas varied according to the applied methods. KMC was validated by calculating the Euclidean distance between the different variables (Tanenbaum et al., 2018; Wang et al., 2018) or by conducting Chi-squared tests (Tanenbaum et al., 2018). A few articles (Vosbergen et al., 2015; Zhang et al., 2016; Zhu et al., 2019) qualitatively validated clusters by engaging subject experts as well as users themselves in reviewing the clustering results.

- For **HC**, Miaskiewicz et al. (2008) and Mesgari et al. (2015) validated their results by considering relations between variables within clusters. The formerly calculated cosine similarity of angles between pairs of non-zero vectors; the latter, on the other hand, calculated the Pearson correlation (the extent of a linear relationship between two variables). Holden et al. (2017) determined the statistical significance between different variables as well as tested for a variance with the Kruskal-Wallis test and Welch's ANOVA, respectively.
- All articles that applied **PCA** (N = 5, 6.5%) complemented it with at least one other quantitative method. As a result, validation metrics also varied; they included Cohen's kappa (Brickey et al., 2012; Brickey et al., 2010), Euclidean distances of variables (Wang et al., 2018), Spearman's correlation between two ranked variables (Dang-Pham et al., 2015), and even qualitative review with survey participants (Tu, Dong et al., 2010).
- Similar to PCA, **LSA** was also often combined with other methods, especially hierarchical cluster analysis (Brickey et al., 2012; Brickey et al., 2010; Miaskiewicz et al., 2008). Researchers validated their results through cosine similarity tests.
- For **NMF**, An, Kwak, Salminen et al. (2018) calculated cosine similarity for pairs of personas until the closest pairs were determined. In another study employing NMF (An, Kwak, Jung et al., 2018), researchers used the Kendall rank correlation coefficient to compare the ranking of personas' demographic groups with the ranking of demographic groups in the raw data.

5.3.2. Qualitative and mixed evaluation of DDPD

Qualitative validation was common. In total, 24 articles (31.2%) incorporated qualitative feedback into their persona validation stages. These generally involved gathering a small sample of members from the initially surveyed population to evaluate the personas in open discussion groups. An exception is Dupree et al. (2016), who recruited a mutually exclusive, yet still relevant, subpopulation to evaluate the personas' representativeness. In that study, the validation stage group was tasked with self-identifying with one of the five final personas and rating how realistic they are.

Out of all 24 articles that used mixed quantitative-qualitative methods, 8 (33.3%) incorporated qualitative methods in the validation stage only, while 14 (58.3%) incorporated qualitative methods to both the initial data collection and the validation stages. Figure 3 shows that mixed methodologies in proportion to the total number of articles published per year have consistently been incorporated, with peaks in 2010 and 2015. These peaks may be attributed to rises in the popularity of incorporating qualitative aspects in validation, such as subject experts or user consultations.

6. Periods of DDPD research

6.1. Conceptualization into periods

Based on the results, we synthesize DDPD research into three periods:

(1) **Quantification** (2005–2008), which consists of the first experiments with quantitative methods for DDPD,
(2) **Diversification** (2009–2014), which was a transition period to more pluralistic use of data and algorithms, and
(3) **Digitalization** (2015–present), which is highlighted by a revitalized interest in DDPD research following the abundance of social media and Web analytics data, as well as the rapid development of data science algorithms and frameworks. For this research, "present" is the end of 2020.

6.2. First period: Quantification

The first period is marked by a focus on establishing the basic need for quantitative methodologies in the persona domain (Mcginn & Kotamraju, 2008) and early experimentation with different methods, especially those well-known in quantitative research tradition (e.g., clustering, principal component analysis, factor analysis). There is less interest in combining in-depth qualitative insights with quantitative results. The contextual focus is on software development – in particular, requirements engineering (Aoyama, 2005, 2007), meaning that personas are seen mainly as support for software developers. The primary data source for persona development is survey data, though some experimentation with clickstream data (Zhang et al., 2016) and statistics from gaming also took place (Tychsen & Canossa, 2008). These studies illustrate the potential of personas beyond their initial conception for software developers (Cooper, 2004).

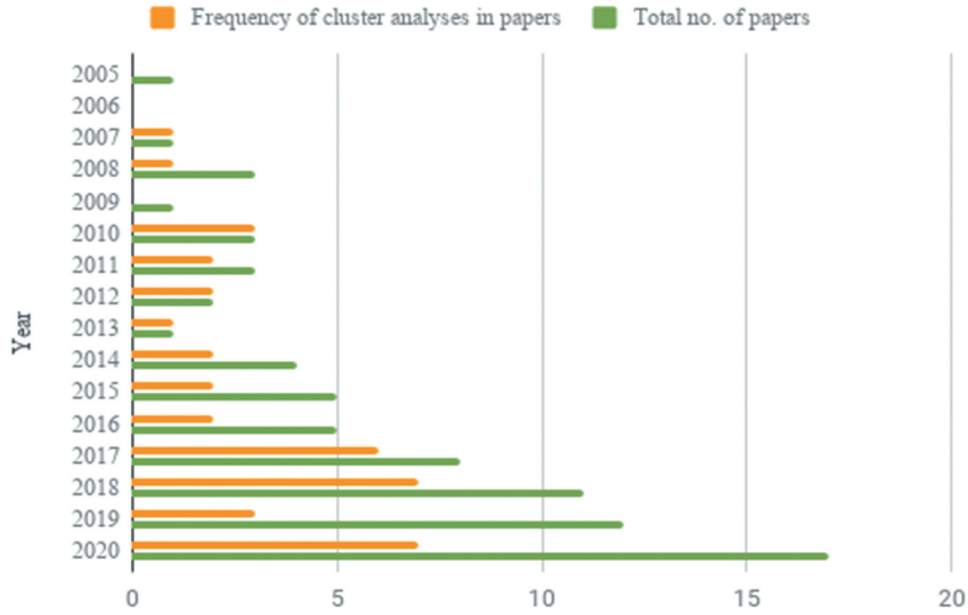

Figure 3. Articles combining mixed methods.

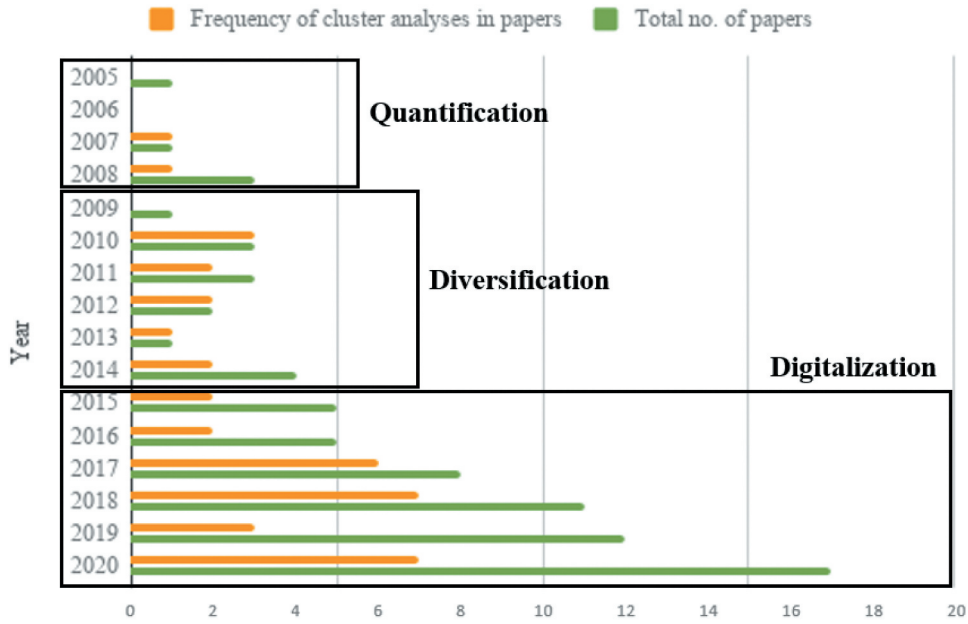

Figure 4. Articles using clustering methods over time. Clustering is consistently popular, but the use of new methods begins to rise in 2014 and throughout the third period (2015 to present).

6.3. Second period: Diversification

In the second stage, persona application contexts expand to new areas (e.g., knowledge management (Brickey et al., 2010), emergency preparedness (Kanno et al., 2011)). The statistical methods remain relatively similar, but more attention is paid to infusing qualitative insights with quantitatively created personas (Tu, He et al., 2010). This orientation for synthesis results in hybrid personas, already previously conceptualized by Pruitt & Grudin (2003). Clustering is the dominant method (see Figure 4), though experimentation with NLP techniques also takes place during this period (Bamman et al., 2013). Researchers introduce behavioral data alongside self-reported data (Dos Santos et al., 2014), representing a milestone in personas quantitatively describing user behaviors. As in the first period, the publication focus is on conference venues, and the number of research outputs is on the modest side. The first and second periods are characterized by the lack of journal publications, whereas the third period shows several journal articles.

6.4. Third period: Digitalization

In the third period, researchers discover social media and online analytics data for persona development (An, Kwak, Jung et al., 2018; Jansen et al., 2019; Salminen, Jung et al., 2019b). Also, "data science algorithms," such as matrix factorization (An et al., 2017), are applied for persona development through frameworks, such as Python's scikit-learn.[2] System-building and the goal of entirely automatic persona development (ranging from data collection to analysis to interactive persona UI) become explicit goals (Salminen, Jung, Jansen et al., 2019a). Researchers expand the notion of behavior, not only using behavioral data for persona development (An, Kwak, Jung et al., 2018) but also applying behavioral theories for interpreting what quantitative personas tell about the world (Jansen et al., 2017). Deep learning neural networks are applied to make personas interactive. Personas using conversational user interfaces are examples of interactive personas (i.e., personas the end-users can interact with) (Chu et al., 2018; Laporte et al., 2012). To date, these approaches are experimental and have not yet resulted in maturity for industry application. Researchers also began to pay attention to the longitudinal aspects of personas evolving and how to address this evolution using computational techniques (Holmgard et al., 2014). Several studies focus on the health context, and new domains are introduced (Holden et al., 2017; Vosbergen et al., 2015).

Dataset sizes (see Table 3) are also increasing across each stage, observed by increases in means, medians, and standard deviations. Though some researchers are still using small datasets in the third era, others are also using bigger datasets with up to 170 K sample sizes. However, these bigger dataset sizes do not necessarily result in more rounded personas, as researchers generated rich, narrative-like personas as early as 2008 (Mcginn & Kotamraju, 2008; Miaskiewicz et al., 2008).

The self-awareness that began in the second period is becoming more commonplace, with researchers acknowledging the challenges of DDPD at a broader spectrum (Mijač et al., 2018; Salminen, Jung, Jansen et al., 2019a). Thus, the third period is characterized both by trust in the potential as well as an urgency to solve the outstanding problems. The accumulated experience of using the methods has painted a more comprehensive picture of the field. Overall, the field of DDPD reached a degree of self-awareness, with literature reviews focused on different clustering methods emerging (Brickey et al., 2012). The introduction of behavioral data took place (Masiero et al., 2011). DDPD also gradually became used for analyzing diverse subpopulations, such as Vietnamese youth (Dang-Pham et al., 2015) and senior European citizens (Wöckl et al., 2012). In such research, the development of personas is applied as a means to an end, rather than being the focus itself. Finally, persona layouts become more sophisticated to include interactive elements provided through Web systems (see Figure 5(a–c)).

Table 3. Survey sample sizes of DDPD. Percentages indicate an increase from the previous era.

	Quantification (2005–2008)	Diversification (2009–2014)	Digitalization (2015–present)
Mean	343	2,034 (493.0%)	12,339 (506.6%)
Max	1,300	12,496 (861.2%)	170,704 (1266.1%)
Median	31	100 (222.6%)	199 (99.0%)
SD	638	4,003 (527.4%)	36,371.1 (808.6%)

6.5. Comparison of the periods

6.5.1. Shifts in the use of data

Even though survey datasets (see blue color in Figure 6) have been the consistently popular format of data, the focus shifted from surveys to web data (gray line in Figure 6) in the third period. Web and social media data sources have risen since 2015, and 2018 marks the first year that web data exceeded survey data. This trend continued in 2020. Also noteworthy is the increase in the data being collected from system logs and interfaces, enabling the creation of personas that represent various user behaviors (Mijač et al., 2018; Wang et al., 2018; Zhang et al., 2016).

6.5.2. Methodological diversity across the periods

More than a third (N = 31, 40.0%) of the reviewed studies combined quantitative and qualitative approaches. Also, 40 (52.0%) employed several quantitative methods. While no specific combination of quantitative methods dominated, combinations often included at least one type of clustering analysis (e.g., k-means, hierarchical). In addition, mixed quantitative-qualitative methods included incorporating qualitative components to the data collection (Hill et al., 2017; Matthews et al., 2012; Tempelman-Kluit & Pearce, 2014), as well as validation stages (Aoyama, 2005, 2007; Dupree et al., 2016; Miaskiewicz & Luxmoore, 2017). In terms of individual algorithms, clustering (particularly, k-means and hierarchical) remains popular throughout the periods. However, it is no longer dominant in the third period, as there is a trend in combining multiple quantitative methods simultaneously (see Figures 7 and 8). The year 2018 saw the least proportion of articles conducting cluster analyses since 2015. This can be attributed to researchers applying new models, including the Dirichlet persona model (Bamman et al., 2013) and non-negative matrix factorization (An, Kwak, Jung et al., 2018; An, Kwak, Salminen et al., 2018).

The methodological diversity taking place between the periods can also be shown quantitatively, using the Gini index (G). This measure reveals the deviation in the use of methods relative to the equal (i.e., evenly occurring) use of methods. The closer G is to zero, the more evenly each method is used. Thus, using the *formula* shown in Equation 1, *we* compute the G for each period P,

$$G_P = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}} \quad (1)$$

where G for a period P is calculated by dividing the sum of absolute differences of each pair of years (i, j) in the period P. N is the number of years. In contrast, x_i denotes the sum of different methods applied in the year i. X-bar is the average number of methods applied in the period. For each period, i is restricted to the number of methods that were deployed in that year (i.e., those conforming to $x_i \geq 1$). This makes the comparison fairer, as some methods in later years might not have been available earlier. The results (see Figure 9) indicate

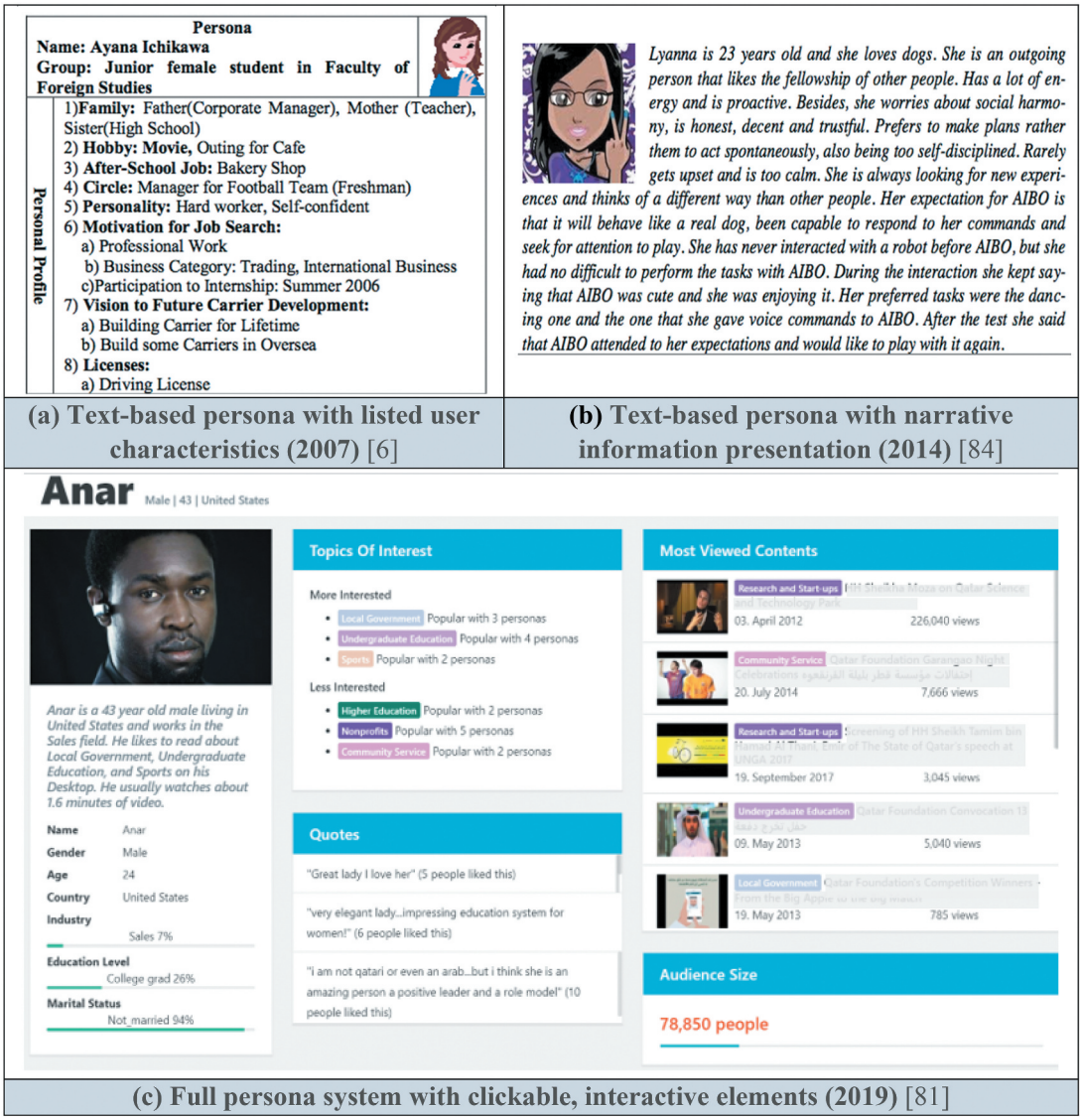

Figure 5. Persona layouts from early text-based approaches to modern persona systems.

a linearly decreasing *G* score that is to be interpreted here as an increase in methods diversity (again, the closer the *G* score is to zero, the more evenly each method is being used).

The diversity in the methods applied by authors can be seen in many articles individually developing and introducing new models, such as the neural speaker model developed by Li et al. (2016), the Dirichlet persona model by Bamman et al. (2013), the Hanako method by Aoyama (2005, 2007), the ego-splitting algorithm by Epasto et al. (Epasto et al., 2017), and more. Several articles also developed their clustering methods based on specific variables selected for their studies (Aoyama, 2007; Bamman et al., 2013; Tu, Dong et al., 2010).

7. Challenges of DDPD

Central themes that arose in discussions of challenges of DDPD involved concerns with (a) data quality, (b) data availability, (c) method-specific weaknesses, and (d) human and machine biases, such as the persistent need for judgment calls ("manual labor") that creates a potential source of bias and obstacles for completely automated DDPD.

7.1. Data concerns

Numerical data is often cited as the advantage of DDPD relative to qualitative-created personas. Nonetheless, many articles identify data-related challenges. Mijač et al. (2018) cite time or cost factors in particular as an obstacle to the amount of data that was able to be collected and analyzed due to the high cost of participant recruitment. While behavioral data is considered an essential advantage of DDPD (An, Kwak, Jung et al., 2018; Dos Santos et al., 2014), the most popular data source for DDPD is self-reported survey data. Survey data has at least two issues. First, Tu, Dong et al. (2010) highlight the potential issues with objectivity when authors select which questions to ask (and therefore, which answers to consider) from surveys. Second, Ford et al. (2017) also highlight the subjectivity of survey participants' answers (self-reporting), as some participants may exaggerate in their answers

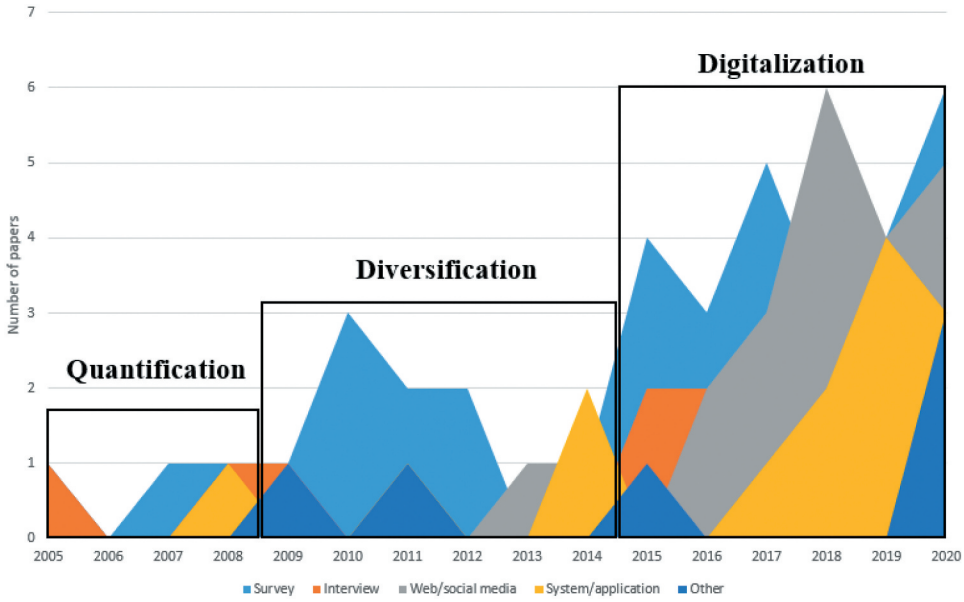

Figure 6. Popularity of data sources for DDPD over time. Surveys (in blue color), Web/social media (green color), and systems/applications (purple color) are the most dominant data sources.

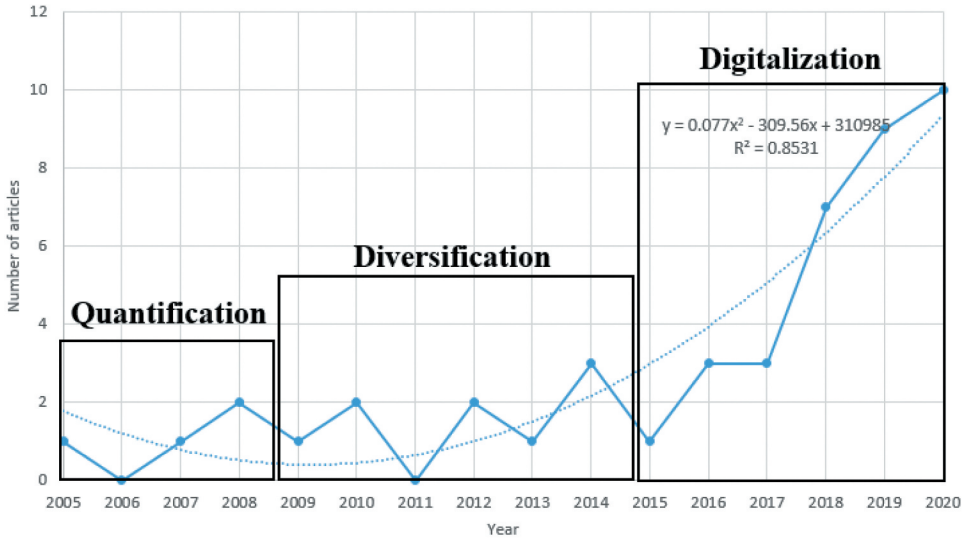

Figure 7. Articles using mixed quantitative methods. The line represents a polynomial growth trend.

depending on the context, such as rating their productivity levels. These limitations obstruct the representativeness and validity of personas. Interestingly, while MPD is often criticized for the lack of quantitative verification (Chapman & Milham, 2006), for DDPD, the lack of qualitative insights was also posed as a similar issue for purely quantitative studies.

Data concerns were related not only to the quality of data but also to the lack of in-depth insights regarding the users. The implication is the "breadth-depth trade-off" of using quantitative versus qualitative data with resulting personas remaining shallow and unable to "*provide the deep narrative understanding that designers often seek*" (Holden et al., 2017, p. 1073). Moreover, data is not always available in the type the researchers want; online platforms impose restrictions on what user data is shared (e.g., by not giving out demographic variables) and how much (e.g., by applying thresholds or sampling). The implication is the restricted availability of persona information. As reported by Wöckl et al. (2012, p. 27), "*Due to numerical constraints, the number of variables used for creating clusters is limited and additional associated variables are needed to allow a more detailed precision.*"

Limited datasets translate to concerns with the applicability and transferability of results to other contexts (Brooks & Greer, 2014; Ford et al., 2017; Kim & Wiggins, 2016; Rahimi & Cleland-Huang, 2014; Watanabe et al., 2017) as the datasets represent only one context (e.g., one educational institution (Kim & Wiggins, 2016)) or users of one digital platform (e.g., YouTube (An, Kwak, Jung et al., 2018)). The implication is

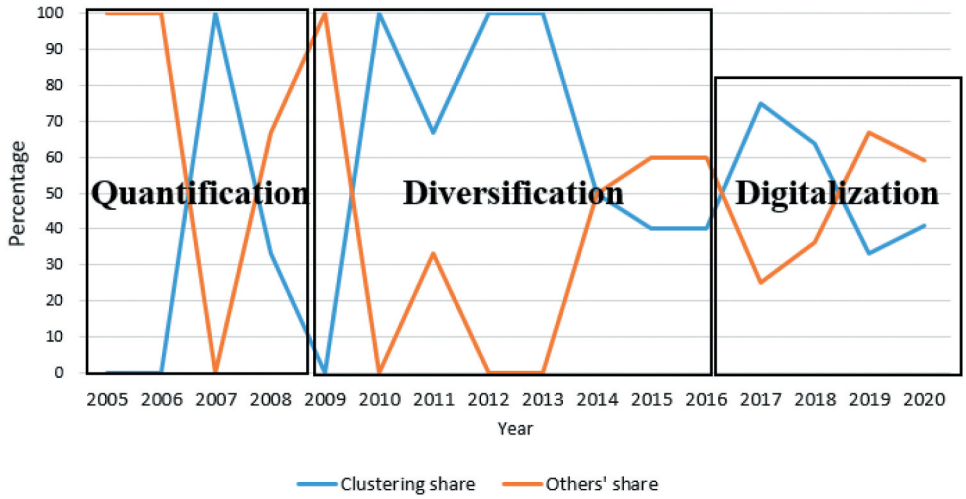

Figure 8. Clustering vs. other methods. The figure shows the percentage of articles using clustering vs. other methods per year. These methods roughly align with the three periods, with clustering being the most predominant during the second period (2009–2014).

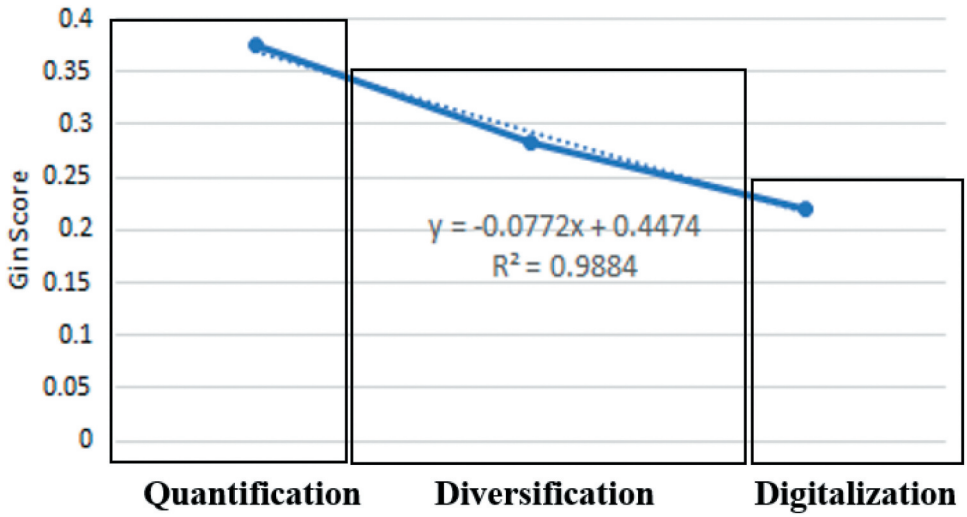

Figure 9. Gini scores in different periods show a decreasing trend (denoted by the dotted line), indicating more even use of methods. Coupled with the notion that the number of methods increases over time, this lends support to the argument that methodologically, DDPD is diversifying over time.

the hindrance for authors to establish the generalizability of their personas to other settings or purposes. For example, An, Kwak, Salminen et al. (2018) note that the lack of demographic attributes in Twitter at the tweet-level makes their DDPD approach incompatible with Twitter. Merging data from multiple sources is also mentioned as a challenge as data types and structures may vary among different online platforms (Mijač et al., 2018). For this reason, the literature lacks "cross-platform" personas.

7.2. Method-specific weaknesses

The literature suggests that each DDPD method has its strengths and weaknesses as several researchers cite method-specific weaknesses. For example, Kwak et al. (2017) noted that a limitation of the k-means clustering is that a single demographic group must fall into one persona; however, in reality, many behavioral segments can be found from one demographic group, as people in the same demographic group can and often do behave differently. Another mentioned weakness of clustering is the "need for specialists to use expert judgment during clustering [to define hyperparameters]" (Minichiello et al., 2018, p. 19). NMF also similarly requires manual parameter setting (An, Kwak, Jung et al., 2018), wherein the parameters are often set using rules of thumb (An, Kwak, Salminen et al., 2018). For LSA, the weakness is in the dependency on text corpora, which is typically missing from online analytics data. As such, the inability to incorporate behavioral data (e.g., user engagement metrics) is a weakness of LSA (Salminen, Jung, Jansen et al., 2019a).

Finally, it can be argued that none of the reviewed methods can independently instill in-depth interpretations of the data into a "rich" persona narrative. The interpretative step highlights the major challenge of "going from data to narrative," an essentially and ultimately creative process that seems to require human interpretation and judgment. A related aspect is that quantitative personas do not automatically project

themselves on the paper (or digital format), but the process of transferring the results into persona profiles requires several manual steps. As expressed by Wöckl et al. (p. 3), "a main challenge when creating personas from quantitative data is the translation of numerical output into text" (p. 3). The citation emphasizes the complicated relationship between automation and manual work in DDPD.

7.3. Human and machine biases

It is essential to acknowledge that the quantity of data does not automatically result in a higher quality of personas. Instead, any biases and errors in the data are passed on to personas. For example, when generating personas from online analytics data, the measurement error is unknown, as the platforms do not share their methods for inferring user attributes or the errors of these methods. The existence of unknown measurement errors means that the data sources should not be blindly trusted. In a similar vein, the use of quantitative data science algorithms, especially when coupled with imbalanced user data, can result in aggravated stereotypes, thus making the output personas unreliable or even harmful for practical decision making.

Our evaluation shows that the challenges of MPD may not disappear when applying DDPD. Manual methods can be used (or are even necessary) for addressing the DDPD challenges. For example, a lack of depth and representativeness can be addressed using qualitative methods to collect and analyze data (Holden et al., 2017). Manual steps typically involved with DDPD include, at least, the following:

- **Hyperparameters**: setting the values for hyperparameters (i.e., manually adjustable parameters) for algorithms (e.g., choosing the "right" number of personas for clustering)
- **Write-up**: writing up the narrative persona descriptions shown to personas end-users (Wöckl et al., 2012)
- **Evaluation**: evaluating if stakeholders adopt the personas for decision-making and that the personas "work" in the sense of being perceived as empathetic, realistic, and useful for user-centric tasks.

Therefore, persona development, even within the application of "data and algorithms," involves some degree of creative effort. Data-driven personas do not automatically "project on paper" (or another form of medium) but require some manual process of refining the user data into rich and meaningful persona profiles that serve end-users' information needs. Thus, the DDPD process involves manual steps such as determining the right number of clusters or underlying patterns and writing the persona description. While there can be automatic techniques for these, their use is not typical or clearly stated in the literature. Again, this supports the notion that DDPD has not reached maturity yet. Indeed, one of the primary consensus points among researchers is the need to combine manual and automatic methods for persona development. A common approach is to use quantitative data to explore user behavior and enhance these behavioral archetypes ("skeletons," "templates," "prototypes") with qualitative insights to create more holistic personas (Minichiello et al., 2018; Salminen, Şengün et al., 2018, p. 77). Quantitative data is thus used for corroborating qualitative personas, while qualitative data enriches them.

7.4. Summary of challenges

Interestingly, the cited concerns of DDPD often overlap with the claimed strengths of DDPD or even mirror those from qualitative research. This observation implies that researchers may not always achieve the idealized benefits of DDPD. Another possibility is that the challenges of DDPD coincide with those of MPD, but they manifest in specific ways. For example, some criticize data collection for MPD as expensive (An, Kwak, Jung et al., 2018), but also survey data collection (the most popular source of data for DDPD) requires recruiting a robust sample of participants and thus carries a high cost.

Many articles start with the premise: *DDPD is great because we do not need manual steps and subjectivity*, and end with the notion that *DDPD could be better with better manual steps and subjectivity*. Not only this, but manual steps are part of the reviewed DDPD methods in terms of setting hyperparameters for the statistical algorithms and finalizing the persona description by means, such as choosing a picture and writing a textual description. These manual steps are necessary to fix the shallowness of the DDPD outputs (Wöckl et al., 2012). Nevertheless, in return, the manual choices along the way, from data to finalized personas, involve many potential entrapments for biased decision making. For example, the choice of picture for the persona is not arbitrary at all, as the picture severely impacts the end-user stereotypes of the persona, along with variables such as race, gender, and age (Hill et al., 2017; Salminen, Nielsen et al., 2018).

Thus, there is a challenge in automation: *How does one make quantitative personas more deep, compelling, insightful?* Attempting this goal results in increased complexity in methods, as more and more computational techniques are needed to discern specific nuances of online audiences. For example, there may be a need for one algorithm to detect demographic attributes, another one for behaviors, and the third one for persona pain points. As each novel technique adds to the cumulative "measurement error" of personas, highly complex DDPD processes are potentially vulnerable to cascading failures. In other words, if one information type in the personas is predicted erratically, this error is reflected in other persona information as well, as the information pieces are interlinked in the underlying database.

8. Research trends

8.1. Human-persona interaction

Interactions between users and personas is a trend reflected by the development of comprehensive systems that create personas in real time and even allow users to control the selected data and persona attributes (An, Kwak, Salminen et al., 2018; Mijač et al., 2018; Salminen, Jansen et al., 2018). Systems can enable end users to create personas and explore in-depth information regarding them. Interaction can be achieved by uploading user data and choosing which persona attributes the output should contain (Salminen, Jansen et al.,

2019), even though these opportunities have not yet materialized into working systems. Nonetheless, the extant research suggests the field is not far from end-user organizations being able to create their own personas on demand by using their datasets. Another notable avenue is the rise of "persona chatbots" in the field of NLP; these bots have distinct conversational styles that reflect different personality types.

The personas for chatbots or dialogue systems react to user inputs interactively by imitating realistic conversations with people (Amer Jid Almahri et al., 2019; Hwang et al., 2019). One more evolving avenue is the use of so-called procedural personas that enable game developers to test how personas (i.e., archetypical player types) react to changes in the game world. These procedural personas simulate real-time decision making under various environmental stimuli, especially in videogame context (Holmgård et al., 2018), and help understand how different player types react to in-game events. Some studies also demonstrate the ability to predict the content preferences of personas (An, Kwak, Salminen et al., 2018) and lifestyle articles (Dhakad et al., 2017). These studies suggest that personas could be coupled with recommendation systems and thus represent an exciting future trend.

Overall, the development of interactive personas (i.e., *Human Personal Interaction* or HPI) is an interesting research direction, but studies are at a very early stage. The essential question is about the benefits: what benefits do persona users gain from "talking to the persona?" This issue could be addressed with user studies to devise requirements for persona developers, not only regarding *what* information should be available for interaction but also *how* to interact with the persona. Researchers have proposed conversational UIs, as well as an interactive persona layout, but thus far, studies testing these interaction techniques remain scarce. Also, there is a lack of tying these personas to real systems and "proper" persona profiles with narrative descriptions and information, such as demographics, goals, attitudes, and so on.

8.2. Fully automated persona systems

Concerning the use of automated data collection, only seven articles (9.1%) used application programming interfaces (APIs) to collect data for persona development. Nonetheless, API usage is an increasing trend, as three of the seven articles using APIs are from 2018. The sources included WeChat user data (Wang et al., 2018), YouTube Analytics (An, Kwak, Jung et al., 2018; An, Kwak, Salminen et al., 2018; Salminen, Şengün et al., 2018, p. 772), Google Analytics (Mijač et al., 2018), Twitter FireHose (Li et al., 2016), and Wikipedia (Bamman et al., 2013). The most common social media platform was YouTube (N = 5). The advantages of automatic data collection via APIs include speed, volume, and cost-effectiveness (De Souza et al., 2004). Using preexisting data is highly lucrative for persona developers due to time and cost benefits (Zhu et al., 2019).

Moreover, data structures of online platforms regarding user attributes are similar, which facilitates the application of replicable methods on different platforms (An, Kwak, Salminen et al., 2018). Because these datasets are typically aggregated (as opposed to individualized), they preserve the privacy of individual users (Wöckl et al., 2012). We expect API-based data collection for personas to become more commonplace in the future. Further, there have been attempts to automate persona generation (An, Kwak, Jung et al., 2018; An, Kwak, Salminen et al., 2018). Several authors (Epasto et al., 2017; Ishii et al., 2018; Miaskiewicz & Luxmoore, 2017) expressed plans to refine further and automate their methods. However, these attempts are still ongoing. As remarked by Mijač et al. (2018, p. 1431): "*Examples of an automatic update of personas are scarce, and even those are not fully implemented but are rather on the level of proof-of-concept.*" Salminen, Jung, Jansen et al. (2019a) provide a research roadmap for completely automatic persona generation.

9. Central research gaps

9.1. Lack of resource sharing

Most authors of DDPD studies do not share their resources, including datasets, code, and algorithms applied. As these are not made publicly available for other researchers, replicating DDPD studies is challenging. From a scientific perspective, this hinders the incremental, evolutionary progress of the field as a whole. Moreover, there is a sense of fragmentation: researchers are developing the methods independently and often repeating the same methods without presenting a case for why and how a particular method is better than another already published. Comparative studies are not conducted, with no clear statements of progress or collaboration to do so. Similarly, authors frequently express a desire to generate personas for different domains in their future work sections, but cross-domain applications are rarely followed through.

9.2. Insufficient evaluation of DDPD methods

The emphasis of the DDPD articles is on reporting the *development* of personas. In turn, researchers evaluate the personas most often using technical metrics that measure how the personas satisfy statistical requirements. While it seems that some external feedback is frequently collected, these attempts tend to be informal and not rigorously described. There is little-to-no information on how persona user feedback resulted in modifications of the personas or how the personas *were used* for real decision making in user-centric tasks, with even limited experimentation in this area (Salminen, Jung et al., 2020). Proper user studies (i.e., external validation with real users) are needed to address the applicability of personas in conjunction with their actual impact on the employing organization.

Practical evaluation is also crucial because the technical sophistication of the methods vary greatly from simple counts to complex combinations of multiple computational models, as well as for establishing applicability, which is one of the reemerging themes in DDPD research. Articles throughout the periods of Quantification (Thoma & Williams, 2009), Diversification (Chapman et al., 2015), and Digitalization (Miaskiewicz & Luxmoore, 2017) dealt with the aspect of generating real value for organizations and individuals with DDPD, as well as struggles with organizational adoption. For this, Thoma and Williams (2009) and Holden et al. (2017)

discussed the need for incorporating more qualitative methods, particularly to validation stages, to ensure representativeness in the personas.

Finally, some authors state plans to test their methodologies on other comparable population groups, such as different countries or universities (Dos Santos et al., 2014; Kim & Wiggins, 2016; Wöckl et al., 2012), while others wish to broaden their existing data samples (Dos Santos et al., 2014; Holden et al., 2017; Tu, Dong et al., 2010) or even explore current methodologies in entirely different industries (Aoyama, 2005, 2007; Chu et al., 2018); yet, such comparative studies do not currently exist.

9.3. Lack of standardization

Due to the divergence of the methods, there are no unified or standard metrics for evaluating the quality of quantitative personas. However, there are some preliminary attempts to create a standardized questionnaire for measuring persona users' perceptions of the personas (Salminen, Kwak et al., 2018). This gap is significant because, in the absence of quality standards, researchers face the challenge of defining the boundaries of quantitative personas. Since DDPD uses statistical methods to create personas, persona creators can verify the methods using quantitative metrics. This potential for standardization, as far as we can see, is an enormous advantage to DDPD in general. However, the lack of a unified metric that would be applicable *across* the different methods erodes this advantage. Authors generalize traits due to the limited number of final personas that they see fit to create. While they can certainly create more personas to capture subtler and esoteric characteristics, this would result in personas that may be too complex to apply in familiar contexts. Authors must thus consider the opportunity cost of including and excluding fringe personas, depending on their goals.

Brickey et al. (2012), Bamman et al. (2013), and Holden et al. (2017) similarly highlighted their limitations in contextualizing personas when it came to unexpected outliers in the clusters, such as deciding which traits are applicable. To alleviate this challenge, Zhang et al. (2016), Tychsen and Canossa (2008), and Miaskiewicz et al. (2008) have suggested incorporating user evaluations of the personas into the validation stages to capture the most relevant yet comprehensive traits in the final personas.

9.4. Lack of consideration for inclusivity

Most of the articles focused on "core users," "representative segments," or other forms of majority users. Further, the articles were limited in resources and creating personas for particular purposes, so identifying outliers was not a priority or were actively removed from the data (Jansen et al., 2016). Statistical data science algorithms tend to represent means and averages, meaning that outliers are considered less critical. The lack of inclusivity in DDPD research is a direct contrast to the HCI research community's ongoing drive toward inclusivity (Goodman-Deane et al., 2018; Hill et al., 2017) through the examination of outliers, deviating behavior, and discriminated groups (Hill et al., 2017; Marsden & Haag, 2016). While many of the articles did pose inclusivity as something to work on in the future, these plans were framed in terms of improving statistical representativeness (i.e., what characteristics are being mistaken as "fringe" but are highly relevant to the key personas) rather than promoting inclusivity.

Only one research article explicitly mentions the concept of "algorithmic bias" in association with personas (Salminen, Jung et al., 2019b). Fixing this gap is compelling for not just the HCI community but any organizations interested in analyzing user behavior; interesting insights can often be found by inspecting outliers and minorities. Thus, new DDPD approaches such as outlier detection for persona development are focal points for future studies. Statistical methods may help "fix" the shortcomings of the methods that are reliant on "means" and "averages" instead of "deviations" and "outliers." The so-called fringe personas lead not only to a statistical question about outliers but also to an ethical question of fairness. More particularly, explicitly designing for the fringe communities (e.g., racial or sexual minorities) can be, in itself, the goal of a persona project, tilting the goal of "eliminating outliers" to "focusing on outliers" (although, as stated, this ultimately depends on the use case of personas).

In some cases, there can be severe limitations for applying DDPD methods because many publicly available datasets may not contain information on these sensitive or "protected" attributes. Harnessing such insights requires special attention to creative data collection approaches as well as collaboration with minority stakeholders to produce the necessary data.

9.5. Risk of losing immersion

Current research is unable to conclude whether persona creators lose something in the process of DDPD relative to MPD. The negligence of this question may be because several authors describe MPD as an iterative, analytical process that, in itself, provides user insights to participants (Cooper, 1999; Long, 2009; Nielsen, 2019). The oft-suggested remedy for this is the co-creation of personas by HCI professionals and users together. This collaborative effort has the potential to not only enhance mutual understanding about users but also drive the emergence of shared mental models among team members. With DDPD techniques typically being drastically different from the workshop-driven, collaborative persona development process, it is worthwhile to ask whether the positive aspects of the shared understanding are lost. The risk is particularly poignant because the researchers applying DDPD methods may overgeneralize traits due to the limited number of final personas to develop.

While the persona creators can undoubtedly increase the number of personas to capture subtler characteristics (and, therefore, more esoteric user groups), doing so can result in personas that may be too complex or even irrelevant to apply. Thus, authors must carefully consider the cost of excluding fringe personas.

10. Implications

The following sections summarize the main implications for stakeholders. We have separated these for researchers and

practitioners, with the former focusing on the development of research practices of DDPD and the latter on applicability.

10.1. Takeaways for researchers

We categorized the evolution of DDPD into three thematic stages: Quantification, Diversification, and Digitalization. In our assessment, to reach Stage 4 ("Maturity"), several action points (APs) are needed from the research community:

- **AP1**: Conduct replication studies that apply the same methods to different datasets or different methods to the same dataset.
- **AP2**: Conduct comparative studies to investigate different methods by their technical merits, as well as the overlaps/deviations of the resulting personas.
- **AP3**: Conduct formal evaluation studies to evaluate both accuracy (internal validation) and impact (external validation) of personas.
- **AP4**: Share resources such as datasets, code, and algorithms to enable others to replicate results.

Furthermore, many of the gaps in the current body of research (shared resources, metrics, standardization) can be filled by building a more robust research community around DDPD. This community-building could take place through workshop organization, networks/meetups, or even a special interest group for data-driven personas. On another note, based on its popularity, k-means clustering could be a baseline technical method for DDPD. Replication of this method and others are crucial steps to address the lack of objectivity for which persona research has been criticized (Chapman & Milham, 2006).

In terms of improving validation, several authors have suggested ways of going beyond mere persona development and testing the aptitude of the developed personas in meeting stakeholder goals (Goodman-Deane et al., 2018; Miaskiewicz & Luxmoore, 2017; Rahimi & Cleland-Huang, 2014; Tanenbaum et al., 2018; Watanabe et al., 2017). Some emerging studies have shown promise in this regard, such as the use of longitudinal data and a standardized algorithmic approach to compare persona sets over time (Jung et al., 2019) and between different organizational units (Zaugg & Ziegenfuss, 2018). In the healthcare sector, researchers design tailored medical interventions to subpopulations represented by the personas and investigate how patient adherence and health outcomes are subsequently affected (Tanenbaum et al., 2018; Vosbergen et al., 2015). Reference studies from the interpretative research tradition, particularly those providing in-depth insights from persona users, include Friess (2012), Matthews et al. (2012), and Rönkkö (2005). Articles summarizing DDPD methods have been published for factor analysis (Kwak et al., 2018), cluster analysis (Brickey et al., 2012), and non-negative matrix factorization (An, Kwak, Salminen et al., 2018). Articles discussing the role of personas amidst the transformative impact of Web and social media analytics include (Jansen, Salminen et al., 2020; Mijač et al., 2018; Salminen, Jansen et al., 2018, 2019).

Overall, DDPD methods represent the best efforts to make use of techniques and processes that are available at a given time. To increase trust in DDPD, authors can (a) apply triangulation by independent samples to corroborate personas and (b) increase algorithmic transparency, including explicit statements of where the data originates, how it was collected, and what were the analysis steps that resulted in the visible persona profiles. Most likely, because it is costly and time-consuming, few DDPD studies follow this best practice.

10.2. Takeaways for practitioners

We recommend the following Persona Guidelines (PGs) to practitioners wishing to apply quantitative methods to create personas:

- **PG1: Clustering is a safe choice**. Clustering techniques are the most common choice among quantitative methods. These include KMC, HC, and others that are well-established, and persona creators can combine them with other methods such as EFA or PCA in the data-incorporation stage or also qualitative methods in the narrative-building stage. Nonetheless, as discussed earlier, clustering does include some fundamental limitations. Other methods, such as NMF, can be applied to address these concerns partially, but each method involves some degree of subjectivity.
- **PG2: Question the numbers**. Practitioners should not blindly rely on the outputs of statistical methods. Additional steps to ensure data quality, such as triangulating the results with other methods like qualitative interviews, are vital. Practitioners with limited knowledge about quantitative methods should "ask stupid questions" to avoid the "mystique of numbers" (Siegel, 2010). Questions include how personas were created, what manual choices the process involved, and how results were validated.
- **PG3: Be conscious of algorithmic bias**. Surveys are the most popular data sources for DDPD. Nonetheless, even when analyzed quantitatively, survey results include several threats to validity, such as social desirability bias. "Data-driven" does not necessarily mean "objective" or "honest." Researchers are becoming increasingly aware of algorithmic biases and the ways in which an algorithmic method may introduce undesired generalizations into the personas (Salminen, Jung et al., 2019b).
- **PG4: Analyze minority subsets**. Relying only on quantitative data and the significant patterns they generate can lead to the exclusion of minority groups, posing challenges for achieving inclusivity (Marsden & Haag, 2016). Consider splitting datasets into "majority personas" and "minority personas" and developing separate personas for both groups.
- **PG5: Iteratively work to increase usefulness**. Both qualitative and quantitative personas should be evaluated for truthfulness vis-a-vis the real user base and usefulness (i.e., whether they serve decision-makers' goals). Persona validation methods mentioned by Minichiello et al. (2018) include on-site visits,

dissemination, feedback from persona users, interviews, surveys, anti-persona comparisons, log file verifications, and persona user and usage observations.

Is DDPD suitable for a given organization? DDPD is not a perfect method for persona development, though it does have its time and place. Therefore, practitioners in organizations should reflect on the following guiding questions (GQs) before initiating DDPD projects:

- **GQ1**: *Does your organization have an extensive offering of products/services/content in online environments?* The scope is important because of the scalability of algorithms for persona creation in multiple domains (e.g., e-commerce, social media, news). With only a few products, there is typically not enough dimensionality for algorithms to separate the data.
- **GQ2**: *Does your organization have a large and diverse user/customer base?* Variety is important for practical reasons: if an organization has a very narrowly targeted customer base in one specific market, understanding this customer base can more easily be achieved using qualitative methods, such as interviews, rather than trying to model the customer base using DDPD.
- **GQ3**: *Is your organization actively collecting digital data on your users/customers?* Active data collection (e.g., CRM system, Web log files, electronic health records, etc.) is important because "data is the fuel of personas." Not all organizations, however, have sufficient data for DDPD, or the data is structured or formatted incorrectly for algorithms to process it.
- **GQ4**: *Is it possible to quantify the user attributes your organization is interested in?* The information needs (e.g., engagement with online content) of the team are important because quantitative information is not always what decision-makers look for in personas. Insights such as pain points, needs, and wants can be hard to quantify and can often be distilled better using a qualitative persona approach.

If an organization's answers to most of the questions above are yes, then DDPD has great potential in enabling insights about its customers. In other cases, MPD may be more feasible, for example, to produce quick "prototype personas" (Gothelf, 2012), as long as the risks of doing so are clear. Nonetheless, we stress the importance of evaluating the applicability of DDPD carefully before committing resources to it. Based on our experience in the field, one can argue that personas can provide value for 90% of all organizations; however, DDPD has a considerably narrower margin than this, possibly only 20% of all organizations or less. The DDPD has a narrow nature because, for DDPD to work well relative to MPD, one needs to satisfy the data requirements of volume, variety, velocity, and veracity (i.e., the Big Data traits (Baig et al., 2019)) for DDPD to be useful.

The fact that the DDPD method has narrower applicability than MPD is not commonly understood. Instead, decision makers tend to assume, roughly speaking, that, as long as they have a social media account, DDPD can be useful. This fallacy is parallel to the phenomenon of "mystique of numbers" reported by Siegel (Siegel, 2010), and it is a misleading thought. Therefore, when it comes to takeaways for practitioners, avoiding conflated expectations of DDPD methods is our primary advice. One ought to understand the stringent requirements for not only data volumes but also how the data is structured and accessible to algorithms. Based on our interactions with practitioners, we maintain that organizations significantly differ by their ability to understand and leverage DDPD methods in a productive manner. We refer to this notion as "DDPD readiness." Organizations should assess their DDPD readiness before the initiation of DDPD projects.

11. Conclusion

Most data for DDPD originates from surveys, but the use of behavioral Web analytics data and textual social media data is gaining momentum. The results indicate that dataset sizes for data-driven personas have significantly increased over the years, while persona development methods have simultaneously evolved to become more diverse and complex. Clustering techniques are the most common algorithms. Researchers often use clustering in conjunction with PCA, EFA, or other data exploration techniques. It is common to combine several quantitative methods and enhance the results with qualitative material. In terms of progress, the literature shows a lack of cumulative milestones, shared resources (e.g., code, algorithms, data), and replicative and longitudinal studies. The lack of quality standards hinders the comparison of algorithms and the establishment of the superiority of one method over others. Ongoing research trends include interactivity between personas and their users and fully automated persona systems. Research priorities include addressing bias from both humans and algorithms, enhancing the transparency of DDPD algorithms, and conducting impact-driven evaluation studies with persona end users to develop systems that serve users' informational needs in professional application domains and use cases.

Notes

1. https://www.dropbox.com/s/7dvewzy4ry7npr6/ONLINE%20SUPPLEMENTARY%20MATERIAL_csur.docx?dl=0
2. https://scikit-learn.org/stable/

Acknowledgements

Open Access funding provided by the Qatar National Library.

ORCID

Joni Salminen http://orcid.org/0000-0003-3230-0561

References

Amer Jid Almahri, F., Bell, D., & Arzoky, M. (2019). Personas design for conversational systems in education.

An, J., Kwak, H., & Jansen, B. J. (2017). Personas for content creators via decomposed aggregate audience statistics. In *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 632–635). IEEE.

An, J., Kwak, H., Jung, S.-G., Salminen, J., & Jansen, B. J. (2018). Customer segmentation using online platforms: Isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining*, *8*(1), 54. https://doi.org/10.1007/s13278-018-0531-0

An, J., Kwak, H., Salminen, J., Jung, S.-G., & Jansen, B. J. (2018). Imaginary people representing real numbers: Generating personas from online social media data. *ACM Transactions on the Web (TWEB)*, *12*(4), 27. https://doi.org/10.1145/3265986

Aoyama, M. (2005). Persona-and-scenario based requirements engineering for software embedded in digital consumer products. In *Proceedings of the 13th IEEE International Conference on Requirements Engineering (RE'05)* (pp. 85–94). IEEE. https://doi.org/10.1109/RE.2005.50

Aoyama, M. (2007). Persona-scenario-goal methodology for user-centered requirements engineering. In *Proceedings of the 15th IEEE International Requirements Engineering Conference (RE 2007)* (pp. 185–194). IEEE. https://doi.org/10.1109/RE.2007.50

Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2019, November). Big data adoption: State of the art and research challenges. *Information Processing & Management*, *56*(6), 102095. https://doi.org/10.1016/j.ipm.2019.102095

Bamman, D., O'Connor, B., & Smith, N. A. (2013). Learning latent Personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (p. 10). IEEE.

Brickey, J., Walczak, S., & Burgess, T. (2012, May). Comparing semi-automated clustering methods for Persona development. *IEEE Transactions on Software Engineering*, *38*(3), 537–546. https://doi.org/10.1109/TSE.2011.60

Brickey, J., Walczak, S., & Burgess, T. (2010). A comparative analysis of Persona clustering methods. In *Americas Conference on Information Systems (AMCIS2010)* (pp. 217). ACM.

Brooks, C., & Greer, J. (2014). Explaining predictive models to learning specialists using personas. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14* (pp. 26–30). ACM Press.https://doi.org/10.1145/2567574.2567612

Chapman, C., Krontiris, K., & Webb, J. 2015. Profile CBC: Using conjoint analysis for consumer profiles. In *Sawtooth Software Conference Proceedings*. Google Research.https://research.google.com/pubs/archive/44167.pdf

Chapman, C. N., & Milham, R. P. (2006). The Personas' new clothes: Methodological and practical arguments against a popular method. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 634–636). SAGE Publications. https://doi.org/10.1177/154193120605000503

Choi, J., Yoon, J., Chung, J., Coh, B.-Y., & Lee, J.-M. (2020, November). Social media analytics and business intelligence research: A systematic review. *Information Processing & Management*, *57*(6), 102279. https://doi.org/10.1016/j.ipm.2020.102279

Chu, E., Vijayaraghavan, P., & Roy, D. (2018). Learning Personas from dialogue with attentive memory networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2638–2646). Association for Computational Linguistics. Retrieved June 20, 2019, from https://www.aclweb.org/anthology/D18-1284

Cooper, A. (1999). *The inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity* (1st ed.). Sams - Pearson Education.

Cooper, A. (2004). *The inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity* (2nd ed.). Pearson Higher Education.

Dang-Pham, D., Pittayachawan, S., & Nkhoma, M. (2015, November). Demystifying online personas of Vietnamese young adults on Facebook: A Q-methodology approach. *Australasian Journal of Information Systems*, *19*(1), https://doi.org/10.3127/ajis.v19i0.1204

De Souza, C. R., Redmiles, D., Cheng, L. T., Millen, D., & Patterson, J. (2004). How a good software practice thwarts collaboration: The multiple roles of APIs in software development. *ACM SIGSOFT Software Engineering Notes*, *29*(6), 221–230. https://doi.org/10.1145/1041685.1029925

Del Vecchio, P., Mele, G., Ndou, V., & Secundo, G. (2018, September). Creating value from social big data: Implications for smart tourism destinations. *Information Processing & Management*, *54*(5), 847–860. https://doi.org/10.1016/j.ipm.2017.10.006

Dhakad, L., Das, M., Bhattacharyya, C., Datta, S., Kale, M., & Mehta, V. (2017). SOPER: Discovering the influence of fashion and the many faces of user from session logs using stick breaking process. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17* (pp. 1609–1618). ACM Press. https://doi.org/10.1145/3132847.3133007

Dillahunt, T. R., Wang, X., Wheeler, E., Cheng, H. F., Hecht, B., & Zhu, H. (2017). The sharing economy in computing: A systematic literature review. *Proceedings of the ACM on Human-Computer Interaction 1* (pp. 38).CSCW.

Dos Santos, T. F., De Castro, D. G., Masiero, A. A., & Junior, P. T. (2014). Behavioral persona for human-robot interaction: A study based on pet robot. In *International Conference on Human-Computer Interaction* (pp. 687–696). Springer.

Drego, V. L., Dorsey, M., Burns, M., & Catino, S. (2010). *The ROI of Personas*. Forrester Research. https://www.forrester.com/report/The+ROI+Of+Personas/-/E-RES55359

Dupree, J. L., Devries, R., Berry, D. M., & Lank, E. (2016). Privacy Personas: Clustering users via attitudes and behaviors toward security practices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)* (pp. 5228–5239). ACM. https://doi.org/10.1145/2858036.2858214

Epasto, A., Lattanzi, S., & Leme, R. P. (2017). Ego-splitting framework: From non-overlapping to overlapping clusters. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*(pp. 145–154). ACM. https://doi.org/10.1145/3097983.3098054

Ford, D., Zimmermann, T., Bird, C., & Nagappan, N. (2017). Characterizing software engineering work with Personas based on knowledge worker actions. In *Proceedings of the 11th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '17)* (pp. 394–403). IEEE Press. https://doi.org/10.1109/ESEM.2017.54

Friess, E. (2012). Personas and decision making in the design process: An ethnographic case study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1209–1218). ACM.

Gaiser, B., Panke, S., & Arnold, P. (2006). Community design-the Personas approach. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 520–525). Association for the Advancement of Computing in Education (AACE).

Goodman-Deane, J., Waller, S., Demin, D., González-de-heredia, A., Bradley, M., Clarkson, J. P., & Clarkson, J. P. (2018). Evaluating inclusivity using quantitative Personas. In *In the Proceedings of Design Research Society Conference 2018*. Design Research Society. https://doi.org/10.21606/drs.2018.400

Gothelf, J. (2012). Using proto-personas for executive alignment. *UX Magazine*, Article No: 821.

Guo, A., & Jianhua, M. (2018, March). Archetype-based modeling of persona for comprehensive personality computing from personal big data. *Sensors*, *18*(3), 684. https://doi.org/10.3390/s18030684

Guo, H., & Razikin, K. B. (2015). Anthropological user research: A data-driven approach to personas development. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction (OzCHI '15)* (pp. 417–421). ACM. https://doi.org/10.1145/2838739.2838816

Hill, C. G., Haag, M., Oleson, A., Mendez, C., Marsden, N., Sarma, A., & Burnett, M. (2017). Gender-inclusiveness Personas vs. stereotyping: Can we have it both ways? In *Proceedings of the 2017 CHI Conference* (pp. 6658–6671). ACM Presshttps://doi.org/10.1145/3025453.3025609

Holden, R. J., Kulanthaivel, A., Purkayastha, S., Goggins, K. M., & Kripalani, S. (2017, December). Know thy eHealth user: Development of biopsychosocial personas from a study of older adults with heart failure. *International Journal of Medical Informatics*, *108*, 158–167. https://doi.org/10.1016/j.ijmedinf.2017.10.006

Holmgård, C., Green, M. C., Liapis, A., & Togelius, J. (2018, February) . Automated playtesting with procedural Personas through MCTS with evolved heuristics. *arXiv:1802.06881 [cs]*. . Retrieved June 20, 2019, from. http://arxiv.org/abs/1802.06881

Holmgard, C., Liapis, A., Togelius, J., & Yannakakis, G. N. (2014). Evolving personas for player decision modeling. In *Computational Intelligence and Games (CIG), 2014 IEEE Conference on* (pp. 1–8). IEEE.

Hou, W.-J., Yan, X.-Y., & Liu, J.-X. (2020). A method for quickly establishing Personas. In *Artificial Intelligence in HCI (Lecture Notes in Computer Science)* (pp. 16–32). Springer International Publishing. https://doi.org/10.1007/978-3-030-50334-5_2

Huh, J., Kwon, B. C., Kim, S.-H., Lee, S., Choo, J., Kim, J., Choi, M.-J., & Yi, J. S. (2016, October). Personas in online health communities. *Journal of Biomedical Informatics*, *63*, 212–225. https://doi.org/10.1016/j.jbi.2016.08.019

Hussain, Z., Milchrahm, H., Shahzad, S., Slany, W., Tscheligi, M., & Wolkerstorfer, P. (2009). Integration of extreme programming and user-centered design: Lessons learned. In *Agile Processes in Software Engineering and Extreme Programming*, P. Abrahamsson, M. Marchesi, & F. Maurer (eds.) (pp. 174–179). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-01853-4_23

Hwang, S., Kim, B., & Lee, K. (2019). A data-driven design framework for customer service chatbot. In *International Conference on Human-Computer Interaction* (pp. 222–236). Springer.

Ishii, R., Ito, S., Ishihara, M., Harada, T., & Thawonmas, R. (2018). Monte-Carlo Tree Search Implementation of Fighting game ais having Personas. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)* (pp. 1–8). IEEE. https://doi.org/10.1109/CIG.2018.8490367

Jansen, A., Van Mechelen, M., & Slegers, K. (2017). Personas and behavioral theories: A case study using self-determination theory to construct overweight Personas. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2127–2136). ACM. https://doi.org/10.1145/3025453.3026003

Jansen, B. J., Jisun, A., Kwak, H., & Cho, H. (2016). Efforts towards automatically generating personas in real-time using actual user data. In *Qatar Foundation Annual Research Conference Proceedings Volume 2016 Issue 1*. Hamad bin Khalifa University Press (HBKU Press). ICTPP3230.

Jansen, B. J., Jung, S.-G., & Salminen, J. (2019). Capturing the change in topical interests of personas over time. *Proceedings of the Association for Information Science and Technology*, *56*(1), 127–136. https://doi.org/10.1002/pra2.11

Jansen, B. J., Jung, S.-G., & Salminen, J. (2020, October). From flat file to interface: Synthesis of personas and analytics for enhanced user understanding. *Proceedings of the Association for Information Science and Technology*, *57*(1). https://doi.org/10.1002/pra2.215

Jansen, B. J., Salminen, J., & Jung, S.-G. (2020). Data-driven Personas for enhanced user understanding: Combining empathy with rationality for better insights to analytics. *Data and Information Management*, *4*(1). https://doi.org/10.2478/dim-2020-0005

Jen Mcginn, J., & Kotamraju, N. (2008). Data-driven persona development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1521–1524). ACM. https://doi.org/10.1145/1357054.1357292

Jenkinson, A. (1994). Beyond segmentation. *Journal of Targeting, Measurement and Analysis for Marketing*, *3*(1), 60–72.

Jung, S.-G., Salminen, J., & Jansen, B. J. (2019). Personas changing over time: Analyzing variations of data-driven personas during a two-year period. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19* (pp. 1–6). ACM Press. https://doi.org/10.1145/3290607.3312955

Kanno, T., Ooyabu, T., & Furuta, K. (2011). Integrating human modeling and simulation with the Persona method. In *Universal Access in Human-Computer Interaction. Users Diversity (Lecture Notes in Computer Science)* (pp. 51–60). Springer Berlin Heidelberg.

Kim, E., Yoon, J., Kwon, J., Liaw, T., & Agogino, A. M. (2019). From innocent irene to parental patrick: Framing user characteristics and personas to design for cybersecurity. In *Proceedings of the Design Society: International Conference on Engineering Design* (pp. 1773–1782)Cambridge University Press.

Kim, H. M., & Wiggins, J. (2016). A factor analysis approach to persona development using survey data. In *Proceedings of the 2016 Library Assessment Conference* (pp. 11). MDPI.

Konstantakis, M., Alexandridis, G., & Caridakis, G. (2020, June). A personalized heritage-oriented recommender system based on extended cultural tourist typologies. *Big Data and Cognitive Computing*, *4*(2), 12. https://doi.org/10.3390/bdcc4020012

Kwak, H., Jisun, A., & Jansen, B. J. (2017). Automatic generation of Personas using youtube social media data. In *Proceedins of the Hawaii International Conference on System Sciences (HICSS-50)* (pp. 833–842). AIS.

Kwak, H., Jisun, A., Salminen, J., Jung, S.-G., & Jansen, B. J. (2018). What we read, what we search: Media attention and public attention among 193 countries. In *Proceedings of the Web Conference*. ACM. Retrieved May 10, 2018, from http://arxiv.org/abs/1802.06437

Laporte, L., Slegers, K., & De Grooff, D. (2012). Using correspondence analysis to monitor the Persona segmentation process. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design (NordiCHI '12)* (pp. 265–274). ACM. https://doi.org/10.1145/2399016.2399058

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., & Dolan, B. (2016). A Persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 994–1003). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1094

Long, F. (2009). Real or imaginary: The effectiveness of using personas in product design. In *Proceedings of the Irish Ergonomics Society Annual Conference*. Irish Ergonomics Society Dublin.

Marsden, N., & Haag, M. (2016). Stereotypes and politics: Reflections on personas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4017–4031). ACM.

Masiero, A. A., Leite, M. G., Filgueiras, L. V., & Aquino, P. T., Jr. (2011). Multidirectional knowledge extraction process for creating behavioral Personas. In *Proceedings of the 10th Brazilian Symposium on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction (IHC+CLIHC '11)* (pp. 91–99). Brazilian Computer Society. Retrieved June 24, 2019, from http://dl.acm.org/citation.cfm?id=2254436.2254454

Matthews, T., Judge, T., & Whittaker, S. (2012). How do designers and user experience professionals actually perceive and use personas? In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (pp. 1219). ACM Press. https://doi.org/10.1145/2207676.2208573

Mesgari, M., Okoli, C., & Ortiz De Guinea, A. (2015). Affordance-based user Personas : A mixed-method approach to Persona development. In *AMCIS 2015 Proceedings*. AMCIS. https://aisel.aisnet.org/amcis2015/HCI/GeneralPresentations/1

Miaskiewicz, T., & Luxmoore, C. (2017). The use of data-driven personas to facilitate organizational adoption–a case study. *The Design Journal*, *20*(3), 357–374. https://doi.org/10.1080/14606925.2017.1301160

Miaskiewicz, T., Sumner, T., & Kozar, K. A. (2008). A latent semantic analysis methodology for the identification and creation of personas. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1501–1510). ACM. http://dl.acm.org/citation.cfm?id=1357290

Mijač, T., Jadrić, M., & Ćukušić, M. (2018). The potential and issues in data-driven development of web personas. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1237–1242). https://doi.org/10.23919/MIPRO.2018.8400224

Minichiello, A., Hood, J. R., & Harkness, D. S. (2018). Bringing user experience design to bear on STEM education: A narrative literature review. *Journal for STEM Education Research*, *1*(1–2), 7–33. https://doi.org/10.1007/s41979-018-0005-3

Minichiello, A., Hood, J. R., & Harkness, D. S. (2017). *Work in progress: Methodological considerations for constructing nontraditional student Personas with scenarios from online forum usage data in calculus*. American Society for Engineering Education.

Mulder, S., & Yaar, Z. (2006). *The user is always right: A practical guide to creating and using personas for the web*. New Riders.
Nielsen, L. (2019). *Personas - user focused design* (2nd ed.). Springer.
Pruitt, J., & Grudin, J. (2003). Personas: Practice and theory. In *Proceedings of the 2003 Conference on Designing for User Experiences (DUX '03)* (pp. 1–15). . ACMhttps://doi.org/10.1145/997078.997089
Radjenović, D., Heričko, M., Torkar, R., & Aleš, Ž. (2013). Software fault prediction metrics: A systematic literature review. *Information and Software Technology*, *55*(8), 1397–1418. https://doi.org/10.1016/j.infsof.2013.02.009
Rahimi, M., & Cleland-Huang, J. (2014). Personas in the middle: Automated support for creating personas as focal points in feature gathering forums. *In Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering (ASE '14)* (pp. 479–484). ACM. https://doi.org/10.1145/2642937.2642958
Rönkkö, K. (2005). An empirical study demonstrating how different design constraints, project organization and contexts limited the utility of Personas. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences - Volume 08 (HICSS '05)*. IEEE Computer Society. https://doi.org/10.1109/HICSS.2005.85
Salminen, J., Guan, K., Jung, S.-G., Chowdhury, S. A., & Jansen, B. J. (2020). A literature review of quantitative Persona creation. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM. https://doi.org/10.1145/3313831.3376502
Salminen, J., Jansen, B. J., An, J., Kwak, H., & Jung, S.-G. (2018, November). Are personas done? Evaluating their usefulness in the age of digital analytics. *Persona Studies*, *4*(2), 47–65. https://doi.org/10.21153/psj2018vol4no2art737
Salminen, J., Jansen, B. J., An, J., Kwak, H., & Jung, S.-G. (2019). Automatic Persona generation for online content creators: Conceptual rationale and a research agenda. In L. Nielsen, Ed.. *Personas - user focused design*. 2nd (pp. 135–160). Springer London. https://doi.org/10.1007/978-1-4471-7427-1_8
Salminen, J., Jung, S.-G., Chowdhury, S. A., Sengün, S., & Jansen, B. J. (2020). Personas and analytics: A comparative user study of efficiency and effectiveness for a user identification task. In *Proceedings of the ACM Conference of Human Factors in Computing Systems (CHI'20)*. ACM. https://doi.org/10.1145/3313831.3376770
Salminen, J., Jung, S.-G., & Jansen, B. J. (2019a). The future of data-driven personas: A marriage of online analytics numbers and human attributes. In *ICEIS 2019 - Proceedings of the 21st International Conference on Enterprise Information Systems* (pp. 596–603). SciTePress, Heraklion.
Salminen, J., Jung, S.-G., & Jansen, B. J. (2019b). Detecting demographic bias in automatically generated Personas. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*(pp. LBW0122:1-LBW0122:6). ACM. https://doi.org/10.1145/3290607.3313034
Salminen, J., Kwak, H., Santos, J. M., Jung, S.-G., An, J., & Jansen, B. J. (2018). Persona perception scale: Developing and validating an instrument for human-like representations of data. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (pp. 1–6). ACM Press. https://doi.org/10.1145/3170427.3188461
Salminen, J., Nielsen, L., Jung, S.-G., An, J., Kwak, H., & Jansen, B. J. (2018). "Is more better?": Impact of multiple photos on perception of Persona profiles. In *Proceedings of ACM CHI Conference on Human Factors in Computing Systems (CHI2018)*. ACM. https://doi.org/10.1145/3173574.3173891
Salminen, J., Şengün, S., Kwak, H., Jansen, B. J., An, J., Jung, S.-G., Vieweg, S., & Fox Harrell, D. (2018, June). From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas. *FM*, *23*(6). https://doi.org/10.5210/fm.v23i6.8415
Siegel, D. A. (2010). The mystique of numbers: Belief in quantitative approaches to segmentation and persona development. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)* (pp. 4721–4732). ACM. https://doi.org/10.1145/1753846.1754221
Spiliotopoulos, D., Margaris, D., & Vassilakis, C. (2020, September). Data-assisted Persona construction using social media data. *Big Data and Cognitive Computing*, *4*(3), 21. https://doi.org/10.3390/bdcc4030021
Stevenson, P. D., & Mattson, C. A. (2019, July). The personification of big data. *Proceedings of the Design Society: International Conference on Engineering Design*, *1*(1), 4019–4028. https://doi.org/10.1017/dsi.2019.409
Tanenbaum, M. L., Adams, R. N., Esti Iturralde, S. J., Hanes, R. C., Barley, D. N., & Hood, K. K. (2018, November). From Wary Wearers to d-Embracers: Personas of readiness to use diabetes devices. *Journal of Diabetes Science and Technology*, *12*(6), 1101–1107. https://doi.org/10.1177/1932296818793756
Tempelman-Kluit, N., & Pearce, A. (2014, September). Invoking the user from data to design. *CRL*, *75*(5), 616–640. https://doi.org/10.5860/crl.75.5.616
Thoma, V., & Williams, B. (2009). Developing and validating Personas in e-Commerce: A heuristic approach. In *Human-Computer Interaction – INTERACT 2009 (Lecture Notes in Computer Science)* (pp. 524–527). Springer Berlin Heidelberg.
Torgerson, C. (2003). *Systematic reviews*. A&C Black.
Tu, N., Dong, X., Rau, P. P., & Zhang, T. (2010). Using cluster analysis in Persona development. In *2010 8th International Conference on Supply Chain Management and Information* (pp. 1–5). IEEE.
Tu, N., He, Q., Zhang, T., Zhang, H., Li, Y., Xu, H., & Xiang, Y. (2010). Combine qualitative and quantitative methods to create Persona. In *2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering* (pp. 597–603). IEEE. https://doi.org/10.1109/ICIII.2010.463
Turner, P., & Turner, S. (2011). Is stereotyping inevitable when designing with personas? *Design Studies*, *32*(1), 30–44. https://doi.org/10.1016/j.destud.2010.06.002
Tychsen, A., & Canossa, A. (2008). Defining Personas in games using metrics. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share (Future Play '08)* (pp. 73–80). ACM. https://doi.org/10.1145/1496984.1496997
Van Laar, E., Van Deursen, A. J. A. M., Van Dijk, J. A. G. M., & De Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in Human Behavior*, *72*(2017), 577–588. https://doi.org/10.1016/j.chb.2017.03.010
Vosbergen, S., Mulder-Wiggers, J. M. R., Lacroix, J. P., Kemps, H. M. C., Kraaijenhagen, R. A., Jaspers, M. W. M., & Peek, N. (2015, February). Using personas to tailor educational messages to the preferences of coronary heart disease patients. *Journal of Biomedical Informatics*, *53*, 100–112. Elsevier. https://doi.org/10.1016/j.jbi.2014.09.004
Wang, L., Li, L., Cai, H., Xu, L., Xu, B., & Jiang, L. (2018). Analysis of regional group health persona based on image recognition. In *2018 Sixth International Conference on Enterprise Systems (ES)* (pp. 166–171). IEEE. https://doi.org/10.1109/ES.2018.00033
Watanabe, Y., Washizaki, H., Honda, K., Noyori, Y., Fukazawa, Y., Morizuki, A., Shibata, H., Ogawa, K., Ishigaki, M., Shiizaki, S., Yamaguchi, T., & Yagi, T. (2017). ID3P: Iterative data-driven development of persona based on quantitative evaluation and revision. In *Proceedings of the 10th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE '17)* (pp. 49–55). IEEE Press. https://doi.org/10.1109/CHASE.2017.9
Williams, K. L. (2006). *Personas in the design process: A tool for understanding others* [PhD Thesis]. Georgia Institute of Technology.
Wöckl, B., Yildizoglu, U., Buber, I., Diaz, B. A., Kruijff, E., & Tscheligi, M. (2012). Basic senior Personas: A representative design tool covering the spectrum of european older adults. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '12)* (pp. 25–32), ACM. https://doi.org/10.1145/2384916.2384922
Wu, I.-C., & Yu, H.-K. (2020, November). Sequential analysis and clustering to investigate users' online shopping behaviors based on need-states. *Information Processing & Management*, *57*(6), 102323. https://doi.org/10.1016/j.ipm.2020.102323

Zaugg, H., & Ziegenfuss, D. H. (2018, August). Comparison of personas between two academic libraries. *Performance Measurement Metric*, *19* (3), 142–152. https://doi.org/10.1108/PMM-04-2018-0013

Zhang, X., Brown, H.-F., & Shankar, A. (2016). Data-driven Personas: Constructing archetypal users with clickstreams and user telemetry. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)* (pp. 5350–5359). ACM.

Zhu, H., Wang, H., & Carroll, J. M. (2019). Creating persona skeletons from imbalanced datasets - A case study using U.S. older adults' health data. In *Proceedings of the 2019 on Designing Interactive Systems Conference - DIS '19*, ACM Press, San Diego, CA, USA, 61–70. https://doi.org/10.1145/3322276.3322285

About the Authors

Joni Salminen is currently a research scientist at Qatar Computing Research Institute, HBKU; and at Turku School of Economics. His current research interests include automatic persona generation from social media and online analytics data, the societal impact of machine decisionmaking (#algoritmitutkimus), and related social computing topics.

Kathleen Guan is a research student in Neuroscience and Psychopathology through a joint graduate program between University College London and Yale School of Medicine. She has a Bachelor of Science in Foreign Service in International Law from Georgetown University, and research training in Public Health from Johns Hopkins University.

Soon-Gyo Jung is a software engineer focused on implementing data analytics systems at Qatar Computing Research Institute. He received a B.E. degree in computer software from the Kwangwoon University, Seoul, Korea, in 2014, and an M.S. degree in electrical and computer engineering from the Sungkyunkwan University, Suwon, Korea, in 2016.

Bernard J. Jansen is currently a Principal Scientist in the social computing group of the Qatar Computing Research Institute. He is a graduate of West Point and has a Ph.D. in computer science from Texas A&M University. Professor Jansen is editor-in-chief of the journal, Information Processing & Management (Elsevier).

Appendix A. List of coded articles

Table A1. Reviewed research articles.

Title	Year	Authors	Publication Venue
Persona-and-scenario based requirements engineering for software embedded in digital consumer products	2005	Mikio Aoyama	*13th IEEE International Conference on Requirements Engineering (RE '05)*
Persona-scenario-goal methodology for user-centered requirements engineering	2007	Mikio Aoyama	*15th IEEE International Requirements Engineering Conference (RE '07)*
A Latent Semantic Analysis Methodology for the Identification and Creation of Personas	2008	Tomasz Miaskiewicz, Tamara Sumner, and Kenneth A. Kozar	*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*
Data-driven persona development	2008	Jennifer Jen McGinn and Nalini Kotamraju	*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*
Defining Personas in Games Using Metrics	2008	Anders Tychsen and Alessandro Canossa	*Proceedings of the 2008 Conference on Future Play: Research, Play, Share*
Developing and validating personas in e-commerce: A heuristic approach	2009	Volker Thoma and Bryn Williams	*Human-Computer Interaction – INTERACT 2009* (Lecture Notes in Computer Science)
A Comparative Analysis of Persona Clustering Methods	2010	Jon Brickey, Steven Walczak, and Tony Burgess	*AMCIS 2010 Proceedings*
Combine Qualitative and Quantitative Methods to Create Persona	2010	Nan Tu, Qiuyun He, Tian Zhang, Haofeng Zhang, Yahui Li, Han Xu, and Yang Xiang	*3rd International Conference on Information Management, Innovation Management and Industrial Engineering*
Using cluster analysis in persona development	2010	Nan Tu, Xiao Dong, Pei-Luen Patrick Rau, and Tao Zhang	*8th International Conference on Supply Chain Management and Information*
Integrating Human Modeling and Simulation with the Persona Method	2011	Taro Kanno, Tomohiko Ooyabu, and Kazuo Furuta	*Universal Access in Human-Computer Interaction. Users Diversity* (Lecture Notes in Computer Science)
Multidirectional Knowledge Extraction Process for Creating Behavioral Personas	2011	Andrey A. Masiero, Mayara G. Leite, Lucia Vilela Leite Filgueiras, and Plinio Thomaz Aquino Jr.	*Proceedings of the 10th Brazilian Symposium on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction* (IHC+CLIHC '11)
Basic senior personas: a representative design tool covering the spectrum of European older adults	2012	Bernhard Wöckl, Ulcay Yildizoglu, Isabella Buber, Belinda Aparicio Diaz, Ernst Kruijff, and Manfred Tscheligi	*Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '12)
Comparing Semi-Automated Clustering Methods for Persona Development	2012	J. Brickey, S. Walczak, and T. Burgess	*IEEE Transactions on Software Engineering*
Learning Latent Personas of film Characters	2013	David Bamman, Brendan O'Connor, and Noah A Smith	*Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*,
Behavioral Persona for Human-Robot Interaction: A Study Based on Pet Robot	2011	Thiago Freitas dos Santos, Danilo Gouveia de Castro, Andrey Araujo Masiero, and Plinio Thomaz Aquino Junior	*International Conference on Human-Computer Interaction*
Evolving personas for player decision modeling	2014	Christoffer Holmgard, Antonios Liapis, Julian Togelius, and Georgios N. Yannakakis	*Computational Intelligence and Games (CIG)*
Personas in the Middle: Automated Support for Creating Personas As Focal Points in Feature Gathering Forums	2014	Mona Rahimi and Jane Cleland-Huang	*Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering* (ASE '14)
Explaining predictive models to learning specialists using personas	2014	Christopher Brooks and Jim Greer	*Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (LAK '14)
Invoking the User from Data to Design	2014	Nadaleen Tempelman-Kluit and Alexa Pearce	*CRL* (College & Research Libraries) 75
Affordance-based User Personas : A Mixed-method Approach to Persona Development	2015	Mostafa Mesgari, Chitu Okoli, and Ana Ortiz de Guinea	*AMCIS 2015 Proceedings*
Anthropological User Research: A Data-Driven Approach to Personas Development	2015	Hang Guo and Khasfariyati Binte Razikin	*Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction* (OzCHI '15)
Demystifying online personas of Vietnamese young adults on Facebook: A Q-methodology approach	2015	Duy Dang-Pham, Siddhi Pittayachawan, and Mathews Nkhoma	*Australasian Journal of Information Systems* 19
Profile CBC: Using Conjoint Analysis for Consumer Profiles	2015	Chris Chapman, Kate Krontiris, and John Webb	*Sawtooth Software Conference Proceedings*
Using personas to tailor educational messages to the preferences of coronary heart disease patients	2015	S. Vosbergen, J. M. R. Mulder-Wiggers, J. P. Lacroix, H. M. C. Kemps, R. A. Kraaijenhagen, M. W. M. Jaspers, and N. Peek	*Journal of Biomedical Informatics* 53
A Factor Analysis Approach to Persona Development using Survey Data	2016	Hae Min Kim and John Wiggins	*Proceedings of the 2016 Library Assessment Conference*
A Persona-Based Neural Conversation Model	2016	Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan	*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*
Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry	2016	Xiang Zhang, Hans-Frederick Brown, and Anil Shankar	*Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16)
Personas in online health communities	2016	Jina Huh, Bum Chul Kwon, Sung-Hee Kim, Sukwon Lee, Jaegul Choo, Jihoon Kim, Min-Je Choi, and Ji Soo Yi	*Journal of Biomedical Informatics* 63
Privacy Personas: Clustering Users via Attitudes and Behaviors Toward Security Practices	2016	Janna Lynn Dupree, Richard Devries, Daniel M. Berry, and Edward Lank	*Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16)
Animal personas: representing dog stakeholders in interaction design	2017	Ilyena Hirskyj-Douglas, Janet C Read, and Matthew Horton	*Proceedings of the 31st British Computer Society Human Computer Interaction Conference* (HCI '17)

(*Continued*)

Table A1. (Continued).

Title	Year	Authors	Publication Venue
Ego-Splitting Framework: From Non-Overlapping to Overlapping Clusters	2017	Alessandro Epasto, Silvio Lattanzi, and Renato Paes Leme	*Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '17)
ID3P: Iterative Data-driven Development of Persona Based on Quantitative Evaluation and Revision	2017	Yasuhiro Watanabe, Hironori Washizaki, Kiyoshi Honda, Yuki Noyori, Yoshiaki Fukazawa, Aoi Morizuki, Hiroyuki Shibata, Kentaro Ogawa, Mikako Ishigaki, Satiyo Shiizaki, Teppei Yamaguchi, and Tomoaki Yagi	*Proceedings of the 10th International Workshop on Cooperative and Human Aspects of Software Engineering* (CHASE '17)
Know thy eHealth user: Development of biopsychosocial personas from a study of older adults with heart failure	2017	Richard J. Holden, Anand Kulanthaivel, Saptarshi Purkayastha, Kathryn M. Goggins, and Sunil Kripalani	*International Journal of Medical Informatics* 108
Personas for Content Creators via Decomposed Aggregate Audience Statistics	2017	Jisun An, Haewoon Kwak, and B. J. Jansen.	*Proceedings of Advances in Social Network Analysis and Mining (ASONAM '17)*
SOPER: Discovering the Influence of Fashion and the Many Faces of User from Session Logs using Stick Breaking Process	2017	Lucky Dhakad, Mrinal Das, Chiranjib Bhattacharyya, Samik Datta, Mihir Kale, and Vivek Mehta	*Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*
The Use of Data-Driven Personas to Facilitate Organizational Adoption–A Case Study	2017	Tomasz Miaskiewicz and Coryndon Luxmoore	*The Design Journal* 20
Characterizing Software Engineering Work with Personas Based on Knowledge Worker Actions	2017	Denae Ford, Thomas Zimmermann, Christian Bird, and Nachiappan Nagappan	*Proceedings of the 11th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (ESEM '17)
Analysis of Regional Group Health Persona Based on Image Recognition	2018	L. Wang, L. Li, H. Cai, L. Xu, B. Xu, and L. Jiang	*2018 Sixth International Conference on Enterprise Systems (ES)*
Archetype-Based Modeling of Persona for Comprehensive Personality Computing from Personal Big Data	2018	Ao Guo and Jianhua Ma	*Sensors* 18
Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data	2018	Jisun An, Haewoon Kwak, Soon-gyo Jung, Joni Salminen, and Bernard J. Jansen	*Social Network Analysis and Mining* 8
Evaluating Inclusivity using Quantitative Personas	2018	Joy Goodman-Deane, Sam Waller, Dana Demin, Arantxa González-de-Heredia, Mike Bradley, and John P. Clarkson	*Design Research Society Conference*
From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas	2018	Joni Salminen, Sercan Şengün, Haewoon Kwak, Bernard J. Jansen, Jisun An, Soon-gyo Jung, Sarah Vieweg, and Fox Harrell	*First Monday* 23
From Wary Wearers to d-Embracers: Personas of Readiness to Use Diabetes Devices	2018	Molly L. Tanenbaum, Rebecca N. Adams, Esti Iturralde, Sarah J. Hanes, Regan C. Barley, Diana Naranjo, and Korey K. Hood	*J Diabetes Sci Technol* 12
Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data	2018	Jisun An, Haewoon Kwak, Joni Salminen, Soon-gyo Jung, and Bernard J. Jansen	*ACM Transactions on the Web (TWEB)* 12
Learning Personas from Dialogue with Attentive Memory Networks	2018	Eric Chu, Prashanth Vijayaraghavan, and Deb Roy	*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,
Monte-Carlo Tree Search Implementation of Fighting Game AIs Having Personas	2018	Ryota Ishii, Suguru Ito, Makoto Ishihara, Tomohiro Harada, and Ruck Thawonmas	*2018 IEEE Conference on Computational Intelligence and Games (CIG)*
Research on the Annual Reading Report of Academic Libraries Based on Personas	2018	Lan Du and Ziqi Wang	*Open Journal of Social Sciences, 6*
The potential and issues in data-driven development of web personas	2018	Tea Mijač, Mario Jadrić, and Maja Ćukušić	*2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*
Creating Persona Skeletons from Imbalanced Datasets – A Case Study using U.S. Older Adults' Health Data	2019	Haining Zhu, Hongjian Wang, and John M. Carroll	*Proceedings of the 2019 on Designing Interactive Systems Conference – DIS '19*
The Future of Data-driven Personas: A Marriage of Online Analytics Numbers and Human Attributes	2019	Joni Salminen, Soon gyo Jung, and Bernard James Jansen	*ICEIS 2019 – Proceedings of the 21st International Conference on Enterprise Information Systems*
Personas Changing Over Time: Analyzing Variations of Data-Driven Personas During a Two-Year Period	2019	Soon-gyo Jung, Joni Salminen, and Bernard J. Jansen	*Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI EA '19)
Automated Playtesting of Matching Tile Games	2019	Luvneesh Mugrai, Fernando de Mesentier Silva, Christoffer Holmgård, and Julian Togelius	*IEEE Conference On Games (COG) 2019*
Detecting Demographic Bias in Automatically Generated Personas	2019	Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen	*Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI EA '19)
Creating Manageable Persona Sets from Large User Populations	2019	Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen	*Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI EA '19)
Personas Design For Conversational Systems In Education	2019	Fatima Ali Amer Jid Almahri, David Bell and Mahir Arzoky	*Informatics, 6*
A Data-Driven Design Framework for Customer Service Chatbot	2019	Shinhee Huang, Beomjun Kim, and Keeheon Lee	*21st International Conference on Human-Computer Interaction (HCII '19)*
Capturing the Change in Topical Interests of Personas Over Time	2019	Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen	*82nd Annual Meeting of the Association for Information Science & Technology*
Deriving Personas Based on Attitudes to Interruption and Information Overload	2019	David Goddard, Paul Mulholland, and Lara Piccolo	*Proceedings of the 17th European Conference on Computer-Supported Cooperative Work*

(*Continued*)

Table A1. (Continued).

Title	Year	Authors	Publication Venue
From Innocent Irene to Parental Patrick: Framing User Characteristics and Personas to Design for Cybersecurity	2019	Euiyoung Kim, JungKyoon Yoon, Jieun Kwon, Tiffany Liaw, and Alice M. Agogino	*Proceedings of the Design Society: International Conference on Engineering Design*
Using personas to exploit environmental attitudes and behavior in sustainable product design	2019	Margaret Carey, Eoin J. White, Muireann McMahon, and Leonard W. O'Sullivan	*Applied Ergonomics, 78*
Explainable Recommendations via Attentive Multi-Persona Collaborative Filtering	2020	Oren Barkan, Yonatan Fuchs, Avi Caciularu, and Noam Koenigstein	*Fourteenth ACM Conference on Recommender Systems*
Does this persona represent me? Investigating an approach for automatic generation of personas based on questionnaires and clustering	2020	Karina da S. C. Branco, Rhenara A. Oliveira, Francisco L. Q. da Silva, Jacilane de H. Rabelo, and Anna B. S. Marques	*19th Brazilian Symposium on Human Factors in Computing Systems*
A Method for Quickly Establishing Personas	2020	Wen-jun Hou, Xiang-yuan Yan, and Jia-xin Liu	*International Conference on Human-Computer Interaction*
From flat file to interface: Synthesis of personas and analytics for enhanced user understanding	2020	Bernard J. Jansen, Soon-gyo Jung, and Joni Salminen	*Association for Information Science and Technology Annual Meeting*
Giving Faces to Data: Creating Data-Driven Personas from Personified Big Data	2020	Soon-gyo Jung, Joni Salminen, and Bernard J. Jansen	*25th International Conference on Intelligent User Interfaces*
Using logistic regression for persona segmentation in tourism: A case study	2020	Rui Kang	*Social Behavior and Personality: an international journal, 48*
A Personalized Heritage-Oriented Recommender System Based on Extended Cultural Tourist Typologies	2020	Markos Konstantakis, Georgios Alexandridis, and George Caridakis	*Big Data and Cognitive Computing, 4*
Creating user stereotypes for persona development from qualitative data through semi-automatic subspace clustering	2020	Dannie Korsgaard, Thomas Bjørner, Pernille Krog Sørensen, and Paolo Burelli	*User Modeling and User-Adapted Interaction, 30*
Developing personas & use cases with user survey data: A study on the millennials' media usage	2020	Mingyu Lee, Jiyoung Kwahk, Sung H.Han, Dawoon Jeong, Kyudong Park, Seokmin Oh, and Gunho Chae	*Journal of Retailing and Consumer Services*
CAUX-Based Mobile Personas Creation	2020	Mo Li and Zhengjie Liu	*International Conference on Computer Engineering and Networks*
Persona Prototypes for Improving the Qualitative Evaluation of Recommendation Systems	2020	Joanna Misztal-Radeck and Bipin Indurkhy	*28th ACM Conference on User Modeling, Adaptation and Personalization*
Visualized Benefit Segmentation Using Supervised Self-organizing Maps: Support Tools for Persona Design and Market Analysis	2020	Fumiaki Saitoh	*Asian Conference on Intelligent Information and Database Systems*
Enriching Social Media Personas with Personality Traits: A Deep Learning Approach Using the Big Five Classes	2020	Joni Salminen, Rohan Gurunandan Rao, Soon-gyo Jung, Shammur A. Chowdhury, and Bernard J. Jansen	*International Conference on Human-Computer Interaction*
Designing Prototype Player Personas from a Game Preference Survey	2020	Joni Salminen, Jukka Vahlo, Aki Koponen, Soon-gyo Jung, Shammur A Chowdhury, and Bernard J Jansen	*2020 CHI Conference on Human Factors in Computing Systems*
Data-Assisted Persona Construction Using Social Media Data	2020	Dimitris Spiliotopoulos, Dionisis Margaris, and Costas Vassilakis	*Big Data and Cognitive Computing*
Data Driven Decision Making to Characterize Clinical Personas of Parents of Children with Cystic Fibrosis: A Mixed Methods Study	2020	Rhonda D. Szczesniak, Teresa Pestian, Leo L. Duan, Dan Li, Sophia Stamper, Brycen Ferrara, Elizabeth Kramer, John P. Clancy, and Daniel Grossoehm	*BMC Pulmonary Medicine, 20*
DAPPER: Learning Domain-Adapted Persona Representation Using Pretrained BERT and External Memory	2020	Prashanth Vijayaraghavan, Eric Chu, and Deb Roy	*10th International Joint Conference on Natural Language Processing*