Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry

Xiang Zhang
Dept. of Statistics
North Carolina State University
Raleigh, NC, USA
xzhang23@ncsu.edu

Hans-Frederick Brown
User Experience
Platfora Inc.
San Mateo, USA
hans@platfora.com

Anil Shankar
User Experience
Platfora Inc.
San Mateo, USA
anilkshankar@gmail.com

ABSTRACT

User Experience (UX) research teams following a user centered design approach harness *personas* to better understand a user's workflow by examining that user's behavior, goals, needs, wants, and frustrations. To create target personas these researchers rely on workflow data from surveys, self-reports, interviews, and user observation. However, this data not directly related to user behavior, weakly reflects a user's actual workflow in the product, is costly to collect, is limited to a few hundred responses, and is outdated as soon as a persona's workflows evolve. To address these limitations we present a quantitative bottom-up data-driven approach to create personas. First, we directly incorporate user behavior via clicks gathered automatically from telemetry data related to the actual product use in the field; since the data collection is automatic it is also cost effective. Next, we aggregate 3.5 million clicks from 2400 users into $39,000$ clickstreams and then structure them into 10 workflows via hierarchical clustering; we thus base our personas on a large data sample. Finally, we use *mixed models*, a statistical approach that incorporates these clustered workflows to create five representative personas; updating our mixed model ensures that these personas remain current. We also validated these personas with our product's user behavior experts to ensure that workflows and the persona goals represent actual product use.

Keywords

Personas, User Experience Research, Data-driven design, Machine Learning, Data Science, User Analytics

1. INTRODUCTION

Cooper originated the *Persona* concept for Design and User Experience (UX) practice [3]. He considers a persona as an archetypal user. A persona comprises this user's behaviors, goals, wants, needs, and frustrations when using a product. While designers and user-experience (UX) researchers use personas as a powerful design tool in practice, a cost-effective automatic way to create a persona that can remain current is still an open question [6].

In Platfora's case, at the product's inception in 2011 we consulted with *Cooper*, a design strategy consulting agency based in San Francisco, to explore the needs of our product's end-users. Platfora is an enterprise focused big data analytics platform that transforms raw data in Hadoop (or in other data silos) into interactive, in memory business intelligence. Four main Platfora personas emerged from Cooper's research efforts; we discuss the two main user personas here and cover the other while describing the data-driven personas later in the paper. The first, *Jeffrey*, a business analyst, whose primary goal is to understand internal client needs and build out reports for them. Jeffrey would explore both the data import and visualization parts of Platfora to be successful in his work role. The second persona, *Marybeth*, a business analyst, unlike Jeffrey focuses only on explaining the story behind the data to her clients. She would primarily spend time in the visualization component of Platfora. With Platfora's increased adoption in data analytics world, we witnessed new users different from the existing Platfora personas begin to leverage our product. Hence, we started to investigate a data driven approach to refine our existing personas.

Many have suggested creating personas based on data-driven methods. Prutti argued that personas are not believable if they are designed by committee (not based on data) or if the relationship to user data is not clear [15]. To ameliorate Prutti's concern Mulder recommends personas generated by quantitative techniques to avoid human bias [13]. Quantitative or qualitative techniques depend on user data and McGinn opined that persona creation suffers from a common problem of not being based on first-hand customer behavior data, and if so, then the size of data is not statistically significant [10]. So, a quantitative, data-driven approach to create behavior defined personas still remains an open avenue for user research.

In this paper, we propose a fully data-driven approach to construct behavioral personas from raw clicks gathered in telemetry data from our product use. Our approach is different from the existing persona creation approaches in that we approach the problem bottom-up from raw clickstreams without top level features. Instead of relying on survey data, user interviews (transcribed or otherwise), we follow a two stage statistical machine learning approach to create data-driven personas that are solely based on user behaviors.

In the first stage, we aggregate 3.5 million clicks from 2400 users into $39,000$ clickstreams to identify ten common workflows. Even though this process is completely data-driven we also validated these ten workflows with user behavior experts in UX research, UX Design, and Product Management within Platfora; these folks were cognizant of existing Cooper personas. The persona experts veri-

CHI'16, May 07-12, 2016, San Jose, CA, USA
© 2016 ACM. ISBN 978-1-4503-3362-7/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858523

fied the clustering of workflows and helped to map a few workflows clusters to existing personas, thereby helping us to refine and update Platfora's archetypal users. Since we focus on the quantitative aspect of creating personas we do not provide additional details about these experts reviewing our personas in this paper.

In the second stage, we use a mixed model to group these representative workflows into five personas. To the best of our knowledge, this is the very first attempt to create data-driven personas from raw clickstreams. Our raw clickstream data gathered via product telemetry spans a duration of two years worth of clicks in *Platfora*'s user interface.

The personas that we generated from clickstreams validated two of the existing Platfora personas while showing that the other two personas had evolved into *Full-stack* users. A Full-stack user generates workflows related to Platfora's four main product pages related to *Datasets* (data import and curation), *Lenses*, (creating in-memory data-marts), *Vizboards* (a board populated by multiple visualizations), and *SystemPage* (monitoring the health and status of various Platfora objects).

We next review related work on data-driven personas. The following section describes our approach to leverage clicks from raw telemetry data gathered in the field (from customers) to create common workflows which are then grouped into personas using a mixed model. The subsequent section details our approach to identify ten common workflows with a hierarchical clustering technique on clickstream information to derive five personas. We conclude this paper by identifying the next steps for data-driven persona work that can not only incorporate the clicks in the raw data but also factor the order of the clicks to create richer, more distinguishable personas.

2. RELATED WORK: DATA-DRIVEN PERSONAS

Researchers have relied primarily on survey data to create data-driven personas. Sinha identified three personas for online restaurant finder by using *principal component analysis* on survey data gathered from 63 respondents. In this survey respondents rated 32 dimensions related to restaurant experience [16]. Instead of principal components, Javahery used *clustering* to construct three personas for guiding the new product design for a bioinformatics visualization application [7]. The data gathered from questionnaire and user observations included attributes from 22 users related to domain experience and background.

McGinn used *factor analysis* on the survey data gathered from 1300 respondents to develop eleven personas for customers of a training organization [10]. Each survey included 18 multiple-choice questions on demographic and behavior data. Miaskiewicz used *latent semantic analysis* to create four personas for users of an Institutional Repository (IR) at a large university based on the transcribed text from interviews [12] . The 20 interviews were audio recorded, next transcribed, and then stored in word-by-interview matrices.

Brickey compared several existing qualitative and quantitative persona creation methods for users of an online knowledge management system [1]. He compared the results from manual experts and found that principal component analysis produced the most promising personas for his target user group.

Tobias mined *Tomb Raider: Underworld* game telemetry to predict when a player would stop playing and also the time this player would take to finish the game [9]. Canossas leveraged game telemetry to generate play-personas and focused on improving the player experience for single player action games [2]. Drachen worked in player modeling for the Tomb Raider: Underworld game [5]. They first identified six gameplay features and then clustered these features using a hierarchical clustering technique in the game telemetry data.

Similar to the existing body of data-driven persona construction, we use a clustering algorithm to create Platfora personas. However, unlike the existing approaches that leverage mostly survey data, we harness raw user behavior data from clickstreams captured during Platfora use by thousands of users. Our approach to use this clickstream information about user behavior to create personas is different in two ways from approaches that rely on survey or interview data. First, our approach is more robust due to the large volume of user-behavior data that we harness to generate personas. Second, we use a bottom-up approach that pops out the personas automatically via hierarchical clustering and mixed models. Our approach overlaps with work on player modeling in games in that we leverage product telemetry to generate Platfora personas. We extend this work in player modeling in two ways. First, our approach starts bottom up from raw clickstream instead of identifying any top level features. Second, instead of clustering top level features we use a mixed model to process workflows to generate personas. We describe how we gather clickstreams and product telemetry data in the next section.

3. TELEMETRY AND CLICKSTREAMS

The spacecraft industry was one of the originators of using *telemetry* to gather data remotely from sensors [14]. This telemetry data collection process was designed for efficiency, transparency, and reliability. We designed Platfora's telemetry architecture based on similar goals – to gather user behavior data with efficiency, transparency, and reliability. At the lowest level our telemetry data consists of a click (user interface event) captured whenever a user accesses a user interface element in Platfora. Platfora's customers can opt-in to send their product telemetry back to us.

A *clickstream* is a sequence of clicks generated by a user in one login session. We gathered such clickstream information for two years from July of 2013 to July of 2015. Figure 1 shows the structure of this clickstream data. This data contains four different levels. Going up from the bottom, at level 4 is a click; this forms the

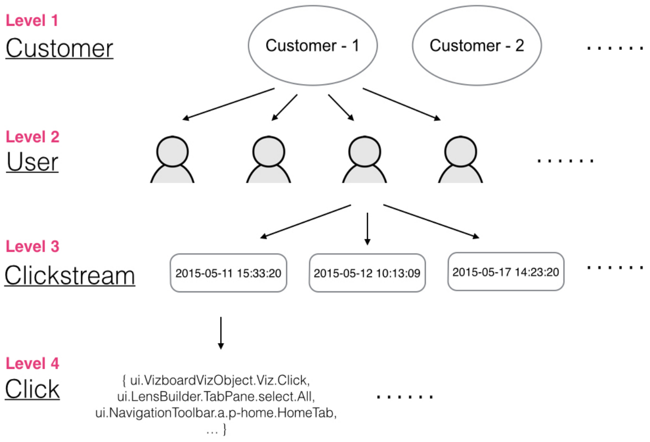

Figure 1: Product Telemetry data with four levels used to create personas.

smallest granularity of telemetry data. These clicks aggregate to form clickstreams at level 3. Higher up, at level 2,we map clickstreams to specific users (whose identification (ID) is anonymized at the customer site). The top most level corresponds to a group of users who represent a Platfora customer site. Since a single user can generate multiple clickstreams at different times, we chunk each clickstream based on the start date and time to create unique clickstream information.

Our initial analysis of clickstream lengths showed that any clickstream with less than 20 clicks was too short to reflect meaningful user behavior during a session. Hence we removed these clickstreams. This resulted in a total of 3.5 million clicks from 39,000 clickstreams generated by 2400 users from 30 customers. The average clickstream length is 92 clicks (3.5 million/39,000). Our persona construction process leverages these clickstreams by first clustering them into common workflows and then characterizing each user by the frequency of common workflows via mixed model.

4. CONSTRUCTING PERSONAS

We construct personas from clickstreams derived from telemetry data using the following algorithm.

Algorithm 1 Data-driven Persona Construction from User Clickstreams

```
Assume there are $P$ users and $N$ clickstreams in total.
*Step 1: Calculate pairwise clickstreams distance*
Calculate the $N \times N$ distance matrix, where the $(i, j)$ entry of the distance matrix is the distance between the $i$-th clickstream and the $j$-th clickstream.
*Step 2: Cluster clickstreams*
Implement hierarchical clustering algorithm on the distance matrix to obtain $K$ clusters of clickstreams. The number $K$ is chosen by the user.
*Step 3: Map clickstreams to common workflows*
Map each clickstream to one of the $K$ common workflow. Record the number of $k$-th common workflow generated by $i$-th user as $y_{ik}$, $i = 1, \ldots, P$, $k = 1, \ldots, K$.
*Step 4: Fit mixed model*
Fit mixed models with $M$ components on the multinomial data $\{y_{ik}\}_{1 \leq i \leq P, 1 \leq k \leq K}$. The number of personas, $M$, is chosen by the user. The mathematical details of this step can be found below.
```

Example data and the corresponding R code are available upon request from authors. We next discuss Step 2 and Step 4 in more detail.

4.1 Step 2: Cluster clickstreams

Platfora users follow a few common workflows to accomplish their goal and our telemetry data captures these workflows as clickstreams. Note that each clickstream (as shown in Figure 1) in our telemetry data has timestamp associated with a series of clicks (user interface events). Therefore all the clickstreams generated by users are unique in our dataset. We use *Ward*'s hierarchical clustering algorithm to cluster clickstreams and to identify common workflows [17]. We chose this hierarchical clustering algorithm as it only requires pairwise similarity between clickstreams and it produces a hierarchical treeplot. This treeplot helps to visualize the clustering process and to determine an appropriate number of clusters.

To summarize the common workflows, we first examine the most frequent clicks within a cluster, the average length and duration of clickstreams in each cluster, and the frequency of clicks that comprise key steps in Platfora in all clusters. Next, user behavior experts (UX researchers, UX Designers, Product Managers) review these clusters to verify the representative behaviors of each common workflow.

In the first stage of data-driven persona creation we cluster common clickstreams to identify common workflows. Here is an example. *Jane* can import a dataset into Platfora in two different ways. The telemetry data captures these two workflows as two different clickstreams, say *data1* and *data2*, respectively. The same user, *Jane* can follow different workflows to visualize the imported data and the resulting clickstreams could be *viz1* and *viz2*, respectively. When we cluster *Jane*'s clickstreams, we assume that *data1* is more similar to *data2* when compared to *viz1*. We use the Jaccard index to measure the similarity between two clickstreams [8]. With this index, we define the similarity between two clickstreams as the ratio of the number of common unique clicks over the number of all unique clicks.

For example, say, *Jane* generates two clickstreams S_1 and S_2; these two clickstreams are simply two sequences of clicks; an alphabet represents clicks in these two clickstreams. Assume $S_1 = \{A, B, B, C, D, B, A, C\}$ and $S_2 = \{E, F, F, B, A, C, B, B\}$. For these these two sequences, there are 3 common unique clicks ($\{A, B, C\}$) and the count of union of all unique clicks is 6 ($\{A, B, C, D, E, F\}$). We therefore calculate the similarity between S_1 and S_2 as 0.5 (3/6).

Figures 2 - 4 demonstrate the process of identifying common workflows. We first cluster clickstreams into separate groups based on click similarity and then summarize the common workflow within each cluster. In Figure 2 each dot represents a clickstream. The colors of the dots denote the similarity between clickstreams. A clustering algorithm groups all the clickstreams into separate clusters using the pairwise clickstreams similarity. Figure 3 shows that dots with similar color gradients are grouped together resulting in six clusters.

After the clustering we summarize each cluster by a common workflow which best represents all the clickstreams in that cluster. This process of identifying a workflow is radically different from survey or interview based approaches that only capture a user's self-report of her workflow. Instead of starting top-down, our data-driven approach is bottom-up; we cluster similar clickstreams into a common workflow.

We treat all similar yet distinct clickstreams as a common workflow. Figure 4 shows this clustering process where the bigger dots denote the derived common workflows, with the best representative color in that cluster. For example, in this figure, the clustering process summarized both dark red and light red by the same red color. This process indicates that a cluster of similar clickstreams represent a common workflow even though each clickstream is not exactly the same. Once we identify the common workflows from the raw data clickstreams we can next examine each user based on the different usage of workflows.

4.2 Step 4: Fit mixed model

Users from the same persona have similar goals and behaviors when using the product [3]. We view this frequency of common workflows as a quantitative signature of each user's behavior. That is, we believe users with similar frequency of common workflows are likely to be instances of the same persona. This assumption is the foundation for our *mixed model* to generate personas.

A mixed model is a widely used statistical approach to detect sub-populations [11]. In our telemetry data each persona represents a sub-population of users. A mixed model assigns each user to one of the personas and characterizes the common workflow frequency

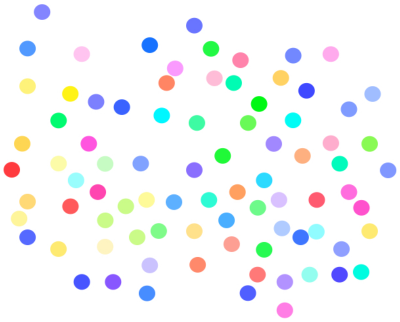

Figure 2: Each dot represents clickstream. Similar colors represent similar clickstreams.

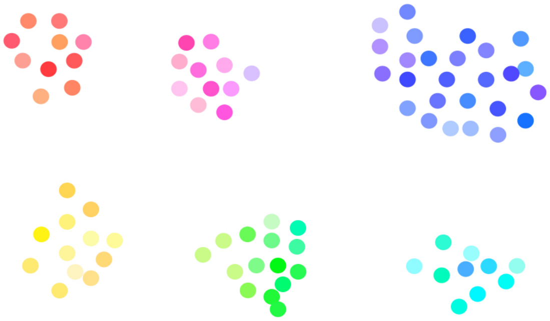

Figure 3: Hierarchical clustering groups clickstreams into six clusters based on their similarity.

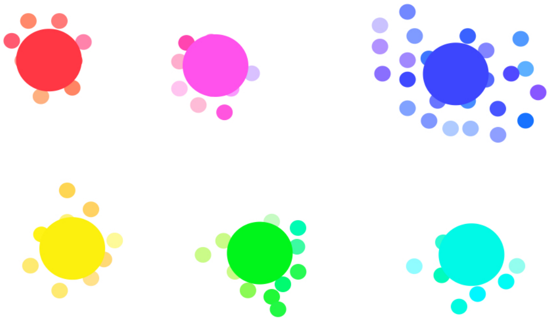

Figure 4: A central big dot depicts common workflow. This workflow summarizes clickstreams in each group.

that uniquely defines each persona. Assume there are P users and K common workflows. Denote y_{ik} as the number of k-th common workflow that are generated by the i-th user, for $i = 1, \dots, P$ and $k = 1, \dots, K$. Write $\mathbf{y}_i = (y_{i1}, \dots, y_{iK})$. Assume there are M Personas. Denote Z_i the label of Persona for the i-th user. If $Z_i = m$, the i-th user behaves as the m-th Persona, $m = 1, \dots, M$. We assume the numbers of common workflows are independently generated from multinomial distribution given the labels of Personas. That is,

$$\mathbf{y}_i | (Z_i = m) \sim \text{Multinomial}(\mathbf{p}_m),$$

where $\mathbf{p}_m = (p_{m1}, \dots, p_{mK})$, p_{mk} is the chance of m-th Persona to generate k-th common workflow, $m = 1, \dots, M$, $k = 1, \dots, K$. Denote $\pi_m = \Pr(Z_i = m)$ and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)$. The full likelihood of the observed data, given the labels of Personas, can be written as

$$L(\boldsymbol{\pi}, \mathbf{p}_1, \dots, \mathbf{p}_M) = \prod_{i=1}^{P} \sum_{m=1}^{M} (\mathbf{I}(Z_i = m)\pi_m \prod_{k=1}^{K} p_{mk}^{y_{ik}}),$$

where $\mathbf{I}(S) = 1$ if S is true and 0 otherwise. The maximum likelihood estimators (MLE) of $(\boldsymbol{\pi}, \mathbf{p}_1, \dots, \mathbf{p}_M)$ is simply the maximizer of $L(\boldsymbol{\pi}, \mathbf{p}_1, \dots, \mathbf{p}_M)$, and we solve this MLE with the Expectation-Maximization (EM) algorithm [4].

Figure 5 shows our mixed model persona creation approach based on the derived common workflows. There are nine users in this example; each user generated multiple clickstreams and these clickstreams are represented by dots. The colors of the dots indicate the common workflows as shown in Figure 4.

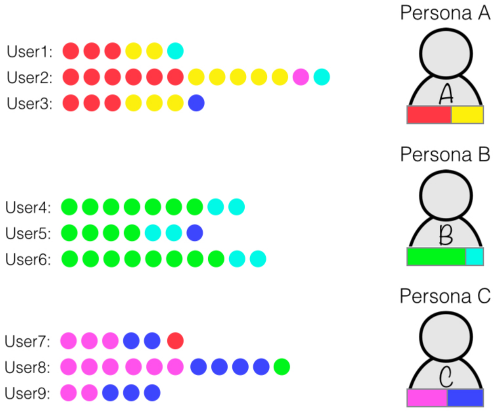

Figure 5: This figure shows the construction of personas based on clickstream types corresponding to each user. The dots following a user id represent clickstreams and colors represent the corresponding workflows. For example, Users 1, 2 and 3 have half of their clickstreams doing *red* workflow and the other half doing *yellow* workflow. Our mixed model summarized this user group as an archetypal user: Persona A.

For example, User1 generates six clickstreams: three *red* common workflows, two *yellow* common workflows and one *blue* common workflow. This generalization from clickstreams to common workflows makes it possible to easily compare the behavior of two users. Consider User2; even though User2 generates more clickstreams than User1, their behavior patterns are similar: about half of their clickstreams are *red* common workflows and the other half are *yellow* common workflows. Figure 5 shows that this pattern is also shared by User3. Thus we group users with common workflow frequency together to create an archetypal user (persona) to represent this user group. In this example, Users 1, 2, and 3, are represented by the same archetypal user, Persona A.

You can now interpret that Persona A is a type of user who generally spends half of her time in *red* workflow and the other half of the time in *yellow* workflow. Similarly, you can now see that Persona B is predominantly spends time in the *green* workflow whereas Persona C splits time between *pink* and *blue* workflows. As in the first stage, we again consulted with user behavior experts to review both the common workflows discovered by the mixed model and the ensuing personas created by the combination of these workflows. These subject matter experts agreed that these data-driven mixed-model constructed personas derived from raw clickstream data represented archetypal users of our product.

5. FIVE PERSONAS VIA HIERARCHICAL CLUSTERING AND MIXED MODELS

This section first describes our approach to cluster clickstreams to identify common workflows. We next show how to use a mixed model to build data-driven personas. This data-driven approach helps identify ten common workflows and five personas for our product users.

5.1 Ten Common Workflows via Hierarchical Clustering of Clickstreams

Figure 6 shows the result of clustering $39,000$ clickstreams. In this figure the vertical axis shows the dissimilarity between the clusters when they merge together. At the lowest level is each clickstream. This clustering process hierarchically merges similar clickstreams as you traverse up to the root node of the tree. Based on expert views and examining the frequency of the top 25 clicks in each cluster, we chose 10 as an appropriate number of clusters. Increasing the number of clusters did not result in meaningful workflows and reducing the clusters below 10 washed out the differences in workflows. The existing Cooper personas guided the development of these clusters. The red boxes in Figure 6 highlight these ten clusters. We excluded a small group of $1,147$ clickstreams (about 3 percent of the total) from the analysis (between Cluster 6 and 7 in Figure 6). This cluster represented isolated feature testing and not the common workflows.

We examine the clickstream clusters to summarize the behavior of the ten common workflows from Figure 6. This summary provides insight into the key steps triggered by each workflow and we believe that these key steps are specific to different personas. With the help of user behavior experts we identified 14 key steps related to four main Platfora product pages (Datasets, Lenses, Vizboards, and System Page). Figure 7 shows the frequency of clicks on these key steps within the ten common workflows. We order the steps such that steps from the same product page are grouped together. This ordering helps to characterize workflows by the frequency of the key steps.

Figure 8 shows clickstream frequency of steps in Cluster 3 (from Figure 6) for a workflow related to datasets and lenses in Platfora. There is a high frequency of actions for dataset and lens related activities for Cluster 3. Specifically, the frequency is high for *Modify Dataset* and for *Modify Lens*; this is common behavior of users who modify a dataset before building a lens in Platfora. We thus summarize this workflow as *Build lens with optional modification to a dataset*.

Figure 9 shows most frequent dataset related activities in Cluster 1 and Cluster 5. To tease out the difference between these two clusters we examined the top 30 clicks in each cluster with a word cloud. The dataset parsing activity was in Cluster 5's top clicks, whereas this activity was absent in Cluster 1. In Figure 9 you can see that Cluster 5 has higher frequency of clicks for *Create Dataset* step. When compared to Cluster 1, Cluster 5's higher frequency of *Create Dataset* along with the parsing activity reflected a common Platfora workflow in which users parse a dataset (curate it) only when creating a new dataset thus confirming the difference between Clusters 1 and 5. In the same figure, when compared to Cluster 5, Cluster 1 has higher activity for *Modify Dataset*, *View Dataset* and *View Lens*. This is again representative of Platfora user behavior where users are likely to examine a lens or a dataset before modifying an existing dataset. Our key contribution to data-driven persona creation is that we detect the difference between these two workflows *only* from the raw clickstream stream data. Table 1 lists the ten common workflows that we detected by clustering clickstream information. We labeled this workflows after consulting with Platfora's user behavior experts who verified the clustered clickstreams as representative user activities noted in various internal user studies and evaluations.

Table 1: The ten common Platfora workflows identified by clickstream clustering.

Cluster	Workflow
1	Modify dataset
2	Check System Page
3	Build a lens with optional modification to a dataset
4	Create a lens and visualization
5	Create a dataset
6	Create a dataset, a lens, and a visualization all in the same workflow
7	Include data segments in visualization
8	Modify visualization data iteratively
9	Export data from a visualization
10	Explore and view a visualization

5.2 Five Data-driven Personas

We described the identification of common workflows from clickstreams in the previous section. Based on these workflows we next construct data-driven personas. Users with similar workflow frequency share similar behavior even though their individual clickstreams may not exactly be the same. These users are also likely to be instances of the same persona. To generate personas, we focus our analysis on the $1,011$ users who generated at least five clickstreams; these clickstreams comprise 90 percent of the total number of clickstreams in our product telemetry. We dropped users with less than five clickstreams as this number was insufficient to depict meaningful high-level user behavior.

Based on discussion with user behavior experts at Platfora we input five personas to the mixed model. Figure 10 shows the frequency of the ten common workflows for the $1,011$ users where the users are ordered by their personas. Here, we assign each user to a persona that this user is most likely to belong to based on the fitted mixed model. In Figure 10 we were now able to detect groupings of users in the heatmap; we show these groupings by the blue blocks. That is, we detected five groups of users who mapped to

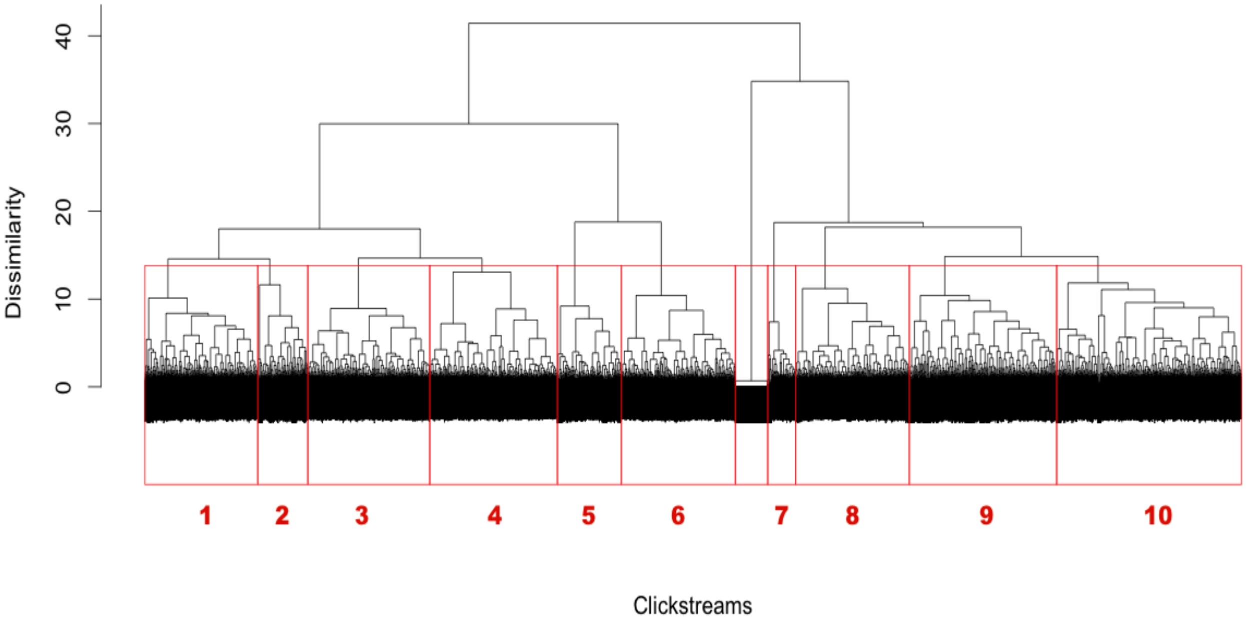

Figure 6: The ten clickstream clusters. At the bottom of the tree is each clickstream. This clustering process hierarchically merges similar clickstreams as you traverse up to the root node of the tree. The red boxes highlight the ten clickstream clusters that we leverage to create personas.

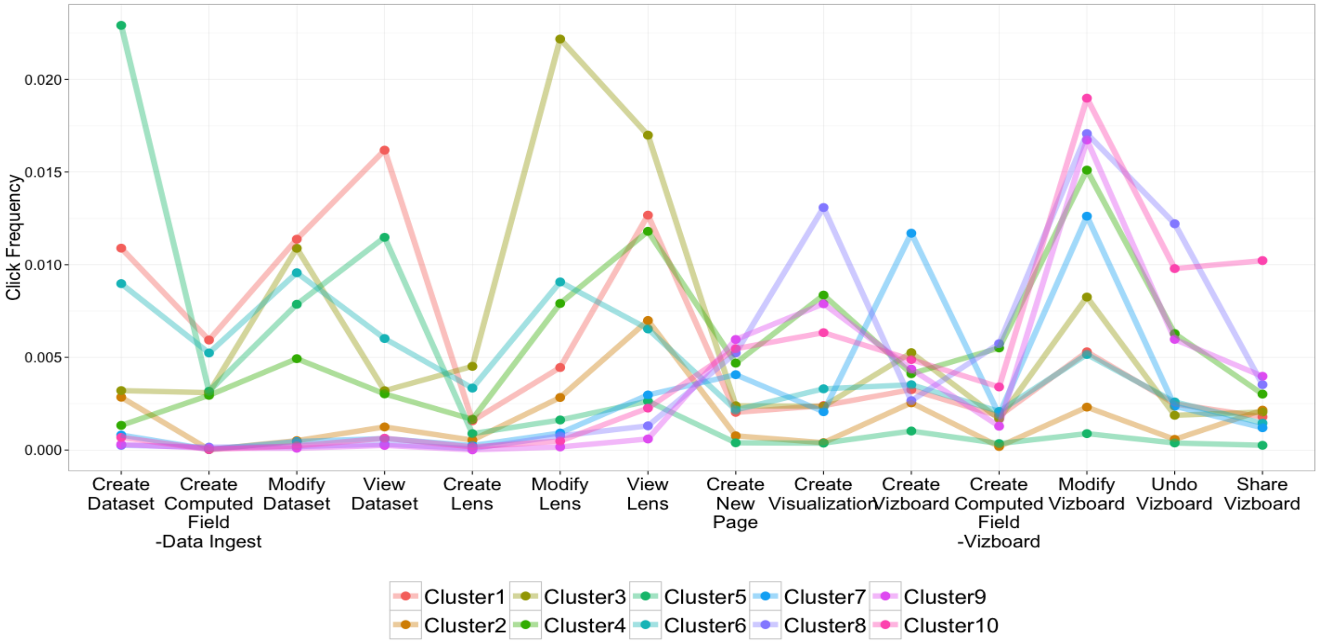

Figure 7: Frequency of 14 key steps in Platfora within each common workflow. You can see the difference in frequency of the key steps across clusters (groups of users) as shown by the different colored lines.

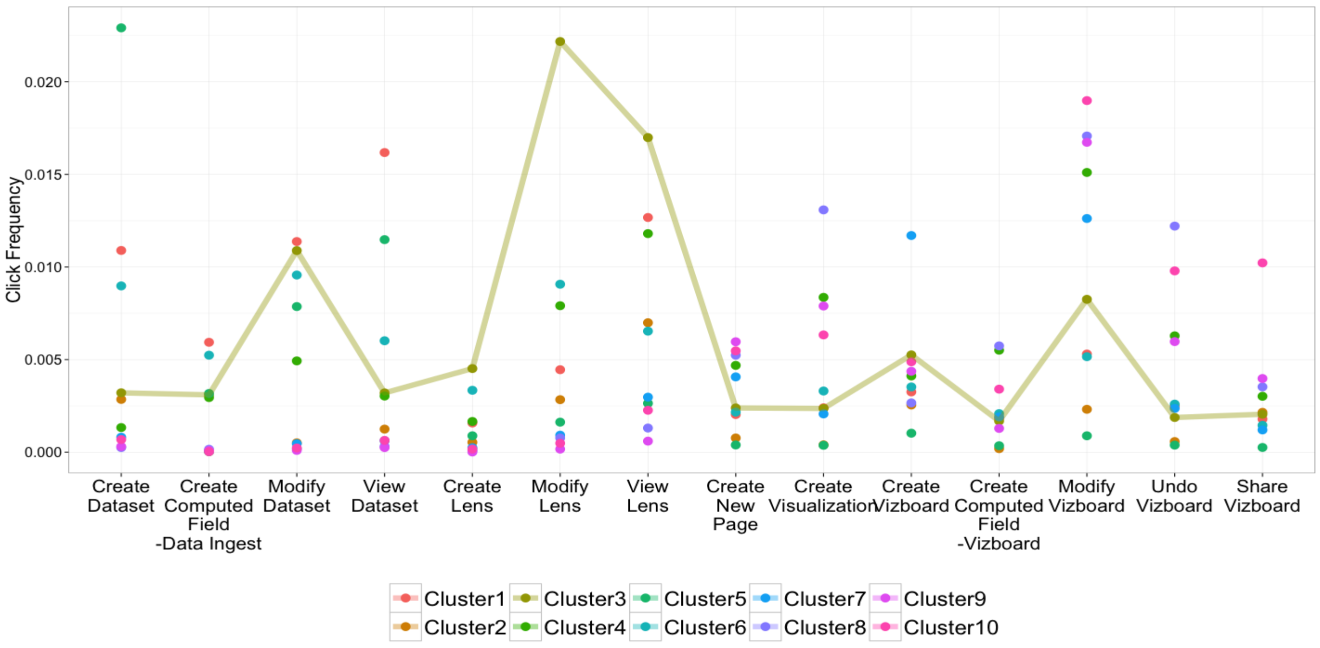

Figure 8: A workflow from Cluster 3: *Build lens with optional modification to a dataset*. Here you can see the high clickstream frequency for *Modify Dataset* and *Modify Lens*.

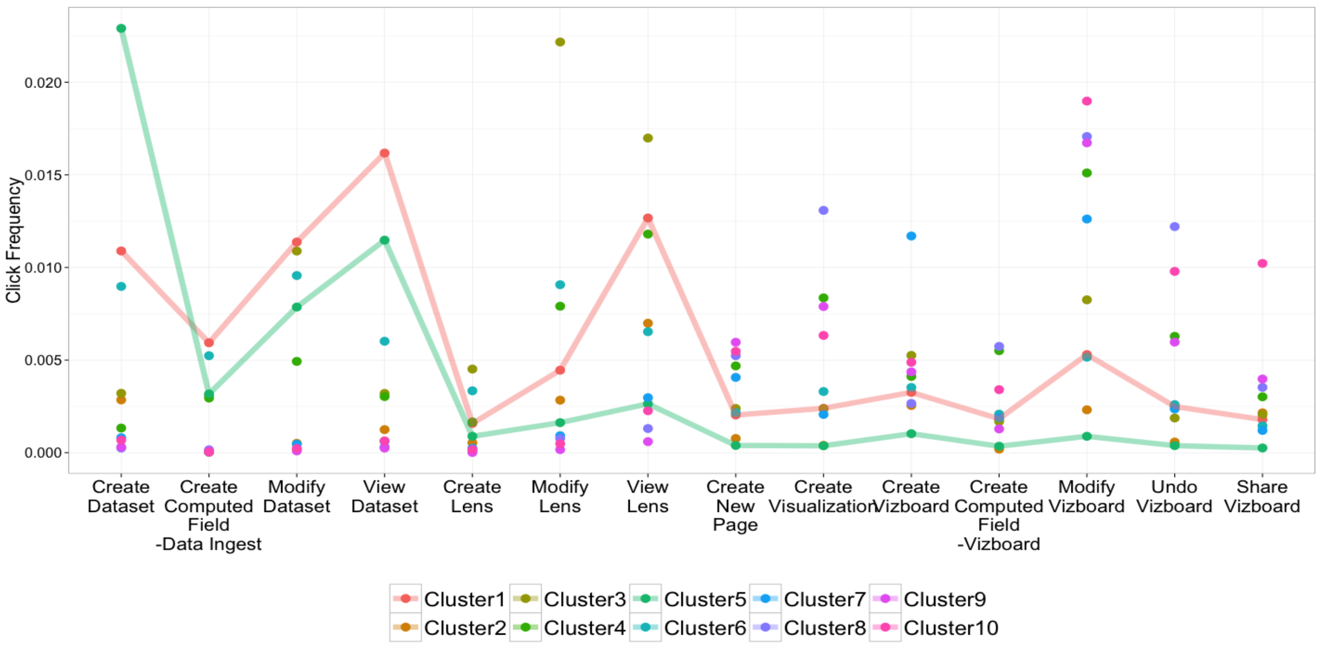

Figure 9: Frequency of dataset related activities in Cluster 1 (red line) and Cluster 5 (green line).

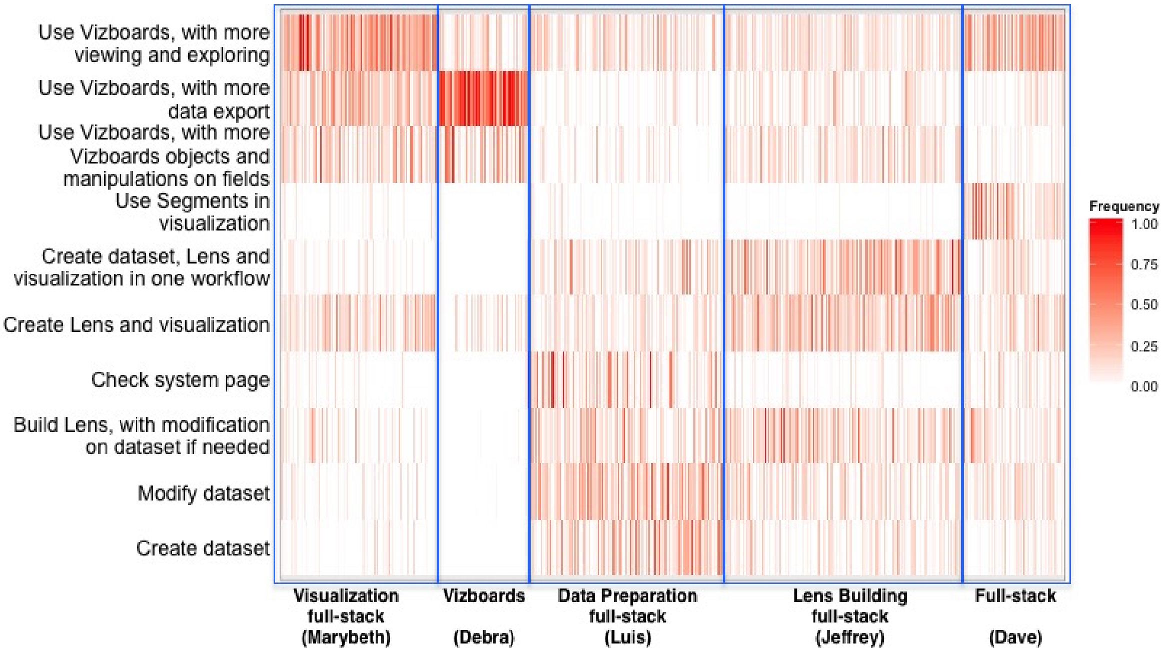

Figure 10: Heatmap of common workflows frequency for each user, ordered by personas. The block of users within the same persona share similar frequency of common workflows.

five personas. We thus harnessed raw clickstream data to identify common workflows that subsequently mapped to personas.

Each persona in our model has a unique probability to generate one of the ten common workflows; this probability is a quantitative signature of each persona. More specifically, these chances are the probability vectors that define the multinomial distribution of each persona. Figure 11 shows these chances of common workflows within each persona; we describe each persona below:

1. *Full-stack user specialized in data preparation (Luis)*: This red line that represents this persona in Figure 11 show that this persona has higher chances for two workflows related to Datasets and Lenses, and for viewing and exploring Vizboards. This persona has the high probability for workflows *Create dataset* and *Modify dataset*; these probabilities add up to 40 percent of this persona's total workflows. Note that these two workflows directly related to datasets appear with the highest chance within the same persona, even though the mixed model does not use the information that these two workflows are related. We believe this co-occurrence does not happen by chance but rather reflects one of the main characteristics of this persona, that is, this archetypal user is similar to a Data Administrator. This user is *Full-stack* with a focus on dataset preparation. By Full-stack, we mean a user who is likely to perform all the workflows related to datasets, lenses, and visualization in Platfora.

2. *Full-stack user specialized in Lens building (Jeffrey)*: This persona has the highest probability related to the following workflows: *Build Lens with optional modification to a dataset*, *Create Lens and visualization* and *Create dataset, Lens and visualization in one workflow*, all of which directly relate to a user's behavior of building a Lens in Platfora. These workflows pertain to actions on multiple objects from Datasets, Lenses and Visualizations within one workflow; these probabilities add up to 50 percent of the workflows within this persona. This persona is interesting as her workflows focus on several main product pages within the same workflow unlike other personas who tend to focus on a single object such a Dataset or a Vizboard.

3. *Full-stack user specialized in Segments (Dave)*: This persona has a 15 percent probability to generate the workflow *Use Segments in visualization*; the other personas have zero chance for the same. We therefore characterize this persona as a full-stack user who focuses on *Segments*; a Segment is an advanced Platfora feature that collects members of a population based on user defined attributes.

4. *Full-stack user specialized in Visualizations (Marybeth)*: This persona has a 70 percent chance that her clickstream is one of the three common workflows related to Vizboards with higher probabilities on leveraging Datasets and Lenses, as well.

5. *Pure Vizboards user (Debra)*: Unlike the previous personas, this persona's clickstreams has a 95 percent probability of representing just one of the three common workflows related to Vizboards. This is the only non Full-stack user among the five personas.

We have thus identified Platfora's personas, bottom-up, based on clustering clickstreams extracted from the product telemetry data.

6. DISCUSSION AND FUTURE WORK

In this paper we proposed a data-driven approach to create personas. Unlike approaches that relied on survey or user interview data to identify workflows, our approach identified workflows by directly incorporating user behavior via clicks and clickstreams. We aggregated these common workflows with a mixed model to

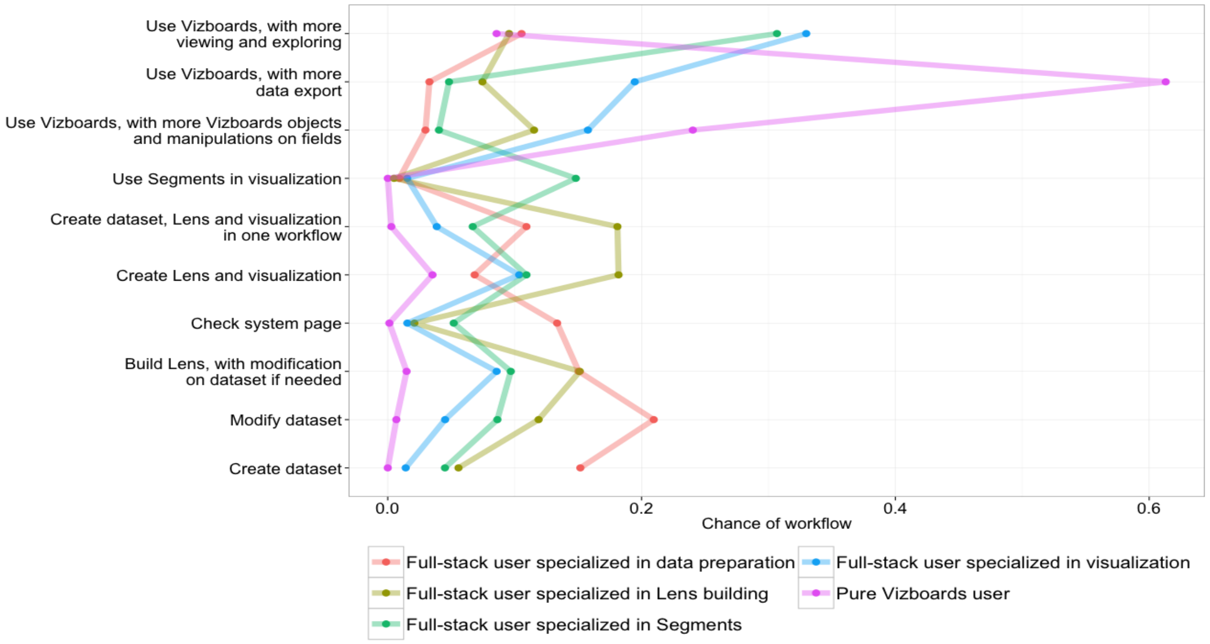

Figure 11: Each data-driven persona has unique behavior as shown by the unique frequency of common workflows as shown on the X Axis.

create data-driven personas. The five personas that we created were also validated by user behavior experts (via interviews) who verified that these personas indeed represented archetypal users of our product. Our approach is not only direct but also low cost as product telemetry is automatically gathered from the field; there is little interference from human interviewers or noisy self-reports.

Clickstream based persona creation based on product telemetry is scalable. Scalability is possible as additional telemetry data can help better understand existing persona characteristics (workflows) either by refining individual personas or by tracking the evolution of personas over time as users increasingly learn and utilize various product features. With 39,000 clickstreams from 2400 users, we also believe that our data is robust enough to generate five personas when compared to existing data-driven approaches that have only relied on a few hundred survey samples or self-reports.

We initiated this data-driven persona work to primarily refine original Platfora personas. The personas that we generated from clickstreams validated two of the existing Platfora personas while showing that the other two personas had evolved into *full-stack* users; we would not have garnered this crucial insight into our product's users without conducting a time-consuming costly persona reconstruction effort via traditional means. For example, we initially assumed that *Jeffrey*, a citizen data-scientist persona, to be *familiar* with datasets, *expert* at data curation and visualization workflows, with not much of an emphasis on sharing of insights via collaborative options (comments or emails). When we mapped existing *Jeffrey*'s to the personas discovered via clickstream clustering, we discovered that this persona was maturing over time in terms of using all four Platfora components eventually morphing into a full-stack user. We believe that it would be tedious and hard to capture such insights into the evolution of a persona over time with surveys or self-reports and that our approach based on clickstreams is a better representation of persona evolution (or maturation) over time.

You can extend our clickstream based approach to incorporate sequence and frequency information of clicks by considering different distance metrics. To use the sequence information, you can consider *n-gram* clicks as the smallest granular element in your current distance metric, where an n-gram click is *n* consecutive clicks in sequence. By considering consecutive clicks, the sequence information is retained as two clickstreams are considered to be similar if they share common clicks and if common clicks appear in the same order. To use the frequency information, you can define the distance between two clickstreams as the Euclidian distance between the two frequency vectors of clicks. With new distance metric, two clickstreams are similar if they share common clicks and the frequency of common clicks are close, so the frequency information is not dropped. The rest of our bottom-up procedure will be exactly the same. The work in this paper was motivated by the need to validate existing Cooper personas with clickstream information. Since calculating the similarity between clickstreams is only one step of our proposed technique, we find it reasonable to use our current simple, yet coarse, distance metric to verify the feasibility of our approach. Incorporating sequence and frequency information is definitely a promising avenue for future research.

We believe that our approach generalizes across applications that capture clickstreams that reflect user behavior. In this work we mined the clickstreams of Platfora users. However, neither the data is limited to clickstream information, nor the underlying system is limited to an analytics platform. Our approach would generalize if applications record a user's behavior in multiple sessions and if each session logs this user behavior as a sequence of actions. For example, our proposed approach would work for an e-commerce website. Here is how: as a user visits the website multiple times, this user is likely to explore a sequence of webpages. Our technique can capture the most common webpage paths that this user explores in the website and build a mixed model for related users based on the frequencies of their most common webpage paths. Thus our approach is domain neutral in all these three critical phases: calculating distance between two sequences of actions, performing hierarchical clustering, and constructing a mixed (user) model.

Our approach to create personas starting with clicks (the atoms of user behavior) can indeed be more rounded by existing approaches that rely on user interviews and surveys. Similar to other research techniques, a triangulation of methods that include data-driven clickstream approaches, surveys, interviews, user observation, and self-reports can ultimately create personas that best represent an archetypal product user.

7. ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments and suggestions that helped us better situate our work and improve the quality of our paper. We also thank the Product Team and the UX experts at Platfora who helped to validate the personas in multiple user interviews.

8. REFERENCES

[1] Jon Brickey, Steven Walczak, and Tony Burgess. 2010. A Comparative Analysis of Persona Clustering Methods.. In *AMCIS*. 217.

[2] Alessandro Canossa and Anders Drachen. 2009. Play-personas: behaviours and belief systems in user-centred game design. In *Human-Computer Interaction–INTERACT 2009*. Springer, 510–523.

[3] Alan Cooper and others. 1999. *The inmates are running the asylum:[Why high-tech products drive us crazy and how to restore the sanity]*. Vol. 261.

[4] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* (1977), 1–38.

[5] Anders Drachen, Alessandro Canossa, and Georgios N Yannakakis. 2009. Player modeling using self-organization in Tomb Raider: Underworld. In *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*. IEEE, 1–8.

[6] Jonathan Grudin and John Pruitt. 2002. Personas, participatory design and product development: An infrastructure for engagement. In *PDC*. 144–152.

[7] Homa Javahery, Alexander Deichman, Ahmed Seffah, and Thiruvengadam Radhakrishnan. 2007. Incorporating human experiences into the design process of a visualization tool: A case study from bioinformatics. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*. IEEE, 1517–1523.

[8] Lieve Hamers and Yves Hemeryck and Guido Herweyers and Marc Janssen and Hans Keters and Ronald Rousseau and André Vanhoutte. 1989. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing Management* 25, 3 (1989), 315 – 318.

[9] Tobias Mahlmann, Anders Drachen, Julian Togelius, Alessandro Canossa, and Georgios N Yannakakis. 2010. Predicting player behavior in Tomb Raider: Underworld. In *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on*. IEEE, 178–185.

[10] Jennifer Jen McGinn and Nalini Kotamraju. 2008. Data-driven persona development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1521–1524.

[11] Geoffrey McLachlan and David Peel. 2004. *Finite mixture models*. John Wiley & Sons.

[12] Tomasz Miaskiewicz, Tamara Sumner, and Kenneth A Kozar. 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*. ACM, 1501–1510.

[13] Steve Mulder and Ziv Yaar. 2006. *The user is always right: A practical guide to creating and using personas for the web*. New Riders.

[14] NASA. 1987. Telemetry: Summary of concept and rationale. *NASA STI/Recon Technical Report N* 89 (Dec. 1987), 13455.

[15] John Pruitt and Jonathan Grudin. 2003. Personas: practice and theory. In *Proceedings of the 2003 conference on Designing for user experiences*. ACM, 1–15.

[16] Rashmi Sinha. 2003. Persona development for information-rich domains. In *CHI'03 extended abstracts on Human factors in computing systems*. ACM, 830–831.

[17] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.