A Machine Learning Model for Understanding How Users Value Designs

Applications for Designers and Consumers

by

Jeremy Bilotti

Bachelor of Architecture
Cornell University, 2018

Submitted to the Department of Architecture and the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degrees

of

Master of Science in Architecture Studies and
Master of Science in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

June 2021

Signature of Author ....

Jeremy Bilotti
Department of Architecture
Department of Electrical Engineering and Computer Science
May 14, 2021

Certified by ....

Terry Knight
William and Emma Rogers Professor of Design and Computation
Thesis Advisor

Accepted by ....

Leslie K. Norford
Professor of Building Technology
Chair, Department Committee on Graduate Students

Accepted by ....

Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Thesis Supervisor

Terry Knight, PhD
William and Emma Rogers Professor of Design and Computation

and Reader

Stefanie Mueller, PhD
Assistant Professor of Electrical Engineering & Computer Science
Joint with MechE

A Machine Learning Model for Understanding How Users Value Designs

Applications for Designers and Consumers

by

Jeremy Bilotti

Submitted to the Department of Architecture and the Department of Electrical Engineering and Computer Science on May 14, 2021 in Partial Fulfillment of the Requirements for the Degrees

of

Master of Science in Architecture Studies and
Master of Science in Electrical Engineering and Computer Science

Keywords: Machine learning, probabilistic models, furniture design, function theory, user research

Abstract

In this thesis, I demonstrate a number of advances toward developing a machine learning (ML) model of how designs are valued by their users. The model can be used to better understand the implications of furniture design decisions, as well as for commercial strategy.

Existing ML systems have been trained on the physical and aesthetic features of completed furniture designs. We consider these methods to be "top-down" because designers and software engineers alone determine which features are considered important to the value of a design. To better capture the nuances of how users actually value the various functions of their furniture, I first develop a framework for ingesting and classifying user feedback. Next, I conduct a user survey to test this framework, generating a "bottom-up", labeled dataset from the feedback, requiring no post-processing. Finally, I develop methods for the computational analysis of this data. The analysis is based on a probabilistic ML model trained on the real user data collected. The model is trained to quantify how users value various features of furniture designs, beyond only physical and aesthetic features. I show how the model can augment existing datasets and produce data visualizations to inform design practice and commerce.

This framework represents a step toward a future in which data sets for furniture—and other design domains—are more accessible. By making user feedback available to designers at scale, and establishing methods for collecting this data, we can accelerate the development of designer intuition and deliver significantly greater value to more users.

Thesis Advisor: Terry Knight
Title: William and Emma Rogers Professor of Design and Computation

Thesis Reader: Stefanie Mueller
Title: Assistant Professor of Electrical Engineering & Computer Science, Joint with MechE

Table of Contents

1. Introduction

People value furniture designs in subjective, multifaceted ways. Understanding these nuances helps designers develop intuition for creating new designs. It also helps businesses make strategic decisions when developing, marketing and selling designed products. And yet, while an abundance of furniture designs already exist, our ability to access and leverage data about exactly how users value them is limited, even with the growing ubiquity of machine learning (ML) for data analysis. To work around this problem, some ML systems for classifying and recommending furniture designs have been developed, such as the product recommender systems used by Wayfair, a technology company in furniture and design e-commerce (Yusuf & Wayfair, 2019). However, these ML models rely on what we define as "top-down" representations of furniture designs: representations constrained by typology, style and other classifications which are assumed to be constant for all users. These representations are called "top-down" because they are constructed on the assumptions of designers, businesses and software developers, and not on user experience. In reality, users may value various furniture designs of similar typology and style quite differently. Each user's conceptual interpretation of "typology" and "style" may also be distinct. Can ML models for design be more useful if they are not built on "top-down" assumptions?

If a primary goal in furniture design and business is to continually improve the outcome of how users value designs, we need a "bottom-up" ML framework for collecting and interpreting user feedback about the value of those designs, rather than a "top-down" recommendation system. In this thesis, I present such a framework. It is built on the foundation of representing designs based on how they are valued, as reported directly by users. The framework helps designers and businesses to understand how users value designs, and includes methods for (1) designing data representations for furniture designs, (2) designing user surveys about the value of furniture designs, (3) distributing user surveys and (4) developing ML models to process and interpret survey data.

This thesis shows steps toward a future in which user feedback is more accessible to furniture designers, furniture businesses and others. I demonstrate how the framework's ML model can be used to show an overview of large data sets of user feedback, model the preferences of new users and augment existing data sets. The model's efficacy is evaluated by comparing results to a test data set separated from the original data set.

By adapting this framework to address other types of designed artefacts (products of the design process), other disciplines such as architecture or user interface design can also benefit. By integrating the framework shown in this thesis, businesses who offer products can reduce the risk of strategic decision making; assumptions about how customers may value product offerings will no longer rely on intuition, but will be informed by "bottom-up" data reported by users. If we adopt frameworks which help disciplines to better understand subjectivity in how users value designs, we have the potential to produce more value for a drastically greater number of users.

2. Background

2.1. Assumptions about Functionality and Use

The framework presented in this thesis relies on a number of key assumptions surrounding the use and valuation of artefacts (in the case of this research, the artefacts are furniture):

1. People use designed artefacts in different ways. As such, artefacts perform different functions for different people.

2. There is little distinction between "functional" and "non-functional" classifications when describing how people use artefacts (Crilly, 2010). For example, a dining chair may be used as a place to sit during a meal, but it may also be used to demonstrate personal aesthetic preference to dinner guests.

3. The value of an artefact is defined differently and individually by each user, based on the utility of that artefact to the user. We assume that individuals can identify their own preferences, and use them to build relationships between alternatives when making decisions (Fishburn, 1970).

The framework shown in this research builds on these assumptions to take advantage of the nuances of how artefacts are valued. To understand how, we must first unpack these assumptions and the work that others have done in relation to them. First, we acknowledge that theory surrounding the design, use, function and value of artefacts has been researched extensively, particularly in the context of the branch of philosophy called Function Theory. Research in Function Theory sometimes accepts that there is no singular definition for functionality, because of its intrinsically subjective nature. (Kroes & Meijers, 2006), (Crilly, 2010) and (Mahner & Bunge, 2001) are three examples which are aligned with this perspective. Contrasting views, as in Pahl and Beitz's "Engineering Design" (1977), argue that functionality can be understood objectively; through a systematic analysis of inputs and outputs, functions are intentionally designed into an artefact. In the context of this thesis, we understand this approach as "top-down", because it regards the process of design as originating and ending with designers (who may also be called engineers, among other titles). In a 2010 article, Nathan Crilly rebuts the top-down view by elaborating on the breadth of definitions of functionality which have historically been both accepted and challenged, establishing the importance of also considering the user's role. For example, it is clear that sofas, while intended for sitting on, are also used ubiquitously for sleeping on. The set of functions defined by a sofa's designer, therefore, is a subset of all possible functions of the sofa. We can further explore this superset of functions by accepting non-physical-spatial functions as valid, such as the role the sofa plays in its user's social interactions, or how it expresses its user's personal identity.

2.2. Classification Systems Which Include More User Perspectives

If a primary goal is to understand the nuanced value of all designs, we need some way to classify and analyze designs and their corresponding artefacts. By working toward accepting an unconstrained definition of functionality, we can let go of the notion of classifying artefacts based on "functional" and "non-functional" features (Crilly, 2010). Instead, we can begin to classify them based on their utility to users. (Bailey, 1994) elaborates on the subjectivity of classification, emphasizing that the quality of a classification relies on the ascertainment of key characteristics that are important to how the classification is used. Because characteristics which determine the value of furniture designs are different for every user (Fishburn, 1970), it is necessary to develop as broad a classification system as possible within an expandable framework if we want to understand most user perspectives.

Machine-learning-based classification systems have been used extensively to drive decision making in the context of business, commerce and even design. These algorithms learn and improve their performance as they are continually trained on more data (Dhall et al., 2019). Their ubiquity has allowed e-commerce and web-based services such as Amazon and Netflix to deliver relevant content to users based on user behavior and preference data (Amatriain, 2013). The set of data used to train a ML algorithm and improve its performance at making predictions is called a training data set. It is used to fit and continually refine parameters of a machine learning model, which define the decision making process of the algorithm (James, 2013). As a result, the capacity of a machine learning model to make accurate inferences, predictions and outputs is a direct result of the quality of its training data. In fact, we have seen that bias, inaccuracy and unrigorous data sourcing in training sets have significant effects on how machine learning algorithms perform in the real world (Turner Lee, 2018); stated simply, our algorithms are only as good as our data. As such, one objective of this thesis is to establish a more rigorous and less biased standard for data sets relating to design domains—and training data sets in particular. This thesis shows a framework for collecting data about furniture design, specifically.

Data collection, however, presents a significant challenge to the development of a machine learning model for understanding how users value designs. Moreover, data collection is often a bottleneck for the field of machine learning, in general (Roh et al., 2019), because large, labeled datasets rarely exist for newly developed applications of ML. In the case of furniture design, existing models have often been applied in the context of e-commerce, as the aforementioned Wayfair system has. For these models, data sets are constructed based on user purchasing and browsing behavior (Yusuf & Wayfair, 2019), which is indicative of a users' preconceptions of designs rather than their experience of designs. To take steps closer to a data set based on user experience, we can begin by collecting user feedback about designs.

2.3. Data Collection and Machine Learning for Furniture Designs

To obtain a high quality "bottom-up" training data set, it is necessary to allow user feedback to directly shape the data. Roh et al. (2019) provide an overview of the state of data collection methods, identifying "data generation" as the process of manually or automatically gathering data when no existing data is available. Roh et al. (2019) discuss crowdsourcing as a means of both gathering and preprocessing data, including labeling. The methods shown in this thesis employ crowdsourced data, which is both gathered from and preprocessed by users. Because nearly every person has interacted with a piece of furniture, we hypothesize that conducting a user survey is an effective way to crowdsource data about furniture designs. Some forms of machine learning, then, may be effective at analyzing patterns and trends in large data sets about furniture design, making learnings more interpretable and accessible (as long as learnings can be shown through visualizations). In other design domains such as architecture, it has been observed that designers develop significant intuition over time simply by interacting with clients and processing their feedback (Luck, 2007). Similarly, by expanding access to user feedback about furniture designs, we hypothesize that furniture designers and businesses can accelerate their development of design intuition to produce more value for end users.

State of the art machine learning models based on matrix factorization and collaborative filtering are typically built using a matrix of Users $\times$ Items, where each User-Item pair is associated with a boolean value of True or False (Adomavicius et al). Collaborative filtering refers to the use of matrix factorization to predict unknown preferences of a user based on the preferences of other users (Su, 2009). For example, Wayfair uses a matrix factorization system to train a machine learning model on user interest in furniture items for their e-commerce catalogue (Yusuf & Wayfair, 2019). Although this method is highly applicable (and demonstrated) in industry for recommending products to users, it does not allow for User-Item pairs to represent data more complex than binary values. In many cases, users value furniture designs based not on one factor, but rather on a combination of factors; often, a user may have both positive and negative opinions about different and unrelated features of a design.

A number of works exist which have taken steps toward machine learning models that reflect the nuance of user preference and utility. (Adomavicius et al.) clearly define the concept of multi-criteria decision making (MCDM) problems in machine learning, whose solutions "model a user's utility for an item as a vector of ratings along several criteria" (Adomavicius et al.). (Sahoo et al., 2012) demonstrate significant improvements to machine learning recommendation quality when they train a model using multi-component data (using the Yahoo!Movies data set). Similarly, (Kouadria et al., 2020) develop a multi-critera collaborative filtering recommender also trained on data from the Yahoo!Movies data set. The recommender generates a "top-N list" (Kouadria et al., 2020) of movie recommendations. Although this type of multi-criteria system is a step toward a more nuanced recommender, we note that the output of the system is a list of recommended items, and does not provide detailed information about the quality of users' preferences which could help designers and businesses to build intuition. Toward producing flexibility in machine learning recommenders, (Alkan and Daly, 2020) present a solution which incorporates time as a variable of influence on model parameters. This produces more accurate predictions for users, as the utility of items may change over time. In the context of design, (Iqbal et al., 2018) present a multimodal recommender that produces recommendations for matching assortments of furniture. While this is a step toward producing more value for users in the context of e-commerce and design, the theoretical foundation for the work is "top-down"; it produces recommendations only on the basis of "style and aesthetic preference". (Xing et al., 2020) make a fundamental contribution to research in this domain by demonstrating how machine learning models can be trained on data sets for which "[user] ratings of...design works are set as the ground-truth annotations" (Xing et al., 2020). Ultimately, their model is used to evaluate only aesthetic preference, but in a "bottom-up" way. If we want to leverage data to truly understand how users value furniture designs and build on these notable works, we need to design the structure of that data to reflect the nuance of user preferences. The first step toward leveraging data, however, is collecting it.

3. A Framework for Understanding How Users Value Designs

3.1. Methods for Designing Data Representations for Furniture Designs

To begin collecting user feedback on furniture designs, we must first understand how the feedback will be used. Because a primary objective of this thesis is to take steps toward a machine learning model of the value of designs, the user feedback will be used as input or training data in a machine learning model; this helps to define the structure of the data, and will inform the design of the user study to be conducted.

In this section, I first describe how I developed a basis for structuring data representations for furniture designs. I then describe the process of engineering specific features comprising the data representation.

One way of working toward capturing the nuance of how users value furniture designs is to define a data representation composed of a broad set of features. An example of data representation for a furniture design is an n-dimensional vector, each dimension of which represents one feature of that design. Each dimension has a magnitude representing something about its corresponding feature. In the case of this thesis, we will use an n-dimensional vector to represent a furniture design and all of its features. The magnitude of each dimension represents a user's valuation of that feature, and can be positive or negative depending on the user's preferences. A positive magnitude indicates that a feature adds value to the design, from the user's perspective. A negative magnitude indicates that a feature takes value away.

These features will be used as inputs in a user study to collect feedback and build a training data set. As a step toward developing a nuanced representation, I first generate a broad range of feature types which work towards expanding the definition of functionality of designed artefacts. An artefact, in this thesis, is defined as a physical object that results from the process of design. For some domains, the functions of an item can be evaluated using a binary value (e.g. "functional" or "non-functional", "good" or "bad", "0" or "1"). However for design domains, and specifically design domains which produce physical artefacts, functionality for users can be understood in many distinct ways; a designed artefact may serve many functions for a user which may not be readily visible to an observer, i.e. a design may serve a sentimental or emotional function which contributes significantly to the user's valuation of the design. This is why we consider an individual function of a furniture design to be a feature of that design, and so we refer to "functions" in this research synonymously with "features". To capture a broad range of possible functions, I structure training data based on six categories of functionality, as discussed in (Crilly, 2010): (1) Proper and system function (2) design, use and service functions (3) manifest and latent functions (4) physical and status functions (5) technical, social and ideological functions and (6) aesthetic and non-aesthetic functions. Figure 1 shows the taxonomy of these categories (top) and specific examples of functions in those categories which apply to furniture designs (bottom). The functional categories are defined below, with reference to function theory research discussed in section 3, and particularly (Crilly, 2010):

1. Proper and system functions relate to a user's perception of how a design will be used, and then how that design is actually used. We can understand the user's perception of function to be distinct from its actual use, because it may differently but significantly contribute to the user's valuation of the design.

2. Design, use and service functions reflect the variance in how an artefact's designer intended it to be used versus how it is actually used. This category also encompases functions that may have been unexpected by the designer, but also contribute to the design's value.

3. Manifest and latent functions are functions that are uniquely beneficial to each artefact's user, and are not readily apparent to observers.

4. Physical and status functions directly reflect the physicality of a designed artefact: its shape, form, material and construction. This category also accounts for functions of an artefact that result from its physical presence, specifically its relationship to the space around it; physical artefacts may often serve to define or mark spaces.

5. Technical, social and ideological functions have to do with an artefact's functionality during interactions. Technical functions have to do with physical touch. Social functions have to do with the artefact's role in social interactions. Ideological functions have to do with an artefact's cultural or personal significance to the user.

6. Aesthetic and non-aesthetic functions are functions that have to do with the way an artefact looks, or the ideas relating to the artefact's design, respectively.

It should be acknowledged that this list of categories is not exhaustive, because a truly comprehensive representation of functionality will consider all facets of a user's relationship with an artefact. Instead, this thesis works to develop a definition of functionality which is broader than the definition typically used when building machine learning systems for functional designs. This thesis provides a framework which can be expanded to include other categories of functionality, but begins with the aforementioned six categories.

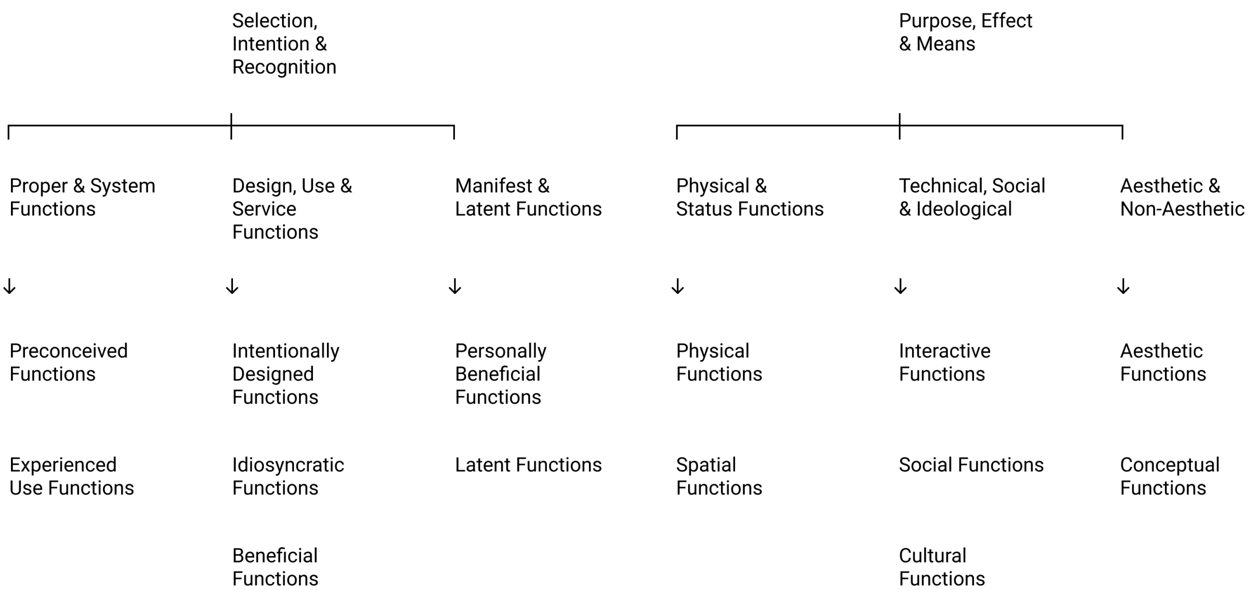

Figure 1: Taxonomy showing categories of functionality (top) and specific functions of furniture designs (bottom).

3.2. Results of Feature Engineering

Each of the six categories describes a different way to evaluate the functionality of an artefact. However, we must further break down these categories to develop a more specific data representation for furniture designs, identifying real features of furniture designs that are encompassed by each category. The reason for further breaking down these categories is so survey respondents can easily evaluate furniture designs they have experienced; the existing six categories are too broad to be easily intelligible both by respondents and those analyzing responses. The features identified in this thesis are described below, listed under their parent categories. Each one refers to a feature of a furniture design which can contribute to a user's valuation of that design:

1. Proper and system functions of furniture designs
    a. Preconceived functions: functions of a furniture design which a user anticipated before owning or using the design.
    b. Experienced use functions: functions of a furniture design which result from its unique context, during its actual use.
2. Design, use and service functions of furniture designs
    a. Intentionally designed functions: the functions of a furniture design that were intended by its designer.
    b. Idiosyncratic functions: the functions of a furniture design that were not intended by its designer.
    c. Beneficial functions: the functions of a furniture design which lead to unintended benefits or positive outcomes.
3. Manifest and latent functions of furniture designs
    a. Personally beneficial functions: the functions of a furniture design which lead to benefits or positive outcomes that are unique and personal to each user.
    b. Latent functions: functions of a furniture design which users are unaware of.
4. Physical and status functions of furniture designs
    a. Physical functions: the size, shape, material and construction of a furniture design.
    b. Spatial functions: the function of a furniture design as a marker or symbol of space, e.g. a table which divides a room.
5. Technical, social and ideological functions of furniture designs
    a. Interactive functions: the ergonomics of the furniture design when a user physically interacts with it.
    b. Social functions: the role of the furniture design in social relationships, and its capacity to affect interpersonal and social dynamics.
    c. Cultural functions: the furniture design's significance to a user's personal identity or culture.

6. Aesthetic and non-aesthetic functions of furniture designs
    a. Aesthetic functions: the way the furniture design looks.
    b. Conceptual functions: the ideas, symbolism, stories or concepts that relate to the furniture design.

Now that we have established a broad range of possible functions that a furniture design may have, we can evaluate whether each one adds or takes away value, from a user's perspective. A straightforward method for evaluating features is to conduct a user survey. The design of the survey is important, because it determines the structure of the survey's resulting data. Survey design also affects respondents' ability to complete a survey without fatigue, thus affecting the quality and accuracy of responses. By distributing bespoke user surveys in this thesis, I show how a survey can also be a tool for feature engineering, offering information on the relevance and comprehensibility of each feature identified above.

3.3. Methods for Designing User Surveys About the Value of Furniture Designs

In this section, I describe the process of designing user surveys based on the data representation described in the previous section. The goal of the user survey is to quantify how users value various features of a furniture design. The output of the survey is a training data set which does not require data cleaning in order to be used for machine learning. It is generated in a bottom-up way, directly from user feedback.

To generate a ready-to-use training data set for machine learning from a user survey responses, we need a user survey which effectively allows respondents to create a data representation of a furniture design simply by completing a survey. User research best practices suggest that collecting speculative data is less useful than collecting known data, because bias may be introduced to the respondent if they are asked to speculate (Lewis, 2006). As such, the survey asks respondents to answer questions about a furniture design artefact they own or have already used. This bottom-up method is an important improvement on state-of-the-art methods for classifying furniture designs, which often use image recognition or simple "style"-based classification. Both of these methods are top-down because they classify designs from the perspective of the machine learning engineer rather than allowing classifications to emerge directly from user feedback.

Figure 2: High level structure, or flow, of the user survey on the value of furniture designs.

The high level survey flow, or the procession of a user through the survey, is shown in Figure 2. First, respondents optionally answer questions about themselves, including demographic questions, in the “User Input” section. Although users input information about themselves, the data cannot be used to identify any respondent, and the survey is completely anonymous. Next, respondents input a furniture design that they have owned or previously experienced, in the “Item Input” section. The next section, “Feature Classification” is where users classify the furniture design by indicating which features add or subtract value from the design. The last section, “Feature Weighting”, allows users to quantify each feature’s influence on their perceived valuation of the design.

Four detailed versions of the survey flow are shown in Figures 3, 4, 5 and 6. The process of designing the survey involved creating iterations, evaluating those iterations for complexity and potential to generate user fatigue, and then testing the least complex iteration with a small sample of respondents. Research from (Aguinis et al., 2019) was also referenced throughout the survey design process. The design iterations are described below.

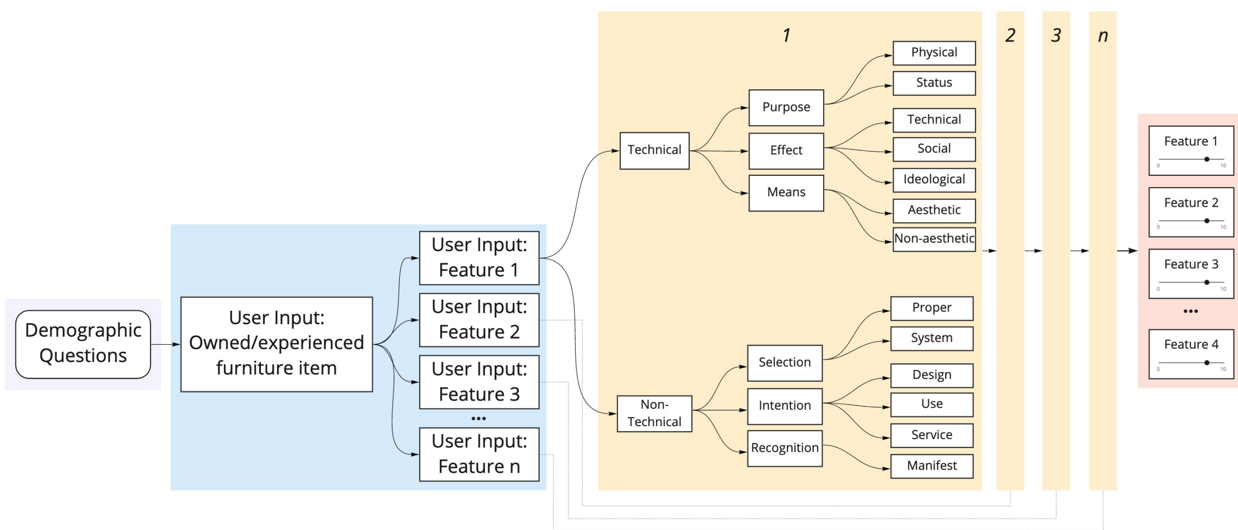

Figure 3: Survey Flow A, in which users generate features and categorize them using drill-down questions.

3.4. Flow A: Drill Down

Users first answer demographic questions about themselves. (Amaya, 2020) provides valuable best practices for asking users about their personal identity. Next, users input a furniture design artefact they have owned or experienced. Next, users use open-ended text inputs to identify features of the design. Next, each feature is classified under a category described in section 3.1. by the user, using drill-down selection menus (multi-step menus which allow the user to select categories and sub-categories). Finally, features are weighted (given magnitude) based on their influence on the user's valuation of the design, on a scale from 0 to 10 points. 10 indicates the strongest influence, and 0 indicates no influence.

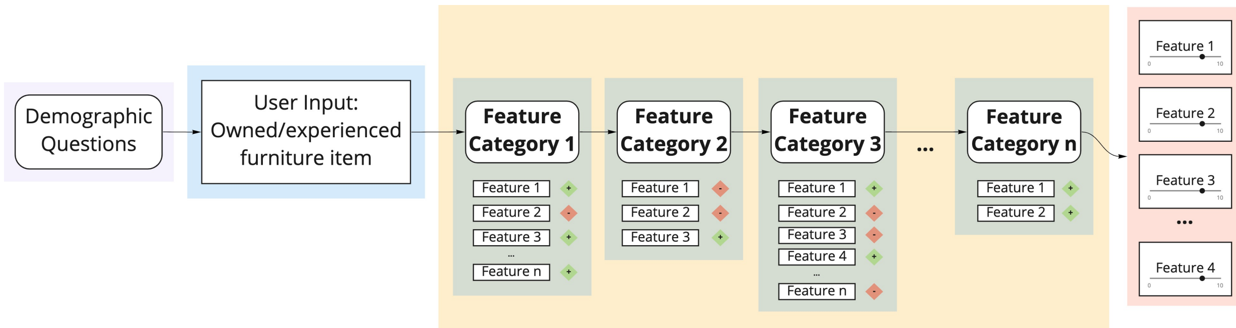

Figure 4: Survey Flow B, in which users generate features which they then sort and rank.

3.5. Flow B: Sort and Rank

Users first answer demographic questions about themselves. Then, users input a furniture design artefact they have owned or experienced. Next, users are sequentially presented with descriptions of each of the six categories of functionality described in 3.1. After each category description, users use open-ended text inputs to identify features of the furniture design which are encompassed by the category that has just been presented. For each feature, users indicate "+" or "-", identifying whether the feature adds value to the design or takes value away. After users input features for each of the six categories from 3.1., they indicate a magnitude for each feature based on its influence on the user's valuation of the design. Users choose a magnitude on a scale from 0 to 10 points. 10 indicates the strongest influence, and 0 indicates no influence.

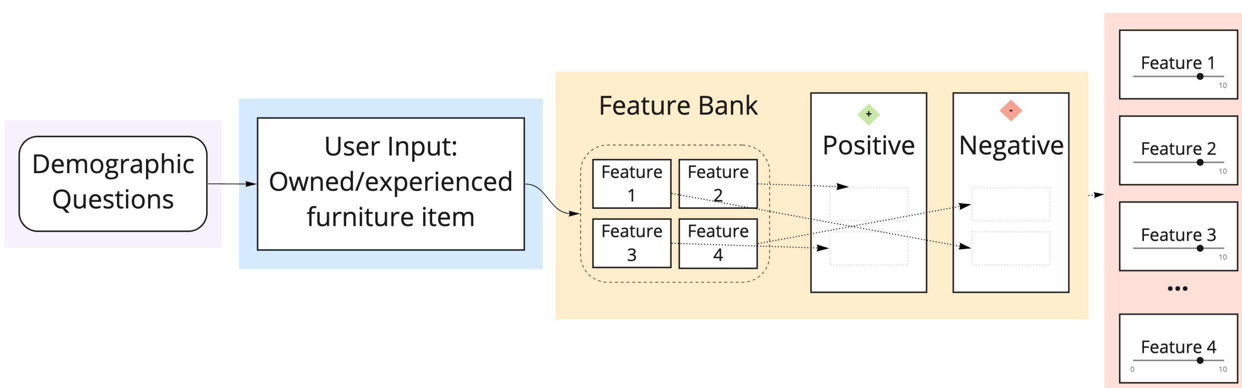

Figure 5: Survey Flow C, in which users classify features listed in a word bank.

3.6. Flow C: Feature Bank

Users first answer demographic questions about themselves. Then, users input a furniture design artefact they have owned or experienced. Next, a list of features specific to furniture design, as described in 3.1., is presented to users in the form of a “feature bank” (a collection of features, written as words, which can be used as responses). Users then sort the features into two categories: features which add value to the design and features which take value away. Users are not required to sort every feature; some may not be applicable. After users sort features from the feature bank, they indicate a magnitude for each feature based on its influence on the user’s valuation of the design. Users choose a magnitude on a scale from 0 to 10 points. 10 indicates the strongest influence, and 0 indicates no influence.

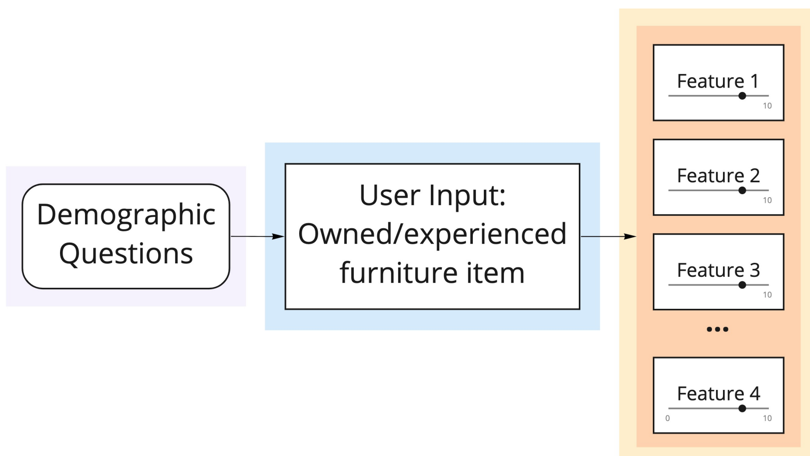

Figure 6: Survey Flow D, in which Feature Classification and Weighting are combined into one step using slider inputs.

3.7. Flow D: Feature Categories Only

Users first answer demographic questions about themselves. Then, users input a furniture design artefact they have owned or experienced. Next, users are presented with descriptions of features specific to furniture design, as described in 3.1. Users are asked to indicate an integer magnitude on a scale from -5 to 5 for each feature, representing its influence on the user’s valuation of the design. -5 indicates that a feature significantly takes value away from the design. 0 indicates that a feature does not add or take away value. 5 indicates that a feature significantly adds value to the design. Users have the option to indicate “Not Applicable” for any feature.

3.8. Survey Complexity Analysis

A quantitative analysis of each survey design is conducted to evaluate complexity. Table 1 shows the evaluation for survey designs A, B, C and D. If we assume t hat respondent fatigue increase linearly with survey complexity, we can make an informed determination as to which survey is likely to yield high completion rates and high quality responses. We define survey complexity as a function of the number of user input questions ($\mathrm{n_u}$), the number of item input questions ($\mathrm{n_i}$), the number of features in the survey ($\mathrm{n_f}$), and the number of features which users indicate as "not applicable" ($\mathrm{n_{f_{N/A}}}$). We define two constants: constant NT to represent a unit of complexity of a "non-text input" survey question (i.e. multiple choice, drop-down selection, etc.), and constant T to represent a unit of complexity of a "text input" survey question (i.e. an open-ended question for which respondents type text). Complexity is calculated as a summation of the complexities of all of the survey sections. Because text input questions cause considerably greater respondent fatigue than non-text input questions (Crawford et al., 2001), T > NT. In most cases:

$$\mathrm{f_{N/A}} \subset \mathrm{f} \;\rightarrow\; \mathrm{n_{f_{N/A}}} < \mathrm{n_f}$$

In the case that a respondent indicates every feature is N/A:

$$\mathrm{f_{N/A}} \subseteq \mathrm{f} \;\rightarrow\; \mathrm{n_{f_{N/A}}} = \mathrm{n_f}$$

Survey D's complexity, in that case, is computed as:

$$\begin{aligned}
&\mathrm{NT}(\mathrm{n_u} + \mathrm{n_i} + \mathrm{n_f} + \mathrm{n_{f_{N/A}}}) + \mathrm{T} \\
&= \mathrm{NT}(\mathrm{n_u} + \mathrm{n_i} + \mathrm{n_f} + \mathrm{n_f}) + \mathrm{T} \\
&= \mathrm{NT}(\mathrm{n_u} + \mathrm{n_i} + 2\mathrm{n_f}) + \mathrm{T}
\end{aligned}$$

Survey flow D achieves the smallest complexity in all cases but the above case, assuming the aforementioned variables and relationships. In the case above, survey D's complexity is equivalent to survey C's. By combining the feature classification and feature weighting steps into one, and significantly reducing the number of text inputs, we determine survey D least likely to cause respondent fatigue, and select it to be used in this research. But, it is important to note that there is a non-trivial tradeoff between reducing text inputs and quality of responses; surveys A and B, for example, allow users to define their own features, which has the potential to produce higher quality results. Surveys C and D pre-define a set of possible features of furniture designs (as in 3.1.), but allow respondents to indicate if a feature is not applicable. To achieve a true, bottom-up classification of furniture designs, users would ideally generate bespoke features using text inputs. However, once distributed, high amounts of user fatigue were observed to considerably reduce the survey's completion rates and response quality, even when using a long version of the simplest design, survey D (see section 3.7.), and so options A and B are determined to be infeasible to distribute. I describe future work which can lead to an even more optimal survey design in section 6.

	User Input		Item Input		Feature Classification		Feature Weighting		Survey Complexity
	Non-text input (NT)	Text input (T)	Non-text input (NT)	Text input (T)	Non-text input (NT)	Text input (T)	Non-text input (NT)	Text input (T)	
A	n_u	0	n_i	$1 + n_f$	$3 \times n_f$	0	n_f	0	$NT(n_u + n_i + 4n_f) + T(n_f + 1)$
B	n_u	0	n_i	1	n_f	n_f	n_f	0	$NT(n_u + n_i + 2n_f) + T(n_f + 1)$
C	n_u	0	n_i	1	n_f	0	n_f	0	$NT(n_u + n_i + 2n_f) + T$
D	n_u	0	n_i	1	$n_{f_{N/A}}$	0	n_f	0	$NT(n_u + n_i + n_f + n_{f_{N/A}}) + T$

Table 1: Quantitative analysis of survey complexity.

3.9. Survey Design Results

3.9.1. Survey Questions

To prompt users to respond in the "user input", "item input", "feature classification" and "feature weighting" steps, we need a set of survey questions. Respondents are not required to give a response for any single question to complete the survey, except for an initial disclaimer in which the respondent acknowledges their voluntary participation in—and option to cease responding at any time to—the research study. If a respondent does not acknowledge the disclaimer, they may not participate. Respondents under the age of 18 also may not participate.

For the user input step, respondents are asked to select their age, personal income, gender identity, racial identity and location (by zip code, city, state or country). Users are also asked a series of questions which provide insight into their exposure to the discipline of furniture design. For example, users are asked where they get information about furniture and design, and are given the option to select numerous media sources, books, museums, or input their own source.

In the item input step, users are prompted to identify a piece of furniture that they picked out, have owned or have lived with. Users identify the furniture design by selecting a typological category (i.e. seating, tables, lighting, etc.) and then a specific type (i.e. bench, dining chair, desk chair, etc.). Users also indicate the brand or manufacturer and designer of the furniture, if known. Although this data can be used to classify furniture designs, we are interested in generating a classification based on how users value designs—not based on typology. So, we collect this information for future reference but do not use it as training data in the development of a machine learning algorithm.

In the combined feature classification and weighting step, users are presented with a list of features that may apply to their furniture, as in 3.1. Users select "not applicable" if a feature does not apply. If it does apply, users indicate a quantitative weight for the feature on a scale of -5 to 5 by sliding a slider.

Physical Qualities - Your furniture's physical shape or form.
Materials - The materials used to make my furniture.
Physical Qualities - The quality of its construction.
Physical Qualities - The furniture's size.

Interactive - The physical comfort of interacting with your furniture.
Interactive - The way it fits your body.
Interactive - Its weight.
Interactive - Its temperature when touched.
Interactive - How often you use it.

Advantages - How others use and react to my furniture.
Advantages - It impresses others.
Advantages - It makes others feel comfortable.
Advantages - It generates income in a business setting.
Advantages - It encourages relaxation.
Advantages - It encourages productivity.
Advantages - It generates discussion.
Advantages - It sparks passion, interest or contributes to a hobby.
Advantages - It teaches something.

Figure 7: Selected survey questions generated and tested with users. See appendix for full list.

3.9.2. Survey Interface

The user survey was built using Qualtrics XM, a web-based software for creating and distributing surveys. The survey does not include any images, because respondents are asked to answer based on a furniture design they already own or have experienced. The survey is completed in a web browser, accessible via an anonymous link. Respondents are programmatically prevented from responding to the survey more than once, however this process does not allow respondents to be individually identified. Respondents proceed through the survey by answering one or more questions, and navigating to the next step using a "next" button. Respondents may also move backward through the survey by using a "back" button. By separating questions into multiple steps, user fatigue is significantly reduced as users are able to better perceive their progress. To further improve this perception, a progress bar is displayed during the survey, indicating at all times the remaining portion of questions to be completed. The survey's interface is designed using a 12 to 14 point sans-serif font on a white or grey background for accessible viewing. When a user selects an answer, that answer's background changes to a high-contrast, dark color to avoid confusion about selection. The survey interface is responsive to browser size; its graphic layout and dimensions change as a function of browser size or device. The survey interface is optimized for viewing on a mobile device, and prevents text from being obscured or occluded by page elements. Button positions are adjusted for physical accessibility on mobile devices. Improvements to the accessibility of the interface include further considerations for respondents who have visual impairment or differences in physical ability, such as interface color options, size options and text-to-speech.

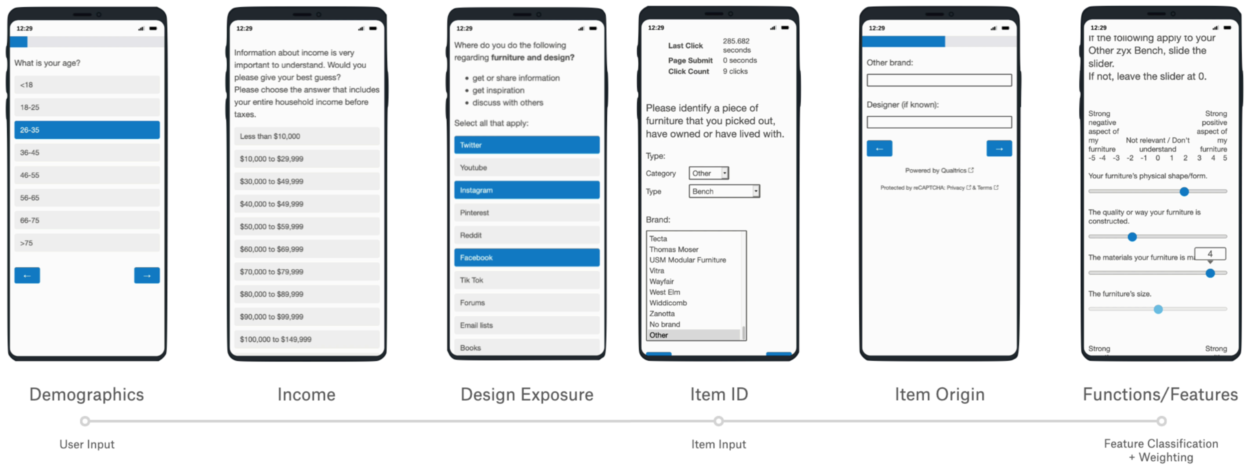

Figure 8: Mobile survey interface and flow, from user input to feature weighting.

3.10. Methods for Distributing User Surveys About Furniture Designs

3.10.1. Survey Sample Population

In this section, I show a process for distributing user surveys about furniture designs. I describe the specifications for the population samples in this thesis, and discuss how the design of the survey affects its completion by users. I also show methods for evaluating the relevance of the data representation of a furniture design described in 3.1., based on feedback from user survey respondents.

The user survey in this thesis focuses on users in the United States. The methods for data collection should be viewed as part of the framework, which is adaptable to other populations but may require modifications or further developments to produce the same results.

The survey in this thesis was completed by respondents in three rounds: The first round was conducted with a sample of five graduate students in Cambridge, Massachusetts (p=5) studying design, technology and business. The second round was conducted with an anonymous general United States population sample of 17 people (p=17). The third round was conducted with an anonymous general United States population sample of 27 people (p=27). In total, there were 49 responses from people in the United States (p=49). In the second and third rounds, respondents were each offered entry into a raffle for a pendant lamp as an incentive (one entry per respondent). Because the population sampled in this research is limited in size, the results of the work should be understood, again, as part of a framework which can be further developed to better represent general populations. With a larger population sample size, or with a sample from other locations, the resulting data will more accurately reflect the views of those populations, including the valuation of furniture designs. This thesis serves to show a process for effectively collecting survey data about the value of designs, which can be immediately used as a training data set for machine learning.

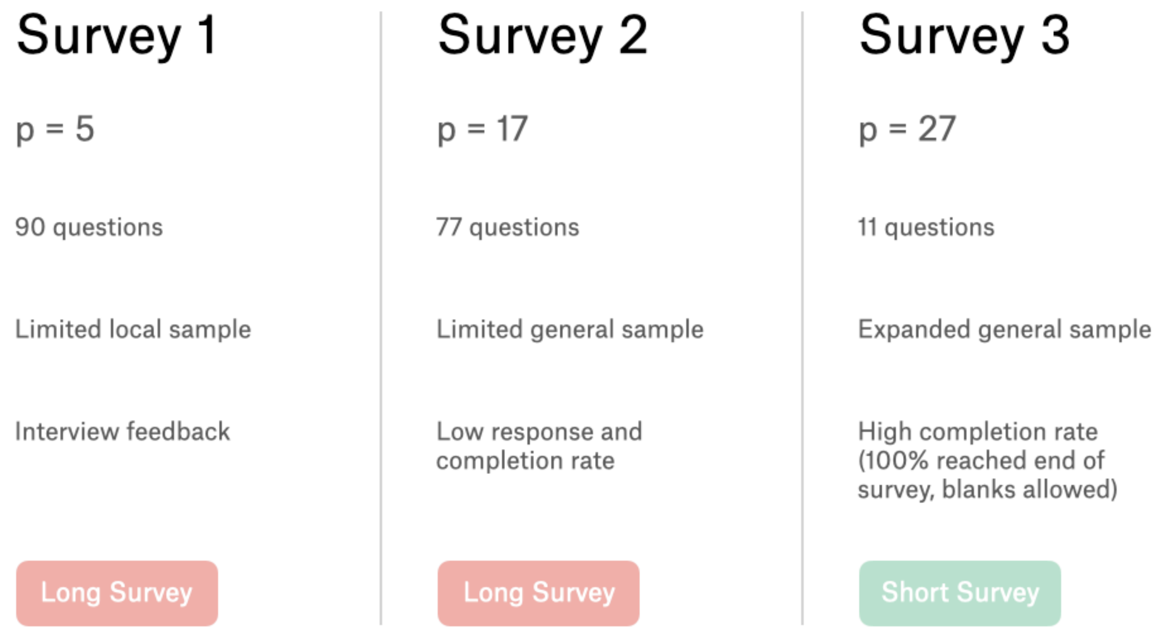

Figure 9: Evolution of user survey throughout distribution, testing and soliciting respondent feedback.

3.10.2. Survey Round 1 Methods and Feedback

The first round of survey distribution was limited to a small sample size (p=5) such that interviews could be conducted after completion. The survey used design D from section 3.7., but weighting and categorization were initially separated into two steps. The survey included a total of 90 questions. Respondents, after completing the survey, were asked to provide feedback about the survey's content, ease of completion, comprehensibility and suggested improvements. Interviews with respondents yielded the following notes:

- Features relating to physical properties were particularly relevant.
- Some features were difficult to understand.
- Completion time was too long; 17 minute completion average. This was a significant issue that every respondent mentioned.
- Demographic question wording could be improved.
- Weighting and categorizing separately was arduous.
- The mobile user interface was broken in some views.

- Some features were redundant or incomprehensible. Latent functions and preconceived functions were difficult to understand and did not receive responses. Intentionally designed functions were understandable, but unknown to respondents (users were not aware of the intentions of furniture designers). Experienced use functions and idiosyncratic functions were perceived as redundant.
- Slider behavior was inefficient.
- A "not applicable" option was useful.
- Slider scale of -5 to 5 was confusing and increased response time.
- Default position of sliders increased response time and needed adjustment.
- Sliders were effective on desktop but less so on mobile, due to aforementioned problems.

3.10.3. Survey Round 2 Methods and Outcomes

The second round of distribution was to a random, anonymous, general United States population with sample size p=17. Survey responses were solicited through social media advertisements, the placement of which was randomized without preference to any particular population within the United States. Survey design D was still used for this round, but a number of changes were made based on interview feedback from round 1. Word count was reduced by 25-50% on every feature question. The behavior and functionality of the sliders was improved on desktop and mobile. The scale of the sliders was simplified from [-5:5] to [-3:3]. The feedback from round 1 also allowed for further feature engineering. Questions about preconceived and latent functions were omitted because they received few responses and were difficult to understand. Intentionally designed functions were omitted because users were generally not aware of the intent of furniture designers and were unable to answer these questions. Experienced use functions and idiosyncratic functions were merged, and some redundant questions from these categories were omitted (Figure 10). The number of questions was reduced to 77. Even with these changes, the completion rate of the survey was low (no user was able to complete the survey in its entirety, and significant drop-off was observed during the feature classification and weighting section).

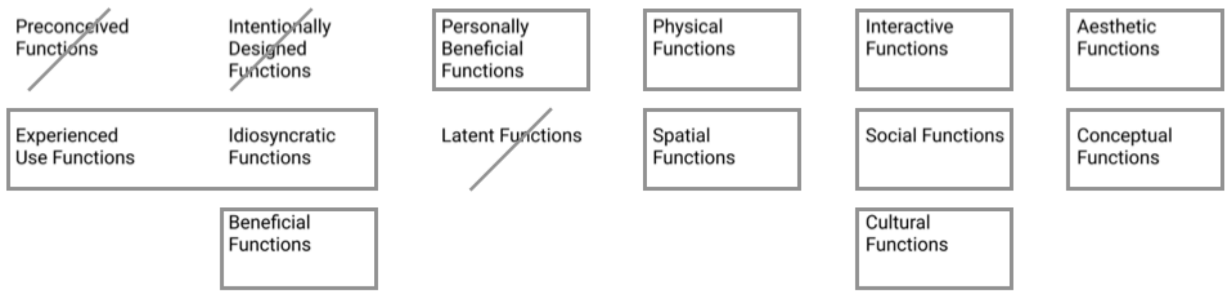

Figure 10: Feature engineering based on survey results and respondent behavior.

3.10.4. Survey Round 3 Methods and Outcomes

The second round of distribution was to a random, anonymous, general, United States population. Survey responses were solicited through social media advertisements as in round 2. In round 3, the design of the survey was maintained, but the 77 questions were reduced to 11. Each question was representative of a category shown in Figure 1. The physical functions category was allocated two questions, because all respondents in round 1 noted it as the most relevant to them. A larger number of users (p=27) responded to the survey with these modifications, and all respondents reached the end of the survey. Less drop-off was observed during the feature classification and weighting section.

3.11. Developing a Machine Learning Model for Understanding How Users Value Designs

I describe methods, in this section, for using bottom-up data to train a probabilistic machine learning model. I first explicate the model's bottom-up training data set, which is a result of the user survey in 3.7. I then describe how the framework in this thesis builds in checks for bias in the data set. Next, I describe the probabilistic model. Then, I describe the procedure for learning probabilistic distributions on user valuations of furniture designs based on the training data (training the model). Finally, I explain a procedure for augmenting datasets using the model's learned probability distributions. This procedure can be adapted to make predictions about new user preferences, for example, in the context of design practice or commerce.

3.11.1. A Bottom-Up Training Data Set

Working toward the goal of establishing more rigorous standards for data sets about design, this thesis demonstrates the use of a data set output directly from user surveys. The user survey described in previous sections was designed to produce detailed representations of furniture designs by asking respondents to indicate which features add or subtract value from the design, and how significantly. The output training data set is structured as a two-dimensional array of

Furniture Designs × Design Features. Each element of the array is quantified with an integer value between -3 and 3, corresponding to how one user values a feature of a furniture design. One column of the array represents all surveyed users' valuations of a single design feature, as it applies to each user's chosen design:

$$\begin{array}{r c}
 & \text{Feature 0} \\
\begin{array}{r} \text{Design } x \\ \text{Design } y \\ \text{Design } z \\ \cdots \\ \text{Design } n \end{array} & \begin{bmatrix} -3 \\ 1 \\ -2 \\ \cdots \\ 2 \end{bmatrix}
\end{array}$$

A single row of the array is equivalent to an n-component vector representation of one user's complete valuation of a furniture design:

$$\begin{array}{r c}
 & \text{Feature 0 ... Feature } n \\
\text{Design } x & [\,-3, -2, 0, -3, 1, 3, 2, 3 \ldots 1]
\end{array}$$

3.11.2. Bias Checks and Equitable Data Science

If we acknowledge that ignoring differences between groups of people produces bias and inequity, the data we have collected about respondents and their furniture typologies can serve as a series of checks that allow us to assess bias present in our data. Groups of rows can be identified by corresponding user data, such as demographics. Similarly, groups of columns can be identified by furniture typology or brand. It is therefore possible to produce useful statistics for groups of users and groups of furniture typologies, such as group row and column averages. We can also observe differences and similarities between groups of user who have had exposure to the discipline of design and those who have not. Typically, designers produce furniture designs without first conducting a study of how their design may or may not be biased, or how it could be more equitable for users. Design data sets which do not ignore demographic information, but rather make it more accessible, provide an opportunity for designers and design retailers to view users' self-reported feedback while considering how their work may serve groups differently.

Data sets which exclude demographic information do not offer this possibility, obscuring biases that are inevitably present in data. While three data types for building in bias checks are presented in this thesis (user demographics, furniture typology/origin and exposure to design discipline),it is possible to expand the framework to include others. Inclusion of these checks should not be a substitute for fair and unbiased population sampling when distributing surveys, however. Although the data set used in this thesis should be viewed as a small preliminary sample not representative of larger populations, but rather as an example in a framework, a breakdown of a demographic data is shown in Figures A1, A2, A3, A4 and A5.

3.12. A Probabilistic Machine Learning Model

3.12.1. Building Probabilistic Models

Beyond simply providing a framework for data collection, this research demonstrates how a simple probabilistic machine learning model can be built to process user feedback. The complete data array described at the beginning of 3.11.1. serves as the model's training set. A test set, or a portion of the training data, is reserved (not included in the training set) so it can be used to evaluate the model's efficacy after training.

The framework presented here should be understood to potentially work with any type of machine learning model, including neural networks, Bayes networks, and others—although some models (such as collaborative filtering-based models discussed in 2.3.) may be more poorly suited to the furniture design domain than others. Because we would like to condition predictions and outputs on existing data from our survey, it makes sense to start by developing a probabilistic model.

In the case of probabilistic models, training data provides a basis for learning probability distributions. A probability distribution is a function which represents the likelihood of possible events given certain conditions (Everitt, 2006). In the case of the framework in this thesis, the existing conditions are the features of furniture designs, and the events are the valuation of those features by users. This means that our goal is for the model, when trained, is to learn the likelihood that users will rate each feature a -3, -2, -1, 0, 1, 2 or a 3. To further emphasize this research as a framework which can be expanded and adapted, we should also resolve that introducing new training data to the model will change its learned probability distributions. An ideal framework for learning how users value designs will be adaptable, and so this type of model is one appropriate choice.

The probabilistic model in this thesis was written using Python 3, a programming language commonly used for data science and machine learning. A number of Python resources and libraries exist for building machine learning models, however, this framework demonstrates building a model without the use of said resources; including pre-trained models in the research introduces more potential for biased results, and often perpetuates bias in machine learning in other contexts (Turner Lee, 2018). The only dependency of this framework's model is Numpy, a library of mathematical operations and scientific computing functions.

Figure 11: Portion of training data set with data sorted by functional category. Horizontal axis: features of designs experienced by users. Vertical axis: survey respondent/user.

3.12.2. High Level Procedure of the Algorithms

The highest level (main) procedure of the algorithm performs all other sub-procedures. First, the training data set is imported. Next, the data set's array is transposed or otherwise transformed, if necessary. Print statements are included at this stage to display a portion of the data for review and debugging. Constants and variables are then defined in the procedure. Next, the model is trained and a probability distribution is learned. The learned distribution probabilities are rounded to two decimal places based on a constant defined earlier, which can be adjusted. Print statements at this stage also help to provide transparency into the algorithm, and help to debug. Following this, procedures for generating augmented data, augmenting the data set, and making predictions about new data are coded. Other output procedures can be included here as well, such as procedures for data visualization.

Algorithm 1 Learn Design Valuations

```
procedure GenerateDistributions(RatingScale, Answers)
    NumberOfQuestions ← Number of survey questions
    Distributions ← an empty array of size (len(Answers), len(RatingScale))
    for i in range(NumberOfQuestions) do
        NonNaNs ← Count if not (answer to question i) == N/A
        ThisDistribution ← An empty array of size len(RatingScale)
        for Rating in RatingScale do
            Occurrences ← Count if (answer to question i) == Rating
            ThisDistribution[Rating] ← Occurrences/NonNaNs
        for Probability ∈ ThisDistribution do
            NormalizeProbability
        Distributions[i] ← ThisDistribution
        return Distributions
```

Algorithm 1, Procedure 1: Training procedure for probabilistic machine learning algorithm, capable of learning a probability distribution based on a training data set.

4.12.3. Training Procedure

The first sub-procedure is the algorithm's training procedure, shown in Algorithm 1, Procedure 1. An important input, or parameter, of the procedure is a rating scale in the form of a Numpy array. In the case of this research, a rating scale was assigned as follows:

```
rating_scale = numpy.array([-3,-2,-1,0,1,2,3], dtype="float")
```

The training data set is also input as a parameter in this procedure. Metrics on the training set and the rating scale are extracted and saved in memory, such as the number of features mentioned in the survey and the shape of the rating scale array. These values are used to instantiate an empty array (an array whose values are all equal to `0`, using `numpy.zeros()`) in which the model's learned distributions will be stored. The 0 values will be overwritten during training. We then iterate over the list of features in the training set, performing a series of operations for each:

1. First, we count and save the number of NaN (Not a Number) values associated with the feature, which represent "not applicable" user responses, or unanswered questions.

2. Next, we instantiate an empty array in the shape of the rating scale using the `numpy.zeros()` function, which will be used to save the probability distribution for the feature.

3. Then, for each of the possible ratings in the rating scale, we count the number of occurrences of the rating in the data associated with the feature. From this value, the probability that the rating will occur can be computed by dividing the number of occurrences by the number of non-NaN answers for this feature.

4. Once the probabilities for all of the ratings are computed for this feature, the distribution is normalized. This means the probabilities are proportionally adjusted such that they sum to 1.

5. The distribution for this feature is then used to overwrite the corresponding row in the array which contains all of the distributions.

This series of operations is repeated for each feature in the training set, producing a set of learned probability distributions conditioned on the training data. This set is returned by the procedure in the form of a Numpy array. The larger the training set, the more accurate these distributions will be. However, this exact method does produce a general distribution for each feature regardless of user demographics or furniture typology. To expand this framework and build on the previous discussion of equitable data science, the algorithm should be modified to compute different probabilities for different groups of users and furniture designs. I further unpack this point in section 6 when discussing future work.

Algorithm 1 Learn Design Valuations

procedure AugmentUserData(*Answers*, *NumberOfNewUsers*)
    RatingScale ← [-3,-2,-1,0,1,2,3]
    Distributions ← GenerateDistributions(*RatingScale*, *Answers*)
    NewDataSet ← GenerateNewData(*RatingScale*, *Answers*, *Distributions*)
    AugmentedDataSet ← AugmentData(*Answers*, *NewDataSet*)
    return *AugmentedDataSet*

Algorithm 1, Procedure 2: Procedure for augmenting existing data sets of user feedback on the value of furniture designs.

3.12.4. Procedure for Generating Augmented Data

Data augmentation is used to expand the volume of data sets based on known data. The added data is also called *synthetic data* (Roh et al., 2019). Because data sets on the value of designs—and specifically furniture designs—are uncommon, this framework includes a procedure for augmenting a dataset. Ideally, the population sample of the framework's survey will be large enough to render augmentation unnecessary, but for small data sets it is useful nonetheless. Augmented data sets, however, cannot be used again to cyclicly train the algorithm, as the data set would produce no significant change in the learned distributions. In the case of this research, an augmented data set is useful for training other algorithms, producing visualizations, or modeling larger populations in the context of practice or commerce. For example, if the survey sample size is adequately representative of a larger population of size x, the dataset can be augmented to dimensions

$$[\mathrm{n_f},\ x]$$

and referenced for quantitative values. Furniture designers will be able to make estimations as to how many users will likely value various design features highly. Furniture retailers will be able to make more informed decisions about production and inventory numbers, regional adjustments to features such as material, shape or size, and decisions about warehousing or operations such as shipping. Both will be able to improve the equity of their practices with access to user-volunteered feedback at scale, rather than relying on intuition to make decisions.

The procedure for generating augmentation data representative of the training set begins with the input of four parameters: the rating scale array, the training set, the learned distributions and an integer value n (the number of new data points to generate). The rating scale array is the same Numpy array referenced in the training procedure. The training set is unchanged. The learned distributions were returned by the training procedure as a Numpy array. Once these parameters are passed into the procedure, the number of features in the training set is computed using Numpy array.shape method. This value is used as the x-dimension of an empty array, instantiated using `Numpy.zeros()`. Its y-dimension is n, the number of new data points to generate. This empty array will be used to save newly generated data, and is named `new_data`. For each feature in the training set, the following series of operations is then performed:

1. First, a local variable `new_points` is instantiated.

2. `Numpy.random.choice` is used to select n number of values from the rating scale. Each value in the rating scale is chosen with a unique probability learned from the training data, reflecting actual user feedback. These values are saved using the variable `new_points`.

3. `new_points` is added to new_data. This operation overwrites the row of zeros corresponding to the feature for which the new points were generated.

4. These operations repeat until n data points are generated.

After all of the new data points are generated, the procedure returns `new_data`.

3.12.5. Procedure for Augmenting a Data Set

The procedure for augmenting a dataset using new data generated in the previous section is straightforward. Parameters or inputs to the procedure are the training data set and the new data set. The procedure first makes a copy of the training data. Next, the new data set is merged with the training data set using concatenation. The `Numpy.concatenate()` function can be easily used.

4. Results of Algorithm Training and Development

This thesis presents a preliminary implementation of a machine learning algorithm for understanding how users value designs. As such, the methods composing the data collection and machine learning framework are the primary results. Outputs from the algorithm itself, trained on the user feedback collected in this thesis, are also secondary results.

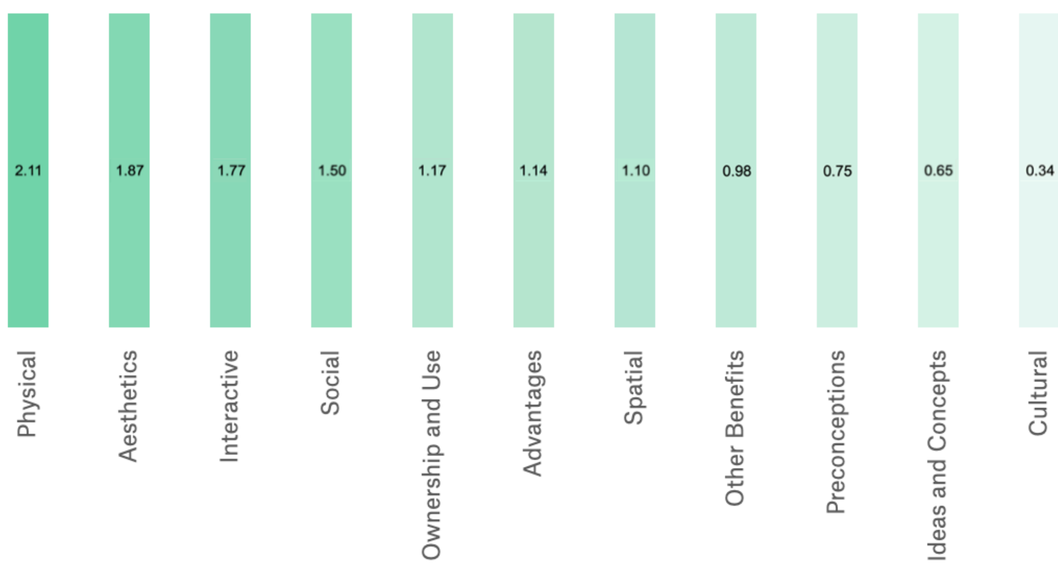

Figure 12: Average user rating for functional categories of furniture designs, based on a user survey.

4.1. Results of Preliminary Training of the Probabilistic Algorithm in this Thesis

The probabilistic algorithm described in section 3.12.2. was trained on user feedback collected using the methods in 3.7. Because the first round of survey distribution led to significant changes in the survey interface, those five responses were omitted from the training set. After this step, the training set contained 44 responses. 27 of these responses addressed only 11 of the design features, as shown in Figure 11. A test data set with 10 users was also reserved from this group. The training set was still usable even with missing responses, as the training algorithm accounts for blanks. Respondents also left some answers blank intentionally. The algorithm was able to generate probability distributions even with missing data points, which positions it as a method similarly useful to matrix factorization, in which missing data is inferred.

The average rating for each category of furniture design feature is shown in Figure 12 above. Corroborating the feedback from the first round of user surveys, physical functions were perceived to be the most valuable, with an average rating of +2.11 out of 3. Aesthetics and interactive functions, two similarly tangible categories, were also rated as having high value. Respectively, users rated them an average of +1.87 and +1.77. The lowest average rating was +0.34, given to cultural functions. The average rating for all categories is positive, although many respondents did indicate negative ratings for features on an individual basis. This phenomenon is important to note and may indicate that the clarity of the survey questions could be improved to help evaluate the bias of this output.

The probabilistic model was effectively trained on the training data. The distribution learned from the training data is shown in Figure 15. Along the horizontal axis of this figure, potential user ratings for features are listed from -3 to 3. Along the vertical axis, each category of feature is listed. The probability distribution for each category is read horizontally in the table. Likelihoods of zero are left blank in the table. The likelihoods are rounded to two decimal places and normalized for each category.

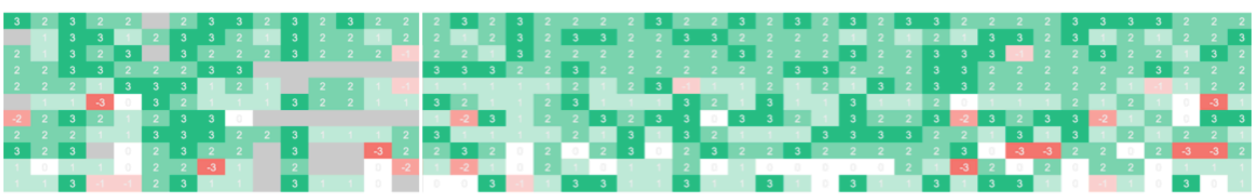

Figure 13: Visualization of average user ratings of functional categories, aggregated with an augmented data set. Horizontal axis: user/respondent. Vertical axis: functional category. Left section: User Data. Right section: Augmented Data.

The trained algorithm was also used to output augmented data. The augmented data set is shown in Figure 13. Grey elements indicate missing user responses. This output serves both as a test of the data augmentation capabilities of the algorithm, but also as an evaluative tool to assess the efficacy of the model. The average probability distributions for each category of features in the augmented data correspond the to the probability distributions learned from the training data, confirming that the augmentation algorithm can accurately produce new data points. A test data set of ten user responses was also used to evaluate the efficacy of the algorithm. The results of this evaluation are shown in Figure 14. The average likelihood of each user rating deviated from the learned distribution by a an average of 0.18 points (on a scale from -3 to 3), indicating that the learned distributions can produce reliable data for the population which was sampled. This does not necessarily indicate that the distributions learned in this study are applicable to other populations, but instead suggests that this framework is effective at predicting how users similar to those surveyed value designs.

	Test Set (points, [-3:3])	Deviation (points, [-3:3])	Augmented Set (points, [-3:3])
Physical	2.43	0.03	2.40
Interactive	2.07	0.03	2.10
Aesthetic	2.00	0.10	2.10
Ownership & Use	2.44	0.14	2.30
Advantages	1.71	0.18	1.53
Spatial	1.14	0.19	1.33
Social	1.56	0.21	1.77
Ideas & Concepts	1.93	0.10	1.83
Cultural	1.73	0.56	1.17
Other Benefits	0.42	0.22	0.63
Preconceptions	1.15	0.18	1.33
	Average Deviation:	0.18	

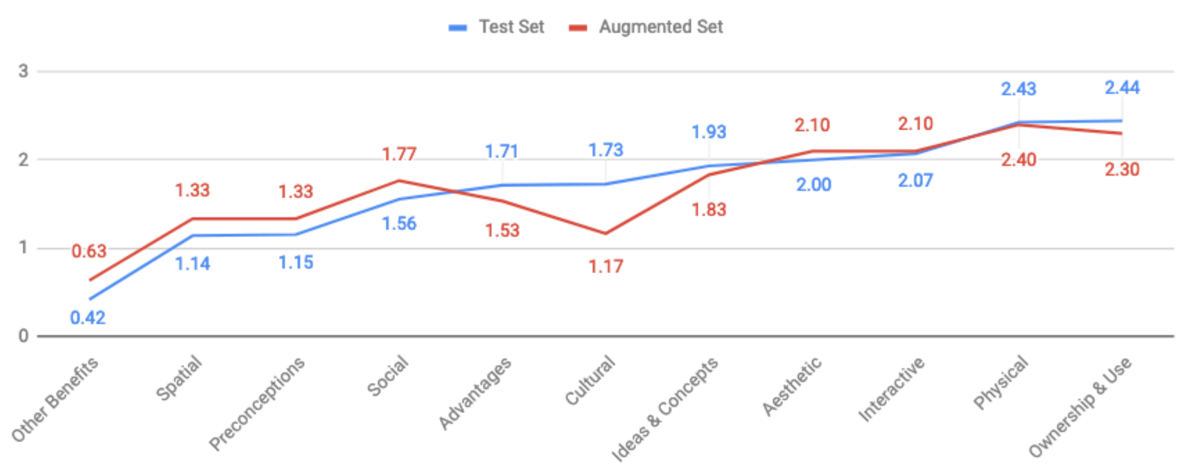

Figure 14: (Top) Table showing the deviation of an augmented data set from a training set based on a user survey, shown by average point rating for each functional category of furniture designs. (Bottom) Average point rating for each functional category as computed for the test set and augmented set.

5. Conclusions

The framework presented here demonstrated a multi-step process for developing data representations for furniture designs, collecting user feedback with surveys, building machine learning models and using them to analyze the feedback. The framework ultimately produced a simple probabilistic machine learning model which was able to use learned probability distributions to augment a dataset with synthetic data. The new data closely reflected the test data set, indicating that the training was effective. Perhaps most notably, the structure of the data on which the model was trained was developed in a “bottom-up” way, originating with user feedback.

5.1. Embracing Data Science in Design

Fields such as finance, technology and medicine have incorporated data science and machine learning into their practices, yet design-related fields have yet to largely adopt data-driven approaches. By showing how this framework can be successfully built using crowdsourced feedback, we take steps toward a future in which all design disciplines—furniture design, architecture, product design, and others—can make more informed decisions about producing work for users. If we are able to construct more nuanced data representations of designs such as furniture, we can naturally envision how similar representations can be constructed about constructed spaces, user interfaces or designed products. For design-related e-commerce, such as furniture e-commerce, this thesis represents steps toward more accurate and effective product recommendations. Current recommender systems work well, but are severely limited in their potential to serve all groups of users, regardless of demographic, exposure to design, or background. E-commerce and technology companies must work closely with the designers of physical artefacts such as furniture, using data science to better understand how users actually value designs.

	-3	-2	-1	0	1	2	3
Physical						0.57	0.43
Interactive					0.29	0.36	0.36
Aesthetic			0.07		0.07	0.57	0.29
Ownership and Use						0.56	0.44
Advantages			0.07		0.29	0.43	0.21
Spatial	0.07			0.07	0.50	0.21	0.14
Social		0.11		0.11	0.11	0.33	0.33
Ideas and Concepts					0.33	0.40	0.27
Cultural	0.09			0.09		0.45	0.36
Other Benefits	0.08	0.08		0.25	0.33	0.25	
Preconceptions			0.15	0.08	0.46	0.08	0.23

Figure 15: Visualization of machine learning algorithm's learned probability distributions, representing the likelihood that a user will assign a rating (horizontal axis) to a furniture design's feature based on the feature's functional category (vertical axis).

6. Future Work

Future work which builds upon this thesis will likely involve adapting the framework to other design disciplines and populations of users. It will be necessary to test the framework's scalability in the process. Part of this involves making continual improvements and updates to the data collection methods, including user surveys. If we work toward as inclusive a survey as possible, we can expand our dataset simply by taking steps to make its interface more accessible, for example. From the perspective of the designer or business utilizing this framework to inform practice, a user interface for visualizing learned data will also be critical. A prototype for this interface is shown in Figure 16, including tools for viewing anonymized user comments, learned features and distributions, embeddings of furniture designs, and examples of designs which users value highly.

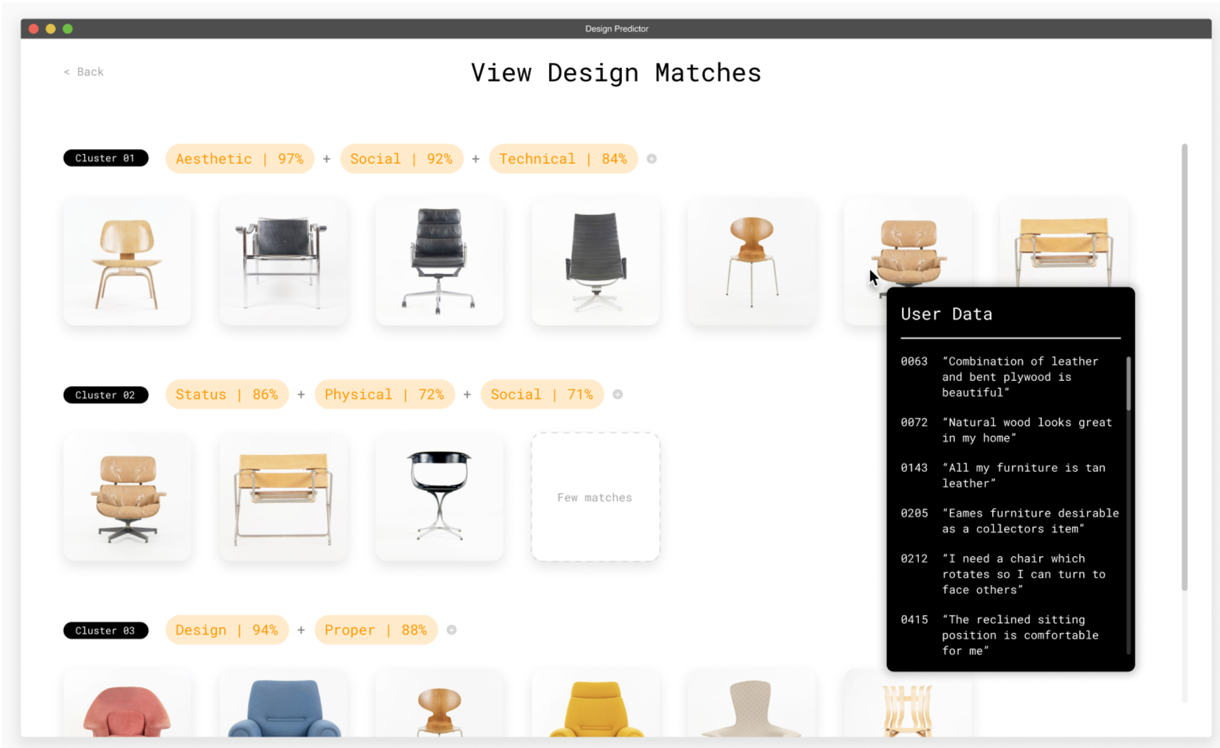

Figure 16: Prototype of user interface for designers, businesses and practitioners to analyze user feedback using machine learning. Furniture photographs courtesy of David Rosenwasser, 2021.

As this framework is further developed, it may also be necessary to consider alternative and state-of-the-art machine learning methods to best serve the businesses and disciplines utilizing the framework. For example, incorporating neural networks may offer an opportunity to generate a larger and more accurate set of features (although interpretability of the results will be an inevitable pitfall). This thesis presented a simple probabilistic model, however more advanced Bayesian adaptations are possible. The feature variables used in this model were considered to be independent of one another (assumed to have no influence on one another). As such, the model's learned probability distributions did produce some level of variance compared to the test set in section 4 (in the case of averages for some features, variance as high as 0.56 points). A future iteration which would improve prediction accuracy will capture the relationships between features instead of assume them to be independent. For example, we can intuitively sketch a directed acyclic graph (DAG) of relationships between features simply by thinking about our own individual experiences with furniture (Figure 17). If a larger population is surveyed regarding its experience of these relationships, we may be able to derive a "bottom-up" DAG for use as a Bayes network in our model. Recommender systems which use Bayes nets and belief propagation have been shown to have high levels of accuracy, interpretability, and computational scalability (Ayday et al., 2012). In addition to improving the computational methods for training and constricting this framework's model, future research should also work to further integrate demographic user data into the computation. In the case shown in this thesis, the distribution learned from user data is only applicable to users similar to the population sample used. If a greater number of users are surveyed, the training set can be clustered according to user demographic information and background. Next, unique distributions can be learned for each cluster, taking a step toward less generalization in our machine learning models. Through this process, this framework can help to diminish echo chambers in recommender systems developed for design domains (Ge et al., 2020). As designers, businesses, and practitioners of all kinds, the more we recognize differences in the way people value designs, the more we can work to deliver the most value to all people.

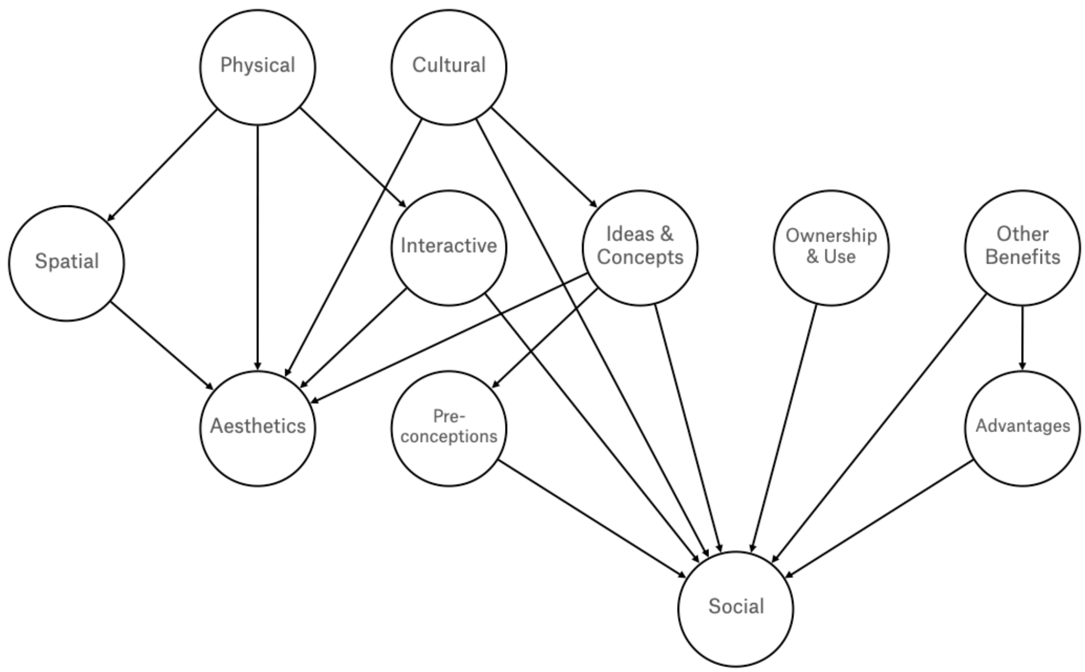

Figure 17: Example of a directed acyclic graph showing dependency relationships between categories of functionality of furniture designs.

7. Acknowledgements

I would like to first thank my family for the love, motivation and support they have given me during the development of this research, and always.

I would like to extend my sincere appreciation to Terry Knight and Stefanie Mueller for their immense generosity, constructive feedback, and unfaltering willingness to share their wisdom with me as my thesis advisors. I am immensely grateful for their mentorship.

I would like to thank David Rosenwasser, (M.Des Technology 2021, Harvard Graduate School of Design) for his friendship and inspiration. I would also like to acknowledge his intellectual contributions to this project as it relates to his Master's thesis and our ongoing creative collaborations. I wish to thank Jacob Riedel for his support and feedback throughout the development of this research, and for his confidence in both me and David.

Similarly, I would like to extend my gratitude to Svafa Gronfeldt, Gilad Rosenzweig, Dennis Frenchman, Norbert Chang and the rest of the MIT designX team for their mentorship, inspiration, encouragement and generosity.

I wish to also acknowledge the feedback, guidance, critique and support so generously given by Skylar Tibbits, Larry Sass, Takehiko Nagakura, George Stiny, Axel Kilian, Kimberle Koile, Jinane Abounadi, Lavender Tessmer, Diego Pinochet, Cagri Hakan Zaman, Tom Silver, Natasha Española, Sasson Rafailov, Daniel Acheampong, Elina Oikonomaki, Wonki (Kii) Kang and Reef Loretto.

8. References

1. Adomavicius, Gediminas, Nikos Manouselis, and YoungOk Kwon. n.d. "Multi-Criteria Recommender Systems," 34.

2. Aguinis, Herman, N. Sharon Hill, and James R. Bailey. 2019. "Best Practices in Data Collection and Preparation: Recommendations for Reviewers, Editors, and Authors." Organizational Research Methods, March, 109442811983648. https://doi.org/10.1177/1094428119836485.

3. Alkan, Oznur, and Elizabeth Daly. 2020. "User Profiling from Reviews for Accurate Time-Based Recommendations." ArXiv:2006.08805 [Cs], June. http://arxiv.org/abs/2006.08805.

4. Amatriain, Xavier. 2013. Big & personal: data and models behind netflix recommendations. In Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications (BigMine '13). Association for Computing Machinery, New York, NY, USA, 1–6. DOI:https://doi.org/10.1145/2501221.2501222

5. Amaya, Ashley. 2020. "Adapting How We Ask about the Gender of Our Survey Respondents." Medium. September 14, 2020. https://medium.com/pew-research-center-decoded/adapting-how-we-ask-about-the-gender-of-our-survey-respondents-77b0cb7367c0.

6. Ayday, Erman, Arash Einolghozati, and Faramarz Fekri. 2012. "BPRS: Belief Propagation Based Iterative Recommender System." ArXiv:1209.5335 [Cs], September. http://arxiv.org/abs/1209.5335.

7. Bailey, Kenneth D. 1994. Typologies and Taxonomies: An Introduction to Classification Techniques. Quantitative Applications in the Social Sciences 07–102. Thousand Oaks, Calif: Sage Publications.

8. Bobadilla, Jesus, Santiago Alonso, and Antonio Hernando. 2020. "Deep Learning Architecture for Collaborative Filtering Recommender Systems." Applied Sciences 10 (7): 2441. https://doi.org/10.3390/app10072441.

9. Bobadilla, Jesús, Fernando Ortega, Abraham Gutiérrez, and Santiago Alonso. 2020. "Classification-Based Deep Neural Network Architecture for Collaborative Filtering Recommender Systems." International Journal of Interactive Multimedia and Artificial Intelligence 6 (1): 68. https://doi.org/10.9781/ijimai.2020.02.006.

10. Capelleveen, Guido van, Chintan Amrit, Devrim Murat Yazan, and Henk Zijm. 2019. "The Recommender Canvas: A Model for Developing and Documenting Recommender System Design." Expert Systems with Applications 129 (September): 97–117. https://doi.org/10.1016/j.eswa.2019.04.001.

11. Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." Journal of Artificial Intelligence Research 16 (June): 321–57. https://doi.org/10.1613/jair.953.

12. ChenHung-Hsuan, ChenPu. 2019. "Differentiating Regularization Weights -- A Simple Mechanism to Alleviate Cold Start in Recommender Systems". ACM Transactions on Knowledge Discovery from Data (TKDD). 13: 1–22. doi:10.1145/3285954. S2CID 59337456.

13. Crawford, Scott D., Mick P. Couper, and Mark J. Lamias. 2001. "Web Surveys: Perceptions of Burden." Social Science Computer Review 19, no. 2: 146–62. https://doi.org/10.1177/089443930101900202.

14. Crilly, Nathan. 2010. "The Roles That Artefacts Play: Technical, Social and Aesthetic Functions." Design Studies 31 (4): 311–44. https://doi.org/10.1016/j.destud.2010.04.002.

15. Dacrema, Maurizio Ferrari, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. "A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research." ACM Transactions on Information Systems 39 (2): 1–49. https://doi.org/10.1145/3434185.

16. Dhall, Kaur, and Juneja. 2020. "Machine Learning: A Review of the Algorithms and Its Applications." Proceedings of ICRIC. 2019. pp 47-63.

17. Everitt, Brian. 2006. The Cambridge dictionary of statistics (3rd ed.). Cambridge, UK: Cambridge University Press. ISBN 978-0-511-24688-3. OCLC 161828328.

18. Ge, Yingqiang, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. 2020. "Understanding Echo Chambers in E-Commerce Recommender Systems." In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2261–70. Virtual Event China: ACM. https://doi.org/10.1145/3397271.3401431.

19. Harambam, Jaron, Dimitrios Bountouridis, Mykola Makhortykh, and Joris van Hoboken. 2019. "Designing for the Better by Taking Users into Account: A Qualitative Evaluation of User Control Mechanisms in (News) Recommender Systems." In Proceedings of the 13th ACM Conference on Recommender Systems, 69–77. Copenhagen Denmark: ACM. https://doi.org/10.1145/3298689.3347014.

20. Hosseinzadeh Aghdam, Mehdi. 2019. "Context-Aware Recommender Systems Using Hierarchical Hidden Markov Model." Physica A: Statistical Mechanics and Its Applications 518 (March): 89–98. https://doi.org/10.1016/j.physa.2018.11.037.

21. Iqbal, Murium, Adair Kovac, and Kamelia Aryafar. 2018. "A Multimodal Recommender System for Large-Scale Assortment Generation in E-Commerce." ArXiv:1806.11226 [Cs], June. http://arxiv.org/abs/1806.11226.

22. James, Gareth. 2013. "An Introduction to Statistical Learning: with Applications in R." Springer. p. 176. ISBN 978-1461471370.

23. Keusch, Florian, Bella Struminskaya, Christopher Antoun, Mick P Couper, and Frauke Kreuter. 2019. "Willingness to Participate in Passive Mobile Data Collection." Public Opinion Quarterly 83 (S1): 210–35. https://doi.org/10.1093/poq/nfz007.

24. Kouadria, Abderrahmane, Omar Nouali, and Mohammad Yahya H. Al-Shamri. 2020. "A Multi-Criteria Collaborative Filtering Recommender System Using Learning-to-Rank and Rank Aggregation." Arabian Journal for Science and Engineering 45 (4): 2835–45. https://doi.org/10.1007/s13369-019-04180-3.

25. Kouki, Pigi, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. "Personalized Explanations for Hybrid Recommender Systems." In Proceedings of the 24th International Conference on Intelligent User Interfaces, 379–90. Marina del Ray California: ACM. https://doi.org/10.1145/3301275.3302306.

26. Kroes, P., & Meijers, A. 2006. The dual nature of technical artefacts. Studies in History and Philosophy of Science.

27. Lee, Yunju, Haram Won, Jaeseung Shim, and Hyunchul Ahn. 2020. "A Hybrid Collaborative Filtering-Based Product Recommender System Using Search Keywords." 지능정보연구 26 (1): 151–66. https://doi.org/10.13088/JIIS.2020.26.1.151.

28. Lewis, James R. 2006. "Usability Testing". IBM Software Group. Boca Raton, Florida.

29. Lobe, Bojana, David Morgan, and Kim A. Hoffman. 2020. "Qualitative Data Collection in an Era of Social Distancing." International Journal of Qualitative Methods 19 (January): 160940692093787. https://doi.org/10.1177/1609406920937875.

30. Luck, Rachael. 2007. "Learning to Talk to Users in Participatory Design Situations." Design Studies 28 (3): 217–42. https://doi.org/10.1016/j.destud.2007.02.002.

31. Mahner, M., & Bunge, M. 2001. "Function and functionalism: a synthetic perspective." Philosophy of Science.

32. Mazeh, Itzik, and Erez Shmueli. 2020. "A Personal Data Store Approach for Recommender Systems: Enhancing Privacy without Sacrificing Accuracy." Expert Systems with Applications 139 (January): 112858. https://doi.org/10.1016/j.eswa.2019.112858.

33. Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi. 2020. "Recommender Systems and Their Ethical Challenges." AI & SOCIETY 35 (4): 957–67. https://doi.org/10.1007/s00146-020-00950-y.

34. Nassar, Nour, Assef Jafar, and Yasser Rahhal. 2020. "Multi-Criteria Collaborative Filtering Recommender by Fusing Deep Neural Network and Matrix Factorization." Journal of Big Data 7 (1): 34. https://doi.org/10.1186/s40537-020-00309-6.

35. Pahl, G., & Beitz, W. 1996. Engineering design: A systematic approach. (K. Wallace, L. Blessing, F. Bauert, Trans.) (2nd ed.). London, UK: Springer. (Originally published 1977 as Konstruktionslehre, Springer, Berlin).

36. Ripley, Brian. 1996. Pattern Recognition and Neural Networks. Cambridge University Press. p. 354. ISBN 978-0521717700.

37. Roh, Y., G. Heo, and S. E. Whang. 2019. "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective." IEEE Transactions on Knowledge and Data Engineering, 1–1. https://doi.org/10.1109/TKDE.2019.2946162.

38. Sahoo, Nachiketa, Ramayya Krishnan, George Duncan, and Jamie Callan. 2012. “Research Note —The Halo Effect in Multicomponent Ratings and Its Implications for Recommender Systems: The Case of Yahoo! Movies.” Information Systems Research 23 (1): 231–46. https://doi.org/10.1287/isre.1100.0336.

39. Si, Luo, and Rong Jin. n.d. “Flexible Mixture Model for Collaborative Filtering,” 8.

40. Sinha, Bam Bahadur, and R. Dhanalakshmi. 2019. “Evolution of Recommender System over the Time.” Soft Computing 23 (23): 12169–88. https://doi.org/10.1007/s00500-019-04143-8.

41. Su, Xiaoyuan, and Taghi M. Khoshgoftaar. 2009. “A Survey of Collaborative Filtering Techniques.” Advances in Artificial Intelligence 2009 (October): 1–19. https://doi.org/10.1155/2009/421425.

42. Turner Lee, N. 2018. "Detecting racial bias in algorithms and machine learning", Journal of Information, Communication and Ethics in Society, Vol. 16 No. 3, pp. 252-260. https://doi.org/10.1108/JICES-06-2018-0056

43. Vabalas, Andrius, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. 2019. “Machine Learning Algorithm Validation with a Limited Sample Size.” Edited by Enrique Hernandez-Lemus. PLOS ONE 14 (11): e0224365. https://doi.org/10.1371/journal.pone.0224365.

44. Xing, Baixi, Kejun Zhang, Lekai Zhang, Xinda Wu, Huahao Si, Hui Zhang, Kaili Zhu, and Shouqian Sun. 2020. “And the Nominees Are: Using Design-Awards Datasets to Build Computational Aesthetic Evaluation Model.” PLoS ONE 15 (1): 1–19. https://doi.org/10.1371/journal.pone.0227754.

45. Yusuf, Samuel. 2019. “Wayfair DS Explains It All: Samuel Yusuf on Collaborative Filtering.” https://tech.wayfair.com/data-science/2019/08/wayfair-ds-explains-it-all-samuel-yusuf-on-collaborative-filtering/

46. Zheng, H., H. Hu, and Z. Han. 2020. “Preserving User Privacy for Machine Learning: Local Differential Privacy or Federated Machine Learning?” IEEE Intelligent Systems 35 (4): 5–14. https://doi.org/10.1109/MIS.2020.3010335.

9. Appendix

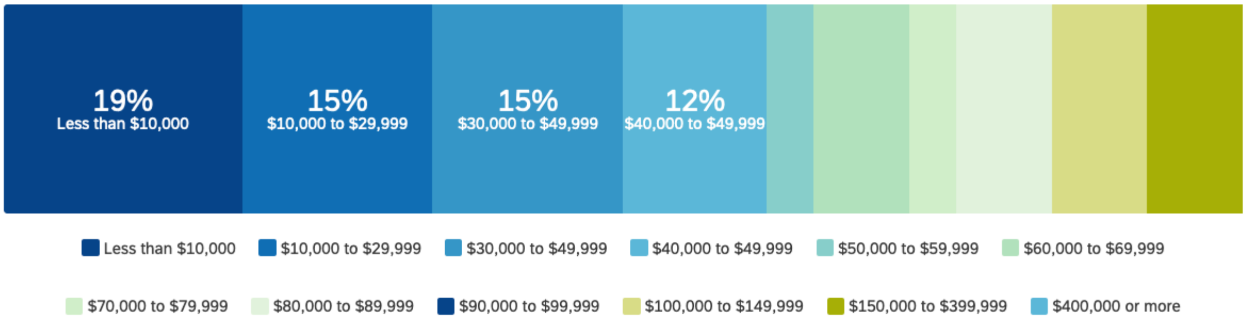

Figure A1: Personal income of survey respondents shown as percentage of all respondents.

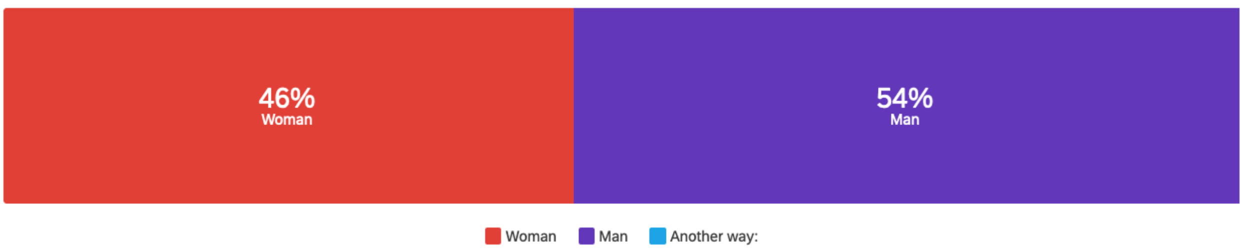

Figure A2: Gender identity of survey respondents shown as percentage of all respondents.

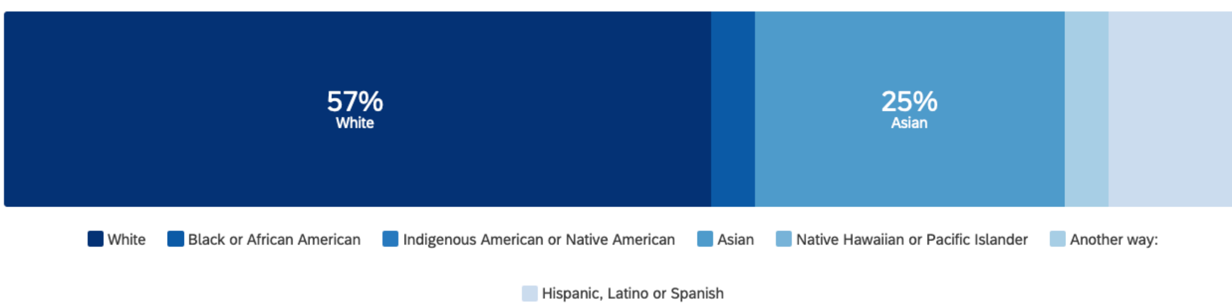

Figure A3: Racial identity of survey respondents shown as percentage of all respondents.

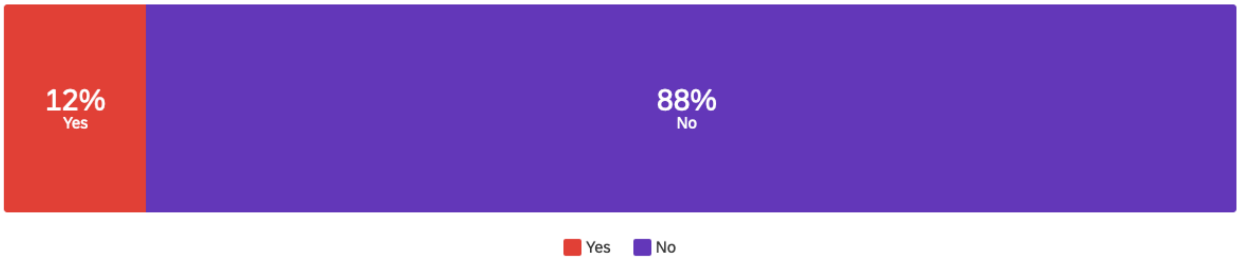

Figure A4: Response by survey respondents to a question asking "Have you used interior design or an architect's services?" shown as percentage of all respondents.

Instagram	15.44%	Architizer	2.68%
In-person galleries	10.74%	Reddit	1.34%
Books	10.74%	Houzz	1.34%
Museums	10.07%	Email lists	1.34%
Pinterest	6.71%	Design-Milk	1.34%
Dezeen	6.71%	DesignBoom	1.34%
ArchDaily	6.71%	Tik Tok	0.67%
Architectural Digest (AD)	4.70%	Other:	0.67%
Dwell	4.03%	None	0.67%
Youtube	3.36%	Forums	0.67%
Medium	2.68%	Design Addict	0.67%
Facebook	2.68%	Twitter	0.00%
Design Within Reach (DWR)	2.68%	Dexigner	0.00%

Figure A5: Where survey respondents get information about furniture and design, shown as percentage of all responses.

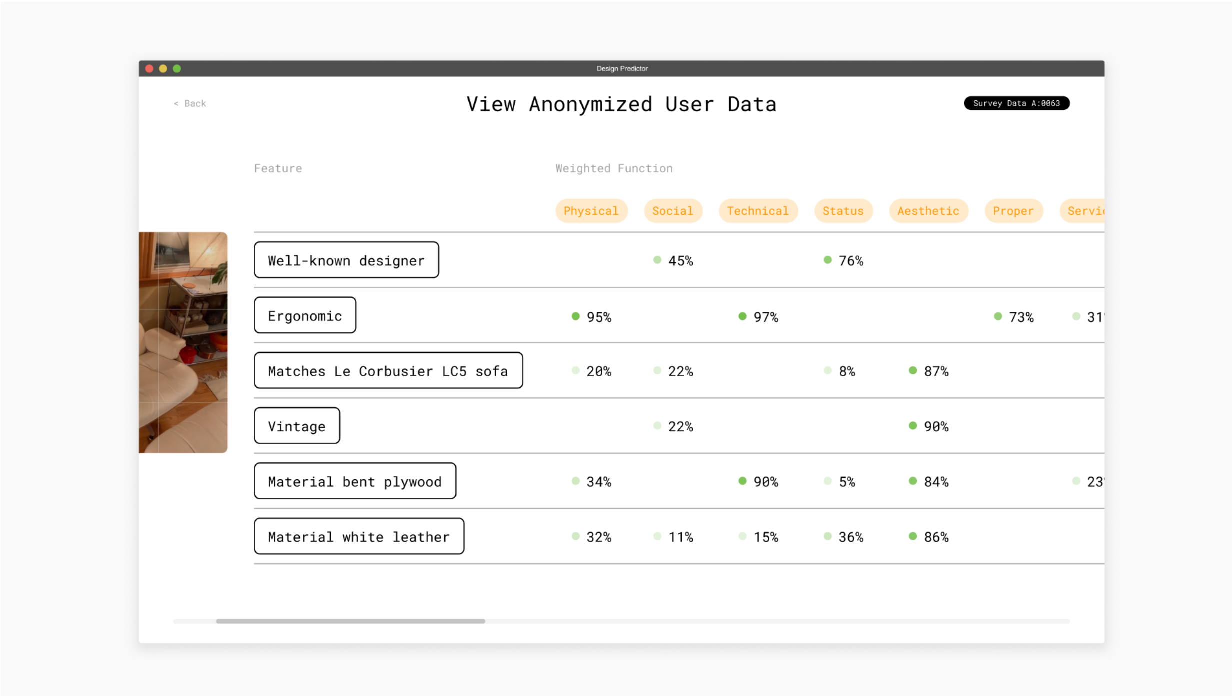

Figure A6: Prototype of user interface for designers, businesses and practitioners to analyze user feedback using machine learning, showing analysis of user valuation based on features of furniture designs.

Physical Qualities - Your furniture's physical shape or form.
Materials - The materials used to make my furniture.
Physical Qualities - The quality of its construction.
Physical Qualities - The furniture's size.

Interactive - The physical comfort of interacting with your furniture.
Interactive - The way it fits your body.
Interactive - Its weight.
Interactive - Its temperature when touched.
Interactive - How often you use it.

Aesthetics - How the furniture looks; its aesthetics.
Aesthetics - The furniture's finishes/quality of its surfaces.
Aesthetics - Its color.
Aesthetics - How its materials look, to you.
Aesthetics - How light hits it.
Aesthetics - How you feel when you look at it.
Aesthetics - How it looks in its surroundings.
Aesthetics - How simple it looks.
Aesthetics - How complex it looks.

Ownership and use - How well your furniture functions.
Ownership and use - The way you use it has changed over time.
Ownership and use - The way you use it is unique because of the way it looks.
Ownership and use - It needed (or needs) some modifications.
Ownership and use - The way you use it is a result of experimentation.
Ownership and use - How it functions with other pieces of furniture.
Ownership and use - How it looks with other pieces of furniture.
Ownership and use - It serves a function when in a new orientation.
Ownership and use - You are discovering new uses for the furniture.
Advantages - Your furniture makes others feel welcome.

Advantages - How others use and react to my furniture.
Advantages - It impresses others.
Advantages - It makes others feel comfortable.
Advantages - It generates income in a business setting.
Advantages - It encourages relaxation.
Advantages - It encourages productivity.
Advantages - It generates discussion.
Advantages - It sparks passion, interest or contributes to a hobby.
Advantages - It teaches something.

Figure A7: First half of full list of feature questions from user survey.

Functions in space - How your furniture defines or creates spaces around it.

Functions in space - Your furniture divides a space.

Functions in space - It marks an entrance or exit.

Functions in space - It marks the center of a space.

Functions in space - It hides or obscures a view.

Functions in space - It creates a view.

Functions in space - It creates a place to do an activity.

Social qualities - How your furniture reflects your identity or taste.

Social qualities - It seems to gather people together.

Social qualities - It seems to separate people.

Social qualities - It is a part of your daily routine.

Social qualities - How often guests use it.

Social qualities - How special it seems to others.

Social qualities - What it says about how well-off you are.

Social qualities - It is very neutral, or doesn't say anything about you.

Ideas and concepts - The idea, concept, or history of my furniture's design.

Ideas and concepts - The "style" of the furniture.

Ideas and concepts - The story of how it was made or manufactured.

Ideas and concepts - Its design is common, and has no single designer.

Ideas and concepts - It was made for you, specifically.

Ideas and concepts - It has been customized.

Ideas and concepts - It is symbolic.

Ideas and concepts - It represents a position you take on an issue.

Ideas and concepts - It represents a goal you have, or achieved.

Ideas and concepts - It is unique.

Ideas and concepts - It represents that you are unique.

Ideas and concepts - Owning it supports the people who designed or made it.

Cultural - Its significance to your personal or cultural identity.

Cultural - The furniture's family significance.

Cultural - It plays a role in a ritual or practice.

Cultural - It represents a personal value.

Cultural - It connects you to a community.

Cultural - The significance of its designer, or maker, to your cultural identity.

Other benefits - Its significance as a part of a collection, or as an investment.

Other benefits - It has monetary value as an investment.

Other benefits - It was a gift.

Other benefits - You made or designed it yourself.

Other benefits - It was recycled, salvaged or re-used (including "vintage").

Preconceptions - The impression you had of the furniture before you owned it.

Preconceptions - The way it was advertised.

Figure A8: Second half of full list of feature questions from user survey.