An Exploration into the Application of Human-in-the-Loop Technologies for Personalised Music Recommendation

ELLIS CHRISTOPHER

To my family without whom this thesis would never have got written.

Acknowledgements

I would like to thank my supervisors Steve Benford, Max Wilson and Genovefa Kefalidou for their continued support and guidance throughout this process as well as my family for the many nights and hours of proof reading.

I would also like to give special thanks to the following friends whose support has been invaluable:

- Dana Fincham Lucas
- Mathew Farenden
- Hazel Webb
- Ronnie & Zara
- Sarah O'Donovan
- Susanna Silva

  whose music has been a constant survival mechanism throughout this thesis. I strongly urge any reader interested in music to check out her YouTube channel https://www.youtube.com/channel/UCkGa9Yamyof6Jsm6l9vgvaQ
  .

Abstract

The evolution of the internet over the last 30 years has drastically changed the way we find and consume music. Today we can get near-instantaneous access to vast libraries of music with streaming services like Spotify offering archives in excess of 30 million tracks. Faced with such overwhelming choice it can be easy to become paralysed by the possibilities. We need some automated and effective means of navigating this sea of content to identify the music we want.

The currently accepted solution to this problem is the music recommender system. At their core, modern music recommenders are computer programs which suggest music to users by attempting to accurately predict their taste preferences and identify corresponding appropriate tracks to recommend from a digital musical archive. Unfortunately, in recent years it has been increasingly found that predicting music in this way often produces accurate but obvious, impersonal and uninteresting recommendations that are not necessarily useful or desirable to users. This has lead to the rise of a problem which has become known within the industry as the personalisation problem. In essence, systems are producing recommendations which may be accurate but which are perceived to be impersonal.

In this thesis, we consider how allowing the user to manually engage with and influence the outcome of these automated systems could mitigate this problem and lead to more personal and better-received recommendations. We advocate a human-in-the-loop (HITL) approach to music recommendation that puts the user back in control of their recommendations.

The core contributions of this thesis are:

1. An explanation as to the dangers of solely pursuing predictive accuracy in music recommendation

2. A deconstruction and exposition of the personalisation problem for music recommendation.

3. An evaluation as to the role and significance of considering the intended purpose for which a recommendation is being sought when producing recommendations

4. The development and initial validation of a novel HITL strategy for combating the personalisation problem

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Motivation: The Importance of Music Recommendation

Almost 40 years after ‘Video Killed the Radio Star’, internet streaming and download services stand to monopolize our means of media consumption across the board from the written word to the radio and the big screen. The drastic improvements and increased accessibility to fast reliable internet over the last 10 years have facilitated a complete revolution in how we find and consume media.

Today E-readers are increasingly being favoured as a means of consuming books. Music streaming services are rapidly overtaking radio listening, and services like Netflix and Amazon Video have led to a decline in television and cinema viewing as well as the virtual extinction of traditional video rental stores like Blockbuster.

These new services provide us with immediate access to all the content we might desire, from the breaking news to the latest best-seller and must-see films. With such vast amounts of information being made available to us through these services, it would be very easy to become overwhelmed without some assistive means of filtering the relevant information from the irrelevant.

To put it into context, Spotify, as a single streaming service, boasts a library of over 30 million songs. Assuming an average song length of 3:30, that equates to 1,750,000 hours or approximately 200 years of music. That’s enough music to fill more than two lifetimes from the cradle to the grave.

One technical solution to this information overload problem is modern recommender systems; digital algorithms designed precisely to match users with desired content. As such, recommender systems have now become essential to the construction, use and enjoyment of online media services. Indeed, it was the rise in popularity of media streaming services like Netflix around 2006 that led to the recommendation systems' revival and truly established recommender systems as a dedicated field of research distinct from information retrieval (Adomavicius & Kwon, 2011; Bennett & Lanning, 2007).

By 2010 many of the accuracy and scalability issues that had plagued earlier recommender systems had been solved and top-N style recommenders[1] had become commonplace for e-commerce sites. Unfortunately, this is when the problems truly began for the field of music recommendation. Just as movie streaming services were finding widespread adoption and recommender systems were starting to enjoy commercial success and public praise, a backlash began to emerge against music stream services. Users started to complain that they were only being presented with mainstream top chart obvious recommendations ("Slave to the algorithm? How music fans can reclaim their playlists from Spotify," 2016). Niche and personalised tastes were not being reflected in the generic recommendations people were receiving. Some papers even suggest this popularist filter bubble effect might be contributing to the extinction of entire niche genres like Jazz and Classical music (Donnat, 2018). Clearly, something had gone wrong. This was the start of the personalisation problem and is the focus of this thesis.

1.2 The Problem: Personalising Music Recommendation

There is an old joke in the recommender community about a man who walks into his local supermarket and declares he will build them a product recommender system which is more accurate than anything else on the market. What's more, he says he will build it right there and then that very afternoon. He then proceeds to write a program which simply tells every customer between the ages of 18-69 to buy bread and milk on Sundays. Of course, this is highly accurate, most adults doing their pre-week shop on a Sunday are likely to buy staples like bread and milk for the week, but it is not useful as they were going to buy bread and milk anyway.

It's not a very funny joke, but it does highlight an issue at the heart of the personalisation problem; the disjunct between utility and accuracy. In the period known as the recommendation revival from 2006-2010, recommender research focused almost exclusively on predictive accuracy. This was due in large part to the motivating forces for the revival being commercial entities focused on product sales. The thinking was if we can accurately predict what people want, we can sell it to them.

[1] This is a type of recommender system which recommends a ranked list of n items which best match a customer's perceived taste preferences. More detail on them is provided in chapter 2.

In certain applications where products fit into neat categories, this philosophy can work very well. Instead of recommending a person a specific product in a repetitive manner as in the supermarket example above, you seek to predict or identify a category of similar items they like and then proceed to recommend them items from that collection. This works very well for movies which can be grouped relatively easily and non-contentiously by categories like artist, director or release decade.

Competitions, like the Netflix challenge, helped rapidly advance the field but also cemented this way of thinking and established predictive accuracy as the paragon of recommender success (Adomavicius & Kwon, 2011; Jia Rongfei, Jin Maozhong, & Wang Xiaobo, 2007). Overnight predictive accuracy became the sole metric by which recommender systems were assessed. Competing systems were simply judged by comparing how often each system accurately predicted a user's response to a given set of items like movies.

Part of the problem, as Oscar Celma identifies in his thesis 'Music Recommendation & Discovery', is that music is not the same as movies (Celma, 2008). People may listen to the same tracks or playlist several times in succession, but they are unlikely to watch the same film repeatedly (Celma, 2008).

People also frequently disagree when classifying music by categorising like genre. Furthermore, musical genres tend to evolve and become more niche over time. 100 years ago, blues music could be divided into country blues and urban blues. Today there is an infinite array of hotly debated sub-genre complexities which interlink and overlap with one another like electric, piano, jazz, Louisiana, New Orleans and Chicago blues to name a few. The fine-grained and heavily subjective nature of music classification makes it far more difficult to identify categories and similar item pools a given person might like.

In addition people can have different taste preferences at different times, depending on what they are doing. They often don't want to listen to the same non-objectionable vapid pop tunes. They may not actively dislike them, but they are seeking something unique or new which speaks to them and suits the mood or the situation they are in. At the same time, people don't want recommendations so wildly and shockingly different from their pre-existing tastes as to be offensive to them.

The problem can be summarised as the complex task of balancing predictive accuracy against recommendation novelty whilst being mindful of the purpose and time period in which a recommendation is being sought.

Throughout the thesis, especially within chapter 2 and 3, we go into greater depth explaining why the personalisation problem consists of balancing these attributes.

For now, it can be explained as follows:

If the objective is to design a system to recommend content (in our case music) that people want to consume then it stands to reason that we need to be capable of distinguishing what they would like from what they wouldn't like. Meeting this challenge makes up the predictive accuracy component of the personalisation problem. In short, a personal recommendation is one which the intended recipient specifically can be expected to like.

However, as Chapter 2 will explain, as systems got better at accurately predicting peoples likes and dislikes, recommendations became more obvious and less useful. People became less satisfied with their recommendations especially within the domain of music where they were being recommended safe mainstream recommendations. Whilst it was often true that they may to some extent like the tracks being recommended they didn't love them either and were not surprised by them in any regard.

It became apparent that a degree of serendipity, novelty and non-obviousness was needed for people to consider a recommendation personal. They needed to feel that they weren't being recommended the same top of the charts track as everyone else but something which surprised them to some extent. This is where the serendipity aspect came into the personalisation problem. It should be noted that we have used the terms serendipity, novelty and non-obviousness interchangeably and without qualification here. Later in chapter 3, we look at the subtle differences between these terms as they are used in the literature. Here we are simply signposting the problem of recommendations be too obvious and lacking any aspect of surprise. Serendipity research may provide one route of combating this issue but not the only one. In chapter 3 we argue that novelty based research can be more effective.

Finally as revealed in the first study presented in chapter 3 of this thesis people have tastes which vary across short time spans (often hours) depending on the purpose or scenario for which they are intending to listen to music. For instance, a person might wish to listen to fast tempo pop music when running in the morning and smooth jazz when relaxing with a glass of wine after work. This is where dynamism and purpose enter into the personalisation problem. For a recommendation to be personal it ought to take account of the recipients taste preferences during the specific time span that they are seeking recommendations for a given purpose.

1.2.1 The Academic Perspective

Aspects of the personalisation problem, in particular novelty or serendipity, have seen increased recognition within academia over the last 5 years in several areas from Human-computer interaction (HCI) to Business Studies. We consider this issue or serendipity in chapter 3.

Within recommender research, predictive accuracy has started slowly to give way to serendipity as the new metric for constructing and assessing recommenders. The goal of recommenders in the academic sector, especially within music, is shifting from producing the most accurate recommendations to producing the most serendipitous, novel, well-received recommendations.

Whilst this is a promising development, serendipity is only one part of the personalisation problem. To truly produce better received, pleasantly surprising recommendations, systems need to balance serendipity or novelty against accuracy and take account of people’s dynamic tastes. Framing the problem and exploring how these factors combine to produce the personalisation problem is a key aim and major component of this thesis.

1.2.2 The Commercial Perspective

Over the last 3 years, within the commercial sector, personalised recommendation services have become the holy grail pursuit of streaming services like Spotify, Tidal & Apple Music. Companies are continually vying with one another to produce the most personal and tailored service to suit each individual.

In July 2015, Spotify launched Discover Weekly, a human-in-the-loop (HITL) recommendation service designed to produce a weekly 2-hour radio-like stream of music uniquely targeted at individual users.

In September 2016, Apple launched its first attempt to address the personalisation problem in the form of a ‘My New Mix’ service which provides individual users weekly with 25 new tracks in an attempt to pleasantly expand their musical based on their previous listening habits.

Whilst these services show promise, they typically only address a single component of the personalisation issue rather than combating it as an overall unified and nuanced problem. Neither Discover Weekly nor My New Mix takes into account the dynamic nature of people's tastes in that they may fancy different music for different scenarios or purposes.

1.3 The Solution: A Dynamic Human-in-the-loop Approach

The solution proposed in this thesis focuses on incorporating existing human practices and behaviours into automated systems to mitigate the issues of accuracy, novelty and purpose mentioned above. This technique is known in the literature as a human-in-the-loop (HITL) approach. HITL systems allow users to interfere with the recommendation process in some way, typically by pre or post-filtering the content pools which are either presented to or produced by a classic automated recommender algorithm.

At the broadest level within computer science, a HITL system is any system that employs a model which depends upon human interaction (Karwowski, 2006). More narrowly in the domain of recommender systems, it can be defined as any system which allows users to affect the outcome of a recommendation engine by pre or post-filtering the input or output of its recommendation algorithm. Within this thesis we can narrow this definition even further to capture our usage which is as follows:

A HITL recommender is one which enables users to restrict the userbase or content pool that is input to the automated recommendation algorithm.

This very narrow technical definition may seem a little opaque now but it will become clear as the thesis develops and more information is provided as to what a userbase is and precisely how recommendation systems work.

The approach pursued in this thesis differs from previous work by adopting a holistic view of the personalisation problem. Throughout this thesis, we explore the interconnected issues of accuracy, novelty and dynamicity which make up the personalisation problem. We show how these issues interact to produce unique user demands for music recommendation services which often only exist in a reduced capacity or not at all for product recommenders and other media-based recommendation systems.

In the penultimate chapter (chapter 5) of the thesis, we design a HITL recommender which adheres to the narrow HITL definition above and builds on the findings of the thesis. In constructing this design we consider the benefits and dangers of using metaphors to guide HITL design. We highlight how considering everyday human curation and recommendation practices used by professionals like librarians and record store owners can help us to build user-friendly systems.

1.3.1 Research Questions

Identifying and investigating the issues above lead to the formation of the following research questions in the thesis:

1. What are the tenets of making personalised music recommendations?
2. How can human-in-the-loop practices allow users to inform an automated music recommender of their requirements for personalised recommendations?

These questions can be combined to pose the overarching research objective of this thesis, namely:

How can human-in-the-loop techniques be applied to reincorporate the core tenets of making personalised music recommendations into modern recommender services?

1.4 Theoretical Approach

This thesis aims to identify and explore the many facets which make up the personalisation problem and propose a HITL approach of addressing them together.

The core design philosophy behind this thesis centres on learning from everyday human practices and reintroducing them into automated systems. We demonstrate throughout this thesis that including humans in the recommendation process facilitates serendipitous discovery and enhances the novelty of recommendations. It also has the potential to simultaneously allow users to update the system using their current taste preferences in real-time. Incorporating immediate explicit and directed responsiveness to each user’s musical preferences is something which to our knowledge no current commercial or academic music recommendation system achieves.

1.4.1 Core Issues

Within this thesis two central issues are identified as the core tenets of the personalisation problem:

1. The apparent dichotomy between the predictive accuracy and novelty of recommendations
2. The dynamic nature of individuals' musical tastes

Defining and addressing these issues makes up the major work of this thesis. Here we provide a brief explanation of what they are and how they threaten the personalisation of music recommendation services.

1.4.1.1 Predictive Accuracy vs Novelty

Predictive accuracy and novelty initially appear to run counter to one another. If a system produces recommendations by very accurately predicting the songs a person would pick or rate highly from a given content pool, then the recommendations it produces are unlikely to be novel, serendipitous or in any way surprising to that user. Conversely, if a system produces its recommendations for their ability to be surprising and novel then it is highly probable the recommendations will be too esoteric and not accurately match a user’s taste preferences.

In this thesis, we endeavour to understand and address this apparent dichotomy by learning how people confront or avoid it when manually curating playlists. Learning the human practices involved in successful manual music curation allows us to determine which features and components a HITL system should incorporate to address the issue in automated systems.

1.4.1.2 Dynamic Nature of Tastes

The dynamic nature of people’s musical tastes is a complicated issue in its own right. People’s tastes vary for a wide variety of reasons including the activity they’re engaged in, the company they’re with and the mood they are in. Each of these reasons has spawned areas of research such as recommending music for sports or recommending music for mood. In this thesis, we draw a cross-comparison between these fields and explore how they can be applied to a HITL system to facilitate users in ensuring that their tastes are continually being accurately gauged and well served.

1.5 Thesis Methodology

In section 1.4 we described the overarching theoretical approach this thesis pursues in addressing the personalisation problem for music recommendation. Broadly speaking this is a human-in-the-loop approach which endeavours to increase personalisation by enabling users to interact and interfere with the recommendation process. Here we narrow our focus and summarise the core methodological techniques used in the studies and design exercises throughout the thesis to explore, test and pursue a human-in-the-loop approach to the problem of personalisation.

As Oscar Celma observed, music recommendation is an inherently interdisciplinary field of research (Celma, 2010a). Consequently, this thesis has also had to be interdisciplinary and has therefore made use of a variety of methodologies combining qualitative analysis techniques from the fields of psychology and human-computer interaction (HCI) with quantitative techniques from information retrieval and machine learning. Employing a multitude of methods in this manner helped to provide a broad perspective on the personalisation problem and facilitated a complete and holistic response.

An explanation of each of the methodologies is provided in detail in the relevant chapters (3,4 and 5). Here we provide a precursory overview of the key methods used as a means of introducing the thesis narrative and structure.

The exploratory study presented in chapter 3 of this thesis uses a technique known as emergent thematic analysis which is often found in psychology and human-computer interaction research. The method is used in the chapter to identify the core aspects, themes and practices involved in human music curation.

The study presented in chapter 4 uses the methodologies of crowdsourcing and data-analysis common to computer science and machine learning research as a means of further exploring the nature of personalisation and gaining insight into the role of purpose within music recommendation. There is some methodological novelty in the application of this approach both in its use to gain insight into the personalisation problem for music recommendation and in its use to mitigate the impact of a problem known as the WEIRD[2] bias on this thesis (Henrich, Heine, & Norenzayan, 2010). This is discussed in the chapter itself and reflected upon in the conclusion of this thesis.

In chapter 5 the common human-computer interaction approach of scenario-based design is used (Saiedian, Kumarakulasingam, & Anan, 2004; Sutcliffe, RE'98, 1998, n.d.; Sutcliffe, Gault, & Maiden, 2004). It is used to draw together the insights from the preceding chapters and develop a set of guidelines and design principles for building HITL music recommender systems to combat the personalisation problem.

1.6 Thesis Contributions

The main contributions of this thesis are:

1. To explain the dangers that arise from the common practice of using predictive accuracy as the primary (or even the sole) mechanism for constructing and assessing music recommendation systems
2. To reveal how several developments in the wider domain of recommendation systems have negatively impacted on the field of music recommendation and led to a problem defined in this thesis as the personalisation problem
3. To propose a novel human-in-the-loop (HITL) approach to solving the personalisation problem
4. To demonstrate how recommendation purpose (a factor which has largely been ignored outside of sports research) is critical to the effective design of a music HITL recommendation system if it is to solve the personalisation problem

In addition to the above contributions, this thesis also presents a series of innovative methodological approaches and procedural designs for investigating and assessing music recommendation systems with respect to the personalisation problem.

[2] The WEIRD bias is an over generalisation criticism discussed in chapter 4.

1.7 Thesis Structure

This thesis comprises 6 chapters and is structured in the following way:

Chapter one begins by presenting a history of recommendation systems and introduces the personalisation problem. Then it provides a summary of the current state of commercial systems. The final part of the chapter focuses on defining the research question for this thesis and orienting it in the context of existing research.

Chapter two explores the nature of personalisation in inter-personal recommendations between friends and family to develop a detailed understanding of what leads to a recommendation being regarded as personal in normal daily life. The insights gained from the formative exploratory study conducted helped sharpen the focus of this thesis by identifying key factors such as recommendation purpose and the significance of having HITL. This inspired the subsequent approach and HITL design exercise conducted later in the thesis.

Chapter three builds on the findings of chapter two and investigates how the purpose for which a recommendation is being sought affects what information users of a HITL system require to use it effectively to produce the best recommendations.

Chapter four revisits the research question posed in chapter two and demonstrates how the literature and work presented in the intervening chapters can address parts of the question.

Chapter five brings together the findings of this thesis and uses a scenario-driven design exercise and user validation test to explore and develop a series of HITL music recommender design principles.

Chapter six concludes this thesis by highlighting the main contributions and summarising how the HITL approach advocated in this thesis addresses the personalisation problem. The final part of the chapter presents several avenues for future research.

Chapter 2: Literature Review

2.1 A History of Recommendation Systems

This chapter is focused on providing a history of the field and introducing the historic problems and motivations which shaped its development and led to the formation of the personalisation problem with which this thesis is concerned.

This chapter consists of the following sections:

- Section 1 describes the origins of recommender systems, introduces the various types of recommender system and explains the origins of the personalisation problem
- Section 2 reviews the state of commercial systems and how they have impacted on the personalisation problem
- Section 3 defines the personalisation problem
- Section 4 presents the overarching research objective and constituent questions for this thesis

2.1.1 The Origins of Recommender Systems

2.1.1.1 Relationship to information retrieval

As an academic discipline, recommendation systems are said to have emerged from the field of information retrieval in the early 1990s (Ekstrand, Riedl, & Konstan, 2011). It is typically framed as an inverse field of study which operates on the mirrored assumptions and aims of information retrieval (Ekstrand et al., 2011). Information retrieval endeavours to deliver accurate responses when presented with a multitude of diverse queries against a relatively static content base. By contrast, it is often thought in recommendation systems that the content base is constantly changing whilst the tastes and preferences of users remain relatively stable, only changing slowly over time (Ekstrand et al., 2011). Additionally, the directionality of information flow in recommendation systems runs counter to that of information retrieval. Rather than the user requesting content from the system, the system suggests content to the user. Throughout the course of this thesis, the validity of these assumptions will be examined in the context of modern music recommendation systems.

2.1.1.2 Manual Recommender Systems

The first manual recommender systems to emerge in the early 1990s were little more than enhanced mailing lists and bore little resemblance to the complex automated systems used today. Nonetheless, understanding how and why these early systems evolved into the automated systems of today is essential to fully comprehend the limitations of modern systems.

These manual systems aimed to enable users to express an interest in receiving content that had been viewed and/or rated or labelled by other users.

One of the earliest examples of such a system was the Tapestry mail filtering system developed by Xerox. Users of the system were able to filter content based on other users' impressions of that content. The system even had a query language enabling advanced users to construct complex queries like the one below:

(m.sender = Smith OR m.date < April 15, 1991) AND m.subject LIKE %Tapestry%.

This query selects messages that were either from Smith or else sent before April 15, and whose subject field included the word Tapestry (Goldberg, Nichols, Oki, & Terry, 1992).

Another manual system that emerged around this time was the active collaborative filtering system developed by David Maltz and Kate Ehrlich (Maltz & Ehrlich, 1995). This system allowed users to distribute interesting documents by sharing pointers to those documents with others. The most interesting feature of this system was the way it supported the emergence of 'expert nodes' in a community. Within a network of colleagues, it allowed valuable trusted curators to emerge and be followed by other colleagues.

Whilst these systems were powerful, it was quickly found that they became impractical to use efficiently across medium to large scale infrastructures with a lot of content and many users (Ekstrand et al., 2011). This was largely because these systems required users to be familiar with both the content and the other users in the system. It also required users to have an awareness of one other's tastes in order to retrieve useful recommendations. Such a requirement becomes unrealistic once systems contain hundreds or even thousands of users and items.

The realization of this problem led to the development of modern automated recommendation systems. It is these systems that the term recommender system will be used to denote from here onwards in this thesis.

2.1.2 The Evolution of Automated Systems: Types of Recommender

The development of automated recommender systems began in the early to mid-1990s with the invention of automated collaborative filtering followed shortly by the development of content-based filtering. It is interesting to note that the first collaborative systems emerged from academia whilst the first content-based systems emerged from industry. By the mid-2000s, encouraged by industry, researchers begun to explore a third category of recommender, now known as hybrid recommenders, which sought to improve recommendations by combining multiple algorithmic approaches. Today automated recommender systems can broadly be divided into these three main categories: collaborative filtering, content-based and hybrid systems.

Each of these categories is introduced below along with the historic problems that motivated their development. Within each category, the most significant algorithmic approaches are explained with small scale examples using the fictitious characters Alice, Bob, Carol, Dave and Erin to demonstrate how they work.

2.1.2.1 Collaborative filtering

At their core, collaborative filtering systems work by matching content to users based on how it has been perceived by other users with similar tastes or preferences (Ekstrand et al., 2011). However, as the field of recommender systems has matured, several interpretations and distinct approaches to implementing this core idea have been developed making it necessary to identify sub-categories of collaborative recommenders.

At the top level, collaborative filtering recommendation systems can be divided into two categories; memory-based and model-based (Breese, Heckerman, & Kadie, 1998). Unfortunately, it has long been recognized that multiple subtly varying definitions exist within the literature for these terms (Breese et al., 1998). In recent years, however, the following formulation by Badrul Sarwar seems to be widely accepted as a de facto standard:

> *Memory-based Collaborative Filtering Algorithms.*
>
> *Memory-based algorithms utilize the entire user-item database to generate a prediction. These systems employ statistical techniques to find a set of users, known as neighbors, that have a history of agreeing with the target user (i.e., they either rate different items similarly or they tend to buy [sic] similar set of items). Once a neighborhood of users is formed, these systems use different algorithms to combine the preferences of neighbors to produce a prediction or Top-N recommendation for the active user. The techniques, also known as nearest neighbor or user-based collaborative filtering, are more popular and widely used in practice.*
>
> *Model-based based Collaborative Filtering Algorithms.*
>
> *Model-based collaborative filtering algorithms provide item recommendation by first developing a model of user ratings. Algorithms in this category take a probabilistic approach and envision the collaborative filtering process as computing the expected value of a user prediction, given his/her ratings on other items. The model building process is performed by different machine learning algorithms such as Bayesian network, clustering, [sic] and rule-based approaches. (Sarwar, Karypis, Konstan, & Riedl, 2001)*

More formally, memory-based recommender systems can be defined by the following equation:

$$r_{c,s} = \underset{c' \epsilon \hat{C}}{aggr\, r\, c', s,}$$

(Adomavicius & Tuzhilin, 2005)

Where $r_{c,s}$ denotes the predicted rating for the active or target user c for some item s. The value of $r_{c,s}$ is calculated by running an aggregate function $aggr$ across all the ratings r for item s by those users c' who belong to the set of users $\hat{C}$ most similar to c. A range of aggregate functions can be used but the most common approach is to use a weighted sum (Adomavicius & Tuzhilin, 2005). A weighted sum approach dictates that the more similar a user c' is to the target user c, based on some similarity metric like Pearson's correlation, the more heavily weighted their result becomes in calculating the predicted result $r_{c,s}$.

2.1.2.1.1 User-user or k-NN nearest neighbour collaborative filtering

The first automated collaborative filtering systems employed a relatively simple memory-based technique known as User-user or k-NN nearest neighbour collaborative filtering (Ekstrand et al., 2011). The first known system to employ this approach was the GroupLens Usenet article recommender (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994) (Resnick et al., 1994) (Konstan et al., 1997). Other systems around the same time which also used this approach are the Ringo music recommender (Shardanand & Maes, 1995) and the BellCore video recommender (W. Hill, Stead, Rosenstein, & Furnas, 1995).

In user-user collaborative filtering, a similarity function is used to identify sets of users (known as neighbours) who have demonstrated similar tastes preferences (i.e. they have given similar ratings to the same items). For a given user, predictions are then created for the items they have yet to rate by taking a weighted average of the ratings supplied for those items by other users of the system. The ratings provided by those users judged to be most similar to the target user will be weighted more highly in the prediction generation process. These predicted ratings will then be ranked to generate a recommendation list for the target user.

To better understand how this works, consider a small music recommendation system in which the users rate the items they like using a scale of 1-5 stars. The users and items in these systems can be represented in the ratings matrix below:

	Tears In Heaven	Summertime	Hey Jude	Stairway To Heaven	La vie en rose
Alice	3		3	4	1
Bob	4	4	2		5
Carole	5	5		3	4
Dave	3	5	2	1	4
Erin		1	4	4	1

Figure 1: User ratings matrix

Imagine you wanted to discover if 'Summertime' might make a good recommendation for Alice using a user-user K-NN approach. The first step would be to decide on your configuration, i.e. which neighbourhood size, similarity function and average weighting you would like to use. For this example, we will keep things simple and choose the following basic parameters (more details on the benefits and limitations of different similarity and weighting options is provided after the example):

Neighbourhood size: 3
Similarity Function: Pearson's standard correlation
Average Weighting: Offset mean

Having chosen these parameters, the next step is to compute the similarity between each pair of users using Pearson's correlation. See results below:

User pairs	Pearson's Correlation
Alice - Bob	-0.755929
Alice - Carole	-0.327327
Alice - Dave	-0.923381
Alice - Erin	0.944911
Bob - Carole	-1.000000
Bob - Dave	0.718185
Bob - Erin	-0.944911
Carole - Dave	0.764471
Carole - Erin	-0.866025
Dave - Erin	-0.948683

Figure 2: Pearson's correlation scores

After calculating the similarity between users, the next step is to use the offset mean to calculate a series of predictions of each user's ratings for the items that they have not yet encountered. The following equation can be used to compute the offset mean:

$$p_{u,i} = \bar{r}_u + \frac{\sum_{u' \in N} s(u,u')(r_{u',i} - \bar{r}_{u'})}{\sum_{u' \in N} |s(u,u')|}$$

(Ekstrand et al., 2011)

The constituent parts of this equation are defined in the tables below:

Term	Explanation
u	The user we wish to compute a prediction for.
i	The item we want to produce a predicted rating for.
P ui	The predicted value for user u for item i
ru	The mean average of user u ratings.
N	The set of n neighbouring users.
u’	A user other than u belonging to the set of neighbouring users N
s(u,u’)	The similarity between the chosen user u and some other user u’.
Ru’,i – ru’	The mean rating for a user u’ is subtracted from their rating for a specific item i to obtain a normalised rating of item i for that user.
|s(u’,u)|	The magnitude or absolute similarity between the given user u and some other user u’ in the set of neighbouring users.
Top of equation	The sum of the similarity rating for user u and neighbouring user u’ multiplied by normalised rating for item i by user u’ for all u’.
Bottom of equation	The sum of the absolute similarity between user u and every u’ belongs to the set of neighbouring users.

Figure 3: Offset mean equation components

Using this equation, we can complete the ratings table shown above to include predicted values for all users for all items they have not yet rated producing the table below:

	Tears In Heaven	Summertime	Hey Jude	Stairway To Heaven	La vie en rose
Alice	4	2	3	3	1
Bob	4	4	2	2	5
Carole	5	5	3	3	4
Dave	3	5	2	1	4
Erin	2	1	4	4	1

Figure 4: Completed Ratings Matrix

The weighted average table above shows that Alice would likely give a rating of 2 for ‘Summertime’. Therefore, it would not appear to be a good recommendation for her.

2.1.2.1.1.1Similarity Functions

The example above used the Pearson’s correlation coefficient to gauge the similarity between users. In this section, the benefits and limitations of this approach are explored in more detail and several other popular similarity metrics are introduced.

2.1.2.1.1.1.1 Pearson’s Correlation

One of the most common similarity metrics used is Pearson’s correlation coefficient. Examples of user-user CF systems that use this method include the GroupLens Usenet recommender (Resnick et al., 1994) and the BellCore video recommender (W. Hill et al., 1995).

In the context of CF systems, Pearson’s correlation is calculated by first computing the sum of the product of the difference between a user’s rating r for item i and their mean rating value $\bar{r}$ for two users u and v for all items i belonging to the intersection of the set of items rated by u, Iu and by v, Iv. This is then divided by the product of the square root of the square of the sum of the difference between a user’s rating for i and mean rating $\bar{r}$ for both users for all items i belonging to the intersection of the set of items rated by u, Iu and by v, Iv. Formally this calculation is expressed as:

$$s(u,v) = \frac{\sum_{i\in I_u \cap I_v}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i\in I_u \cap I_v}(r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_{i\in I_u \cap I_v}(r_{v,i} - \bar{r}_v)^2}}$$

One of the problems with this method is that it typically produces artificially high similarity estimates for users who have only rated a few items in common (J. L. Herlocker, Konstan, Borchers, & Riedl, 1999; J. Herlocker, Konstan, & Riedl, 2002). In the following papers this issue was addressed by scaling the final similarity estimates whenever a user pair were found to have less than a certain number (often 50) of mutually rated items (J. Herlocker et al., 2002; J. L. Herlocker et al., 1999).

Another variant on this similarity measure proposed by Shardanand and Maes is known as Pearson's constraint correlation. Unlike the standard formulation which computes relative correlations, this method computes absolute correlation values. A constrained Pearson's correlation method was used in the Ringo music recommender (Shardanand & Maes, 1995). It is calculated by substituting the absolute average r_z for an individual user's mean ratings in the equation:

$$s(u,v) = \frac{\sum_{i \in I_u \cap I_v} (r_{u,i} - r_z)(r_{v,i} - r_z)}{\sqrt{\sum_{i \in I_u \cap I_v} (r_{u,i} - r_z)^2} \sqrt{\sum_{i \in I_u \cap I_v} (r_{v,i} - r_z)^2}}$$

It should be noted that the threshold dampening method being applied here could be used in other similarity functions like cosine similarity but this has not been widely researched (Ekstrand et al., 2011).

2.1.2.1.1.1.2 Spearman's Correlation

Another lesser-used correlation metric for CF systems is Spearman's rank correlation coefficient. Where Pearson's correlation measures the linear relationship between variables, Spearman's measures monotonic relationships, i.e. relationships that either continually increase or decrease. Generally, this tends to be useful for ordinal data where the values have no external meaning beyond their use to rank data, e.g. a likeability scale from 1-5.

To calculate the similarity between two users of a CF system using Spearman's correlation, you would use the same equation as for Pearson's above substituting in a ranked list of each user's ratings, values for the actual rating values they provided. The highest-rated item for each would be replaced with a rank of 1. Where multiple items have the same rating, they would be given the average rank for their position.

2.1.2.1.1.1.3 Cosine Similarity

Although more popular in content-based and hybrid systems, cosine similarity is another method which can be used in collaborative systems. Unlike the other similarity functions mentioned, cosine similarity is based on linear algebra rather than statistical measures of correlation. Users are represented as |I|-dimensional vectors in a ratings state-space. A target user's neighbours are then identified by taking the cosine of the angles formed between their vector and other users' vectors. The smaller the angle, the closer a user's tastes are to the target user.

Cosine similarity is computed for two users by taking the dot product of their rating vectors and dividing by the product of their Euclidean norms. Formally this is written as:

$$s(u,v) = \frac{r_u \cdot r_v}{||r_u||r_v||2}$$

As Michael Ekstrand explains, unrated items are given a rating of 0 which means that they drop out of the numerator (Ekstrand et al., 2011). Additionally, he goes on to state that “if the users’ mean baseline is subtracted from the ratings before computing the similarity, cosine similarity is equivalent to Pearson correlation when the users have rated the same set of items and decreases as $\frac{|I_u \cap\ I_v|^2}{|I_u|I_v|}$ decreases”.

2.1.2.1.1.1.4 Prediction Functions

Once a neighbourhood of similar users has been identified, the next step in user-user collaborative filtering is to produce a series of predicted ratings for each user for the items that they have not yet rated. In early systems like Ringo, predictions were computed by simply taking an average of the rankings given by a user's neighbours. (Shardanand & Maes, 1995). Another early method included using multivariate regression on user neighbourhoods to generate predictions (W. Hill et al., 1995).

It was quickly discovered, however, that predicted ratings could be improved by weighting how important a user's rating for an item should be in producing a rating prediction for that item for a given target user (Ekstrand et al., 2011). This led to the development of weighted averaging which is by far the most popular prediction function as it is simple to calculate and has proven to work very effectively in practice (Ekstrand et al., 2011). Furthermore, weighted averaging is the only prediction function that can be shown to be consistent with social choice theory (Pennock, Horvitz, & Giles, 2000a).

A weighted average is computed by multiplying each neighbouring user's normalised rating for an item by the similarity score they were given against the target user. The weighted ratings produced by all neighbours in the target user's neighbourhood are then averaged to produce a prediction. Formally this is calculated as follows:

$$p_{u,i} = \bar{r}_u + \frac{\sum_{u' \in \mathrm{N}} s(u,u')\left(r_{u',i} - \bar{r}_{u'}\right)}{\sum_{u' \in \mathrm{N}} |s(u,u')|}$$

In the equation above, users' mean ratings r_u are subtracted from their item ratings $r_{u',i}$ to compensate for any variation between users in their use of the rating scale, i.e. to correct for instances where users tend to rate higher than most, even on the content they dislike. The equation above can be further improved by normalising users' ratings by converting them into z-scores. This has the advantage over the method above of also compensating for instances where users have above average spread in the rating tendencies, not just higher than average ratings overall (J. Herlocker et al., 2002). The equation for this further optimised weighting function is:

$$p_{u,i} = \bar{r}_u + \sigma_u \frac{\sum_{u' \in \mathrm{N}} s(u,u')\left(r_{u',i} - \bar{r}_{u'}\right)/\sigma_u{}'}{\sum_{u' \in \mathrm{N}} |s(u,u')|}$$

User-user collaborative filtering is very effective for small userbases (with a few hundred users) with active members who gave a lot of ratings (Linden, Smith, & York, 2003). However, it is very computationally expensive and suffers from several data-sparsity issues (Linden et al., 2003). For instance, when systems have a rich set of items but comparatively few users, it is difficult to build a similarity measure between users as they are unlikely to have rated a sufficient subset of the same items to produce an accurate similarity score. Additionally, whenever a user-user system fluctuates in popularity and more users join or leave the system, the entire similarity matrix between users has to be recalculated which is often computationally expensive.

2.1.2.1.2 Item-item collaborative filtering

Recognition of these scalability problems led to the development of another type of memory-based collaborative filtering system known as Item-item based collaborative filtering. Item-item collaborative filtering was first pioneered by Amazon (Linden et al., 2003) in 1998 as a means of efficiently producing product recommendations to customers (Linden, Jacobi, & Benson, 2001).

At its core, Item-item collaborative filtering uses the same nearest neighbour algorithms as user-user systems but modifies them to identify neighbourhoods of similar items rather than neighbourhoods of similar users.

To an extent, Item-item collaborative filtering, although still a memory-based technique in its purest form, can be seen as an early move towards model-based recommendation systems which attempt to separate the task of categorising content and/or users from the task of producing recommendations.

Item-item collaborative filtering can be explained by considering the same music recommendation example used to explain user-user collaborative filtering. As with generating a user-user system, the first step is determining the recommender system parameters. For continuity, we will use the same parameters as we used in the user-user example shown below:

Neighbourhood size: 3
Similarity Function: Pearson’s standard correlation
Average Weighting: Offset mean

It is important to note that in this example our neighbourhood parameter refers to neighbourhoods of items, not groups of users. Having chosen the parameter, the next step is to calculate the similarity between each pair of items using our chosen similarity function. The results of this calculation are shown in the table below:

Item pairs	Pearson's Correlation
Tears in Heaven - Summertime	0.000000
Tears in heaven - Hey Jude	-0.500000
Tears in heaven - Stairway to heaven	0.188982
Tears in heaven - La vie en rose	0.502519
Summertime - Hey Jude	-0.970725
Summertime - Stairway to heaven	-0.755929
Summertime - La vie en rose	0.864159
Hey Jude - Stairway to heaven	0.866025
Hey Jude - La vie en rose	-0.886621
Stairway to heaven - La vie en rose	-0.816497

Figure 5: Item-item Pearson's Ratings

Having calculated the similarity between tracks, the next step is to produce a set of recommendations for Alice by obtaining predicted ratings for the items Alice has not yet listened to. To achieve this, we can use the item similarity scores above to calculate the weighted mean average of Alice’s ratings for the 3 most similar tracks to each shown below.

	Tears In Heaven	Summertime	Hey Jude	Stairway To Heaven	La vie en rose
Alice	4	2	3	3	1
Bob	4	4	2	3	5
Carole	5	5	3	3	4
Dave	3	5	2	1	4
Erin	2	1	4	4	1

Figure 6: Completed Item-item matrix

As before, Alice is only predicted to give ‘Summertime’ a rating of 2, making it a poor recommendation choice.

An important distinction between the user-user approach and Item-item approach which can be seen from our examples is that Item-item filtering cleanly separates the tasks of computing similarity and producing predictive ratings to reveal recommendations. One advantage this approach has over user-user filtering is that we do not have to recalculate our similarity matrix every time a new user is added.

Furthermore, we can obtain recommendations much faster as we only need to consider a given individual’s rating when attempting to produce predictions for them. Cleanly separating the similarity scoring stage from the recommendation stage in this way has clear benefits for speed and scalability. The computationally expensive similarity scoring step which requires an n by m matrix calculation can be computed offline and only needs recalculating infrequently when new items are added. The less computationally expensive step of producing predictions can be done quickly in real-time as it only needs to consider a single user's ratings which are likely to be a comparatively small dataset.

Essentially, Item-item collaborative filtering systems can precompute the similarity between items and then in real-time perform the much simpler task of looking up those items which are most similar to a given user’s most highly rated items. This can drastically increase the speed and efficiency of the recommendation system (Linden et al., 2003).

Additionally, Item-item based collaborative filtering can perform better than user-user systems when a given user has only rated a small number of items. This is because the item-based recommender only needs to select items similar to those the user has rated to produce a recommendation. By contrast, a user-user system has to first identify similar users to the target user by comparing their item ratings against other users. Obviously, when the target user has only rated a few items this task becomes difficult as little can be inferred about the target user’s global tastes or their similarity in taste to other users from only a handful of ratings.

The first Item-item style recommenders like the kind devised by Amazon could sometimes suffer from a problem known as overfitting. By 2005 a new class of recommenders was introduced specifically to deal with this problem known as slope one recommenders (Lemire & Maclachlan, 2005).

Whilst Item-item scaled better than user-user systems with regard to large userbases, they still failed to scale well with regards to large item catalogues. Additionally, they still faced several data sparsity issues such as the cold-start problem for items and the Long Tail problem.

Memory-based recommenders have been very popular as they tend to be easy to implement and don’t require an in-depth knowledge of the content to make recommendations (Su & Khoshgoftaar, 2009).

2.1.2.1.3 Model-based collaborative filtering

Model-based collaborative filtering systems attempt to improve upon the accuracy and scalability of the aforementioned memory-based techniques by using users' ratings to produce models of types of users or items which can then be used to generate recommendations. In model-based collaborative filtering, the recommendation problem is framed as the issue of constructing a model which can successfully predict how a user might rate an item given their past rating preferences.

Although a vast number of different model techniques have been tried, most model-based recommenders can be shown to fit into one of two categories: Bayesian classifiers (Condliff, Lewis, Madigan, & Posse, 1999) (Robles, Larranaga, Menasalvas, Pérez, & Herves, 2003; Su & Khoshgoftaar, 2006) (Miyahara & Pazzani, 2000), and Markov Decision Process classifiers (Shani, Brafman, & Heckerman, 2002).

Let's return to our music recommendation example to explore each of these model-based approaches to collaborative filtering works.

2.1.2.1.3.1Markov Decision Process Classifiers

Markov decision process (MDP) classifiers use maximal reward-based reinforcement learning techniques to solve optimisation problems. A MDP classifier is a five-tuple <S,A, Tr, R, disc> system where S is a set of states, A is a set of actions, Tr is a state transition function, R is a rewards function and disc is a discount factor placed on future rewards.

The set of states seeks to encapsulate all relevant information about the world, whilst actions work to trigger transitions between states (Shani et al., 2002). The tr function provides a probability distribution for all state-action pairings such that it is possible for some state-action pairs to calculate the probability that any state will be transitioned to. For instance, tr(s,a,s') provides the probability that state s` will be reached if action a is performed from state s. The reward function provides an immediate numeric reward or punishment often in the range of -1 – 1 for being in a given state. Finally, the discount factor provides a means of weighting future rewards by a small increment. This allows for the modelling of delayed gratification; a small or large discount factor is used respectively when a model does or does not care about the future.

Viewed as an optimisation problem, an MDP music recommendation service might be constructed as follows. Let S capture the set of states representing all possible variations of the last three tracks a user could have purchased. Let A capture the set of recommendation actions available, that is actions of representing any possible track. Tr(s,a,s`) then provides the probability that given a user's recent purchase history s, a certain track (shown as the last item in s`) is purchased when a particular track is recommended via action a. R can be viewed as the function which returns the profit value for selling a track and can be set to a low value to account for the fact the recommender should look to maximise net profit. Once this model has been constructed, a variety of maximal optimisation techniques such as value iteration or policy iteration can be used to produce an optimal policy, which is to say a set of optimal tracks to recommend given any potential purchase history.

In 'An MDP-based recommender system' Shani writes that, "In an MDP, the decision-maker's goal is to behave so that some function of its reward stream is maximized-typically the average or discounted average reward. An optimal solution to the MDP is such a maximizing behaviour". The important characteristic to note is that it is impossible to model the recommendation process using MDP without seeking to optimise some reward.

This is a problem because the reward factor most easily quantified and typically chosen by commercial entities is net profit. In seeking only to maximise net profit, a recommender system is indifferent as to whether it sells one item to 1000 users or 1000 items to one user. Its optimisations are solely concerned with profit maximisation and as such it will prioritise items that can sell for a greater price over items that may be more appropriate or favourably received by users. In an extreme case, an MDP recommender might determine the optimal strategy is to recommend a set of exceptionally expensive items which 99% of users will dislike or cannot afford but which a small 1% will purchase. The profit maximising strategy here leads most users to be dissatisfied with the service. However, the company using the service is likely to remain happy even if it loses all but that 1% of clients as that 1% results in maximising the company's profit margins.

Each of the model-based approaches shown above is a rich area of interest in its own right, each possessing unique strengths and limitations. However, reviewing them in detail remains outside of the scope of this thesis. Collectively, however, model-based recommenders, regardless of type, can be acknowledged broadly speaking to have the following advantages when contrasted with pure memory-based systems: scalability, speed and avoiding overfitting.

Unfortunately, model-based systems also share several limitations. Whilst they often improve on speed by having pre-computed models, they often lack generality or flexibility. This can result in their producing lower quality recommendations as they fail to adapt or account for pervasive global changes in user behaviour which were not encoded in the original attempt to model a given system's users.

Additionally, they can suffer from transparency and synonymy problems which lead to a decrease in personalisation and user satisfaction. This is explored in greater depth in the next chapters, as it emerges as one of the themes in the study presented in chapter 3.

2.1.2.2 Content-based

Content-based recommenders represented an entirely different approach to the recommendation problem. In content-based recommenders, users and items are typically represented as keyword vectors in a state space. Recommendations are created for a given user by calculating the Euclidean distance between their vector and surrounding item vectors within the state space. Item vectors with the shortest Euclidean distance from the target user are offered as likely recommendations for that user. Examples of content-based systems can be found in the following papers (Balabanović & Shoham, 1997; Ben Schafer, Konstan, & Riedl, 2001; Popescul, Pennock, & Lawrence, 2001; Smyth, 2007).

Content-based filtering has its roots very much within information retrieval where practices such as term frequency inverse document frequency (TFIDF) had been used to model archived documents as early as 1972 (Spärck Jones, 1972).

Unlike collaborative systems, content-based systems manage to avoid a wide array of data sparsity problems.

One of the limitations with many content-based systems is their lack of transparency. Often, they employ black-box machine learning algorithms, like those discussed in model-based collaborative filtering, to perform feature extraction and clustering. This has the unfortunate consequence that it can be very difficult to understand why particular users are receiving particular recommendations. This, in turn, makes it harder to replicate conditions under which recommendations are favourably received, and conversely to mitigate conditions under which recommendations are poorly received. A bit later on in this chapter this problem is explored in greater depth to understand its role as a contributing component to the personalisation problem.

Content-based recommenders typically take much more time to construct than their collaborative counterparts and are typically capable of recommending a significantly reduced set of items. The main reasons for this are the time it takes to determine how the data should be modelled and the additional time it takes to encode the items to be recommended.

Speed and ease of construction are probably the most commonly cited advantages of collaborative systems over content-based systems. The reason collaborative filtering systems are often easier and quicker to construct is that they are content-agnostic. The developer of a memory-based collaborative filtering system does not have to understand or model the content that their system is supposed to be recommending.

2.1.2.3 Hybrid Systems

By the mid-2000s, researchers had begun to observe that each of the fundamental approaches to the recommendation problem reviewed above had its own unique strengths and limitations (Balabanović & Shoham, 1997; Burke, 2002; Torres, McNee, Abel, Konstan, & Riedl, 2004). At the same time, the cold-start problem was taking centre stage as the dominant problem within the field. To address it, researchers started to experiment with combining different recommender systems to produce hybrid systems that attempted to combine the strengths of their component systems whilst mitigating their weaknesses.

Providing a standardised definition of what it takes for something to be considered a hybrid system can be difficult since the term has been used to label different categories of systems over time. The earliest systems that were labelled as hybrid systems tended to be collaborative filtering systems which combined memory and model-based techniques in an aggregated way to produce recommendations. Hybrid systems of this sort can be found in the following papers: (0012, de Vries, & Reinders, 2006; Al-Shamri & Bharadwaj, 2007; Pennock, Horvitz, Lawrence, & Giles, 2000b).

More recently, the term hybrid system has increasingly been used to identify systems which employ multiple types of recommenders and combine both collaborative filtering and content-based systems to produce recommendations. Hybrid systems of this sort can be found in the following papers: (Degemmis, Lops, & Semeraro, 2007; Iaquinta, Gentile, Lops, de Gemmis, & Semeraro, 2007; Jalali, Gholizadeh, & Hashemi Golpayegani, 2014; Lekakos & Caravelas, 2008) (Ghazanfar & Prugel-Bennett, 2010).

By 2015 it had become widely accepted within the community that the cold-start problem could be solved or at least sufficiently mitigated by using hybrid systems (Isinkaye, Folajimi, & Ojokoh, 2015).

2.1.2.4 Optimisations & Dimensionality Reduction

As recommender systems grew in size and complexity during the 2000's, optimisation techniques started to be applied to enable these systems to function in practical applications. Two of the most common optimisation techniques are described below and their impact on personalisation is considered.

2.1.2.4.1 Latent Semantic Analysis & Single Value Decomposition

Latent semantic analysis is a popular technique within information retrieval and natural language processing for capturing the semantics of a document (Deerwester, Dumais, Landauer, & Furnas, 1990). As a technique within information retrieval, it is commonly thought to have been invented by Scott Deerwest & Susan Dumais in the late 1980s (although its roots can be traced back much further to the 1930s and a statistical technique known as correspondence analysis).

Within the domain of information retrieval and recommendation systems, it is often used to reduce the feature state-space used to represent items and/or users in vector model-based systems, e.g. (Billsus & Pazzani, 1998; Hofmann, 2004). In this context, it is typically referred to as Latent Semantic Indexing (LSI).

It was pioneered by Deerwest primarily as a means of combating "a fundamental problem that plagues existing retrieval techniques" (Deerwester et al., 1990). Specifically, the problem of users attempting to retrieve information by using high-level conceptual search-terms in a system that has no understanding of the underlying semantics of the information being searched. As Deerwest writes:

> *The problem is that users want to retrieve on the basis of conceptual content, and individual words provide unreliable evidence about the conceptual topic or meaning of a document. There are usually many ways to express a given concept, so the literal terms in a user's query may not match those of a relevant document. (Deerwester et al., 1990)*

LSI works by first constructing a term-document matrix X in which each entry indicates the number of times a given word occurs in a particular document. When applied to recommender systems, the term-document matrix is typically a feature-entity matrix where entities represent users and/or items and features represent the words used to describe those entities.

Once this matrix has been constructed, an algebraic method known as single value decomposition is then applied to the matrix to produce three separate matrices T, S and PT.

Formally:

$X=T*S*P_{T}$

Although LSI can be computationally expensive to perform, once it has been run on a sufficiently representative subset of the user-item state space for a given recommendation application, it is usually unnecessary to recompute. Furthermore, once computed, the resultant reduced dimension user-item state space is typically constructed to be substantially smaller (approximately 50 times smaller in most language applications) thereby making any subsequent model-based recommendation algorithms much faster to compute. This is perhaps the biggest advantage of using LSI in recommender systems as it enables increased throughput and allows systems to function efficiently whilst handling a far greater set of users and items than would otherwise be practical.

A limitation of LSI, however, is that it inevitably simplifies the state-space and reduces the granular detail or resolution at which items and users are represented. This in turn can have a negative impact on both, in terms of how accurately users and items are represented within a system and how tailored or personal recommendations can be. With reduced granularity, there is also scope for problems like the Long Tail problem to emerge whereby users with niche interests find their preferred items and nuanced conceptual tastes to be poorly represented in the new matrix, thereby reducing the likelihood of them receiving accurate and appropriate recommendations.

Another disadvantage of using LSI is that it often results in a reduced level of transparency as the resultant feature-entity matrix produced by LSI is usually difficult to interpret in any easily understandable sense.

Examples of latent semantic models can be found in the following papers: (Landauer, Littman, Bell Communications Research, Inc., 1994)

2.1.2.4.2 Clustering

Another popular approach towards dimensionality reduction is known as clustering. Clustering is often applied to collaborative filtering-based solutions to avoid computing similarity metrics prediction values across the entire user-item matrix which is likely to be sparsely populated. The accuracy and speed of computing recommendations can be improved by grouping users into subsets reflecting those who have viewed similar items and by grouping items into subsets reflecting those that have been liked by the same set of users (Ungar & P Foster, 2000). Examples of recommenders which use clustering include (Basu, Hirsh, & Cohen, 1998; Breese et al., 1998; Ungar & P Foster, 2000) (Chee, Han, & Wang, 2001)

Providing a complete analysis and overview of all available clustering methodologies is outside of the scope of this thesis. However, for the reader who wishes to know more in this area, a good overview of clustering techniques can be found here: (Mamunur Rashid, Karypis, & Riedl, 2006)

2.2 Recommender Revival & The Origins of the Personalisation Problem

It is commonly acknowledged that the launch of the Netflix Challenge in October 2006 directly contributed to the renewed commercial and academic interest recommender systems research has seen over the past decade (Adomavicius & Kwon, 2011; Ekstrand et al., 2011; Jia Rongfei et al., 2007). In the three years for which competition ran, exponential leaps in progress were made in many areas of recommender research including scalability, hybrid system development and user modelling.

Unfortunately, the narrow aim of the Netflix challenge (improving upon Netflix's own predictive rating accuracy by at least 10%) reinforced the historic accuracy metric against which recommender systems are frequently accessed. Consequently, much of the research in recommender systems over the past decade has focused on predictive accuracy as the sole metric for success.

Initially, it might not seem clear what the problem is. After all, assessing a predictive system by measuring how often it correctly identified items that users purchased or interacted with makes sense. The problem emerges in the realisation that accuracy can come at the expense of utility. There is an old joke in the recommender community which highlights this issue. A systems developer walks into his local supermarket and informs the manager that he can build them a product recommender which is better and more accurate than anything on the market. What's more, he can build this system in 5 minutes. He then proceeds to create a system which simply recommends that everyone between the age of 18-69 buys bread and milk on a Sunday. Now, of course, this is likely to be highly accurate since the vast majority of adults doing supermarket shopping on a Sunday will buy bread and milk for the week, but it is certainly not useful as people were doing this already.

This notion of accuracy as a false or incomplete metric has emerged in several areas of recommender research but is most vocally discussed within the domain of music recommenders. It has been argued that the tendency of automated recommender systems to recommend safe obvious mainstream content is contributing to the extinction of entire more niche genres like Jazz and Bluegrass (Donnat, 2018; "Slave to the algorithm? How music fans can reclaim their playlists from Spotify," 2016). This increased frustration and the public perception of recommenders as just recommending Taylor Swift and Red Hot Chilli Peppers has motivated a shift in the focus of music recommender research over the last three years towards non-obvious or serendipitous recommenders.

The goal for this new breed of recommender is no longer basic predictive accuracy, but personalisation gauged via user satisfaction. One crude but automated way that user satisfaction might be assessed is the number of times a user skips recommendations.

2.3 Other Core Areas of Research

2.3.1 Context-awareness

As researchers sought to address the cold-start problem, they looked at different ways to accumulate information on users and their taste preferences to provide them with accurate recommendations even before they had an established usage history (Adomavicius & Tuzhilin, 2011). This led to a new set of recommenders being developed which have become known as context-aware recommenders. These systems incorporate information about the user such as the activity they are doing or their LastFM profile or their web history in order to build a more accurate model of the user and ascertain their taste preferences.

In the definitions section of chapter 3 we revisit the notion of context and context-aware recommenders and emphasis that they are not of central importance to this thesis. We included the brief passage to context-aware recommenders above for the sake of completeness and to position our own work as distinct from them.

In this thesis and our own research, we make a point to use the term purpose in our own research to signify an element of context that we wish to focus on whilst distancing our work from previous context-aware research. The justification behind this is that the term context has become overload leading to a confusion between the layman's notion of context and a research definition. The research definition discussed later in chapter 3 is tightly bound to the importance and pursue of predictive accuracy. Context-awareness in this sense is used to signify the inclusion of all and any information that can help better profile a user to produce more accurate recommendations. This is almost the exact opposite of what we are seeking to do. The work in this thesis arguably reduces strict predictive accuracy (by allowing users to interfere with the recommendation process) in an effort to increase personalisation.

2.3.2 Privacy & Risk-awareness

Data sparsity and cold-start problems have dominated the field of recommender systems since the mid-2000s. Consequently, much of the research during this period focused on ways to better model users and gain more knowledge about their preferences in order to provide them with ever more accurate recommendations. This has resulted in many commercial organisations which use recommenders like Netflix and Amazon amassing large scale datasets of detailed information about their customers.

Over the last 10 years, several high profile confidentiality breaches and inappropriate recommendations have prompted researchers to examine the need for companies and creators of recommender systems to be aware of the potential risks involved with owning and using these datasets to make recommendations, and the need to protect users' privacy.

One such case occurred in 2008 when Arvind Narayanan and Vitaly Shmatikov, two researchers at the University of Texas, published a paper detailing how they were able to de-anonymise users from the dataset provided for the Netflix challenge (Narayanan & Shmatikov, 2008).

Another interesting case occurred in 2012 when the US discount store Target made the headlines after its automated product recommender correctly identified that a teenager was pregnant from recent changes in her purchase records. Controversy was caused as the recommender initiated that pregnancy literature and adverts be sent to the teenage girl's home, with the result that her family discovered she was pregnant before she had decided to tell them (K. Hill, 2012).

In September 2017 Amazon was publicly criticised after it was discovered that an item-based product recommender was recommending groups of products for making bombs to users who clicked on certain otherwise innocuous items like ball-bearings and cleaning supplies ("Amazon's algorithm "suggests bomb-making recipes"," 2017).

Cases like the ones mentioned above have led to a new small but growing sub-domain of research within the recommender community focused on risk-aware, privacy-conscious recommenders.

2.4 The State of Commercial Music Recommendation Systems

In this section, we review the current state of commercial music recommendation systems and reflect on how they impact on the personalisation problem.

2.4.1 Pandora

Pandora was the first large industrial organisation to follow a uniquely human-led content-driven approach to music categorisation and recommendation. It rejected the more popular user-based collaborative filtering approach to recommendation on the belief that “each individual has a unique relationship with music – no one else has tastes exactly like yours”. In 2000 it launched the Music Genome Project, to build a large descriptive database of music. Each track in the database was assessed by musicologists and represented using an ontology of over 450 musical characteristics.

2.4.1.1 Music Genome Project

The Music Genome Project makes up the core of Pandora’s recommendation system. Individual tracks are represented at a very fine-grained detailed level as vectors of keywords. These keywords describe the music in several different ways including the presence of different instruments, male and female vocals, genre, loudness etc. A recommendation is formed using an Item-item collaborative filtering approach whereby songs are assessed with regards to their similarity to songs a particular user has already played. The songs with keyword vectors that are very close to the vectors of songs a user has already played and liked are put forward as good recommendations.

Pandora’s approach has several distinct advantages over traditional user-user collaborative filtering systems or even machine learning clustering content-based systems. First, the music is classified by a semantically meaningful and relatable taxonomy which makes it easier to understand why certain tracks get recommended and classified as they do. Second, recommendations are made to users based on their taste preferences alone without attempting to classify them as similar to other users. This has the advantage that recommendations can be more individualistic and more accurately representative of a user’s taste preferences.

One of the major problems with Pandora's Genome Project is scalability. It is probably true that you can get the most accurately representational database using real people to classify your music. However, it is a slow and time-consuming process. The entire Genome Project as of 2016 holds less than 2 million songs. By comparison, Spotify's Echonest library has data on over 20 million songs. The smaller pool of content to be recommended reduces the number and diversity of recommendations which can be made which inevitably decreases the extent to which recommendations can be tailored to individuals. Furthermore, in combating the issue of slow indexing, Pandora has to prioritise the songs it indexes focusing on the most popular songs first as this allows it to cater to the needs of most users to varying degrees. The downside of this is that it increases the likelihood that it will face Long Tail criticisms of not reflecting niche tastes and of making generic and mainstream recommendations because the majority of its indexed content is mainstream popular music. Indeed, it has yet to index any classical music at all, lending credence to the criticism that music recommenders and online stream services are contributing to the extinction of entire genres.

Additionally, there is always the concern that the 450-word ontology developed and used by the musicologists in the project is ill-fitting to categorise the music in a way that reflects how non-expert users might want to categorise it. That is to say, a user might define a playlist or view a collection of songs as similar for entirely personal reasons not relating to formal features about the music or they might prioritise some attributes like genre over other attributes like tempo. This means that each user could likely have their own slightly different assessment of which tracks are similar to other tracks and would hence make a good recommendation. Pandora takes no account of users' unique experiential connection to music and seeks to make its recommendations purely on factual aspects about the music without ever consulting the users to see if these aspects are useful for recommending. Is similarity even necessarily a desired aspect of a playlist? This question will be explored in more detail in the study presented in chapter 3.

Another barrier to personalisation in Pandora's system is that it bases its recommendations purely on similarity and predictive accuracy forming one slowly evolving monolithic profile of a given user. It fails to account for the dynamism of different users' tastes, and how in different circumstances and at different times of day, their taste preferences may vary radically. It also fails to introduce any notion of novelty into its recommendation process meaning people could quickly get frustrated at not being exposed to any new or challenging content which might broaden their musical horizons.

Pandora has also stated that although it has not solved it yet, it recognises that accuracy seems to conflict with surprise whilst both features are desired aspects of personal recommendations.

For its various merits and weaknesses, Pandora’s unique approach to music recommendation does highlight several key considerations for the personalisation problem. It suggests that in order to overcome scalability issues, some combination of human input and machine learning will likely be required. This is evident by its recent efforts to use machine learning techniques to reduce duplicate entries in music catalogues, so reducing the catalogue size, giving the human experts less to index, whilst also decreasing the risk of over-representing/recommending particular tracks. By their own admission, focusing on similarity means that currently most users deepen their musical horizons in a given area rather than broaden their musical horizons. The notion of broadening musical horizons seems to play a particular role in well received personal human-to-human recommendations as is shown in my first study in chapter 3. A final concern for personalisation revealed by reactions to Pandora’s approach is the potential for culturally biasing recommendations by prioritising the most popular western content during indexing, and additionally using western musical terminology and concepts when representing the tracks as feature vectors.

2.4.2 Spotify

Spotify is the largest and probably most well-known commercial organisation in the music recommendation industry. Traditionally it has followed a collaborative filtering approach to music recommendation, but over the last 5 years, it has increasingly been exploring hybrid recommendation approaches and human-in-the-loop based recommendation.

Spotify was founded in April 2006 and officially launched its streaming service in October 2008.

2.4.2.1 Echonest

Echonest is a music intelligence and data platform developed by MIT Media lab and purchased by Spotify in March 2014. One of its core components is a REST accessible database of machine analysed and categorised music represented as keyword vectors. In many respects, Echonest represents the opposite philosophical position to Pandora’s Genome Project. The Echonest project is about using machine learning techniques to fingerprint, identify, classify and recommend music with minimal human intervention.

One of the advantages of Echonest is that, due to its limited reliance on people, it scales very well and can quickly incorporate new content. To date, its core database contains keyword vectors for 20 million songs making it 10 times the size of the Genome Project.

2.4.2.2 Discover Weekly

Discover Weekly is Spotify’s latest recommendation system and its first foray into the world of human-in-the-loop recommendation. Discover Weekly is a system which produces recommendations for users by picking tracks from the manually curated playlists of other users. This has the advantage of incorporating a degree of novelty and non-obviousness into recommendations. This is perhaps the first major commercial attempt to address novelty within recommendation.

Unfortunately, Discover Weekly still fails to account for the dynamic and changing nature of users’ tastes. It still relies on having one slowly evolving but monolithic view of who users are. Furthermore, users have no direct control over what is recommended nor even the ability to feedback on the appropriateness of the recommendations. They are also limited to a single weekly playlist which, although ever-changing, is only at best trying to suit their most general taste preferences rather than catering for their desire to listen to different content at different times throughout the day or when they are with different company or in different settings.

Additionally, the complexity and hybrid nature of Discover Weekly often means it is difficult to see why certain tracks are recommended. This has led to criticisms that the service is convoluted and lacks transparency.

A final limiting factor of Discover Weekly, however, is that its core methodology and philosophy focused on a notion of similarity. Users are presented with what are designed to be novel tracks liked by other users with similar tastes. Whilst this may not be a major problem, it may potentially limit the use of Discover Weekly in aiding users to broaden their musical horizons. It lacks the flexibility to allow them to play a selection of music which is entirely new to them on a whim. Again, as will be seen in chapter 3, this is a real strength of human-to-human recommendations between friends. They can adjust their recommendations to suit a person’s taste preferences in the moment and potentially introduce them to an entirely new genre or collection of music that they otherwise might not have discovered.

2.4.3 LastFM

LastFM was launched in 2002 as an online music database, streaming service and curation platform for storing detailed profiles about people’s listening habits and preferences. It incorporates a collaborative filtering-based recommender engine called Audioscrobbler which began as a computer science project at the University of Southampton.

Since its launch, LastFM has gone through several changes, most notably the discontinuation of its streaming component in 2014. Instead of relying on users’ listening to content with the LastFM website, LastFM now relies on users importing the track titles and metadata associated with the content they listen to. This is either done manually or by using one of the many plugins which allow users to automatically scan and incorporate their listening data from third-party services like Spotify, iTunes and Deezer.

As of 2017, LastFM can be viewed as a rich metadata acquisition platform which can be a useful tool for assessing popularity trends and shifts in global listen patterns. Since dropping its streaming service, however, its contributions within the space of music recommendation and in particular the personalisation problem, have been limited.

2.4.4 Apple Music & Beats Music

Apple Music was first launched in June 2015 as Apple’s first foray into the world of online music streaming and recommendation. From its initial launch, it emphasised the desire to create a highly personalised service.

Apple’s music service was formed after its acquisition of Beats Music, a music stream service created by Dr Dre in 2014.

2.4.4.1 For You

For You is the human-in-the-loop recommendation core of Apple’s music streaming service. Although little is known about precisely how it works, it is reported to combine machine-driven taste preference analysis with human-curated playlists, thereby targeting users with carefully chosen curated content.

Interestingly, Apple has chosen not to alter how much a user appears to like a song based on their tendency to skip that song, “*Skips aren't really taken into account, because there are so many reasons you may skip a song--maybe you're just not in the mood for it right now*”. This raises an interesting question and problem about user behaviour interpretation.

2.5 The Personalisation Problem for Music Recommendation

The first section of this chapter was concerned with introducing the different types of recommender and considering the central problems that shaped the development of the field. Reflecting upon this section, it is clear to see that the core motivational problems have tended to turn on issues of scalability or predictive accuracy. The more significant insight that was repeatedly highlighted throughout the section and that can be made explicit here is that this singular driven pursuit of scale and predictive accuracy came at a cost to personalisation. This is especially clear when one contrasts the high degree of user control and personalisation possible in the manual systems presented at the start of section 1 with some of the content-based systems described at the end of the section where it becomes virtually impossible to see why a specific recommendation is made, let alone purposely influence how recommendations are made.

In section 2 of this chapter, the issue of personalisation was explored specifically by considering how the major commercial music recommender services have developed their products to account for personal preferences to a greater or lesser extent. The two main insights here were that first, within music recommendation, personalisation is seen as something of the holy grail. Virtually all of the commercial outfits reviewed made achieving personalised music recommendations a core part of their agenda. The second interesting insight which appears to conflict with the findings of section one is that all commercial outfits are still viewing predictive accuracy as the answer to the personalisation problem and not a contributing factor. To a greater or lesser extent, they all seem to follow the assumption that personalised recommendations can be made better by more accurately predicting whether a user would like a given song, artist or playlist.

This section defines the personalisation problem, specifically as it impacts music recommendation, by revisiting the central problems and solutions considered in sections 1 and 2. In doing this, it is possible to draw out the subtleties of the problem and to reveal that it is not, in fact, a singular problem. Instead, it is a complex multi-layered issue which has arisen as an unintended consequence of rapid development steered by industry towards a singular objective: accuracy: perceived as the means towards profit.

After defining the problem, this section finishes by presenting the core human-in-the-loop approach which is developed and explored in the subsequent chapters of this thesis as a means of addressing the personalisation problem for music recommendation.

2.5.1 Defining the Problem: ‘The Times They Are a-Changin’3

In section 1.3 the fallacy of pursuing predictive accuracy at all costs was highlighted. However, it remains to be seen which other aspects contribute to the personalisation issue and how specifically this affects music recommendation.

Sections 1 and 2 revealed the following tenets within recommender research from the early 1990s through to the late 2000’s:

1. Recommenders ought to recommend content to users that they would like
2. This is best achieved by recommending items to users that they would rate favourably according to some predictive model designed to reflect users’ taste preferences
3. Better recommendations are thought to be those where the predictive model most closely reflects a user’s response once presented with a recommendation, e.g. the model predicted that Jones would like songs by Bob Dylan and Jones then proceeds to play, purchase or otherwise positively rate songs by Bob Dylan thereby validating the predictive model
4. The major barriers to widespread use and adoption of recommender systems are seen to be either their failure to scale effectively or their poor performance when incorporating new content or users

Obviously, none of these tenets mentions personalisation or user response posing the question as to why it was not considered of core importance then and why it has suddenly become an issue now? Part of the reason for this, as suggested in section 1.2 of this chapter, is that there is a misalignment in the expectations of recommender systems between users and service providers. Users want highly tailored results whilst service providers want to maximise economy of scale and target the most users possible thereby maximising net profit. An individual may prefer a bespoke hand-tailored suit to an off the rack suit, but most high street retailers sell off the rack suits as these will be good enough, if not perfect, for most clients allowing them to increase turnover and maximise profit.

[3] Reference to the Bob Dylan song ‘The Times They Are a-Changin’ released in 1964 from the album of the same name.

Understanding this aspect of the problem helps to reveal why the field of music recommendation has been the first and most vocal field to question the traditional tenets of recommender research and demand that personalisation be considered in a systems design. In straightforward product recommenders, a profit maximising predictive accuracy approach is likely to be reasonably useful and crucially inoffensive. If Amazon's recommender provides a series of recommendations to a user, several of which they have already purchased or own and hence do not need, they will simply ignore those recommendations and move on to any of the recommendations which may be of use to them. By contrast, music recommender systems often make their recommendations more forcefully by starting to play the tracks they recommend until a user opts to skip the track or otherwise stop it from playing. This has several negative consequences. First, it forces users to listen to recommendations making it more difficult for them to avoid or ignore inappropriate or imperfect recommendations, e.g. tracks they already own, are tired of or simply don't like. Second, it seeks by design to make only safe popular recommendations. Again, this is a reasonable thing to do for generic product recommenders where users may want to see a list of possibly useful items. In a sense, this can be thought of as the digital equivalent of placing stamps by envelopes in a store or offering to sell Sellotape when people purchase wrapping paper. It's passive, marginally useful and unobtrusive. A similar safe approach in music recommendation is far more intrusive and less pleasing. Users quickly get tired of being recommended and therefore often played, tracks which they don't want to hear because they are too obvious or popularist. Always playing tracks which feature in the charts would be a fairly safe recommendation approach as the charts are formed from popular music most users of a general music streaming service will like. Such an approach is quickly going to be objected to by users, however, as they grow tired of hearing the same content repeatedly. This line of enquiry is pursued further and backed up in chapter 3 which explores the nature of personalisation in-depth.

Personalisation has also become more of an issue in recent years as many of the earlier problems have been solved in ways which constrain users in their ability to direct how a recommender systems functions for them. To an extent, music recommender systems have become victims of their own success. Initially, they were small systems catering for hundreds of users with only thousands of songs. Today, as shown in section 2, systems have millions of users and songs. Reducing this states space to make systems which can be practically useful to any degree, has necessitated reducing the granularity with which users and content can be represented. Even if accuracy were an acceptable and sufficient sole means of producing the best recommendations, it now cannot be obtained as easily as it could in the earlier smaller systems which could afford to represent users and content in more detail. This point also reveals why the problem of personalisation was not so important in the 1990s and early 2000's since problems of scalability were preventing systems from being used on scales which impacted the level of granularity they could employ to pursue predictive accuracy.

Considering the above characteristics, the personalisation problem for music recommendation can be defined as the issue of balancing recommender system demands and taking account of users' responses to recommendations to produce results which are both accurate and appropriate. The system needs to be capable of accommodating a range of different demands and purposes for seeking recommendations. In the next section, the personalisation issue is deconstructed a little further to reveal the avenues which need to be pursued to address the problem.

2.5.2 The Proposed Solution: A Dynamic Human-in-the-loop Approach

In defining the problem above, it became apparent that modern music recommender systems fail to take account of what users demand of recommender systems. They have become too concerned with tackling technical issues of scalability and delivering a more generic service to as many users as possible. If personalisation is to be viewed as a key goal or a new tenet for the next generation of music recommender systems, then the lack of user control and input needs to be addressed. Throughout this chapter, it has been shown that systems have become less personalised as they have grown in complexity and departed from the vision and goals of the first manual systems. In a manner, the approach suggested by this thesis is to return to these systems, to reintroduce the user to the recommendation process. To put them in the driving seat and allow them to steer how a recommender system views them and their taste preferences.

To be clear, the approach advocated in this thesis is not that we return entirely to manual recommenders; clearly, issues of scale prevent this from being feasible. Rather this thesis argues for a human-in-the-loop approach. Human-in-the-loop systems work by incorporating human practices? into existing automated workflows.

The benefits of incorporating human elements into automated systems have been recognised in a wide range of fields including information science (Pontis et al., 2015), human-computer interaction (Kefalidou & Sharples, 2016) and artificial intelligence (Fiasco, 2018).

The benefits of this approach are starting to be explored already in commercial products like Discover Weekly and Pandora's genome project, but these projects have both sought to involve humans in the curation and modelling portion of the recommendation process and not in the recommending aspect. In this aspect, similarity and predictive accuracy remain unchallenged.

The approach in this thesis is to consider how it might be possible to allow users to interfere with the recommendation process by altering how the systems perceive them. In this way, they can dynamically tailor the types of recommendation they get by changing how the system profiles them. This has the potential to allow users to personalise the recommendations they receive by placing different requirements on a recommender system in different situations, perhaps when they are doing certain activities or are with particular company.

By giving users control of how the system perceives them, it opens up possibilities for individuals to find their own answers and to achieve a balance between obtaining accurate but also novel recommendations. Indeed, a given user might want more or less accurate recommendations under different circumstances.

2.6 The Research Question

Having identified the core aspects of the personalisation problem for music recommendation and proposed a route forward, it remains to outline the specific research question for this thesis. This section is concerned with achieving this task and introducing the core definitions and potential dichotomies which will be confronted throughout this thesis.

How can human-in-the-loop techniques be applied to reincorporate the core tenets of making personalised music recommendations into modern recommender services?

To fully understand and effectively address this question, it is necessary to break it down into the following sub-questions which are addressed in chapters 3-5.

1. **What are the tenets of making personalised music recommendations?**

2. **How can human-in-the-loop practices allow users to inform an automated music recommender of their requirements for personalised recommendations?**

The first part of the question concerns the nature of personalisation and relies on understanding what makes certain music recommendations feel personal. This first issue is addressed in chapters 3 and 4 and can be posed by the question:

The second aspect of the question then concerns how human-in-the-loop techniques can be used to facilitate the human practices, which make music recommendations feel personal, within automated systems. This issue is focused on in chapter 5.

2.7 Chapter Summary

This chapter has introduced the field of recommender systems and presented the personalisation problem in the context of the academic and commercial milestones which shaped it.

The chapter started by highlighting how the field emerged from information retrieval in response to the information overload problem which had arisen in the late 1990s with the invention of the internet. The chapter proceeded by introducing the first manual recommender systems and explaining how they were able to produce very personal recommendations since users remained central to the recommendation process and defining what they wanted to get out of the system. The beginnings of the personalisation problem were hinted at in the realisation that users were cut out of the recommendation process as a result of the transition from manual to automated recommender systems which occurred in the mid to late 1990s in response to the scalability problems which had been discovered with manual systems.

The next section of the chapter discussed the evolution of automated recommenders. The three fundamental types of automated recommender (collaborative filtering, content analysis and hybrid systems) were introduced along with the core problems and algorithms which shaped their development. The most pervasive of these problems was revealed to be a data sparsity problem known as the cold-start problem.

The next section of the chapter addressed the impact of the Netflix challenge on the development of the field and the rise of the personalisation problem. The positive aspects of the challenge in revitalising the field and leading towards the solution to the cold-start problem were highlighted. However, it was also shown that several unforeseen negative consequences of the challenge contributed heavily to the development of the personalisation problem facing recommender systems today. The section revealed how the Netflix challenge helped to cement predictive accuracy as the sole metric against which systems were constructed and measured for a decade. This was shown to detract from the personalisation of recommendations as too accurate recommendations can become obvious and be poorly received by users.

This lack of personalisation due to generic recommendations has been particularly strongly voiced in the area of music recommendation where users have expressed a desire for highly personalised non-obvious recommendations. This aspect of personalisation is considered in more detail in the next chapter.

The next section considered the state of commercial music recommender systems and reviews how they have impacted the personalisation problem. Although there have been some attempts within commercial environments to pursue more serendipitous recommendations and address personalisation - even recommended by mood - predictive accuracy remains the driving metric. Part of the problem here stems from a misalignment in the demands of a recommender system. From the commercial side, net profit is an important, often the most important, objective. As such, a recommender system is ambivalent as to how well-received its recommendations are so long as they lead to purchases or plays that maximise profit. By contrast, users wish for novelty, often seeking different things at different times.

The final section introduced the overarching research question for this thesis: **How can human-in-the-loop techniques be applied to reincorporate the core tenets of making personalised music recommendations into modern recommender services?**

Chapter 3: The Nature of Personalisation

In order to address the personalisation problem, it is first necessary to unpack what is meant by personalisation in the context of music recommendation. To this end, this chapter focuses on gaining insight into what aspects of a recommendation lead a person to regard it as personal.

The first section of the chapter introduces several domain specific terms and key areas of related research which feature heavily in defining and exploring the nature of personalisation throughout the chapter and the wider thesis.

The next section of the chapter explores the multiple facets of personalisation identifying several dichotomies within the literature which appear to contribute to the difficultly in generating personalised recommendations.

The third section of the chapter presents a case designed to reveal the core thematic aspects which lead people to regard a music recommendation as personal, as well as identify their frustrations with existing music recommenders in how they fail to achieve these features of personalisation.

The final section of the chapter brings together the literature and case study to identify areas where they agree or conflict and bring insight into the nature of personalisation in the domain of music recommendation.

Succinctly the chapter is structured as follows:

- Part 1 defines key terms relating to personalisation used throughout this thesis
- Part 2 explores the multiple facets of personalisation presented in literature
- Part 3 presents a case study to identify common thematic elements of personalisation and personalised music recommendations
- Part 4 discusses the finding of the study against the backdrop of the literary themes identified

3.1 Definitions

The research presented in this thesis is multi-disciplinary by nature touching on areas of research from: human-computer interaction, human factors, machine-learning, recommender systems, musicology and information retrieval. Many of these areas of research at times deal with similar or related problems. In order not to conflate issues or overload definitions which appear across multiple disciplines and contexts, it is important therefore to narrowly define and differentiate several terms which are used throughout this thesis.

The section below defines several commonly used terms within this thesis such relevancy, appropriateness, novelty, serendipity, purpose and context. These terms are introduced and defined in the process of highlighting two differentiations which are central to the remaining chapters of this thesis. They are as follows:

1. Novelty ≠ Serendipity
2. Purpose ≠ Context

3.1.1 Novelty ≠ Serendipity

Novelty: NOUN (plural **novelties)**

- **[***mass noun]* The quality of being new, original, or unusual. *'the novelty of being a married woman wore off' – Oxford English Dictionary*

Serendipity: NOUN

- [*mass noun*] The occurrence and development of events by chance in a happy or beneficial way. *'a fortunate stroke of serendipity' – Oxford English Dictionary*

Above are the Oxford English Dictionary definitions for novelty and serendipity. In common parlance, novelty, as the definition states, is concerned with the quality of newness. By contrast serendipity is concerned with the occurrence of favourable events by chance: happy accidents.

In several research fields including HCI and recommender research, these definitions have been subverted somewhat and distinguishing them both from their common usage and from one another is a tricky but essential endeavour if the research in this thesis is to be clearly interpreted and understood.

Serendipity centred research has seen a vast expansion in interest over the last 7 or 8 years across a wide range of fields ranging from engineering to human factors to HCI and recommender systems. Human factors and HCI are perhaps the two disciplines which have seen the biggest growth in this area of research. Within these domains the term serendipity is typically used to denote systems or frameworks for connection making and encouraging non-obvious interactions (Kefalidou & Sharples, 2016). A way of thinking about this is that these systems aim at promoting encounters, events and occurrences, which had they occurred by chance, would fit the conventional dictionary definition of serendipity provided earlier.

Within recommender-based research, serendipity has tended to be defined in a slightly narrower sense still. Rather than focusing on connection making and encounters, serendipity centred recommender system research typically seeks to promote novel yet relevant recommendations (Celma, 2008). Unfortunately, both of these terms, relevant and novel, leave quite a lot of room for interpretation and subjectivity causing further clarification to be necessary.

In most recommender research, relevant recommendations are generally held to be those which are both appropriate in the sense that they do not cause undue offence, and likeable, i.e. to the taste of the person seeking the recommendation (Celma, 2008). In this thesis this domain specific definition is departed from slightly and narrowed a little further in order to emphasise the point that an appropriate recommendation does not necessarily have to be liked by person seeking it but rather it has to be suitable to the purpose they envisage using it for. If an event planner is seeking music recommendations for a 70's disco night but themselves do not like disco music, then they still might consider 'Rock the Boat' by The Hues Corporation to be a relevant and appropriate recommendation. This subtle distinction is discussed more in the case study of this chapter and the importance of considering a recommendations purpose, i.e. the event or situation a recommendation is intended for, is explored further in chapter 4.

In common usage, the term novel is often used inter-changeably with the word new to describe something which has not been seen before. It is used to indicate originality and newness whilst serendipity is used to emphasise pleasant yet unexpected occurrences. If a product is described as novel, a person instantly conjures connotations of newness and often excitement. However, novel in its fuller sense can also be used to denote something being given a new and often surprising application. A well-known example of this is the use of Teflon as a non-stick cooking surface. Originally Teflon was discovered by accident in 1938 by research chemist Roy Plunkett who was at the time working on creating a new refrigerant for DuPont. For about 10 years after its initial discovery, its applications were limited virtually exclusively to industry where it was used as a coating for screws and bolts. Subsequently in 1950, a man named Marc Gregoire developed a way of applying it to his fishing tackle to stop it tangling. His wife then thought of the idea of using it for its non-stick properties on her cookware. Hence the first Teflon coated pan was created 12 years after the material had a rich and established usage in industrial applications. It was predominantly used for one purpose but was subsequently found to be more useful and to have a novel application in another previously unconsidered area.

Within recommender systems research, novelty is often used in a highly restricted sense, typically simply to denote systems which promote and aid the discovery of new content. Occasionally novelty is used slightly more broadly to denote new and surprise recommendations, but often at this juncture an author will switch to using the term serendipity, leading there to be some overlap between novel and serendipity focused research.

In other areas of the literature, a novel recommender research is used in a yet further restricted way to label systems which prioritise or exclusively recommend niche specialised content known as long-tail content (Celma, 2010b; Celma & Herrera, 2008). This is broadly unpopular content about which little information in the sense of ratings and user feedback exists.

In this thesis, the term novel is favoured over serendipity to mark a departure from connection making research and to distance common place associations of accidental discovery. It is important to acknowledge that the term novel is used in a broader sense in this thesis than is typical in the recommender community to recognise that a novel recommendation does not need to be strictly new and unknown to the person seeking it but rather needs to be new or unconsidered by them for the application they are wishing to use it for. This largely absent consideration of purpose within the literature was observed in the previous chapter and its importance is supported by the findings of the case study presented later in this chapter. The full role and significance of purpose for personalising music recommendations services is explored later however in chapter 4. As a final comment on the technical use of novelty within this thesis, novelty where it is used is also intended to indicate an aspect of pleasant surprise which is closely associated with the non-technical common place definition of serendipity.

3.1.2 Purpose ≠ Context

Purpose: **NOUN**

- **1**The reason for which something is done or created or for which something exists.

‘the purpose of the meeting is to appoint a trustee’

Context: NOUN

- The circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood.

‘the proposals need to be considered in the context of new European directives’

In everyday usage, context is about the circumstances which shape and form the background or setting for a situation or event. Within the domain of recommender systems, context is often used more narrowly to refer to information about a user which cannot be deduced from their usage of a system and which is not directly given in their search parameters.

Context-aware recommender systems were introduced in the previous chapter as a relatively new area within recommender systems research which typically aims at improving the predictive accuracy of recommendations by incorporation of additional peripheral data about a system’s users. The core tenet of the domain centres around the philosophy that incorporating more data about a user will lead to a clearer picture of who they are and their taste preferences which will in turn help to predict their ratings on items and lead to more accurate and therefore better recommendations. Within this field, some context-aware research has interpreted context as explicitly trying to gauge the mood or emotional state of a user who is seeking a recommendation (Kaminskas & Ricci, 2012). The significance of this research for this thesis and the personalisation of music recommendation is considered in the related research for the case study presented later in the chapter.

Often in this thesis the term purpose is used, and it is important that the concept of purpose be differentiated from the notion of context in recommender systems research. Context as mentioned focuses on peripheral user information, whilst purpose, as it is commonly understood, refers to the reasons for which something exists or is done (see the example usage of purpose in the Oxford English Dictionary above). In this thesis we use the term recommendation purpose to refer to the reason for which an individual or group are seeking a recommendation. Specifically, recommendation purpose is intended to denote the situation for which an individual or group seeking a recommendation intends it to be used and enjoyed. Purpose is used in place of context to denote this concept for two reasons. First, it is intended to signal a strong departure from the philosophy of context-aware recommender research that systems can be improved and personalised by gaining an increasingly accurate profile of users and obtaining ever higher predictive accuracy. Second, it is intended to shift the focus away from user centric profiling towards situational profiling. This shift is made to encourage the exploration of a core investigative tenet of this thesis: that the purpose for which a recommendation is being sought is more significant in the personalisation of music recommendations than knowing everything about the taste preferences and behavioural attributes of the person or persons seeking a recommendation.

Understanding the importance of recommendation purpose and its role in improving the personalisation of music recommendation systems is the focus of the next chapter in this thesis.

3.2 Dichotomies & The Multiple Facets of Personalisation

This section introduces the role of purpose and the accuracy novelty dichotomy as the key components of the problem.

3.2.1 Accuracy vs Novelty

Most advancements in recommender systems over the last decade within the commercial and academic spheres have come from the development of movie recommender systems like Netflix (Ekstrand et al., 2011). For a long time this perpetuated the notion of predictive accuracy as the sole and all-important metric for improving and assessing recommender systems, although this has started to be challenged in recent years (Adomavicius & Kwon, 2011) (Anna, 2016; McNee, Riedl, & Konstan, 2006b; Zhang, Séaghdha, Quercia, & Jambor, 2012).

Chapter two highlighted the negative side-effects and consequences of pursuing accuracy at all costs. Music recommenders which only recommend to people obvious mainstream content, or content they naturally would have thought to play, are of limited value for several reasons. First, no matter how accurately a given recommendation or set of recommendations might fit with a person's overall musical tastes it is unlikely they are going to want to listen to it all the time and never hear anything different. Often people turn to music recommenders as a means of discovering new content they haven't heard before, or simply to hear content they might not have listened to in a while. That being said, people don't always wish to discover new music and they typically want their novel recommendations to fall within some parametrised notion of accuracy such that they don't find the recommendation offensive or unpleasant to listen to.

Accuracy and novelty, at least initially, might appear counter or contradictory demands to place on a music recommendation service. On the one hand, you appear to be demanding that a recommendation reflects what a user would typically want and choose for themselves, whilst, on the other, you are requesting that recommendations be new, unexpected and surprising. Furthermore, this difficulty is compounded by the fact that different people will have different ideal balances between novelty and accuracy (Celma, 2010a).

The delicacy and real skill in producing personalised recommendation then, it appears, turns at least partially on balancing the novelty and unexpectedness of a recommendation against the parametrised accuracy of a person's tastes so as to please rather than offend them with a novel or surprising recommendation they would not have thought to listen to themselves (Adomavicius & Kwon, 2011).

3.2.2 Mood & The Dynamic Nature of Taste

Mood is another important aspect in music recommendation. At different times people have different moods, and this affects their taste preferences as to what they might wish to listen to at a given moment in time.

3.2.3 Dynamic ≠ Over-a-lifetime

Whether content-based, collaborative or hybrid, modern automated music recommenders all share the trait of monolithic profiling. They proceed to form a single impression of a given user and slowly update this impression over time to gain an increasingly more accurate picture of who that user is in terms of their taste preferences.

As will be seen in the case study presented later in this chapter, this is neither a useful nor accurate way of representing people. Sometimes people want to discover new content and at other times they want to remember times from their past and so play familiar songs. Depending on who they are with and the time and event they are intending to use the music for, their requirements can change still further. In reality, people's tastes often vary drastically as a consequence of a wide range of factors including their mood, emotional state, demographic, environment and the company they might be in (B.-J. Han, Rho, Jun, & Hwang, 2009). Properly represented then, a user of a recommender system is not just a single taste profile but a multi-faceted changing collection of inter-relating requirements.

The non-static nature of people's tastes has been periodically recognised in the field of music recommender research (Park, Yoo, & Cho, 2006). However, it has seen little direct attention, often being mentioned in passing only in a wider discussion about contextual recommendation. Additionally, where it has been mentioned many researchers have taken a long-term view of time discussing how users' tastes change over a lifetime (Celma, 2010a). Discussions have tended to focus on stable long-term changes in people's feelings regarding a particular style of music over time.

By contrast, dynamic in this thesis is used explicitly to refer to the momental changes which affect a person's taste preferences over a far shorter period, like a week or even a day. Changes which affect the type of music they wish to listen to or seek during a given searching session.

Although this type of dynamism has been relatively unexplored, a few researchers have started to investigate it and recognise its significance. For instance, in 2012 Negar Hariri wrote a paper in which he introduced a recommendation system which used time-frame bracketed recent user behaviour to infer current contextual parameters and taste preferences which might help to guide a user's recommendations(Hariri, Mobasher, & Burke, 2012). Additionally, Hariri used human curated playlists in his system introducing a further HITL aspect to his system. In the next chapter we discuss the benefits of human-in-the-loop systems in-depth and consider how and why a HITL approach might be able to better accommodate users' dynamic tastes.

3.3 Case Study: Understanding the User's Needs & The Nature of Personalised Recommendations

In chapter 2 it was shown that as recommenders evolved to cater for ever more users and items, they became increasingly complex in order to handle scalability problems and data sparsity problems like the cold-start problem. A negative and unforeseen consequence of this development was a lack in the overall perceived quality and personalisation, particularly regarding music recommendations where users readily criticised systems as being impersonal and only recommending safe mainstream popular content. Music recommenders had suddenly become highly accurate but not very useful.

In chapter 2 it was also acknowledged that the earlier manual recommenders which preceded today's automated systems often achieved a higher level of personalisation, perhaps due to their being user-led and incorporating real human insight. Out of this observation emerged the central tenet or defining question of this thesis. Specifically, how might human insight be reincorporated into automated music recommenders so as to produce a better class of more personalised music recommendations?

In several other areas of human factors and HCI research, a technique known as human-in-the-loop system design has been widely acknowledged in recent years as being an effective method of incorporating serendipity into automated systems (Kefalidou & Sharples, 2016; Pontis et al., 2015). On this basis it has been selected as the approach to pursue the question above within this thesis.

Human-in-the-loop systems are built upon taxonomies or ontologies which enable users to meaningfully interact with the system and influence its output. Typically, this interaction entails users pre or post filtering the content which is made available to or produced by a system. For example, a HITL collaborative filtering music recommender might allow users to filter which users are made available to the collaborative filtering algorithms or it might allow them to filter the recommendation list produced by the algorithms to remove the recommendations which came between a certain time period or were by a particular artist.

The first step in designing such a system is constructing the framework or taxonomy for facilitating the human intervention, and the first task in constructing the taxonomy is understanding what it is that makes people consider certain recommendations or curated sets of music to be personalised and others not.

The aim of this study then can be understood as gaining an insight into the natural process by which people personally curate music with the objective of using this insight to incorporate and facilitate the essential human aspects of personal music curation in automated music recommender systems.

3.3.1 Related Work

3.3.1.1 Human-in-the-loop: Moravec's paradox, Chicken Sexers & The importance of human intelligence

The is a well-known paradox in computer science called the Moravec Paradox which basically stipulates that that which computers accomplish trivially, humans find difficult and vice versa (Moravec, 1988). Standard Von Neumann computational systems are highly efficient at rapidly solving quantifiable problems. Where they struggle is in handling or understanding qualitative problems. Machine learning is a field in computer science of which recommender systems is a subdomain which deals frequently with this issue. Often the question asked is how we take a real world qualitative problem with all its ambiguity and shapelessness and translate it into a strictly defined quantifiable problem that a computer can deal with.

Sometimes, however, this is simply too difficult a task to achieve. The human process of accomplishing the task is simply too varied, subjective or poorly understood. A classic example of this is the case of chicken sexers (Horsey, 2002). In the farming industry, individuals are assigned the task of sorting out the sex of chickens. In training they are not told any specifics on markings or things to look for but simply made to watch someone who can already do the task. After a while, seemingly instinctively or intuitively, they themselves are then able to correctly classify the baby chicks and yet remain unsure about precisely how they are able to do it.

In certain situations that appear to require this poorly quantifiable aspect of human behaviour or judgement, a relatively new approach has emerged known as HITL. As has been discussed earlier in this thesis, these systems work by combining subjective human judgements or intelligence with computational systems. Typically, by allowing human users to filter and correct the input or output to a particular computational algorithm.

In several applications, it has been found that this approach can produce superior results to either purely human or purely computation approaches. One such example is in chess which has long been a test bed of computational efficiency and capability. In 1997 IBM created a computer known as Deep Blue which was capable of beating the then current world champion at chess, Garry Kasparov (Goodman & Keene, 1997). This inspired Kasparov to develop a keen interest in the application and development of computational systems and specifically in the differences between human and machine capabilities. Recently he released a book and hosted a TED talk revealing his findings that the best chess engines and expert human players can be beaten by only moderately powerful computational systems which have been combined with moderately skilled human oversight.

In human factors the advantage of combining human and machine efforts has also been recognised for its ability to enhance connection making and encourage pseudo-serendipitous encounters and interactions in academia (Kefalidou & Sharples, 2016). Furthermore, its potential for enhancing recommender systems has begun to be recognised too (Pontis et al., 2015).

In order to work effectively, HITL recommenders have to be constructed upon an underlying framework or taxonomy for facilitating the human intervention and practice they are attempting to introduce into the automated system (Montaner, López, & la Rosa, 2003). This aspect is where the current limitations arise in current HITL music recommender solutions. The problem is that currently each new paper or system being developed is developed in isolation to address a specific narrow question. This means that each paper or system ends up defining its own unique quasi-frame or taxonomy making it very difficult to effectively compare systems or measure their relative strength and weakness in their abilities to provide personalised results.

To this end, this thesis endeavours to take a step back and first develop a deep understanding of personalisation which can serve as a standard definition or criteria against which future HITL systems can be designed and assessed.

3.3.1.2 Context, Purpose & Scenario-bound Recommendations

Earlier in this chapter the distinction between context and purpose was explained with regards to music recommendation, with purpose being a sub-domain of context which focuses on the reason why a given group or individual is seeking a music recommendation and the event or scenario they are intending to use it for.

Whilst context has been widely explored, purpose has yet to be investigated in-depth outside of a few choice topics such as sports and workout augmentation and situated design and event augmentation.

In the situated design space, much of the work that has been done on purpose has tended to focus on the results of humans curating or recommending music for a given event rather than investigating the process by which those individuals or groups formed their curations or recommendations. To date it appears no research has been done on how people set about curating music for a given purpose.

The study below addresses this and reveals the extent to which people consider purpose when engaging in the task of curating music for particular moods of their choosing.

3.3.1.3 Taxonomies & Ontologies in Music Recommendation

As discussed, earlier HTL systems require a taxonomy or framework to help facilitate a human intervention in the automated system. Existing taxonomies in music recommendation however haven't been designed towards the construction of HITL systems. In general they have concerned the classification of digital assets or the automated classification of music for the purposes of supporting content-based recommendation (“ISMIR 2008,” 2008a) (Raimond, Abdallah, Sandler, & Giasson, 2007) (Levy & Sandler, 2009; Raimond, Abdallah, Sandler, & Giasson, 2007).

Other areas of research in the music taxonomy space have looked at classifying music with respect to mood and users' emotional state (H. H. Kim, 2013) (Rho, Song, Hwang, & Kim, 2009; Song, Kim, Rho, & Hwang, 2009) (Han, Rho, Jun, & Hwang, 2010). Whilst many of these studies are interesting and show promise in that they reveal non-standard, non-genre-based means of classifying and thinking about music, they rarely consider personalisation or subjectivity in their construction and, perhaps even more concerning, focus heavily on predictive accuracy.

Unfortunately, however, there is very little research on the design of taxonomies or frameworks for promoting personalised music recommendation. Indeed, as mentioned earlier, HITL endeavours to explore personalised music recommendation have been held back by the lack of any agreed upon scheme for building or assessing these systems or the taxonomies they rely on.

A final consideration in this area as well as the structure and design of the taxonomy is the influence of the taxonomy designers. If you get experts to design your taxonomy will it be sufficiently relatable to be useful for non-expert users of a system and when you are dealing with inherently subjective demarcations and classification, how do you account for different user base-lines or perspectives?

Several papers have begun to investigate issues investigating things like the level of agreement in experts and non-experts in genre-based classification (“ISMIR 2008,” 2008b) (Sordo, Gouyon, & Sarmento, 2010.

3.3.2 Method

3.3.2.1 Participants

We recruited 11 participants (5 males and 6 females) over 18 years of age. From those, 6 participants were University of Nottingham attendees. Of those 6, 4 were PhD students, 1 was an undergraduate student and 1 was a post-doctoral researcher. Of the rest of the participants, 2 were primary school teachers, 1 worked for the University of Oxford and 2 worked for IBM.

Participants’ demographics	N (%)
Age	
18+	11 (100%)
Gender	
Male	5 (45.45%)
Female	6 (54.55%)
Profession	
PhD Students	4 (36.36%)
Undergraduate Students	1 (9.09%)
Post-Doctoral	1 (9.09%)
Researcher	2 (18.18%)
Primary School Teacher	3 (27.27%)
IT & Computing	

Figure 7: Participants demographics table

All participants were compensated for their time upon completion of both parts of the study with a £10 amazon voucher. The study was approved by the School of Computer Science ethics committee and conducted in accordance with its ethical guidelines.

Whilst 11 participants seems like a small number in this sort of study it is not too great a limitation as we are not seeking to make generalisable claims but rather to conduct an initial exploratory investigation. The objective of this investigation is not to test a hypothesis but rather to identify key themes and ideas some people have about personalised music recommendation. This then provides a starting point for subsequent investigations with larger sample sizes.

Performing an initial investigation like this mitigates the chance of biasing or basing subsequent studies solely from my individual preconceptions. It also inspired the design of the study in chapter 4 (which has 400 participants) and provided material for the final user experience HITL validation study in chapter 5.

Finally when it comes to opinion gathering exploratory exercises it is often found that after a small sample the general range of themes are identified (Nielsen, 1989). This is known in the literature as reaching saturation point. After asking only a small number of people a question you often find you begin to hear the same general ideas, responses and complaints. In the domain of usability testing Jacob Nielsen has gone as far as saying that “with 5 users, you almost always get close to user testing's maximum benefit-cost ratio” (Nielsen, 2012). Whilst we are not doing a pure usability study here the point stands that when it comes to opinion based subjective feedback it often only takes a few people to identify the core themes before you start getting repeat responses. Indeed this phenomenon is why FAQ sections are so common in manuals, websites and instructional media.

3.3.2.2 Design & Procedure

The study was conducted across 23 days between 17th March and 8th April 2015. Prior to the study, participants were given a participant information sheet and consent form detailing their rights, the nature of the study and the type of data which would be collected during the study.

The study consisted of two stages. In the first part of the study, participants were asked to create two playlists to suit two different moods of their choosing. The only additional criteria placed on the participants was that the playlists had to be at least 10 tracks in length.

When describing the playlist creation task to participants, purposely vague or ambiguous language was used like the term ‘mood’ without further qualification or examples. This was done to give participants the maximum creative freedom by providing a degree of flexibility in how they chose to interpret the task. Participants were also allowed to create and record their playlists in an environment of their choosing using any medium they preferred, e.g. Spotify, iTunes, CD or simply keeping a written list of tracks. This was intended to provide a degree of normalcy, by allowing the participants to complete the task in the most natural way, as they might ordinarily do when they were not being studied. Participants were also given the freedom to design their playlists for any demographic and listening scenario they preferred. Interestingly, all 11 participants choose to design their playlists primarily for their own use.

The second part of the study consisted of semi-structured interview sessions in which participants were asked to describe the process by which they created their playlists and discuss any opportunities and limitations they experienced when interacting with the music recommendation tools they used to assist them.

We made sure that participants were interviewed within 24 hours of having created their playlists to ensure that the process of creating the playlists was still fresh in their minds. To facilitate this and help ensure that participants were relaxed, we chose to conduct the interviews in opportunistic settings that were agreeable and convenient for the participants such as their home or place of work.

Interviews lasted for approximately 15 minutes. During the interviews, participants were asked a series of open-ended questions to encourage them to discuss the process by which they created their playlists and also talk about any frustrations they had with the task in general or any of the assistive tools and or recommendation systems they may have used. A typical question was: ‘Describe the process by which you approached the task of creating your playlists?’ A complete set of the guideline interview questions are attached in the appendices at the end of this paper.

3.3.3 Results & Thematic Analysis

We conducted an informal thematic analysis of our semi-structured interviews. This analysis revealed the following overarching themes as being important to most participants regarding a recommendation as personal:

3.3.3.1 Emergent Themes

At the highest level of granularity, five common themes emerged from participants’ interview responses surrounding the personalisation of music playlists. The themes were as follows:

- The intended purpose of the playlist – what audience and setting it was intended to be enjoyed in
- The role of personal memories and emotions in constructing and personalising playlists
- The need for human involvement in the curation process
- The importance of structure and order within a playlist
- The novelty of content within a playlist

The pie charts below provide a first look and high-level overview as to how important each of these themes were to participants. Each pie chart shows the proportion of participants that judged a given theme to be strongly, moderately or not important. A participant was judged to strongly support a theme if they explicitly brought it up in their interview and emphasised its importance for them in creating playlists. A participant was judged to only moderately support a theme if they only hinted at its importance and did not directly discuss how it impacted their process in completing the playlist creation task. A participant was judged as not supporting a theme if they explicitly disagreed with it or otherwise discussed how it was not an important characteristic for them when creating their playlist.

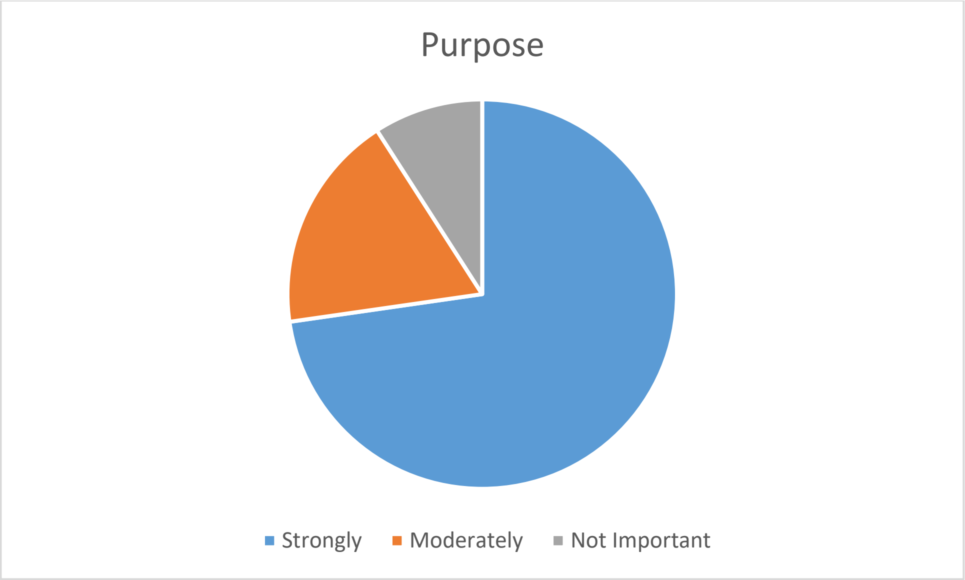


Figure 8: Significance of purpose pie chart

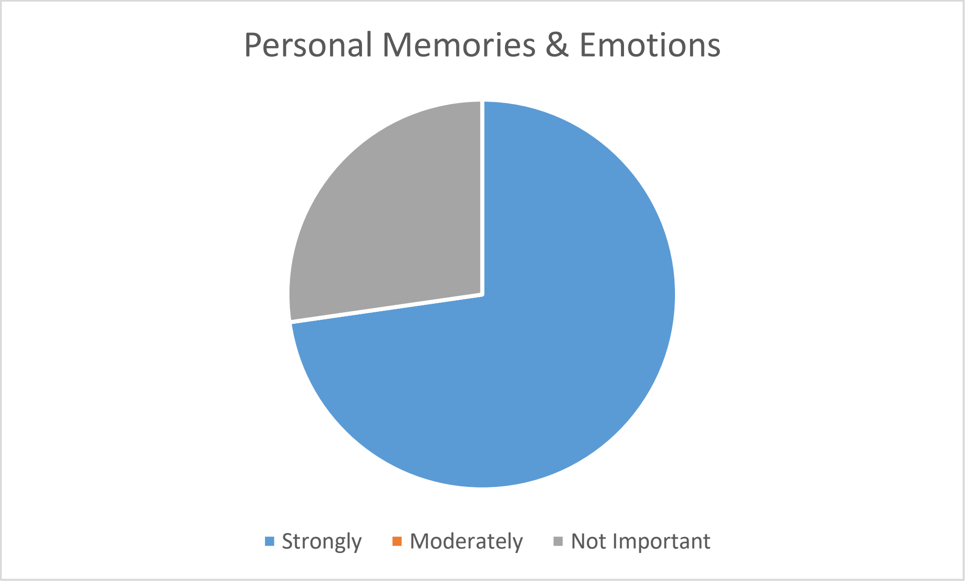


Figure 9: Significance of memories pie chart

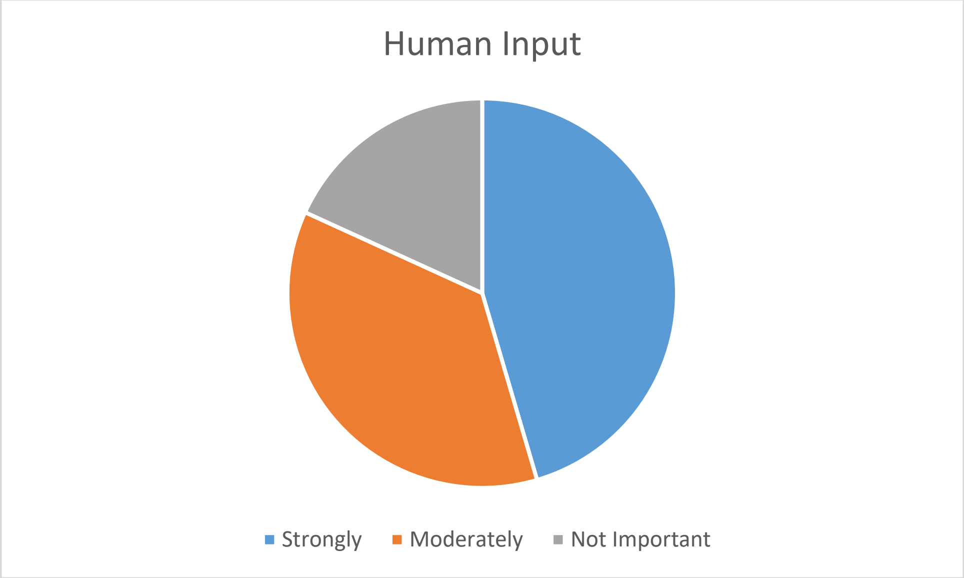


Figure 10: Significance of human input pie chart

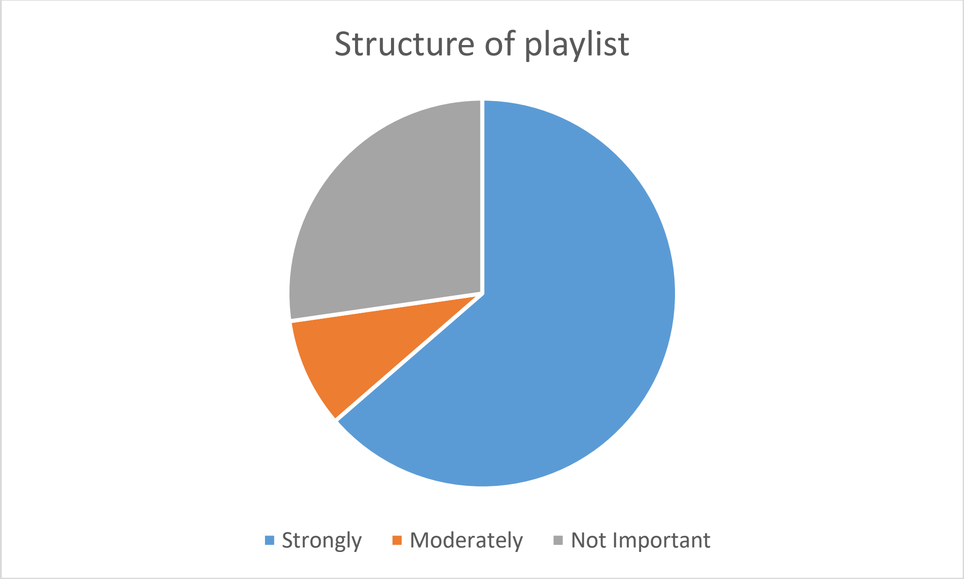


Figure 11: Significance of structure pie chart

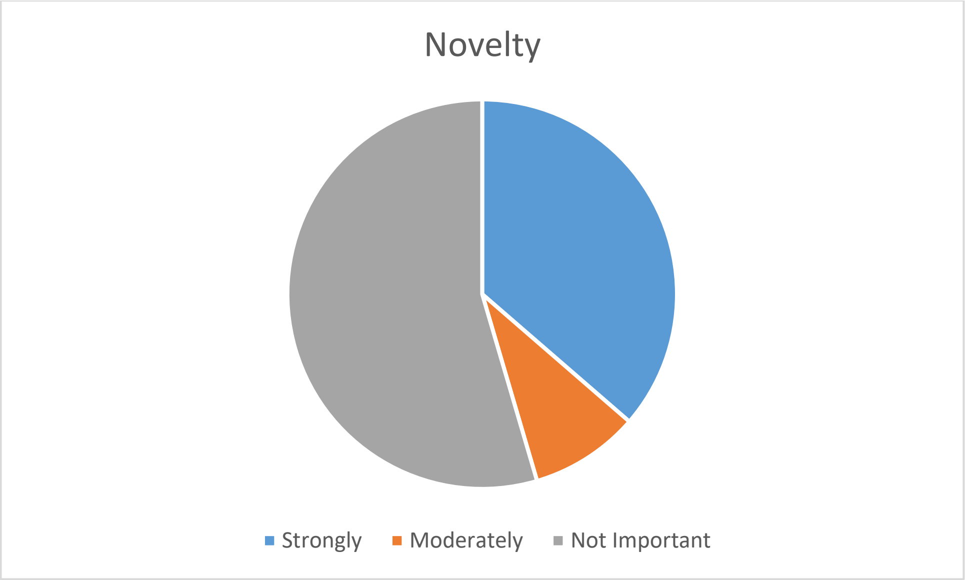


Figure 12: Significance of novelty pie chart

The importance of considering the intended purpose of a playlist along with personal memories and connections to various tracks emerged as the two most important and strongly supported themes with 73% of participants strongly agreeing that they were important for creating their personalised playlists. The next theme to emerge was the importance of structure when creating playlists which 64% participants strongly agreed was very important. The importance of human involvement in the curation process emerged as the next most important with 45% of participants thinking it was strongly important. Finally, 36% of participants strongly felt that novelty was important to them when creating their playlists.

Below, each of these themes is interpreted and presented using exemplary representational quotations from participant interviews. Participants quotes have been anonymised and codified in the form of P(n) where n represents a unique participant number. Participants' codes are given after each quote.

3.3.3.1.1 Purpose

Virtually all participants (10 of the 11) talked about how they had created their playlists for a particular purpose. Often, they spoke about how they might use their playlists to support them either in achieving some task or alternatively simply shifting their mood and cheering them up. In the quote below, a participant discusses how the tracks they curated for one of their playlists was guided by their desire to focus and work. This intended purpose shaped the type of music they sought to include in the playlist by steering them away from songs with lyrics that they might find distracting.

> *Well um I thought of two moods and to get that I thought of two situation that I'd be in so um doing work and wanting to relax and not being distracted by the music, so I picked a relaxingie worky one for that. Which is what I do at school so if if I want to ah settle down and do someone work and be relaxed but not be distracted by the lyrics or the mood of the tunes then I've made a relaxing one. And made a dancy one by thinking about what I'd like to the more upbeat one may be when I'm with my friends or want to dance around. #00:00:49-1# - P(4)*

Another interesting aspect of purpose is revealed in this quote when the participant starts to talk about their second playlist and talk about finding more upbeat songs for when they are with their friends. The intended purpose for their second playlist is linked to a particular audience.

The most common purposes participants tended to mention were work or going into work, cheering themselves up or going out with friends. Evidence of this can be found in the following quotes.

> *But it it is rather nice I mean it would be nice if you were working or something just wanted this in the background and could have all of these grouped together. #00:05:16-5# - P(9)*

> *yeah I think so I had quiet like a there's one that I used to try and fall asleep to so it's like a sleepy mood and there is one that is a bit more upbeat like a bit kind of getting ready to go out or kind of … when I trying to be in a bit of a happier mood #00:00:42-2# - P(5)*

Another interesting characteristic which emerged from several participants' interviews was that their playlist often had the purpose of supporting them by lifting their mood or helping them get through a difficult or boring task like making the journey to work.

> *" Initially I actually just went to play a few songs and it turned into this list um sort of initially it was just I'll put some music on to cheer me up um and this was all sort of music that I would associate more with a partyesque attitude or sort of you know the music I would play if I had friends around so it is a bit more upbeat a bit more orientated to pop music um so there is a bit more in there but not all of it is pop music I'm just quickly going down it. Yeah, I guess it is music that when I look back at it I sort of say well it is sort of music that is good feeling music may be. #00:02:10-9# - P(6)*
>
> *"'*
>
> *yeah so um for ah for one playlist I considered the song that when I'm working or doing some relaxation just want to relax I listen to those kind of musics. And the second one is when I'm in the car or on a trip and I can concentrate more on the words or something then I will choose the songs from the second playlist. #00:01:21-4# - P(8)*

It was interesting that when most participants spoke about purpose, they typically associated it with a particular physical setting and scenario in which they could envisage themselves using the playlist. In addition to associating the playlist's purpose with a particular event, participants frequently considered the audience that would be at the event listening to the playlist. So, purpose became not just the intended event or the environmental setting but also took account of the audience. It often seemed to be the scene or internal scenario the participant had in mind and was intending to use the music in the playlist to augment.

> *Well it is it is actually it is incredibly personal so it is not really a genre if you look at it some of it is old fashioned some of it is newer. But it is a mood in this instance a mood in the sense that these are all songs that I find quite stirring and that I would listen to if I was having a glass of wine on my own just to feel good and enjoy. I mean they are not all happy, but they are just songs that um just stir something in me. #00:03:52-3# - P(9)*

Again, in the quote above, a participant suggests that the playlist is intended to change their mood and associates the playlist with a particular event and scenario they have in mind for which they would be intending to use it.

An important aspect of curating music and an aspect of purpose that most participants agreed upon, was the importance of matching a playlist or curated set of music to the audience that it was intended for. In the quotes below a participant explicitly mentions how they tailored the playlist and deleted tracks that they didn't think their friends would like or recognise.

> *Oh, the relaxing one was for myself and the dancing one was mostly for myself, but I made I put to many on it so the ones I deleted are the ones I think if I was with my friends I'd mostly likely delete those ones cuz they like they'd prefer the others #00:01:26-4# - P(4)*

> *Oh right … ah ones that I know my friends like or that ones I know that they like or may be more recent ones or more mainstream ones I kept in because I knew that they'd recognise them. #00:03:27-3# - P(4)*

> *cuz if you want to dance around and you're with your friends you want them to dance around too not just you (laughing) #00:03:34-1# – (P4)*

Interestingly, the participant returns to the intended purpose of the playlist and suggests that they tailored their playlist to make it more appropriate for its intended purpose since they would want their friends to dance and enjoy it. In a similar vein, the participant below discusses how they chose the tracks in one of their playlists based in part on whether or not they thought their family would object to them playing the songs in the car.

> *'Exactly exactly so I chose those songs in a way that I know not me enjoying that but no one in the car tells me tells me that "ah change change that song" or "we're bored" or "what kind of song are you listening to?" so the second one I was think about I am with friends and family really. #00:03:42-9# - P(8)*

-

3.3.3.1.2 The Role of Mood

The role of mood in music curation was another aspect which seemed to polarise participants. Some seemed indifferent to it whilst others thought it was a really useful way to curate music. The participants below talk about how they circumvented the criteria to create their playlists according to moods and instead opted to create playlists reflecting themes from movies and video games.

> *'ah well basically I didn't chose a particular mood because most of the times that the music that I'm listening to is either music from video games or music from the films that I like. So, I don't know if these two things are considered mood but these are the themes that I choose to create the playlists for. So, I created two playlists one with my favourite themes from movies and another one with my favourite things from video games.' –#00:00:44-5# P7()*

Although very few participants went this far, most (8 out of the 11) strongly tied their playlists to particular situations and purposes rather than abstract moods or emotions. In fact, only one participant opted to evoke specific abstract moods. In the quote below, the participant mentions how they actively enjoyed curating music by mood and explains how they interpreted the task as the job of finding songs to evoke certain moods within them.

> *Participant 3:yeah so … before I did this task I didn't think it [mood] was that important. I didn't pick playlists for certain moods but now doing this task I think it is very interesting and I think it is something I am likely to do again just to pick songs and put them in a playlist for certain moods#00:08:40-6# - P(3)*

> *" I really liked the fact that you could pinpoint songs down to evoke certain moods and to um yeah to be put into those categories, so I enjoyed doing it. #00:08:18-0# - P(3)*

When most (10 of the 11) participants designed their playlists, they tended to pick moods which tied in with situations, for instance songs to get them in the mood to go out, or songs to get them in the mood to work as can be seen from the quotes in the previous section on recommender purpose.

3.3.3.1.3 Genre is not very important

Several participants comment how genre was not the most natural way to group or curate playlists. In the interview segment below, the participant describes how they prefer to group music by artists that sound similar although they may be in different genres.

> *Participant 3: Um I don't know if that is because it just groups it into particular genres and I'm [not] a kind of person that listens to a particular genre #00:04:418#*

> *Interviewer: yeah #00:04:43-0#*
>
> *Participant 3: I just listen sometimes I just like the sound of the music and I like artists that sound similar or though they may not be in the same genre um #00:04:54-6# - P(3)*

In the quote below, another participant comments on how they make playlists of similar artists and make a subset of their favourite tracks from these artists regardless of what genre those artists or specific tracks happen to be in.

> *'I find it quite yeah I don't have much time for that so I kind find an artist I like search them and save all their songs as a playlist for them and then just save my favourite ones into my collection which is a mixture of all genres but ... it depends #00:02:25-4# - P(4)*

In the quote below, another participant also comments on how genre is not so important for them when choosing music. For them the overriding criteria is not that the songs are from the same genre or even time period, but whether or not they help to support the purpose and scenario for which they are creating the playlist. This was a common sentiment amongst many participants that a playlist served and supported a particular purpose, whether that be a specific event or emotional state of mind.

> *'Well it is it is actually it is incredibly personal so it is not really a genre if you look at it some of it is old fashioned some of it is newer. But it is a mood in this instance a mood in the sense that these are all songs that I find quite stirring and that I would listen to if I was having a glass of wine on my own just to feel good and enjoy. I mean they are not all happy, but they are just songs that um just stir something in me. #00:03:52-3# - P(9)*

3.3.3.1.4 Memories, Emotions & Nostalgia

The importance of personal memories and nostalgia emerged as a very common and often strongly felt theme for the majority (8 out of 11) of participants. In the quotes below, participants are explicit about how memory and nostalgia helped focus and shape their playlists.

> *Umm and my other one is much more musicals what I tend to listen to in my down time when I want something a bit kind of slower quite a bit more memory based #00:00:50-6# - P(2)*
>
> *it is so memory based #00:02:11-0# - P(2)*

> *the moods I picked where nostalgic and motivated… so, for the nostalgic one I was just going through songs that I used to play when I was like 18 19. Like when I was younger and yeah and ones that reminded me of like good times' #00:00:46-2# - P(3)*

In the quote below, a participant goes into slightly more depth and admits that they use nostalgia and personal memories to shape their playlist as it helps provide them with memory cues? to times in their youth and acts as a sort of comfort blanket.

> *'Participant 3: yeah so the nostalgia one that purely came about because I was thinking what do I get out of um listening to music and it it's because I like to old songs and I thought .. I am obviously searching for some sort of nostalgia when I am listening to them and when I am driving in the car it does like I don't know it evokes memories of when I was younger and so I find that comforting so that's how the nostalgia one came about #00:12:05# - P(3)*

This quality of personal memories to evoke feelings of comfort and nostalgia was elaborated upon further by another participant in the quoted passage below. The participant explains that certain songs can trigger memories for them and actually get bound up in the memories such that their appreciation of both the song and the memory become linked for them. Furthermore, they echo the previous participant's ideas about looking back to childhood and similar happy times where problems (at least from the point of hindsight and looking back) seem trivial and the world seemed less daunting and, as they put it, "everything seems wonderful".

Another interesting aspect about this participant's interview is how, again like many other participants, they intertwine mood, memories and events. The songs serve as a particular cue into a nostalgic view of a simpler past and are then used or intended to be used in the present moment to create a particular setting or scenario. Interestingly, the participant describes an archetypical television series scenario in which a character may be relaxing in a bath with chilled but optimistic music in the background and describes that this is what they are seeking to replicate in their own feelings in the present. This particular interview segment emphasises this point very clearly but it features to some extent the majority of interviews. The overriding notion or ideal being that songs could be used to recapture the past and bring comforting memories forward to accentuate or set a backdrop for the present moment or a specific event, often with the aim of having a calming effect.

'Participant 6: so probably most of the songs do have a memory of something else attached to them ah which probably shapes how I appreciate the memory actually I guess. Because it does sort of you know the Mmbop you know I remember being a child and being 7 years old and sort of when you 7 everything seems wonderful #00:05:43-2#

Interviewer: yeah #00:05:43-2#

Participant 6: it is one of those and may be that is just why I find it such a sort of cheery song in that approach because it is sort of from the time when peoples biggest fear is not doing their maths homework on time #00:05:54-8# "' " started putting in the playlist music from video games that I used to play whilst I was young then I started building up based on the age. So first I started from the age of let's say 10 years old and then I start putting music of the video games that I was listening whilst I was 15 20 years old or even now. And for me that felt like you know a journey into my gaming experience let's say how I started to play video games what games did I used to play back then and what games am I playing right now because these two games these two kinds of games are completely different and that can be seen in the music as well. #00:03:22-6#'

'eah so the reflective one um again actually a lot of it is quite cheery music in some ways there is a few of them on their which ah I sort of waiting for a star to fall sticks in my head for weeks upon end every time I listen to it but it is quite a nice melodic tune um and sort of it just gets me sort of umm sort of chilled but in a good chill mood. It is almost like how I think on TV shows in a stereotypical way people have a bath with nice positive relaxing music it is sort of to me that sort of mood where it is and upbeat things but relaxing as well at the same time and I kind of like that I guess. Also, it has nice memories attached to for me so I kind of stick to it. I think in this playlist actually in both playlists there are songs which I can attribute to when I've listen to them previously. #00:04:49-1#

Participant 6: Yeah actually I think pretty much all of them I know when with most of these songs I can pinpoint a time not necessarily the first time I've heard it but a time when I listen to it may be with friends or something #00:05:10-0# ' - P(6)

In the interview extract below, another participant also reflects on the importance of memories and nostalgia and explains how songs can again serve as memory cues and act as anchor points to the past, allowing them to access and revisit these times. By this participant's description, the playlist serves to enable a journey through the past, a nostalgic trip down memory lane. This is interesting as their motivation is markedly different from the previous participant's and indeed from most participants, many of whom wish to evoke memories of the past but often to produce certain feelings of calm in the present. By contrast, for this participant although there is an aspect of evoking feelings in the present the overall sense is that the objective is simply to remember and go on a journey through the past for its own sake, not with the goal of augmenting or finding calm in the present moment.

> *'Participant 7: Um I it was a really nice experience because in the beginning I was thinking logically for like when I was 10 years old what was my favourite game. So, you know that was a very logical thing to do my favourite game surely there is going to be music that I like from that game. But then I was thinking the composer of that game is that so what other things has he written has he composed? And I started to a lot of games that I completely forgot surfaced started coming to the surface and I was like you know I completely forgot of that game and I started listening to the music and I started thinking of all the things that were happening during that when that music was playing. And most of the time the games that I am playing have really deep and complicated storyline. And ah there was a particular game that was based on Carl Jung and Nietzsche's philosophies philosophy. And ah while listen to the music I reminded it reminded me of all the things I was trying to find out and all the things that I was reading while I was listening to that music and playing the video games. So you know a lot of memories kind of came to the surface. And the same can be said about the the film as well because the new. The thing is that when I when I listen to the music of a film I do not only remember the film itself the scenes of the film you know what was happening in the film #00:07:59-5#*
>
> *Interviewer: yeah #00:08:01-1#*

Participant 7: but I also remember the setting where I was in when I was watching the film Jules and Jim the one that I mention before and I saw that film while I was at York doing my masters and that was really profound because I it was the first time I was abroad living by myself and so you know while listening to the music I was recalling all the things that happened in York and all my experiences through that year. So ah remind yourself of the film but remind yourself of all the experience that revolved around the first time you watched that film. So, I don't know if I answered you question. #00:08:48-6#'

Participant 7: for me playlist is more like as I said before a journey in my memory so I want it to be something that I am going to … #00:09:19-8#

Interviewer: to be a gradual thing #00:09:21-3#

Participant 7: to be a gradual thing yeah exactly so I am going to keep these playlists and start building on them as the time passes by and yeah I am definitely going to use them in the future because I it is something that I am constantly doing actually I am using them quite often um and especially when it comes to the video games my fiancé is also a fan of the same video games I am and she knows these tunes so we are going to video games concerts together and we are listening to them together and also I share this music with my brother and some other fellow games from forums and other online communities. So, it is something that I use in many different ways… believe the most important this is the thing that I said about memories and um on another thing when it comes to the films it is easier to rematch a film instead of replay a game because games most of the times take I don't know from an hour to 50 hours to be completed but a film is only about 2 hours long most of the time. So, having a playlist of video games acts like a quick reference to video games you know I listen to the music I kinda have the same experience as playing the game. But when it comes to having a reference to the films it really helps me to keep track of what I have seen and what I want to rematch sometime in my life... the first thing that came into my mind was um looking back memories I trying to think what I called it now I'll pull it up. Actually, the second of the two lists was was all about things I remembered from my past growing up. And um just songs that resonated thing that my parents and just brought back memories. So, once I got the theme it was really quick to just kind of skip through your life and think of things that you remembered… the the other playlist was just songs that meant something to me. So, it is all the songs that I think is full of great um emotions or just really stir me or have been important to me probably as an adult not looking back into childhood. So once I got the theme basically from the themes then I started the songs came really quickly. It would be quite easy to make the list long, but this was an initial list. #00:01:11-9# - P(7)

3.3.3.1.5 Human Involvement

The importance of having humans involved in the curation process was something that most participants (9 out of 11) considered to be important to some degree. In the quote below, one participant explicitly comments that they would like a recommendation system that could point them towards human experts. This is particularly interesting because it is of course what the first manual systems like Tapestry were designed to do. This certainly gives support to the notion that modern systems might be able to be improved upon by incorporating some of the popular human aspects of earlier systems in a HITL model.

> *if I could have a recommendation system that would link to me to someone who really new a kind of music #00:11:55-3*
>
> *Yeah it was a guy who had expert yeah that's exactly it it was a guy who had expert knowledge and also his taste #00:12:55-3*
>
> *Interviewer: and how would you define expert taste #00:13:15-3#*
>
> *Participant 1: hmm that's a tough one cuz its different for everybody I mean what I think of as expert taste probably a lot of people wouldn't you know but I suppose it would be like somebody who picks stuff I like (laughing)*
>
> *Participant 1:so not just somebody who like is awesome to trying to think of it trying to think of a way to do it like I've got a friend in town who knows a lot about metal like a lot a lot about metal. and it is fantastic and talking metal with him is great #00:14:06-0#*
>
> *Participant 1:for metal yeah but for other kinds of music his tastes are horrible (laughing) #00:14:15-4# - P(1)*

Another interesting element that participants often hinted at was the ability of people to qualify and respond to people's taste preferences as opposed to attempt to quantify them in some manner. It is interesting in the quote above that when asked, the participant cannot really give a rigid definition of what a human expert even is, but instead stressed the subjective nature of assessing expertise when it comes to matters of taste. This is a very important consideration for designing a HITL system which is explored in greater depth in the next chapters.

> *Um so it would be nicer just to sort of find a way of more capturing how I feel and letting the system … yeah, the biggest issue for me and I think a lot of people will have may be or maybe I just feel like this is ah quantifying as opposed to qualifying how I feel #00:14:32-9# - P(6)*

The final point this participant makes regarding feeling quantified rather than qualified by the system is interesting and lends support to the notion that incorporating a human qualitative aspect into the process could improve the personalisation of music recommendation systems.

In the passage below, another participant also comments how facilitating the collaborative but direct interaction between people could be interesting in aiding discovery. In essence here, the participant is describing the earliest type of manual recommender system which worked by facilitating precisely these sorts of interactions. This again lends support to the notion that music recommendation systems might be able to be improved by allowing direct human intervention with the automated process.

> *Yeah if they know me they would. But also I it is like um I think if you like a lot of similar songs then you. If I I think of someone else I know like a friend who likes some of the same song as me but they might they wouldn't have made exactly the same playlist so there would be some overlap but it would show them some other songs that they might add to their playlist, so it would be useful. #00:03:20-2# ' – P(9)*

> *'I think building a collaborative list would be fun to because it would stretch you further you would think of one song and them the people you were with would think of others and you list would be more diverse, but I would still be nice to put it together. So, I think you could do both. #00:08:03-7#' - P(9)*

3.3.3.1.6 Structure & the importance of order

The importance of order or structure within playlists was an interesting theme which emerges as participants often had strong but polarised opinions on its significance.

In the quote below, a participant comments on the developing and opportunistic structure of their playlists whereby they don't impose a strict structure but find one naturally develops as the “springboard off of one [track] that leads to another”.

> *'No so actually um I mean one of them was ah, so I get a song in my head and usually when I create a playlist I'll start playing one song and then through whatever way serendipity or whatever I'll start thinking of another song based on that one and one that has a similar feeling to me. And then basically I normally add it to a queue and then what ends up happens is I just then sort of keep going through this way so I just spring board of off one ah that leads to all the others sort of flowing. So, I think with this one one of the first ones I picked was Rich Girl by Hall & Oats and sort of that lead to actually sort of the nature of this playlist taking shape sort of where it started was then followed on from that. Ah it was sort board I mean both playlists have a mix of genres in them. Um because it sort of I don't know it just felt as the mood was playing through um. #00:08:48-9#' – P(6)*

In a similar way to the previous participant, the participant below comments on how their playlist had an opportunistic structure as particulars songs they remembered evoked particular memories which in turn reminded them of other songs associated with another moment in time or personal memory.

> *"Participant 11: Well I mean in some way both these lists represent stories in my life. So, I kind of enjoyed creating them because once I remembered one song in my mind it leapt to that moment in time which then transferred me to another moment in time which lead to another song which transferred to another moment in time which lead to another song. And it was not really about the song but it was all about my life and what was happening at time. It is just that these songs are connected to certain moments. #00:04:48-0# - P(11)*

It is interesting that the structures in both of these participants' playlists appear to have little directly to do with the content of the songs themselves such as tempo, rhythmic structure or even genre and more to do with the memories or emotions they evoked for the person creating the playlist.

In contrast to this, the order and structure of the playlist for the participant quoted below is very much driven by the content and rhythmic structure of the music. They talk about the importance of having one song flow into another both musically and even within the titles of the tracks.

> *you kind of pick songs where the ending and beginning link to one another. So that you get a nice sort of segue in and out, so you sort of get a nice flow withe in the music or in the titles or in the bands #00:04:33-6# - P(1)*

> *yeah yah to make them kind of so they work so that they work fairly well together so sort of like take this song drop out by gong. See its sort of dropping in that keyboard myth right at the end of it and then it goes from that to blind in the family stones stand that's got really similar kind of melody when the pianos come in and then I go to stand out from love, so you get the title link #00:05:27-9# - P(1)*
>
> *and stand out was a protest song from Vietnam so that goes into coming out of the rain which is another protest song from Vietnam. and it sort of like like for me the fun is making a playlist is figure out what all those linkages are #00:05:39-7# - P(1)*

In contrast to the above participants, the ones quoted below explicitly comment on how they would play their list on shuffle as they didn't find structure to be very important.

> *Um that is actually something I was thinking about this morning was that I would quite happily put these on shuffle ' #00:12:22-6# - P(2)*
>
> *I do create playlist but what I tend to do is I create a playlist then put all the songs in it rather than creating playlists and then I'll shuffle the songs through that #00:01:03-7# - P(3)*
>
> *No, I usefully put it on shuffle or something like that so it is kind of it will play a bit randomly anyway so … #00:04:16-9# - P(5)*
>
> *Um no I didn't I didn't because I didn't have the time and ah not actually that I didn't have the time but ah right now the playlist is composed of 11 songs, so I do not find a reason to do that and ah this kind of playlist is something that I want to build up procedurally. I do not want to sit in front of my computer for like an hour two hours three hours and make a playlist of like 3 songs. It's like it is something that I want to build a up procedurally in the course of I don't know months or even years. So yeah, I didn't and to tell you the truth I don't believe I would do that I would put it on shuffle. #00:05:27-2# – P(7)*
>
> *No I think it was as they came to me so. Which would be fine actually cuz they are just. Whilst they have a whilst they all mean something to me they are all kind of quite unique and I'd quite enjoy jumping from one to the other, so I don't think the order that they are played would bother me too much. #00:05:45-7# - P(9)*

> *because for me a playlist or more something that you put on your phone or on your mp3 player and you listen without having a particular order order in mind. If I want to have something that has order I would put and album or a classical concert or something like that that has a predefined order for a particular purpose. A playlist for me is more like you know a journey in my memories so I wouldn't like to have an order in my memories I prefer serendipity let's say. you know things to come up ah randomly #00:06:16-6# - P(7)*

For the participant below, the importance of structure seemed to vary according to the intended purpose of the playlist. They state that they would probably give order and structure more consideration if they were using the playlist for hosting a house party.

> *no. no I didn't actually I'd never thought about order but when it comes ... yeah no I hadn't thought about that but that is interesting (laughing) I think yeah if I was properly thinking about it and doing hosting a party or something I might think about the order #00:06:04-2# - P(4)*

3.3.3.1.7 Novelty

Although none of the participants expressly used the word novelty, it seemed to be a key theme with many participants emphasising the need to keep playlists updated so that they didn't get bored. Additionally, participants often desired an element of surprise and discovery to be involved in creating their playlists. In the quote below, a participant comments on how they believe records are better than modern recommender systems because they are artist generated playlists which can facilitate discovery and happy accidents. They bought an album because they liked the cover art or the title track but subsequently found a whole load of tracks they liked.

> *I bought a copy purely I bought a record way way back of the ah album cover for the harder they come… purely because that is an amazing album cover alright … so I found all this great music purely on the I just wanted it because I'm describing the album cover but it is this fantastically 70's illustration*
>
> *thats like one of the reasons its better… it opens up the doors for a happy accident like that's not going to happen to you with Spotify #00:19:06-0# - P(1)*

3.3.3.1.8 Frustrations with Existing Recommenders & Other Themes

In addition to the main themes surrounding personalised music curation, several other minor themes including some frustrations with existing recommender technologies were mentioned in multiple interviews. This section is concerned with highlighting these observations. Below, a user comments on how they find the recommendation process unnatural and get frustrated by the “jarring experience” and the fact that they constantly find themselves skipping songs and subjected to the start of tracks they don’t wish to hear.

> *Participant 6: Sorry yeah I usually I end up using Spotify's um sort of related artists so I end up getting of related artists of related artists of related artists so if you click on one you can keep going down the chain and sort of you look at their top tracks that are played and think yeah may be this one um sometimes I see related artist and I think yeah I remember this song I haven’t heard it in a while. But it is quite a frustrating process because I think is sort of the I think the process are very much do one thing assess do one this assess do one thing assess um and it doesn’t feel like it flows particularly well into the spontaneity of what I think music should be about really, I think. A lot a lot of the time music to me should be just a bit more sort of it follows naturally and perhaps the finding music experience is a quite jarring. I have also used sort of the Spotify radio quite a lot but then you do get quite frustrated with having to skip songs when you sort of it starts playing the song and you think no this doesn’t really match me I’m going to skip it and that can be quite an annoying process at times again because you are constantly hearing the start of a song you don’t want. #00:13:54-6# - P(6)*

3.3.3.1.8.1 Repetition Annoying

One of the commonly mentioned frustrations with existing recommenders was the tendency for repetition. In the quote below, a participant describes how YouTube recommendation algorithms often recommend multiple versions of the same song which have been uploaded by different users or differ in some small way like having lyrics displayed or not displayed on the screen.

> *I only used YouTube really and my head umm instead of sometimes when you type when you're on YouTube and it comes up with the recommended ones it will come up with like different people so like say I so one of my songs is on my way by Phil Collins. And the the recommended half of the recommended are like the same like on my way Phil Collins with lyrics video. Um but by different people who have put it up so if it only came up once and then all the other recommended thing could be different songs which would give me more options which would be nice #00:01:43-1# - P(10)*

Interestingly, however, another participant comments below how YouTube recommendation algorithm enables them to avoid irritating repetitions. They state that any playlist they create is often short and so songs end up getting repeated as the content in the playlist is quickly exhausted. They explain that YouTube allows them to avoid this problem by presenting them with much longer recommended playlists that are not exhausted as quickly.

> *Um I don't usually do it cuz um usually there is sort of like I say YouTube is sort of there ready and you don't even have to click on it it now had a thing where you can just click on something and it will just play the next recommended one for you and um so I feel like YouTube is pretty good like already um I just find I don't really like putting the time in to it. #00:03:49-1# - P(10)*

3.3.3.1.8.2 Existing Recommenders Are Not Thematic

Participants also criticised existing recommenders for not being thematic and relying too heavily on genre or similarity rather than thematic continuity. It is difficult to pin down precisely what is meant by theme but the quote below gives some indication that it goes beyond genre or musical similarity.

> *sometimes they're [traditional recommenders] pretty good … but they're always genre related its never about the character of the song … so its sort of like if you pick … the sonics it recommends you this the seeds the monks the mc5 the kingsmen all kind of like soul-ish kind of bands. but weirdly its like the kingmen playing louis louis and the sonics playing stuff. louis louis is like a fratboy song*

> *the sonics are totally subversive they're totally punk rock the mood of them is totally different so its like they fit in the same genre but what they're about is so different and that is the bit that I find disappointing it doesn't give you thematic recommendations it gives you genre recommendations and that can be handy but there are other ways to link music together other that genre you know? #00:11:10-1# - P(1)*

The thematic aspect of the curation task and the notion of playlists was one that many participants commented on and it often served to juxtapose how they might normally listen to music or receive it from conventional music recommenders. Typically most systems provide recommendations as a sequential set of tracks that although similar to one another (or at least liked by people with similar tastes) collectively ascribe to no particular class, theme or structure. There is generally no notion of internal coherence or significance to the order in which the tracks are recommended.

It should be acknowledged that increasingly modern recommenders have begun to produce broadly thematic recommendations albeit without any internal structure or ordering. However these thematic recommendation attempts are still fairly ridged in using fixed broad themes like genre or era. Manually produced playlist made by real people can be built around much subtler themes. Several participants commented on this and expressed enjoyment in using thematic playlists. They remarked how thinking in this way freed them from the stricter similarity or genre groupings used by traditional recommenders.

> *Spotify … just groups it into particular genres and I'm [not] a kind of person that listens to a particular genre #00:04:41-8# - P(3)*

3.3.3.1.8.3 Existing Recommenders Don't Aid Discovery

Another common criticism of existing recommender systems is that they have no inbuilt mechanism to aid discovery. They typically recommend similar tracks to seed songs or other users but don't give users the ability to vary or filter what is recommended to reflect how adventurous they might be feeling. This criticism is an interesting one as it highlights a strength of the earlier manual systems which relied heavily on such mechanisms and gives credence to the notion that recommenders might be improved by including humans in the recommendation process.

3.3.3.1.8.4Time to Create Playlists

One of the fairly common comments made towards the end of an interview was that participants had enjoyed the study as it forced them to take the time to make playlists, something which they often typically didn't feel that they had time to do. A participant expresses this clearly in the quote below:

> *" I think I would like to have certain playlists but then usually the playlists you make yourself are so short that it is just get repeats and then you get annoyed with them and um and it takes time to make it any longer and it is just one of those things I don't really have time for #00:03:49-1# - P(10)*

3.3.4 Discussion: Personalisation & Considerations for Human-in-the-loop Systems

At a high level, this study revealed that recommendation purpose, personal memories and human involvement were very important for participants in curating personalised music playlists. At a more fine-grained level, the central findings can be broken down and summarised as follows:

- Unimportance of Genre
- Existing Systems Fail to Aid Novelty or Discovery
- The Need for Automated Assistance with Human Oversight
- Reduced Importance of Predictive Accuracy
- The Importance & Characterisation of Mood in Music Curation
- The Importance of Personal Memories & Nostalgia
- The Role of Purpose

In isolation these results highlight the importance of several key factors in personalising music curation: purpose, nostalgia and memories and human involvement. When considered in the context of the personalisation problem and contrasted against a backdrop of existing literature, the result yields far more insight into likely design parameters of an effective HITL solution. The following section is divided into those findings which are supported by the literature, those which conflict with it and those which are entirely absent from it.

3.3.4.1 Findings Which Support the Literature

3.3.4.1.1 Unimportance of Genre

Most participants at one point in their interview stated that their playlist contained tracks from a multitude of genres. Some went further to say that they actively didn’t find genre to be a particularly useful way of categorising or curating music.

The limitations of genre-based classification and the desire for more flexible classifications does seem to find support from the literature. By the mid-2000s researchers had begun to comment on the limitation of genre (Shirky, 2005). In a paper in 2006 Cory McKay and Ichiro Fujinaga acknowledged that one of the difficulties with rigid genre classification is that genres often change and refine over time. The need for more flexible categorisation started to become accepted as the consensus view by (Baccigalupo, Plaza, & Donaldson, 2008).

One of the popular techniques that has emerged to try and provide a more fine-grained and useful means of classifying and recommending music is social tagging in which users of a system simultaneously submit tags to reference content as they see fit (Nanopoulos, Rafailidis, Symeonidis, & Manolopoulos, 2010). These groups of tags can express genre related concepts or content-based concepts concerning rhythm or any other arbitrary means of labelling, but they can be used collectively to reference an item. Furthermore, the cloud of tags associated with an item will be updated and kept current by the community of users doing the tagging over time. As Paul Lamere states, “the real value of these tags emerges when the tags are aggregated into a single, shared pool, sometimes referred to as a folksonomy” (Lamere, 2008).

3.3.4.1.2 Existing Systems Fail to Aid Novelty or Discovery

Another area where the study appears to support the findings from the literature is in the limitations of existing systems to aid novelty and the discovery of new content (Yu-Shian Chiu, Kuei-Hong Lin, & Jia-Sin Chen, 2011). Although for a long time this was not considered important, the role of surprise and the importance of facilitating discovery has begun to be increasingly recognised over the last 10 years (Adomavicius & Kwon, 2011; Celma & Lamere, 2008; Iaquinta et al., 2008; McNee, Riedl, & Konstan, 2006a).

3.3.4.1.3 The Need for Automated Assistance with Human Oversight

The requirement for human involvement in music recommender systems was acknowledged as early as 2008. In the passage below, Byrd comments on some of the unique aspects of human music creation, curation and discovery which were reflected in the findings of this study. Chiefly that humans unavoidably combine objective facts about the music they are seeking with their own contextual knowledge including their personal memories and feelings which current computation systems have no means of accounting for in any truly semantic sense.

> *Music is created by humans for other humans, and humans can bring a tremendous amount of contextual knowledge to bear on anything they do; in fact, they can't avoid it, and they're rarely conscious of it. But (as of early 2008) computers can never bring much contextual knowledge to bear, often none at all, and never without being specifically programmed to do so. Therefore, doing almost anything with music by computers is very difficult; many problems are essentially intractable (Byrd, n.d.).*

Several participants commented on the importance of having people provide them with references as they felt existing systems attempted to quantify them rather than gain any qualitative appreciation for their requirements. Traditionally within the literature since the first automated systems, the goal has invariably been to minimize human involvement. However, in recent years the need for human oversight and the increased value which can be obtained from such systems, particularly in regard to facilitating discovery and surprise, has begun to be recognized (Nanopoulos et al., 2010).

Recent papers have begun to stress the need to incorporate humans to provide a degree of surprise (Kefalidou & Sharples, 2016), with some going as far as to suggest that the objective of recommender systems needs to be reevaluated and extended beyond simple content suggestion to facilitating human interaction (Pontis et al., 2015).

3.3.4.2 Findings Which Conflict with The Literature

3.3.4.2.1 Reduced Importance Predictive Accuracy

One area where the results of this study appear to conflict with the prevailing view of both music recommender literature and commercial recommender design, is on the importance of accuracy. As has been seen in the literature review, most existing systems maintain that predictive accuracy is the most important factor in improving a recommender system. In more recent years this view has been challenged to some degree by the realisation that accuracy needs to be tapered by appropriateness in order to be useful (Kefalidou & Sharples, 2016) (Herlocker, Konstan, Terveen, & Riedl, 2004; McNee, Riedl, & Konstan, 2006a; Yu-Shian Chiu et al., 2011).

Although at times participants commented on being frustrated by inaccurate recommendations, more frequently they were frustrated by the inability of a system to provide them with non-obvious recommendations that they might like.

3.3.4.2.2 The Importance & Characterisation of Mood in Music Curation

The results of this study suggest that mood might be over emphasised within the research community in its importance in music recommendation or at least mis-understood. Most of the current research in this area focuses on signal processing and categorising music by abstract moods (J. Kim, Lee, & Yoo, 2013; Sasaki, Hirai, Ohya, & Morishima, 2013; Song, Kim, Rho, & Hwang, 2009).

Only one participant chose to think of mood in this way. Most participants instead sought to find tracks to match or enhance a specific event they had in mind. Mood was important in so far as the tracks they chose were invariably supposed to make them feel a certain way; safe, nostalgic or relaxed. However, the primary aim seemed to be matching or enhancing a given event and altering or affecting their mood was thought of as a result of this specific aim.

3.3.4.3 Findings Which Are Absent from The Literature

3.3.4.3.1 The Importance of Personal Memories & Nostalgia

Considering the role of personal memories and nostalgia was so important to 8 of the 11 participants, it is surprising that this theme seems entirely absent from commercial recommender systems and the music recommender literature.

Perhaps this reflects the division between human qualifying and computational quantifying that was discussed earlier. If so, this provides further support for the idea that a HITL system might be able to bridge this division and provide a best of both worlds solution to the problem of personalised music recommendation.

3.3.4.3.2 The Role of Purpose

Although a fair amount of research has been conducted into context-aware music recommendation in recent years, much of this research has focused on broad notions of context that focus on the user of a system (Adomavicius & Tuzhilin, 2011; Hariri et al., 2012; X. Wang, Rosenblum, & Wang, 2012). This study revealed that a narrower part of context, specifically the intended usage of a series of recommendations or playlist, seemed to be more important than attempting to gain more information about the user seeking the recommendation. To date, very little research seems to have been done precisely on how the intended purpose of a recommendation might alter the way in which someone attempts to get a recommendation or the information they might want when looking for a recommendation (Ricci, 2012). This topic will be explored further in the next chapter.

3.3.5 Limitations & Future Research

The objective of this study was to gain insight into the personal human process of music curation for the benefit of furthering the development of HITL music recommender systems. To this end the specific playlists users created were not directly relevant as the resulting tracks do little more than serve as useful reference sheets in the interview when it comes to gaining an insight into the human processes being described by the participants. This be said, in a future investigation it could be useful to perform a meta-analysis of the interviews and playlists created to explore any overlap between participants. Did those who chose similar moods or described similar processes have any overlapping content in the playlists – similar titles, tracks, rhythmic patterns or approaches towards playlist structure for instance?

Future research could also be conducted by choosing a different personalised category for participants to curate their playlists other than by mood. The results could then be contrasted to see if different curation criteria altered either the factors or the proportional significance of the factors revealed as important to personalisation in this study.

3.3.6 Conclusions

Overall our findings from this chapter seem largely to be in agreement with the emerging consensus of the field that personalisation is achieved through the right balance of appropriateness and novelty (Celma, 2008; Zhang et al., 2012). Where this thesis differs from Celma's conclusions is in emphasising the importance of recommendation purpose, i.e. what a recommendation is intended for. Finally, this thesis also takes a short-term view of dynamism, suggesting that for a particular individual their reason for seeking a recommendation will often affect their immediate taste preferences.

3.4 Chapter Summary

This chapter has explored the nature of personalisation and identified that personalisation requires a delicate balance between novelty, appropriateness and an appreciation for recommendation purpose. The study also highlighted how people's tastes are dynamic and can vary over a short time frame depending upon their reasons for wanting to listen to music. The study also highlighted the importance of personal memories in shaping the way people naturally approach the task of music curation. This latter point, in conjunction with the literature, suggests that incorporating a certain level of human oversight could prove essential in personalising music recommender systems by giving people the room to bring their personal memories and individualistic search criteria to bear on the results of the system.

The next chapter expands on the findings of this study and delves deeper into the role of purpose and explores precisely what constitutes purpose and how the expected purpose of a playlist affects people's judgement as to the musical knowledge of others.

Chapter 4: The Role of Purpose

4.1 Introduction

The previous chapter revealed that an appreciation of purpose plays a significant role in making music curation and recommendation personal. Virtually all participants set about the task of creating their personal playlists by first imagining a scenario or setting in which they intended to use their playlist. This envisaged scenario then seemed to heavily influence the songs that participants went searching for and how they structured their playlists.

This chapter provides an in-depth exploration of the role of purpose by investigating how the purpose for which a recommendation is sought can affect the judgement people make as to the musical knowledge of others. Answering this question is important since it reveals what information users need access to in order for a HITL system to be able to accommodate this requirement of purpose in influencing music recommendations.

The chapter is broken down into the following structure:

- Case study: Investigating the role of recommendation purpose
- A Human in the Loop Approach: Outlining how a HITL music recommender might work and how it could account for recommendation purpose

4.2 Case Study: The Role of Purpose

4.2.1 Rationale

In the previous study, participants often reported that they felt frustrated by the repetitive nature of recommendations and the lack of user control. For instance, one participant commented that it would be advantageous to be able to express a preference for non-explicit lyrics or to filter music recommendations to remove explicit tracks if they were intending to play their music around children.

> *so, it would be useful to have something more specific that you can just click, and you know that stuff either appropriate for children or appropriate for your mood is going to come up. P(4) #00:07:01-6#*

This feedback highlights the importance of two key concepts that are underrepresented in modern music recommendation, the notion of recommendation purpose (i.e. for whom and under what circumstance is the music intended to be played) and the notion of user control and intervention.

The importance of these concepts seems to be increasingly recognised and reflected in a recent shift towards researching serendipity rather than just predictive accuracy within recommender systems. This is discussed further in the related work section of this chapter. The importance of user control gives support to the idea that a HITL systems might be effective at combating the personalisation problem as they are designed precisely to enable user involvement and control. To facilitate the design of such a system however more detail is required on the significance of purpose.

The aim of this study is to explore whether people’s estimation as to the music knowledge of others about a specific type of music altered if they were informed that the music was being sought for a specific purpose.

This question is very important in guiding the design of a HITL music recommendation system. In the preceding chapters, we saw from the literature and exploratory study, that the consideration of purpose was often central to whether or not a recommendation was well received and/or thought to be personal. In order to design a HITL recommendation system that capitalises on this, we need to know whether different sorts of purpose affect users differently.

A HITL system allows users to pre or post filter the content and/or userbase of a given recommendation system. If part of making this strategy successful in providing personalised recommendations requires supporting the consideration of purpose during this filtering step then we must know whether and how different purposes affect people.

For instance, if it turns out that people consider an individual's listening history to be more useful in generating their recommendation when they are looking to curate music for a social setting like a house party then the system might provide a way of tagging or labelling that user as a good candidate for inclusion in house-party recommendations.

4.2.2 The Experimental Conditions

Three different experimental conditions representing distinct reasons or purposes for curating music/seeking recommendations were investigated in this study:

Temporally bounded purpose: Looking for music from a particular time
Content bounded purpose: Looking for music with particular instrumental characteristics
Socially bounded purpose: Looking for music to suit/be used in a particular social setting

Note the first two purposes for seeking recommendations address features about the music and its origins. The third purpose for curating a set of music and seeking recommendations is distinct from the other two in that has nothing directly to do with the music itself but rather how it is intended to be used.

These three types of purpose were chosen based on the previous study which indicated that people often take account of one or more of these factors when curating music manually. Since HITL music recommenders aim to facilitate a similar manual curation element it seemed natural to consider these same factors when investigating how purpose might affect people's musical judgements.

Of course, the work done in this study could be extended in future work by considering other possible types of purpose. The results could be made more robust by also re-running the study using other specific examples for the three categories (temporal, content and social bounded purpose) that we choose to investigate in this study.

4.2.3 Method

The study was run as a crowdsourcing task in which participants were asked to assess the musical knowledge of six individuals (whom they had been given limited information about) and rank their usefulness in aiding an individual in finding a music recommendation for a specific purpose. It was conducted between 7th July and 3rd September 2016, running for a total of 8 weeks.

4.2.3.1 Participants

The study was conducted on 400 participants over the age of 18 from 56 countries spanning 6 continents. All participants were recruited via the crowdsourcing platform, Crowdflower.

Participants’ demographics	N (%)
Age	
18+	400 (100%)
Continent	
Africa	3
Antarctica	0
Asia	113
Australia	2
Europe	154
North America	35
South America	93
Country	
Albania	1
Algeria	1
Argentina	6
Australia	1
Bolivia	1
Bosnia and Herzegovina	9
Brazil	21
Bulgaria	2
Canada	10
Chile	1
Colombia	2
Croatia	4
Czech Republic	2
Denmark	1
Estonia	1
Finland	1
France	4
Germany	5
Greece	6
Hong Kong	2
Hungary	2
India	71
Indonesia	3
Ireland	1
Israel	1
Italy	8
Japan	1

Lithuania	3
Macedonia, the former Yugoslav Republic of	1
Malaysia	4
Malta	1
Mexico	9
Nepal	1
Netherlands	1
New Zealand	1
Pakistan	1
Philippines	5
Poland	6
Portugal	7
Romania	5
Russian Federation	30
Serbia	12
Slovakia	1
South Africa	1
Spain	19
Suriname	1
Switzerland	1
Taiwan, Province of China	1
Tunisia	1
Turkey	8
Ukraine	14
United Kingdom	6
United States	25
Uruguay	1
Venezuela, Bolivarian Republic of	51
Viet Nam	15

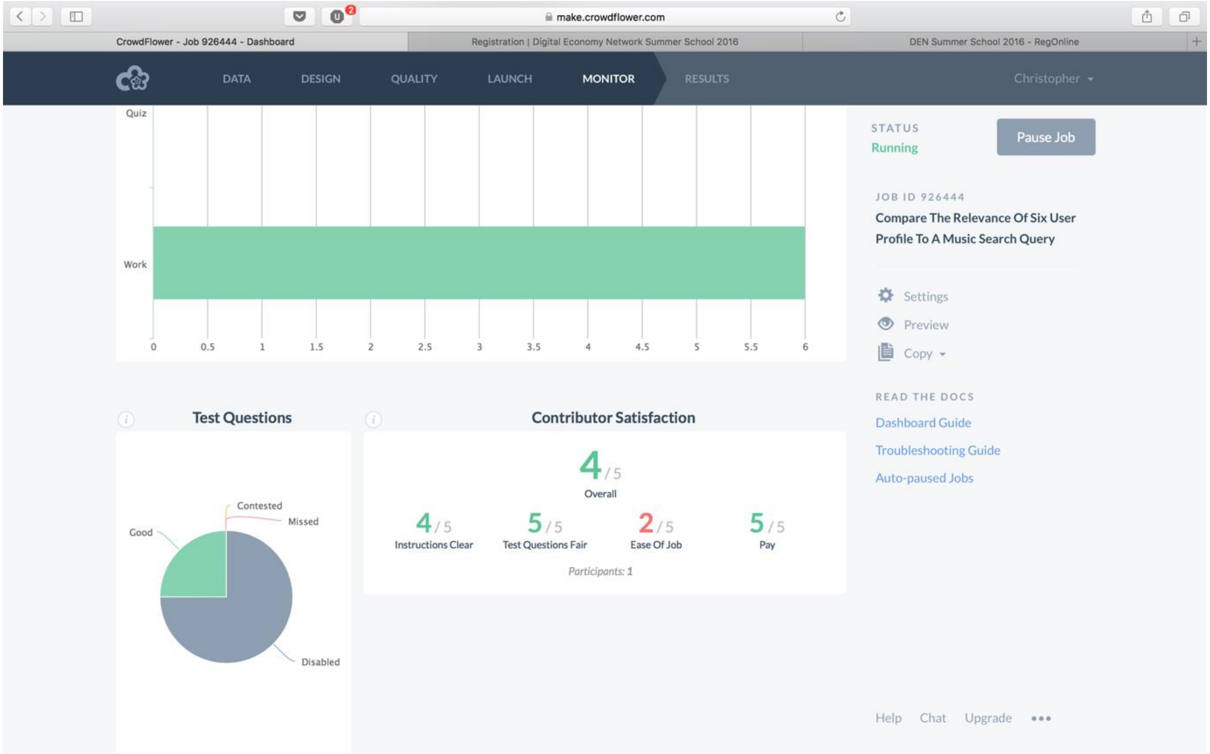
Figure 13: Participants demographics and country table

Participants were compensated for their time in accordance with the standard payment model adopted by Crowdflower. The model compensates participants for their time by paying them 10 cents for each page they complete, where a page consists of a limited number (typically 3 to 5) of short tasks or questions. In the case of this study, each participant completed a single page containing 3 questions and 1 quality control question and hence were paid 10 cents.

4.2.3.2 Materials

4.2.3.2.1 Crowdsourcing Platform: Crowdflower

The crowdsourcing platform Crowdflower was chosen as the framework upon which to construct and distribute this study. There were several reasons for this. First, it allowed us to rapidly conduct several pilot studies until we were confident in the phrasing of the questions and design of the task. Second, since the data returned from these pilots was in the same form as the data returned from our actual study, we were also able to prototype, develop and fine-tune our data analysis approach. During the piloting stage, we were also able to view feedback from participants about the design, ease and perceived fairness of the study (i.e. did they feel adequately compensated for their time). This was very useful in allowing us to improve the design of the final study. A screenshot of the feedback mechanism is shown below:



STATUS
Running

JOB ID 926444
Compare The Relevance Of Six User Profile To A Music Search Query

4/5 Overall

4/5 Instructions Clear
5/5 Test Questions Fair
2/5 Ease Of Job
5/5 Pay

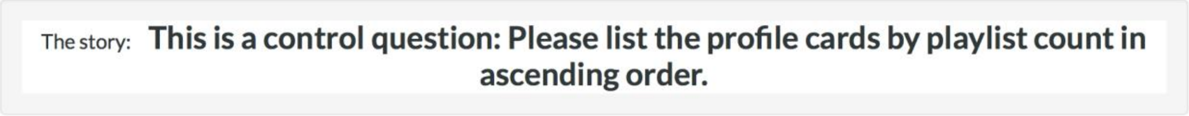
Participants: 1

Figure 14: Crowdsourcing feedback mechanism

Using Crowdflower also gave us access to a large and far more diverse set of participants than would have been feasible if we had chosen to run the study locally as a laboratory experiment. This in turn aids the experimental validity and representational nature of our findings. To an extent it mitigates the impact of the white educated industrial rich demographic (WEIRD) bias to our findings. The WEIRD bias is a common criticism first made against psychology research but in more recent years against human-computer interaction (HCI) research. The criticism centres around the idea that a lot of psychology and HCI studies are run on small (< 30) opportunistically sampled participant groups which typically consist of other university members and academics that make up a narrow white educated industrialised rich demographic (WEIRD). The criticism states that there is a danger of over generalising results from this non-representational demographic to a wider population group, or otherwise neglecting to research other important demographics. Since the findings of this study are intended to inform the development of a general framework for facilitating human intervention in music recommenders, it is important that the results be as representational as possible.

Crowdflower was used in preference to more widely known Mechanical Turk systems as currently Mechanical Turk is not available outside of the US. Furthermore, Crowdflower is slightly more diverse in the range of crowdsourcing tasks it supports.

Finally, Crowdflower was chosen for its built-in accuracy and bias prevention features. Specifically, it allowed us to construct a series of verifiable control tasks and automatically rejected participants who failed to get 99% of these correct. The control tasks were constructed in such a way as to make it evident if participants answered questions at random without reading the profile cards. A typical control task is shown below:

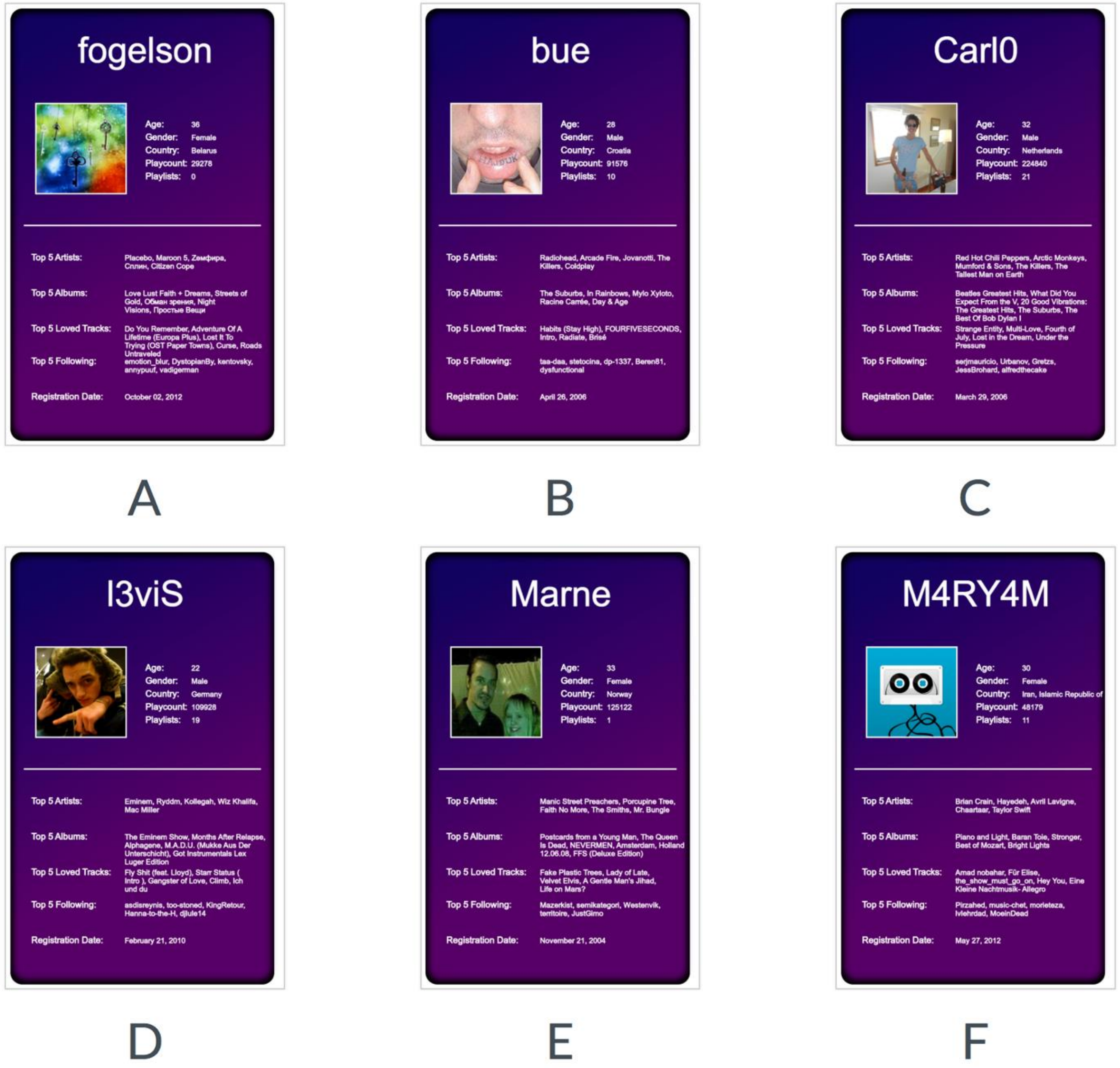


The story: **This is a control question: Please list the profile cards by playlist count in ascending order.**

Figure 15: Control question example

4.2.3.2.2 User Profile Cards

The study consisted of a ranking task in which participants were asked to sort six individuals according to their perceived musical knowledge in a specific area. Information about these individuals was abstracted from randomly selected LastFM profiles and presented to participants in a trading card format shown below.



A

fogelson

Age: 36
Gender: Female
Country: Belarus
Playcount: 29278
Playlists: 0

Top 5 Artists: Placebo, Maroon 5, Земфира, Сплин, Citizen Cope

Top 5 Albums: Love Lust Faith + Dreams, Streets of Gold, Обман зрения, Night Visions, Простые Вещи

Top 5 Loved Tracks: Do You Remember, Adventure Of A Lifetime (Europa Plus), Lost It To Trying (OST Paper Towns), Curse, Roads Untraveled

Top 5 Following: emotion_blur, DystopianBy, kentovsky, annypuuf, vadigerman

Registration Date: October 02, 2012

B

bue

Age: 28
Gender: Male
Country: Croatia
Playcount: 91576
Playlists: 10

Top 5 Artists: Radiohead, Arcade Fire, Jovanotti, The Killers, Coldplay

Top 5 Albums: The Suburbs, In Rainbows, Mylo Xyloto, Racine Carrée, Day & Age

Top 5 Loved Tracks: Habits (Stay High), FOURFIVESECONDS, Intro, Radiate, Brisé

Top 5 Following: taa-daa, stetocina, dp-1337, Beren81, dysfunctional

Registration Date: April 26, 2006

C

Carl0

Age: 32
Gender: Male
Country: Netherlands
Playcount: 224840
Playlists: 21

Top 5 Artists: Red Hot Chili Peppers, Arctic Monkeys, Mumford & Sons, The Killers, The Tallest Man on Earth

Top 5 Albums: Beatles Greatest Hits, What Did You Expect From the V, 20 Good Vibrations: The Greatest Hits, The Suburbs, The Best Of Bob Dylan I

Top 5 Loved Tracks: Strange Entity, Multi-Love, Fourth of July, Lost in the Dream, Under the Pressure

Top 5 Following: serjmauricio, Urbanov, Gretzs, JessBrohard, alfredthecake

Registration Date: March 29, 2006

D

l3viS

Age: 22
Gender: Male
Country: Germany
Playcount: 109928
Playlists: 19

Top 5 Artists: Eminem, Ryddm, Kollegah, Wiz Khalifa, Mac Miller

Top 5 Albums: The Eminem Show, Months After Relapse, Alphagene, M.A.D.U. (Mukke Aus Der Unterschicht), Got Instrumentals Lex Luger Edition

Top 5 Loved Tracks: Fly Shit (feat. Lloyd), Starr Status (Intro), Gangster of Love, Climb, Ich und du

Top 5 Following: asdisreynis, too-stoned, KingRetour, Hanna-to-the-H, djlule14

Registration Date: February 21, 2010

E

Marne

Age: 33
Gender: Female
Country: Norway
Playcount: 125122
Playlists: 1

Top 5 Artists: Manic Street Preachers, Porcupine Tree, Faith No More, The Smiths, Mr. Bungle

Top 5 Albums: Postcards from a Young Man, The Queen Is Dead, NEVERMEN, Amsterdam, Holland 12.06.08, FFS (Deluxe Edition)

Top 5 Loved Tracks: Fake Plastic Trees, Lady of Late, Velvet Elvis, A Gentle Man's Jihad, Life on Mars?

Top 5 Following: Mazerkist, semikategori, Westenvik, territoire, JustGimo

Registration Date: November 21, 2004

F

M4RY4M

Age: 30
Gender: Female
Country: Iran, Islamic Republic of
Playcount: 48179
Playlists: 11

Top 5 Artists: Brian Crain, Hayedeh, Avril Lavigne, Chaartaar, Taylor Swift

Top 5 Albums: Piano and Light, Baran Toie, Stronger, Best of Mozart, Bright Lights

Top 5 Loved Tracks: Amad nobahar, Für Elise, the_show_must_go_on, Hey You, Eine Kleine Nachtmusik- Allegro

Top 5 Following: Pirzahed, music-chet, morteteza, Mehrdad, MoeinDead

Registration Date: May 27, 2012

Figure 16: Profile cards

The 6 profile cards used in this study were randomly selected from a deck of 60 cards representing a virtual userbase to be used and explored throughout the thesis. The process below details how the cards were produced.

The first step in producing the cards was compiling a list of potential usernames which could be run against LastFm's public API to extract profile data. To generate this list I used a random wordlist generation application called crunch. The command shown below was used to generate a plain text file called NAMES.txt containing 146829 random alphanumeric words between 3 and 8 characters in length.

```
crunch 3 8 abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789 -o NAMES.txt
```

After generating the potential usernames the next step was to construct a python script to run these potential usernames against LastFM's database using their public api. If the request returned a valid profile the data was saved in a JSON file named after the found user. If no data was found the script would move on and try another potential username.

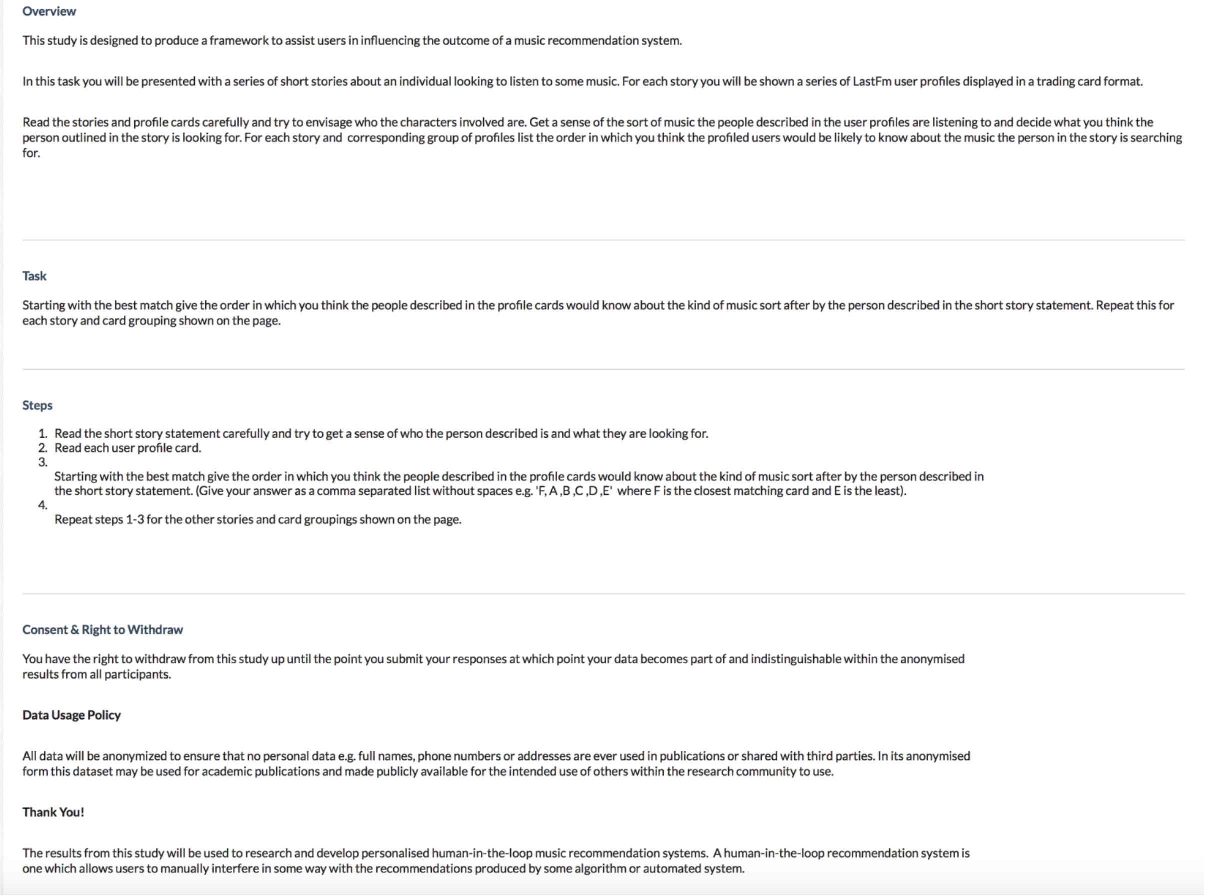
Once 60 profiles were found the next stage was to clean and reshape the data and add randomly allocated ages between 18-80 to the selected profiles. The age field was randomly generated as LastFM no longer supports ages in its user profiles. Cleaning involved checking for and removing incomplete profiles if any were returned. As it happens all our profiles were fully populated. The reshaping step got the data in the correct flat format to be used in the desired fields for our profile cards.

Finally, once the data had been cleaned and shaped another python script was created to generate trading cards from the JSON username files and save each card as a PNG graphic. The scripts mentioned in this section are provided in the appendix.

4.2.3.3 Procedure

4.2.3.3.1 Briefing Participants

Upon selecting our study in the Crowdflower list of tasks, participants were presented with the information sheet shown below:



Overview

This study is designed to produce a framework to assist users in influencing the outcome of a music recommendation system.

In this task you will be presented with a series of short stories about an individual looking to listen to some music. For each story you will be shown a series of LastFm user profiles displayed in a trading card format.

Read the stories and profile cards carefully and try to envisage who the characters involved are. Get a sense of the sort of music the people described in the user profiles are listening to and decide what you think the person outlined in the story is looking for. For each story and corresponding group of profiles list the order in which you think the profiled users would be likely to know about the music the person in the story is searching for.

Task

Starting with the best match give the order in which you think the people described in the profile cards would know about the kind of music sort after by the person described in the short story statement. Repeat this for each story and card grouping shown on the page.

Steps

1. Read the short story statement carefully and try to get a sense of who the person described is and what they are looking for.
2. Read each user profile card.
3. Starting with the best match give the order in which you think the people described in the profile cards would know about the kind of music sort after by the person described in the short story statement. (Give your answer as a comma separated list without spaces e.g. 'F, A ,B ,C ,D ,E' where F is the closest matching card and E is the least).
4. Repeat steps 1-3 for the other stories and card groupings shown on the page.

Consent & Right to Withdraw

You have the right to withdraw from this study up until the point you submit your responses at which point your data becomes part of and indistinguishable within the anonymised results from all participants.

Data Usage Policy

All data will be anonymized to ensure that no personal data e.g. full names, phone numbers or addresses are ever used in publications or shared with third parties. In its anonymised form this dataset may be used for academic publications and made publicly available for the intended use of others within the research community to use.

Thank You!

The results from this study will be used to research and develop personalised human-in-the-loop music recommendation systems. A human-in-the-loop recommendation system is one which allows users to manually interfere in some way with the recommendations produced by some algorithm or automated system.

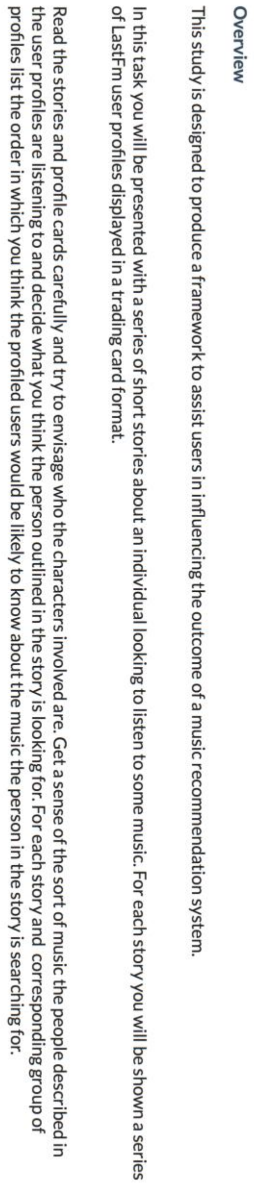
Figure 17: Information sheet

The information sheet consisted of a study overview, task description and consent section. Each section is presented in turn below:

4.2.3.3.1.1 Study Overview

The study overview shown below informed participants of the aim of study and made them aware that the results were intended to be used to construct a framework for enabling users to influence the results of music recommendation systems.

The overview concluded by providing a high-level description of what would be expected of them in the study which can be read below:



Overview

This study is designed to produce a framework to assist users in influencing the outcome of a music recommendation system.

In this task you will be presented with a series of short stories about an individual looking to listen to some music. For each story you will be shown a series of LastFm user profiles displayed in a trading card format.

Read the stories and profile cards carefully and try to envisage who the characters involved are. Get a sense of the sort of music the people described in the user profiles are listening to and decide what you think the person outlined in the story is looking for. For each story and corresponding group of profiles list the order in which you think the profiled users would be likely to know about the music the person in the story is searching for.

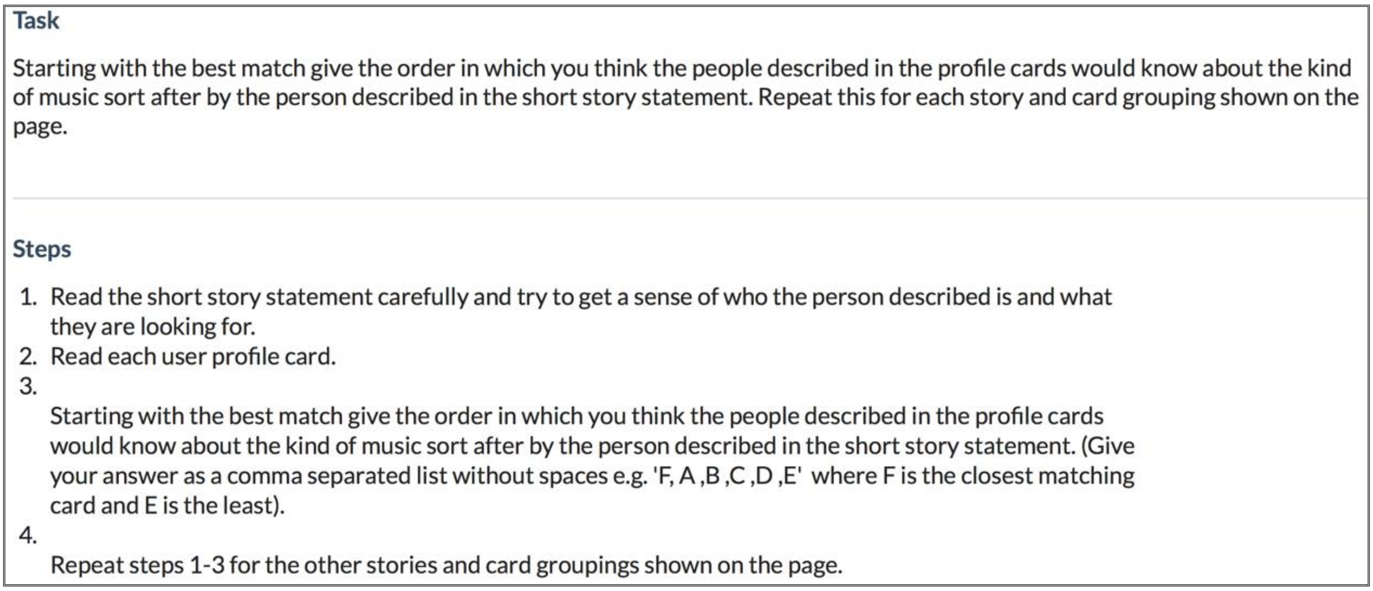
Figure 18: Crowdsourcing overview excerpt

4.2.3.3.1.2 Task Description

For each of the experimental conditions, participants were presented with a narrative statement about an individual who was seeking music for a specific purpose (the specific purposes used are presented in full later in the procedure).

Along with the narrative statements, participants were shown the same six profile cards for each condition and asked to provide an order as to how knowledgeable they thought the people on the cards would be about the type of music the person in the statement was looking for.

The task description is shown below:



Task

Starting with the best match give the order in which you think the people described in the profile cards would know about the kind of music sort after by the person described in the short story statement. Repeat this for each story and card grouping shown on the page.

Steps

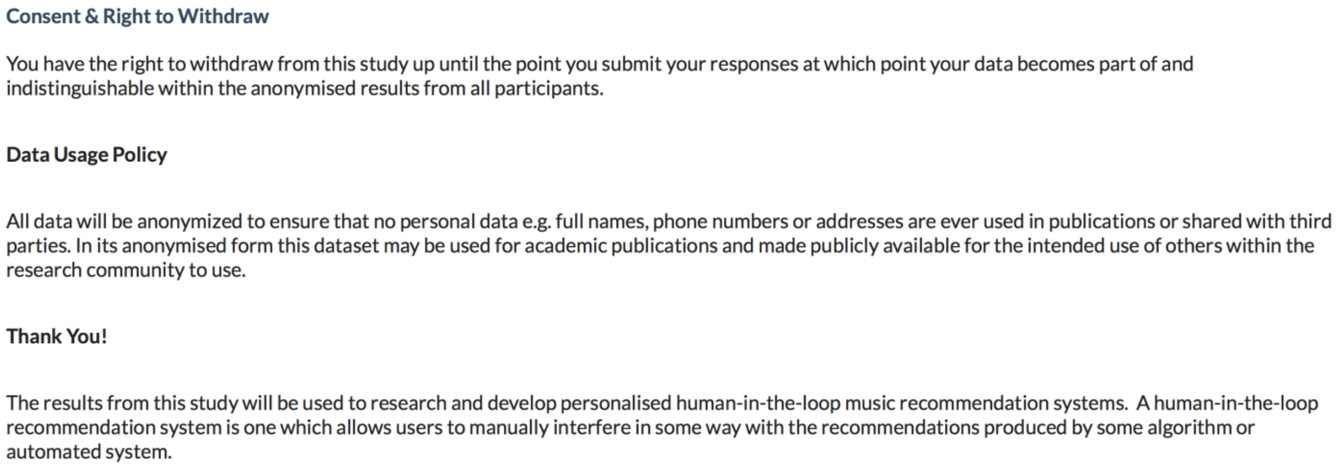
1. Read the short story statement carefully and try to get a sense of who the person described is and what they are looking for.
2. Read each user profile card.
3. Starting with the best match give the order in which you think the people described in the profile cards would know about the kind of music sort after by the person described in the short story statement. (Give your answer as a comma separated list without spaces e.g. 'F, A ,B ,C ,D ,E' where F is the closest matching card and E is the least).
4. Repeat steps 1-3 for the other stories and card groupings shown on the page.

Figure 19: Crowdsourcing task description

Consent

The consent statement informed participants of their right to withdraw from the study at any time up until they submitted their results at which point, due to the anonymous nature of Crowdflower’s data gathering approach, it would no longer be possible to remove their results from the aggregate of collected data.

Additionally, participants were informed that no deanonymized or personal data including full names, phone numbers or address details would be used in publications or shared with third parties.



Consent & Right to Withdraw

You have the right to withdraw from this study up until the point you submit your responses at which point your data becomes part of and indistinguishable within the anonymised results from all participants.

Data Usage Policy

All data will be anonymized to ensure that no personal data e.g. full names, phone numbers or addresses are ever used in publications or shared with third parties. In its anonymised form this dataset may be used for academic publications and made publicly available for the intended use of others within the research community to use.

Thank You!

The results from this study will be used to research and develop personalised human-in-the-loop music recommendation systems. A human-in-the-loop recommendation system is one which allows users to manually interfere in some way with the recommendations produced by some algorithm or automated system.

Figure 20: Crowdsourcing consent notification

4.2.3.3.2 Procedural Walk-through: What the Participants Did

Having described the study task above here we provide a walk-through of what the participants did.

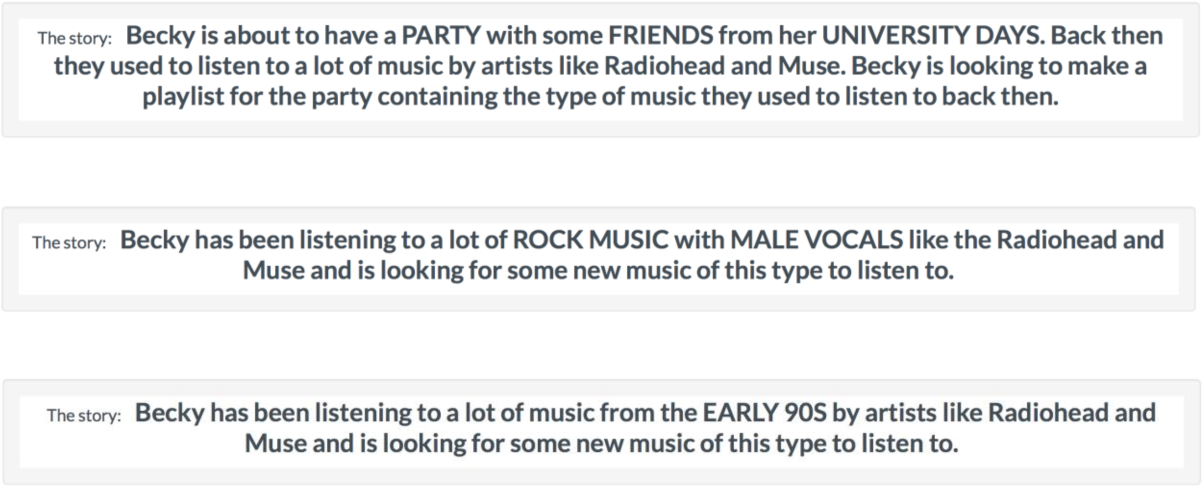
In completing the study the participants did the following:

1. Read an overview of the study
2. Read and signed the information and consent form
3. Read two test questions designed to ensure that participants had read the user profile cards before providing their responses
4. Read each of the three experimental conditions:
    1. Read the narrative statement to discover the type of music that was being looked for and purpose for which it was being sought
    2. Assessed how much knowledge each of the individuals on the profile cards might have about the music being sought and form an order list from most knowledgeable to least
    3. Submit their list as a comma-delimited set of the letters, A-F used to label the profile cards

Below we provide a little more detail about what participants did and present each experimental condition.

Once participants had reviewed the instruction / consent form and agreed to take part in the study they were presented with a screen showing the three experimental conditions in a randomised order.

For each condition participants were presented with one of the narrative statements shown below and the six profile cards shown earlier (the profiles were randomly ordered and then labelled from A-F for each condition).



The story: **Becky is about to have a PARTY with some FRIENDS from her UNIVERSITY DAYS. Back then they used to listen to a lot of music by artists like Radiohead and Muse. Becky is looking to make a playlist for the party containing the type of music they used to listen to back then.**

The story: **Becky has been listening to a lot of ROCK MUSIC with MALE VOCALS like the Radiohead and Muse and is looking for some new music of this type to listen to.**

The story: **Becky has been listening to a lot of music from the EARLY 90S by artists like Radiohead and Muse and is looking for some new music of this type to listen to.**

Figure 21: Narrative statements

In each of the three statements Becky is described as looking for some music for a specific purpose reflecting one of the three experimental conditions; social, content and temporal. In the first statement, Becky is seeking a music recommendation for a social setting (a party with university friends). In the second statement, Becky is interested in the content of the music (she wants her recommendations to be rock music containing male vocals). In the final statement, Becky is seeking music from a particular era (the 1990s).

For each experimental condition, participants were asked to read the narrative statement and determine how knowledgeable they thought each of the individuals described in a series of profile cards would be about the type of music being sought after by the person in the narrative statement. The participants were then asked to sort the 6 profiles according to how knowledgeable they considered each profile was in respect of the music Becky is looking for. Finally, participants were asked to submit their rankings as a comma delimited list of the letters A-F corresponding to the labelled profile cards.

4.2.4 Data Quality Measures

4.2.4.1 Anti-biasing, Order Effects & Randomisation

To avoid ordering bias, we randomised the order in which participants were told to complete the experimental conditions. Additionally, we randomized the order of the profile cards for each condition. This was done to deter participants from simply copying and pasting the same sequence of letters between each condition and to make it easy to see if they did so.

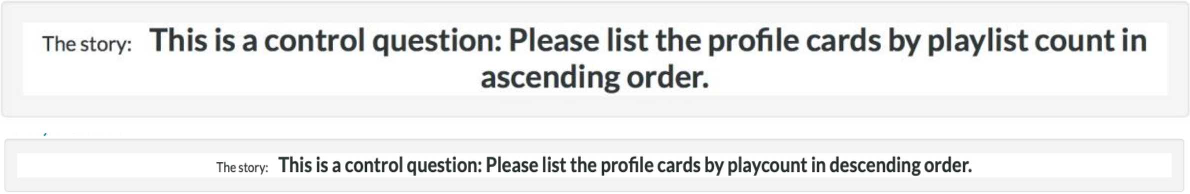
Finally, for anonymity purposes, unique numeric ids were used in place of participant names to identify them.

4.2.4.2 Participant Accuracy & Concentration

We took several measures to ensure our participants were concentrating and engaged with the study at all times to help ensure the high quality of our results.

The first step we took was to only allow the highest tiered Crowdflower contributors to attempt our study. High tier Crowdflower contributors are ones who have repeatedly and routinely demonstrated to have completed tasks in the manner requested by the task designer.

The second step we took to ensure our participants were engaged was to incorporate two control questions which quizzed potential participants on the content of the user profile cards. The control questions used are shown below:



The story: **This is a control question: Please list the profile cards by playlist count in ascending order.**

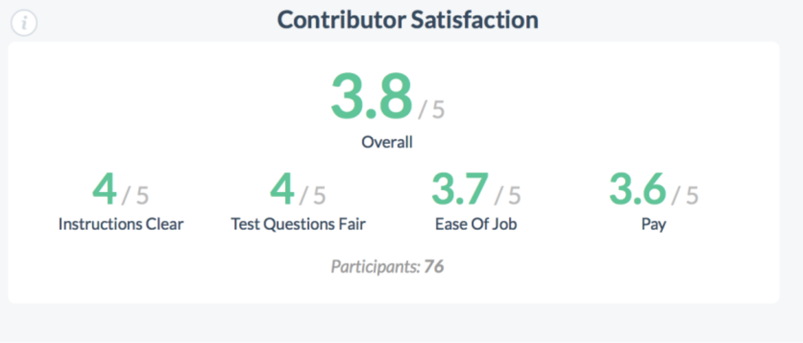
The story: **This is a control question: Please list the profile cards by playcount in descending order.**

Figure 22: Control questions

All potential participants were made to complete the control questions before proceeding to the main study questions. If they failed to answer the control questions correctly, then they were omitted from the participants pool and rejected from the study. Using these control questions also helped to ensure that participants had read the profile cards and weren't just responding with a random order of letters.

To ensure participants gave responses in the correct form, we used regular expressions to validate the entries before allowing them to submit. This made sure that the responses they gave were indeed a comma delimited list of the letters A-F in some order.

We also ran five pilot studies in which we gathered participants for feedback on the clarity of the study, ease of the task, fairness of the questions and compensation. This helped to ensure that when we ran the final study, participants were invested because they felt the task was clear and their compensation appropriate to the level of work they were being asked to do. The final results for user satisfaction are shown in the Crowdflower screenshot below:



Contributor Satisfaction

3.8 / 5
Overall

4 / 5
Instructions Clear

4 / 5
Test Questions Fair

3.7 / 5
Ease Of Job

3.6 / 5
Pay

Participants: 76

Figure 23: User satisfaction ratings

Finally, we enforced a minimum time limit of 30 seconds before participants could submit their responses to help ensure they read the study and thought about their responses rather than just submitting as fast as they could in order to get compensated for their efforts.

4.2.5 Design Justification

Having outlined the study procedure and design it remains to justify why this design pattern was chosen. There were several reasons for this some of which were touched upon earlier in the materials sections. In bullet form there are 5 main reasons why this study was designed as it was:

1. Using Crowdflower provides access to a large international pool of participants something which is not possible using a laboratory study design
2. Crowdflower participants closely approximate the technical savvy userbase that a HITL music system based on this research would be designed for
3. Crowdflower supports automated quality control and anti-biasing methods to ensure quality data acquisition
4. Crowdflower facilitates semi-automated rapid prototyping something which again is more time consuming and less structured in most laboratory study design patterns
5. Finally once designed the automated process of recruitment and execution allows for a vast amount of data to be gathered in a relatively short period of time

Within the narrative of this thesis, this study aimed to further our understanding of purpose to allow us to draw out some conclusions and design principles for creating a HITL music recommender. To aid in the validity of these principles and attempt to avoid generating design guidelines based on regional patterns it was important our participant base closely resemble the vast international technically proficient demographic that might make up the userbase of a real music recommender service like Spotify. Making use of Crowdflower helped us achieve this to a greater extent than a smaller local laboratory design would have.

The automated measure of quality control also helped us to increase the chances that participants engaged in the study. This is generally harder to achieve as a study scales as you cannot afford to spend as much time guiding or monitoring the attentiveness of any one participant.

Finally deploying our study via Crowdflower means every process from design to deployment and data gathering is documented and recorded. This has the advantage that it could be repeated accurately and easily in the future to validate results. Alternatively, it could be modified with little effort to gather more data using more examples of purpose than the three chosen.

4.2.6 Analysis Approach

Since the study was rather large and the methodology fairly complex, we have opted to explain our analysis approach by presenting a small synthetic example. In this example, which is presented below, we go through the procedure of gathering and analysing data from five fictitious participants.

4.2.6.1 Example Participants

Let the five fictitious participants for our example be called Alice, Bob, Carol, Dave and Erin.

4.2.6.2 Example Profile Cards

In the actual study, as previously mentioned, we swapped the ordering of the profile cards between conditions such that the card showing the profile card for Carlo might have been labelled as card A in one condition and card B in another. We have chosen to omit this quality assurance measure from our example here to aid clarity.

For this example, let the blank cards labelled A-F below represent our lastFM profile cards for Fogelson, Bue, Carl0, L3viS, Marne and Ma4y4m respectively and let the ordering remain static across each of the three conditions.

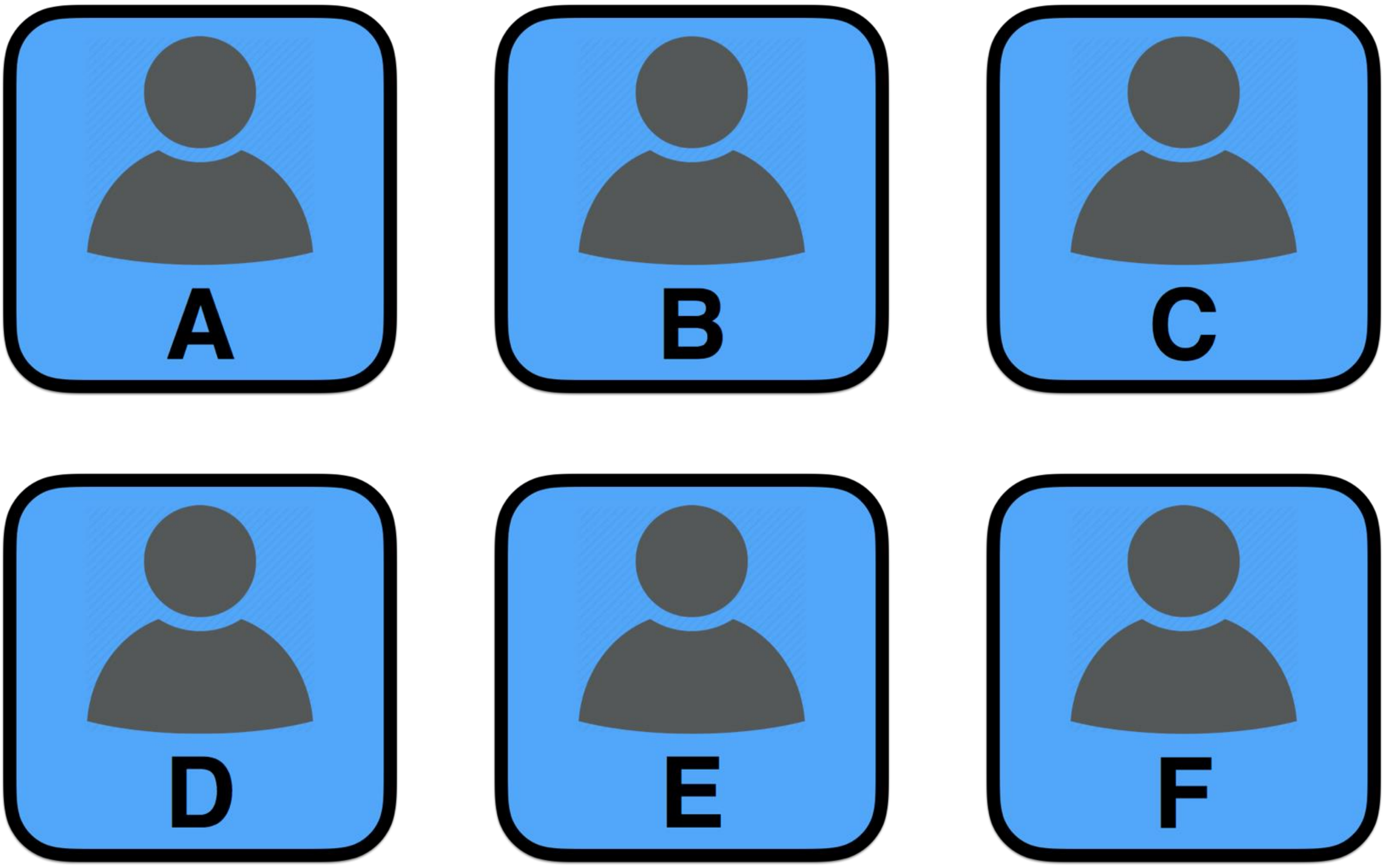


Figure 24: Mock profile cards

4.2.6.3 Raw Data

Give the above setup our raw answer dataset for our synthetic example might look like the table below:

Participant ID	Condition 1	Condition 2	Condition 3
Alice	ABCDEF	DABCEF	ABCDEF
Bob	ABCDEF	ABCDEF	ABCDEF
Carol	BCADEF	ABCDEF	DACBFE
Dave	ABCDEF	AFEDCB	ABCDFE
Erin	ABCDEF	AFBCDE	FBCDEA

Figure 25: Mock results table

Each row displays a given participant's responses with each cell uniquely identifying their response for a given experimental condition. Remember from above that the letters A-F stand for the profile cards of Fogelson, Bue, Carl0, L3viS, Marne and Ma4y4m respectively. Therefore Alice's response for experimental condition 1 can be interpreted as follows: From most to least knowledgeable, Alice's judgement is: Fogelson, Bue, Carl0, L3viS, Marne and Ma4y4m as given by the order 'ABCDEF' above.

4.2.6.4 Analysing the Results – Similarity Metrics

To assess whether the different experimental conditions affected our participants' responses we needed to measure the similarity between their individual responses across the conditions.

4.2.6.5 Cards in Same Position

To get an initial indication of this we chose to record how many profiles each participant put in the same position in their orderings across each experimental condition. In the case of our example dataset, this finding is presented below:

Participant ID	Positions occupied by same profile across all conditions						Frequency
	First	Second	Third	Fourth	Fifth	Sixth	
Alice					X	X	2
Bob	X	X	X	X	X	X	6
Carol							0
Dave				X			1
Erin							0

Figure 26: Mock profile positioning results

Looking at our synthetic data, we can see that Alice placed two cards (E, F) in the same positions (fifth and sixth respectively) for each condition. Most of the participants placed very few profiles in the same position across the conditions with only one participant (Bob) placing all six profiles in the same positions across all conditions. This implies that our experimental conditions affected most of our participants' assessments of the musical knowledge of others to some extent.

However, this result alone fails to provide any insight into the degree to which participants varied their overall rankings across conditions. The significance of this can be better understood by contrasting Carol and Erin's responses. Neither participant placed a single profile card in the same position across the conditions. However, it would intuitively appear that Erin's responses are far more alike than Carol's. In Erin's responses, the cards A, B, C and D are always grouped in the same relative order whilst Carol's results seem radically different for each condition (this can be seen clearly in the figure in 2.4.3).

Counting the number cards participants placed in the same position is useful for identifying whether participants changed their responses at all. However, as contrasting the results of Carol and Erin has shown, a subtler measure is needed to assess the degree to which individual participants varied their responses overall. To obtain this measure we had to construct a normalised means of assessing the extent to which participants altered their orderings across conditions. We opted to use a form of a string distance metric known as the Damerau–Levenshtein score.

4.2.6.5.1 Damerau–Levenshtein Score Explained in Context

The Damerau–Levenshtein score is a string comparison metric commonly used in linguistics and computational biology to measure the variation between genetic sequences. It works by counting the number of alterations, additions, deletions and transpositions it takes to convert one text string to another. For instance, changing the word ‘cool to ‘cook’ requires one alteration (changing l to k) and so obtains a Damerau–Levenshtein score of 1.

An important limiting feature of the Damerau–Levenshtein score for this study is that it doesn’t allow transpositions between the first and last characters of strings. Consider the string sequences ‘1324’ and ‘1234’. To convert ‘1324’ to ‘1234’ using the basic Levenshtein score (which doesn’t allow transpositions), we would have to make 2 alterations (changing 3 to 2 and 2 to 3). Using the Damerau–Levenshtein score, we can simply transpose the middle two characters of the first string which is counted as a single edit. However, if we consider the strings ‘2341’ and ‘1234’, the Damerau–Levenshtein score will not allow us to transpose the first and last characters and instead forces us to make two edits by deleting the ‘1’ at the end of the string and re-adding it to the beginning of the string. Whilst this restriction makes sense for certain applications like spell checking, it does not make sense when comparing the variance of our participants’ responses across conditions. To understand why, consider Carol and Erin’s answers from our synthetic data:

Participant ID	Condition 1	Condition 2	Condition 3
Carol	BCADEF	ABCDEF	DACBFE
Erin	ABCDEF	AFBCDE	FBCDEA

Figure 27: Mock results excerpt

Neither participant places a single card in the same position across all three conditions. However, on inspection, it appears clear that Erin’s answers are more similar than Carol’s. Erin places the profiles BCDE in the same relative order across all conditions. The only change she makes to her ordering is to the profiles A and F. Her estimations as to the relative knowledgeability of the other profiles remain unchanged. By contrast, Carol changes the relative ordering of profile cards in her responses entirely across the conditions. This suggests that, unlike Erin, the different conditions did cause her to re-evaluate the relative level of knowledge of each profile for each condition.

To account for this type of occurrence in our data, we modified the Damerau–Levenshtein score to allow transpositions between the first and last characters of strings.

We then used this modified metric to produce a pairwise similarity matrix of our participants' responses to gain an understanding of the level of similarity between the responses across the conditions.

Using this technique on our synthetic data produces the similarity matrix shown in the table below:

Participant ID	Condition 1 – Condition 2	Condition 1 – Condition 3	Condition 2 – Condition 3
Alice	2	0	2
Bob	0	0	0
Carol	2	4	4
Dave	4	1	4
Erin	2	1	2

Figure 28: Mock similarity matrix

Each cell in the matrix contains a score from 0-6 reflecting the number of alterations required to convert a participant order in one condition to their order in another.

4.2.6.6 Assessing Participant Response Similarity Across Conditions

From the matrix above, it is possible to construct the following frequency table and reveal the degree to which participants varied their responses across each condition. Remember that a score of 6 would represent total ordering change, whilst a score of 0 would represent no change in order.

Pairwise Conditions	Damerau–Levenshtein score						
	0	1	2	3	4	5	6
Condition 1- Condition 2	2	0	2	0	1	0	0
Condition 1- Condition 3	2	1	1	0	1	0	0
Condition 2- Condition 3	1	0	2	0	2	0	0

Figure 29: Mock data Damerau-Levenshtein scores table

Graphing the similarity matrix for our synthetic data as a bar chart reveals the proportion of participants who varied their responses from the minimum to maximum degree (0-6 respectively) across each pairwise condition comparison.

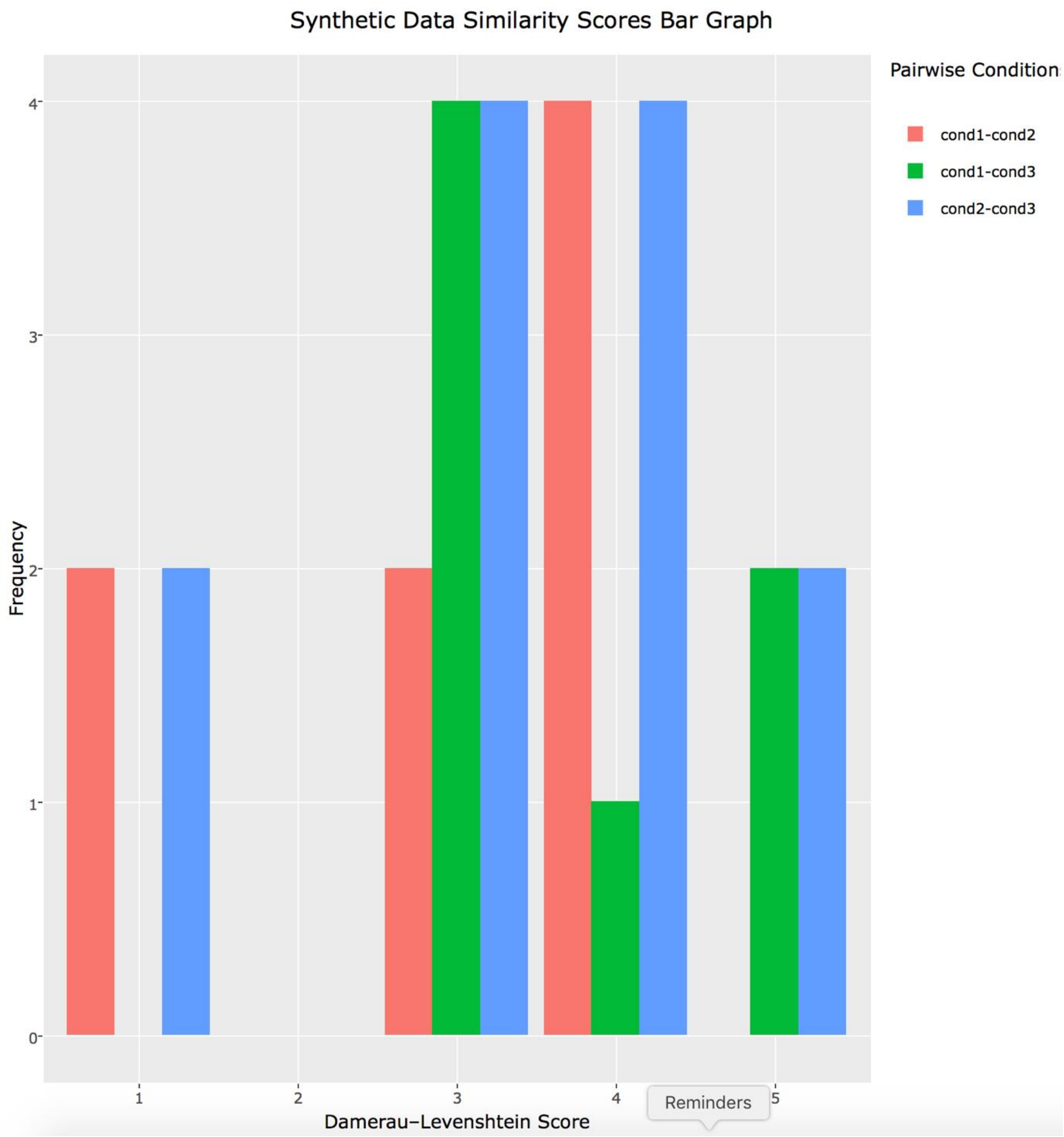


Figure 30: Synthetic data similarity scores graph

From the graph above, it can be observed that all participants varied their responses a moderate degree placing three or four profiles in different positions across the conditions. In this mock example it can also be observed that participants varied their responses less between conditions 1 and 2 than either 1 and 3 or 2 and 3. This might suggest that conditions 1 and 2 have more in common with each other than with the other experimental conditions.

4.2.7 Results

Having explained our analysis approach, we now proceed to present the findings from our study data.

4.2.7.1 Initial Approximation of Participant Similarity

Graphing the percentage of participants who placed 1-6 profile cards in the same position for each condition provides an initial estimate as to the level of variability in participants' responses across the conditions. From the chart below we can infer that there is quite a high level of variability in participants' responses across conditions as only 15.3% or 61 of the 400 participants placed all six profiles in the same position for each condition. By contrast, 39% or 156 participants didn't place a single card in the same position across each condition.

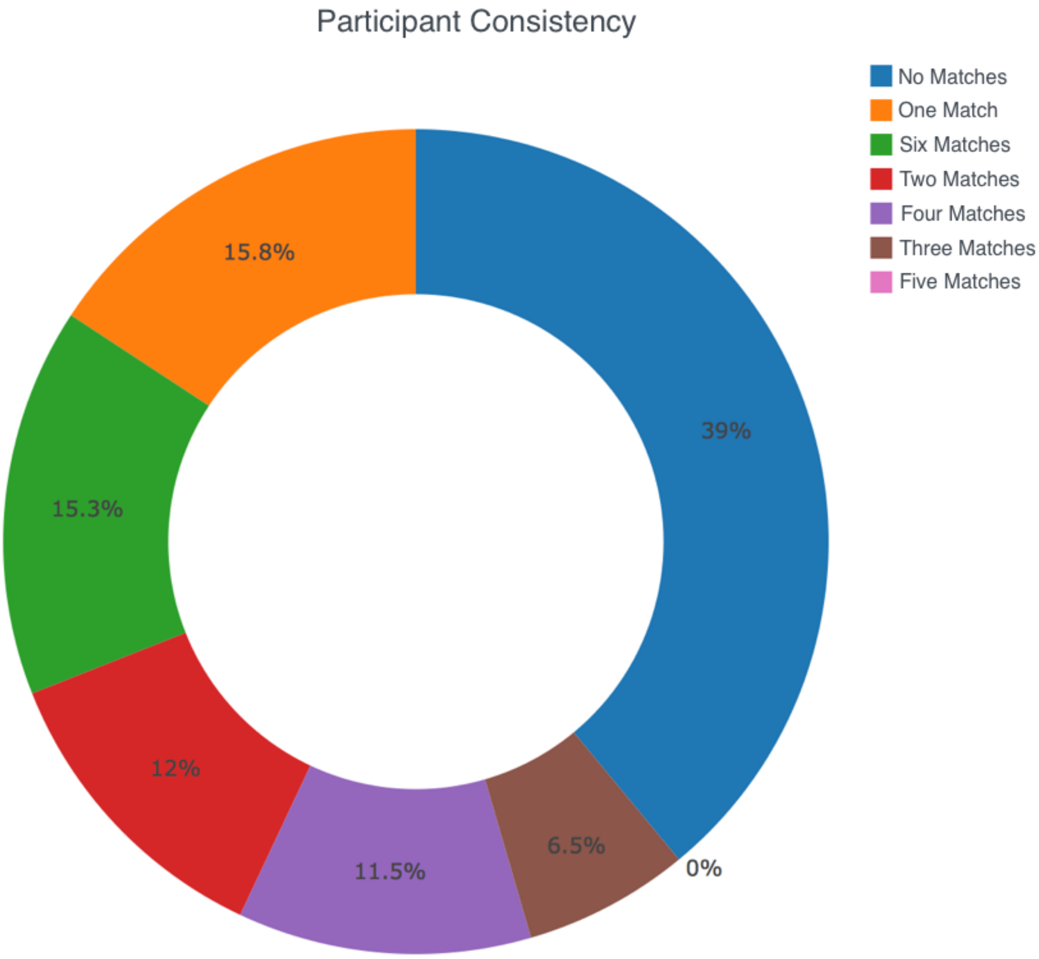


Figure 31: Participant consistency graph

4.2.7.2 Levenshtein Similarity Results

Applying our modified Damerau–Levenshtein measure of similarity to our results matrix produced the following pairwise Damerau–Levenshtein Score table (first 10 of 400 rows):

CONTENT-TEMPORAL	CONTENT-SOCIAL	TEMPORAL-SOCIAL
4	4	6
0	0	0
4	2	2
2	2	4
4	0	4
3	4	6
0	0	0
4	0	4
4	4	5
4	5	3

Figure 32: Similarity scores table excerpt

For each pairwise comparison we calculated the number of participants who altered 0-6 characters (table shown below).

	0	1	2	3	4	5	6
CONTENT-TEMPORAL	81	14	57	61	144	36	7
CONTENT-SOCIAL	85	19	67	45	137	37	10
TEMPORAL-SOCIAL	88	14	67	44	70	31	86

Figure 33: Experimental conditions table

We then generated a histogram to display this information.

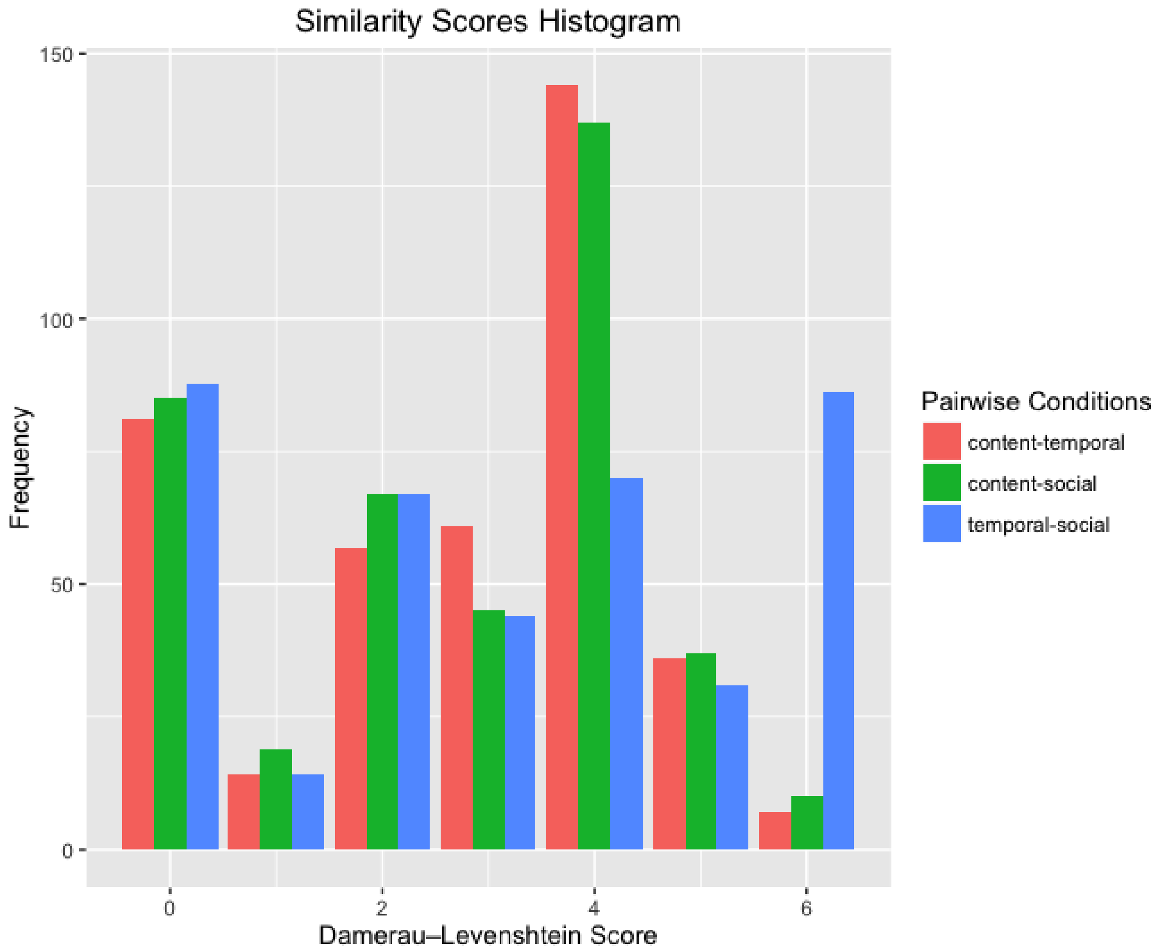


Figure 34: Similarity scores histogram

The above histogram shows a similar rate of agreement between participants' consistency for the content-temporal and content-social conditions with most participants changing over 50% of their rankings.

4.2.7.3 Significance Testing

Having presented the results of our study, it remains to see whether or not these findings have any statistical significance. To assess this, we conducted a series of statistical tests which we now present below:

```
$statistics
    Chisq Df      p.chisq       F DFerror          p.F  t.value      LSD
  34.01731  2 4.104266e-08 17.7196     798 2.954329e-08 1.962941 43.85479

$parameters
      test    name.t ntr alpha
  Friedman condition   3  0.05

$means
                 value rankSum      std   r Min Max Q25 Q50 Q75
content-social   2.675   771.5 1.769804 400   0   6   1   3   4
content-temporal 2.745   752.5 1.712398 400   0   6   2   3   4
temporal-social  3.050   876.0 2.182753 400   0   6   1   3   5

$comparison
                                   difference pvalue signif.     LCL    UCL
content-social - content-temporal        19.0 0.3953          -24.85  62.85
content-social - temporal-social       -104.5 0.0000     *** -148.35 -60.65
content-temporal - temporal-social     -123.5 0.0000     *** -167.35 -79.65

$groups
NULL

attr(,"class")
[1] "group"
```

Figure 35: Significance tests

First, we wanted to establish where there was a statistical significance to the variance we found in participants' similarity scores for each pairing of the experimental conditions. Since our data was non-parametric and no assumptions could be made about the distribution of the data, we determined a Friedman test on the similarity matrix results would be the most appropriate test to perform.

Our hypothesis for this test was as follows:

H0: There is no significant difference in the degree to which participants make similar rankings across each condition pairing.
H1: There is a significant difference in the degree to which participants make similar rankings across each of the condition pairings.

What are the results?

34.017, df = 2, p-value = 4.104e-08 α=0.05

The Friedman test found a significant difference across the pairwise conditions with a p-value of *1.799e-07*. At a 95% confidence level α is equal to 0.05. Since 1.799e-07 < 0.05 we reject the null hypothesis H0 and accept the alternative hypothesis H1. We can therefore state that at a 99% confident level there is a significant difference in the degree to which our participants made similar rankings across each of the condition pairings.

The above result reveals that there was a significant difference in the degree to which participants made similar rankings across the pairwise conditions, but it doesn't reveal which of the pairwise conditions were significantly different from one another. For this we need to perform a post hoc analysis. This is presented below:

```
                              difference pvalue signif.     LCL    UCL
content-social - content-temporal        19.0 0.3953          -24.85  62.85
content-social - temporal-social      -104.5 0.0000     *** -148.35 -60.65
content-temporal - temporal-social     -123.5 0.0000     *** -167.35 -79.65
```

Figure 36: Post hoc analysis results

The post hoc analysis shows that there is a significant difference between the content-social and temporal-social and content-temporal and temporal-social but there was not a significant difference between the content-social and content-temporal.

4.2.8 Discussion

Our results give evidence to suggest that all three aspects of purpose considered, social, temporal and content based seem to have some bearing on how people assess the musical knowledge of others when considering how useful they might be in recommending a particular sort of music. From our results is appears that participants' judgement varied to a similar extent across the content-temporal and content-social conditions with most participants changing over 50% of their rankings in each case. By contrast, the social-temporal condition appeared to polarise participants into two groups; those who entirely changed their rankings, and those who didn't change their rankings at all. This is interesting and provides further evidence to support our finding from the previous chapter, i.e. the importance of the time period or social setting tended to polarise people. They either cared about it greatly, or they did not care about it at all.

This has interesting connotations for the design of a HITL music recommender as it suggests that the requirement to selectively reveal or filter, not just individual attributes about users and their tastes but entire categories, could be helpful. A system which incorporated this level of flexibility could help users quickly access the most relevant profile information to them which would help them to make their decision and interact with the recommendation system.

4.2.8.1 A final note on participant behaviour and interesting patterns

In addition to presenting the results above, given the richness of the dataset, it seems worth asking a few additional questions of the data and commenting on the results. Specifically, what were the most common orderings given overall and as well as for each condition? Secondly, if we take straight mean average of the participants' responses what orderings do we get overall and for each condition then? Below is a set of tables which present the answers to these questions.

4.2.8.1.1 Top 10 Most Common Orders: Overall

Ranking	Number of Participants (out of 400)
bue,Carl0,Marne,fogelson,l3viS,M4RY4M	73
bue,Carl0,fogelson,Marne,M4RY4M,l3viS	60
Marne,Carl0,bue,l3viS,M4RY4M,fogelson	56
bue,Carl0,fogelson,Marne,l3viS,M4RY4M	55
bue,Carl0,Marne,fogelson,M4RY4M,l3viS	54
bue,fogelson,Carl0,l3viS,M4RY4M,Marne	43
Marne,l3viS,M4RY4M,Carl0,bue,fogelson	40
M4RY4M,Carl0,fogelson,l3viS,bue,Marne	39
bue,Carl0,fogelson,M4RY4M,Marne,l3viS	29
bue,fogelson,Carl0,Marne,M4RY4M,l3viS	28

4.2.8.1.2 Top 10 Most Common Orders: Content Condition

Ranking	Number of Participants (out of 400)
M4RY4M,Carl0,fogelson,l3viS,bue,Marne	39
bue,Carl0,Marne,fogelson,l3viS,M4RY4M	33
bue,Carl0,fogelson,Marne,l3viS,M4RY4M	25
bue,Carl0,fogelson,Marne,M4RY4M,l3viS	23
Marne,Carl0,bue,l3viS,M4RY4M,Fogelson	19
bue,Carl0,Marne,fogelson,M4RY4M,l3viS	18
bue,fogelson,Carl0,Marne,M4RY4M,l3viS	12
bue,Carl0,fogelson,l3viS,Marne,M4RY4M	11
bue,Carl0,l3viS,fogelson,Marne,M4RY4M	7
bue,fogelson,Carl0,Marne,l3viS,M4RY4M	7

4.2.8.1.3 Top 10 Most Common Orders: Social Condition

Ranking	Number of Participants (out of 400)
Marne,l3viS,M4RY4M,Carl0,bue,fogelson	40
bue,Carl0,Marne,fogelson,l3viS,M4RY4M	25
bue,Carl0,Marne,fogelson,M4RY4M,l3viS	25
bue,Carl0,fogelson,Marne,M4RY4M,l3viS	23
Marne,Carl0,bue,l3viS,M4RY4M,fogelson	19
bue,Carl0,fogelson,Marne,l3viS,M4RY4M	15
bue,Carl0,fogelson,M4RY4M,Marne,l3viS	14
bue,fogelson,Carl0,Marne,l3viS,M4RY4M	11
bue,fogelson,Carl0,Marne,M4RY4M,l3viS	9
bue,Carl0,M4RY4M,fogelson,Marne,l3viS	8

4.2.8.1.4 Top 10 Most Common Orders: Temporal Condition

Ranking	Number of Participants (out of 400)
bue,fogelson,Carl0,l3viS,M4RY4M,Marne	41
Marne,Carl0,bue,l3viS,M4RY4M,fogelson	18
bue,Carl0,fogelson,Marne,l3viS,M4RY4M	15
bue,Carl0,Marne,fogelson,l3viS,M4RY4M	15
bue,Carl0,fogelson,Marne,M4RY4M,l3viS	14
bue,Carl0,Marne,fogelson,M4RY4M,l3viS	11
bue,Carl0,fogelson,M4RY4M,Marne,l3viS	10
bue,Carl0,M4RY4M,fogelson,l3viS,Marne	10
bue,Carl0,Marne,l3viS,fogelson,M4RY4M	10
bue,Marne,Carl0,fogelson,M4RY4M,l3viS	9

4.2.8.1.5 Averaged Order

Card	Content	Social	Temporal	Overall
fogelson	4.1050	4.1975	4.2150	4.172500
bue	3.9550	3.5825	3.9950	3.844167
Carl0	4.1275	3.4600	3.8150	3.800833
l3viS	3.5725	3.5375	3.5250	3.545000
Marne	2.7600	2.9500	2.6425	2.784167
M4RY4M	2.4800	3.2725	2.8075	2.853333

4.2.8.1.5.1 Rankings

4.2.8.1.5.1.1 Content

M4RY4M, Marne, l3viS, bue, fogelson, Carl0

4.2.8.1.5.1.2 Social

Marne, M4RY4M, Carl0, l3viS, bue, fogelson,

4.2.8.1.5.1.3 Temporal

Marne, M4RY4M, l3viS, Carl0, bue, fogelson,

4.2.8.1.5.1.4 Overall

Marne, M4RY4M, l3viS, Carl0, bue, fogelson

4.2.8.1.6 My Order

Condition	Ranking
Content	M4RY4M, Marne, l3viS, Carl0, bue, fogelson,
Social	Marne,M4RY4M, Carl0, l3viS, bue, fogelson
Temporal	Marne, M4RY4M, l3viS, bue, fogelson, Carl0
Overall	Marne, M4RY4M, l3viS, Carl0, bue, fogelson

Table 1: My Orderings

For interest and comparison, I attempted the study task myself and came up with orderings above. One curious thing I encountered when doing the task was that depending on the purpose I found myself looking at different aspects of the profile cards. For instance, when it came to the social party setting I found myself looking at the age of the people on the profiles with the general guiding principle that people of similar ages or generations tend to have more of a shared musical background. When it came to the content setting however the age category on the card had no relevance for me. In designing a system it may be worth considering how profile cards could be filtered to display or hide different attributes at the user's request to account for this self-filtering behaviour.

Overall my orders can be seen to be fairly close to that of my participants giving me some confidence that people were indeed engaged with the task and not just picking randomly. The fact that Fogelson, for instance, is consistently ranked as being less knowledgeable suggest that whilst there were individual disagreements, in general, there were also common strongly held similarities in people’s assessments. If participants had just been selecting at random the average ranking scores would have been much closer together with all users having a roughly equal chance of being placed in any one position.

The fact that bue is often placed in the first position for each category also suggests that participants read the cards as bue is the only profile which expressly lists Radiohead (a band mentioned in every search scenarios) as a favourite artist.

4.2.9 Limitations & Suggestions for Future Work

Crowdflower was an excellent platform for recruiting a large diverse sample group for this study but it does have several limitations which are worth mentioning when considering the findings of this study. One of the biggest limitations is that Crowdflower lacks a means of performing a post study debriefing or facilitating a direct dialogue with participants at any point during the study. Whilst this may aid the validity of study design by reducing the possibility for observer / conductor influence, it prohibits the design of longitudinal studies or the conducting of follow up interviews. This limitation can be overcome to a degree by including additional instructions and contact details in the study description and consent form. However, it represents a departure from standard proceedings and requires careful consideration and planning. To keep things simple in this study, a decision was made to conduct it as an isolated as opposed to longitudinal study and accept that this prohibits us from doing follow up work with precisely the same study group. Since our results are intended to be highly representative and generalised to a wide degree, this should not be thought of as a major limitation.

Another complexity with this study that took careful consideration was its multivariable design. Two features being assessed are whether individuals' estimations of the musical knowledge of others tends to converge, and whether different purposes for seeking music recommendations affect people's judgements as to the musical knowledge of others. Incorporating both of these research components into a single study made the data analysis procedure lengthy and complex. Although there is no reason to suppose this impacted the findings of the study, if conducted again it might be more efficient to adapt the design and run it as two smaller studies, each focusing on only one research objective.

A broader criticism that could be levelled against this study is that crowdsourcing participants represent a skewed demographic of the population, specifically those which have regular access to the internet and are fairly technologically savvy. Whilst this is a valid criticism of crowdsourcing more generally, it does not apply here as the intended population demographic we wish to model, i.e. music recommender service users are a group of individuals who have access to the internet and are reasonably technologically savvy.

The results of this study reveal that purpose does have a significant impact on how people assess the musical knowledge of others, but it does not clearly indicate how it affects them, only showing that different individuals appear to be affected in different ways.

Future studies could push the findings of this study further by investigating precisely how the purpose for which a person is seeking a music recommendation affects their assessment of the musical knowledge of others.

In this study three specific scenarios were provided to represent the main purposes for which people sought music recommendations. These scenarios were chosen based on prior studies in conjunction with the findings of the literature within the field. If the study were repeated using the same scenarios but a different sample group and the same results were found, then this would suggest our findings have a high degree of external validity and generalisability. The internal validity of this study could also be improved by rerunning it using a wider range of scenarios to represent each distinct purpose for seeking a music recommendation. If the same results were found for each scenario representing a given purpose, this would go further to suggesting our findings here were not the result of any confounding influence introduced though the choice of scenarios used.

4.3 Chapter Summary

The focus of this chapter has been to explore what purpose means in the context of music recommendations and how different purposes affect people's judgements as to the musical knowledge of others. The results of the study presented in this chapter support the conclusion that recommendation purpose is important for people when seeking recommendations and assessing the musical knowledge of others to assist in those recommendations. However, it also reveals that there is no one universal notion or role for purpose within the recommendation problem. When considering recommendation purpose, different aspects seems to matter to a greater or lesser extent to different individuals. For instance, some care a great deal about the time period of music and people's knowledge of this when seeking recommendations and other don't seem to value this aspect at all. This variation in individuals' considerations of attributes when seeking recommendations is something we address in the design exercise in the next chapter.

Chapter 5: A Design Exercise

5.1 Introduction

This chapter combines the findings from the previous chapters and proposes how they could be incorporated in the design of a personalised HITL system.

5.2 Considerations From This Thesis

The literature review and first exploratory study in this thesis revealed the highly subjective nature of musical taste. It identified that a core aspect of personalisation was the ability of a system to identify and satisfy the varying demands of different users.

The exploratory study presented in the third chapter further showed that individual users also have varying demands and **dynamic tastes** depending on factors like the **time** they are using a system, the **social situation** they are in and their **purpose** for seeking recommendations. The literature review supported this finding and further showed that most conventional systems fail to account for the dynamic nature of tastes by simply representing users as singular monolithic taste or listening history vectors.

The fourth chapter explored the significance of these factors and revealed that there was statistically significant variation across study participants in the extent to which temporal and social factors were important to them in assessing the musical knowledge of others.

With regards to designing a HITL music recommender, this supports the findings in the literature from chapter 2 which revealed that users have differing requirements from one another and will use different categorisations and attributes to one another when categorising their music and asking for recommendations.

To summarise, the core considerations for designing a HITL music recommender capable of creating personalised recommendations identified in this thesis are:

- The highly subjective nature of taste
- The dynamic nature of taste
- The significance of purpose in seeking recommendations
- The differing importance of certain classification attributes like when a track was released or its genre to different people
- The differing importance of particular classification attributes to individuals at different times or in different situations

5.3 Restaurant Analogy

5.3.1 The Personalisation Problem Re-examined

One of the aims of this thesis was to provide a greater understanding of precisely what the personalisation problem has been for music recommendation systems. In previous chapters this has been undertaken at the micro-level by conducting an in-depth investigation of the various technical issues that contributed to users feeling recommenders lacked personalisation. In this chapter we return to the macro-level at which the problem was first outlined in the introduction of this thesis and revisit the restaurant analogy. There are two reasons for doing this. First, it enables us to collate the findings of this thesis and present them clearly in the context of the broader problem originally presented. Second, it provides a design pattern and example of how the findings in this thesis could be adopted and implemented by designers of music recommender systems.

In the vernacular of the food court / restaurant analogy, the personalisation problem for music recommenders can be summarized as follows. Conventional music recommendation systems tend to favour popularist collaborative filtering algorithms like item-item and nearest-neighbour approaches due to their scalability and ease of implementation. From the user's perspective the problem with these algorithms is they generalise and make broad assumptions which don't cater for the individual's variances in taste. People rapidly get bored and frustrated by the inability of these systems to accurately reflect their tastes or respond quickly to their needs and preferences at different times.

It is akin to insisting that a person eat the most popular dish on the menu at the same food court in the same restaurant with the same people every day. Whilst it is true that people have favourite restaurants and favourite meals, they don't want them all the time. Furthermore, certain meals are more appropriate for certain events, e.g. a buffet at a meeting, a roast dinner with family on a weekend. Likewise, people tend to prefer different music for different scenarios depending on the situation they are in.

In short, the problem for music recommendation is characterised by the needs of system to account for the dynamic nature of users' tastes and accommodate their various purposes for seeking recommendations.

5.4 The Analogy Developed

This restaurant analogy can be taken further, as will be shown later in this chapter when it is incorporated in a design exercise. Below we outline how the basic analogy of a restaurant can be mapped onto the music recommender domain.

5.4.1 The Table

A table in a restaurant groups together a set of individuals for a particular event with a particular set of taste preferences and current culinary requirements. For example a group of work colleagues having a Christmas dinner, or a group of friends catching up over lunch or simply people just grabbing a quick bite to eat. Depending on the time, availability and the purpose of their culinary meeting, they are likely to have different requirements of the restaurant, both as individuals and as a collective.

In the case of the music recommender, the notion of a table is used to group together a set of individuals for a particular purpose of which music is to be a part. As with the classic case of culinary taste, their individual and collective music tastes are likely to vary depending on their reasons for wanting to obtain some music recommendations. If a person were seeking music recommendations to play whilst trying to get to sleep, they would perhaps be less likely to be looking for heavy metal even if this is a genre of music they might under other circumstances really enjoy.

In essence, a table can be thought to contain a collection of taste profiles representing specific facets of users and their immediate recommendation requirements and tastes. The set of users, or specifically the collection of user profiles at a given table, can broadly be thought to share some similarities in collective musical tastes. More specifically, they can be thought to share the purpose for which they are seeking recommendations, just as you might reasonably expect those at a specific table in a restaurant to have a shared reason for meeting, e.g. a work lunch or catching up with old friends.

5.4.2 The Restaurant

At the next level up, a restaurant can be thought to represent a particular set of broad taste preferences and user requirements, just as Burger King might be thought to represent the prototypical burger joint for people seeking quick meals or fast food. The specific tables in that restaurant will represent a narrower range of tastes and specific reasons for being there but will likely share the global ones of fancying burgers and requiring quick service. Those at a given table might have narrower preferences like wanting vegetarian burgers and the specific purpose of gathering for a birthday party for instance.

In the case of our music recommender system, a restaurant might share the broad musical taste preferences of being instrumental or unplugged and share the broad scenario of being sought for relaxed events. Individual tables will reflect narrower tastes and scenarios, for example containing piano music and being suitable for a school. Individuals at a given table will have narrower tastes and purposes for seeking recommendations still, for example containing classical orchestras and being appropriate for a school assembly.

5.4.3 Food Court

A food court typically groups together a wide range of restaurants and is geared towards giving people a very broad range of choice whilst still sharing certain characteristics, for example a given food court might have the theme of catering fast food or having a particular price point.

In the case of our music system, a food court holds a collection of restaurants which share a broad notion of musical taste and reflect a broad sort of user requirement. For instance, a given musical food court might contain only those restaurants which offer music from a given time period for the broad purpose of entertainment.

5.5 Cautions and Pitfalls of Analogies

Whenever you are attempting to explain something by reference to something it is not, you will inevitably reach a point where your analogy or metaphor breaks down. This is not necessarily a bad thing. In fact, within the context of this thesis, we would argue that it is a good thing as it helps clarify an aspect of HITL music recommendation that hasn't yet been explicitly touched upon.

You may have noticed a slight gap in the analogy above between the conventional restaurant idea of purpose and the one presented in the music analogy. The gap or issue is one of time and delay. In the case of conventional restaurants, those at a table are immediately gathered for a common purpose - to catch up, to celebrate or simply to eat. In the case of those grouped in a music table, they are not necessarily immediately grouped for a shared purpose beyond that of requiring a music recommendation. Their shared purpose comes later when considering the purpose for which they are intending to use or play their music recommendations.

The significance of this is that the purpose for which a music recommendation is being sought cannot be solely captured or surmised by the system merely from the grouping of tables, restaurants and food courts. A well-designed system needs to find a way for users to express / reflect their reasons for seeking a recommendation which can be made apparent to other users of the system. Perhaps users might be asked to make this information explicit in order to join a table for instance.

Another complication this analogy helps to identify is the distinction between users and profiles. In a conventional restaurant a table is made up of real people who expose a certain facet of their character and culinary taste preferences. In the music example a table is made up of musical taste profiles. These profiles might come from several different users, represent multiple facets of the same user or even represent virtual characterisations of artists. Virtual artist profiles are constructed based on the artists own discography on the assumption that an artist is unlikely to produce music which they don't like, to at least some degree. This is a useful aspect of the system to consider as it can be used to break down the conventional barriers between the content producers and content consumers of a music recommendation system. This, in turn, might help artists to find their audience and audiences to feel more connected to their artists and more personally reflected by the recommendation system.

5.6 Questions to Consider When Designing A HITL Music Recommender System

This section presents a series of questions based on the considerations and findings of this thesis. These questions can be thought of as a template or brief which could be used by those who wish to design a modern personalised HITL recommender system.

Here the questions are presented in list form before being analysed in greater depth below:

1. Who are your users?
2. How many users does the system need to be able to handle?
3. What algorithm are you going to use?
4. How are you going to combine human filtering / curation with automated algorithms?
5. How are you going to expose information to users to facilitate the HITL process?
    a. What information are you going to expose to users to assist in the HITL process?
6. How are you going to design the interface to support the HITL process?
7. What additional features are you going to incorporate on top of the HITL architecture?
8. How will these additional features support and enhance the recommender system?
9. What restrictions or challenges does a HITL framework introduce?
    a. How are you going to account for / mitigate the impact of these?

5.6.1 Question Breakdown

5.6.1.1 Who are your users?

When designing any system, it is important to consider who the intended user / users of that system are. The finding of this thesis, however, reveal that this is of special importance for music recommender systems if they are to be considered personal, effective and useful.

In line with this question is the idea of ***recommendation purpose*** and ***scenario***. The first study in this thesis supported findings from the literature that people have widely different ideas about what makes music personal to them and how they might go about curating a personal list of music. However, it also showed, in conjunction with the second study, that the intended purpose for curating the music and the scenario in which it was intended to be played had a big impact on a person’s requirements when seeking recommendations.

In the first study a few teachers spoke about their requirement to be able to filter out explicit music if they were intending to play music in the classroom around children. Another participant in the study talked about their requirements for different kinds of music depending on whether they were listening to it to make them feel upbeat or send them to sleep.

This point highlights another very important system design consideration surrounding the issue of ***dynamism***. The same person often has different requirements of their recommendation system at different times. This is one of the major limitations and reasons for the lack of personalisation in conventional systems. Regardless of algorithmic approach, conventional modern music recommender systems all have a common flaw. They are monolithic and slow to adapt. In the case of collaborative filtering and hybrid systems, this typically arises as the problem of having a singular digitised representation of a user and their tastes. In the case of more content-driven approaches, this problem tends to arise as the issue of having rigid taxonomies of content which fail to adapt over time to changing tastes, categorisations and patterns of classifying music. These rigid systems then quickly become outdated and a gap emerges between how the user wishes to specify their requirements or search and recommendation parameters, and how the system is set up to meet those demands.

Finally, it is also important to consider users when thinking about interface design. How will your intended users interact with the system? How will the learning curve of a HITL system be presented to them to minimise the inevitable frustrations of learning a new system and approach to music recommendation which they are likely to be unfamiliar with? How will you make it easy for them to hear and appreciate the benefits of the system over more conventional recommenders which, whilst imperfect, have the advantage of familiarity?

5.6.1.2 How many users does the system need to be able to handle?

The size of the userbase of a system can function as a bit of a double-edged sword. On the one hand, the more users a system has, the more data there is to draw upon when generating recommendations. On the other hand, this high volume of users makes it more difficult to generate quick responsive recommendations.

Having an idea of the speed of growth and requirements of a system to expand can help shape the design of a music recommender and provide a timeline or framework for introducing various HITL aspects. For example, it is unlikely that incorporating a social tagging aspect into a music recommender system will be very helpful for generating personal recommendations initially if a system has vastly more content than users.

In fact it could notionally reintroduce popularisation and repetition problems if it were weighted too highly as much of the content would not be commented on and so only a few items might be being pulled into the recommendation pool or otherwise weight artificially highly as a result of sparse social tagging.

It might be more advantageous to either delay introducing social tagging until the userbase has grown to a reasonable size like a hundredth of the content size, or alternatively, introduce it immediately, but increase its weighting in generating recommendations slowly over time as the size and taste diversity of the userbase expands.

5.6.1.3 What algorithm are you going to use?

Considerations regarding the HITL features you wish to use can help determine what type of recommendation algorithm you wish to use.

Prior to the onset of the cold start-problem, historically, collaborative filtering systems were often chosen by system designers especially in the commercial sector, in part because they were comparatively simpler to construct than the often more complex content systems.

With the onset of the cold start-problem, hybrid systems started to take over because of their ability to produce reasonably pleasing initial recommendations to new users prior to them building up a large listening history and detailed taste profile on the system.

HITL techniques can be deployed with any type of recommendation algorithm but appreciating the relative strengths and weaknesses of different approaches as well as your system’s objectives can help determine the best approach. Collaborative filtering may be preferable in a research setting when rapid prototyping and iterative design methods are being used.

However, hybrid approaches may be more appropriate for commercial systems where it is important to minimise user frustrations as a result of old issues like the cold-start problem. If a new HITL music recommender has potential but initially feels like a step backwards from the user’s perspective when contrasted with their more conventional hybrid music recommenders like those employed by Apple and Spotify, then they are likely to abandon it before it ever has a chance to show dividends.

5.6.1.4 How are you going to combine human filtering / curation with automated algorithms?

Once you have determined your algorithmic approach and the HITL features you intend to use, you need to think carefully about how you are going to combine them. There are three main options; pre-filtering, post-filtering and both pre and post filtering.

Pre-filtering would allow users to constrain the content which is exposed to the recommender algorithm you have chosen. Post-filtering works by allowing users to filter a list of recommenders generated by the recommendation algorithm. This approach is most useful when your algorithm is of the top-n variety and produces a weighted list of recommendations.

Combined pre and post filtering works to filter the content pool and the algorithmically generated recommendations. This can produce the narrowest set of recommendations therefore producing the best chance of a highly personalised and tailored final recommendation. However, it also requires the most user intervention. Deciding how niche you wish to make your recommendations vs how complex you wish to make things for your users may help you to decide which approach to employ.

5.6.1.5 How are you going design the interface to support the HITL process?

Human-in-the-loop systems unavoidably require more from their users than conventional passive recommender systems. The primary advantage of this is a higher level of personalisation and tailored recommendations. The disadvantage, however, is it increases the chance of user error and user frustration as it requires them to learn something new and unstandardized.

Therefore once you have decided the algorithmic approach that you intend to use and the HITL features you need to employ, you need to think carefully about how you are going to present your system to your users. The objective here is to present the system to new users in such a way as to minimise the learning curve they have to go through and reduce the chances for human error.

It might be advantageous to structure the HITL system components as a layer, or add them on top of a more conventional recommender system. By allowing the components to be toggled on and off, a user could alternate between what they know, and learn something new when they have the time to do so.

Another useful technique could be to structure your HITL interface around an analogy or metaphor that your users are familiar with. This could be achieved in much the same way that users were originally introduced to the concept of personal computing and graphical user interfaces through the analogy of the work surface or desktop which they were already familiar with. Or later, how saving became synonymous with the floppy disk icon as people knew this as a practice for saving their files.

5.6.1.6 What information are you going to expose to users to assist in the HITL process?

When designing a HITL system, is important to avoid overwhelming users with too much choice and information. Indeed, this was the problem recommenders were first designed to avoid. When determining what information to expose to users, think about who your users are and what their search habits and scenarios are likely to be.

Are you going to provide users with information about specific tracks, or other users, or both? Will you make use of user driven techniques like social tagging? Will this extend to tagging other users, or content, or both? If social tagging is used, will users be able to edit / withdraw or update their tags? What issues and potential problems might surround this?

5.6.1.7 What additional features are you going to incorporate on top of the HITL architecture?

Before deciding what information you are going to expose to users and incorporate into your system, it is important to decide what features your system is going to have as this can help you determine the information users need access to. For instance, will the HITL architecture support multi-user interaction? Will it support sharing features? Can users share their HITL intervention strategies?

How will the results of the HITL intervention be stored or utilised so as to maximise user satisfaction and recommendations? Will users be able to save, edit, share or revisit past results of the intervention?

5.6.1.8 How will these additional features support and enhance the recommender system?

It is important when designing your system not to lose sight of how your HITL techniques and additional system features that are constructed on top of these approaches are going to enhance your system.

Are you going to use HITL techniques to:

- Keep your item classification taxonomy up to date?
- Facilitate scenario-based playlist generation and track retrieval?
- Facilitate rapidly changing taste profiles and cater to users' dynamic tastes?
- Allow for sharing and collaboration?
- Support tailored usage patterns and multiple user requirements?

Once you have decided how your HITL techniques are going to be used, be sure that you understand how the features they are adding affect, promote and improve personalised music recommendation.

It may be helpful to incorporate some form of user facing metric or self-assessment whereby they can get some feedback as to how the HITL features they are using have affected and altered their recommendations. This serves a dual purpose. First, it allows users to gauge that the extra effort they are putting in by using / learning a HITL system is benefiting them in a real and tangible way. Second, it helps increase system transparency, potentially reducing users' discomfort with being profiled in the first place.

5.6.1.9 What restrictions or challenges does a HITL framework introduce?

Whilst a HITL approach to designing a recommender system has many advantages in the realm of personalisation, it has several challenges which need careful consideration during the design process.

Perhaps one of the biggest current limitations is people's lack of familiarity with the technology. Most people with internet access have some familiarity with the concept of conventional recommendation systems whether it is through e-commerce sites or media streaming services. Unfortunately HITL systems are rarely seen outside of academia as of 2019. This means that any current system designer cannot rely on standard practices or user familiarity. They are confronted with introducing new technology and an inevitable learning curve to their users.

The goal then becomes how to minimise this learning curve and reduce the chances of user frustration and abandonment in the early phases adopting of the technology.

An added difficulty is the non-static nature of HITL systems. Issues of scalability notwithstanding, they obtain better results with larger userbases and content pools as there is more information to draw upon when generating recommendations and facilitating human intervention. This means that in a naively designed system users will likely see little benefit in a HITL system until it has become established.

User frustration and confusion can be avoided by managing expectations and informing users that it will get better with use (in much the same way that companies like Spotify and Apple originally pitched their collaborative filtering-based systems). This problem can be mitigated to varying degrees though system design by making sure that whatever HITL features are available to users, draw on populated data sets. For instance, a system could introduce a first set of HITL filters which allow users to interfere with existing recommendations simply by manually selecting or removing certain genres from the content pool. This would influence recommender outcome but not require a vast number of users, merely that some of the tracks and content to be recommended were tacked with genre attributes. As the userbase becomes more advanced, HITL features could be introduced once the system has a sufficient number of users to facilitate them. Implementing a system in this way would present users with ever more personalised results whilst minimising confusion, learning time and frustration.

5.7 Design Exploration

In this section of the thesis we explore the design questions presented in the first half of this chapter by going through a design exercise and presenting one way several of these questions might be addressed. This section begins with a quasi-design brief summarizing several of the key aspects presented in the first part of this chapter.

5.7.1 Design Brief

5.7.1.1 System Description

A web-based HITL music recommendation system geared toward generating highly tailored recommendations to suit a user’s immediate taste preferences.

Users can view basic profile information about real and virtual users in the userbase and filter which users they want to be exposed to the collaborative filtering algorithms in order to produce music recommendations.

5.7.1.2 System Requirements

1. The system should consider its userbase and build upon their pre-existing expectations and knowledge of music recommenders
2. The system should be simple to use and aim at minimising the learning curve involved in introducing users to a new and unfamiliar means of music recommendation
3. The system should take account of the number of users it is likely to obtain and deal gracefully with issues of scalability
4. The system must account for the dynamic nature of users' tastes and how recommendation scenario affects their requirements
5. The system should avoid generating static or monolith user profiles
6. The system should consider what HITL features it aims to incorporate and how these features will aid in the personalisation of recommendations
7. Ideally the system should provide a means by which users can tangibly assess the merits of the HITL system to help justify their investment in learning a new system
8. As a secondary objective, the system should aim towards transparency when contrasted with conventional hybrid and content-based music recommenders

5.7.2 User Considerations

For the purposes of this exercise, we will imagine that our userbase consists initially of 1000 users with stable internet and with a content pool of around 30 million tracks. In essence we are modelling our hypothetic userbase on the sort of figures a large established company like Apple or Spotify might expect to use when trialling new system. Imagining we are one of these larger companies gives us the broadest and most realistic opportunity to explore likely design issues that might be encountered in a real-world situation.

Given this userbase, our design ought to make use of existing practices and expectations; a seamless toggle-based means of enabling and disabling any HITL aspects of the system. This would allow it to be explored at the users' leisure and used alongside the system they already know, thereby reducing the possibility for frustration and early abandonment.

Give that our demographic is relatively tech-savvy, at least to the extent of being able to use an existing only streaming service, we can conclude that it is reasonable, even advisable, to make our system a web-based one. This maximised exposure allows for the broadest range of feedback to incorporate into future iterations and for the improvement of the system to target the global demographic it will ultimately need to work for.

Finally, given that our users are familiar with current streamed services, we ought to utilise this knowledge in our interface design, again minimising the new content they need to learn.

Usability Considerations

In order to aid the usability of our system, we will take several steps. The first is to adopt the conventions of existing collaborative recommender systems using the thumbs up / down icon to allow users to provide feedback on tracks and provide users with the ability to generate playlists and view their favourite artists and tracks.

Second, we will use the restaurant analogy detailed in the first part of this chapter to structure our userbase and content, and also in the user interface design to expose the HITL characteristics of our system to users. This tightly couples our HITL philosophy with our interface design in a manner which will immediately and intuitively hold some familiarity with most users.

5.7.3 HITL Components

Our system will use the following HITL features:

1. Pre-filtering of the content pool thorough user profile selection
2. Exposing user metadata concerning playlists, top played and favourite albums and tracks
3. Staging introduction of social tagging of both content & profiles
4. Saving and sharing of user profiles
5. Collaborative profile grouping based on table analogy

5.7.3.1 Pre-filtering

Pre-filtering the content pool is a nice way of introducing users to the concept of HITL intervention. They are likely to already be familiar with the notion of providing recommenders with start-up information as a lot of recommenders now ask users a few basic profile questions to help avoid cold-start problems. Pre-filtering can be seen and presented to the users as simply an extension of this.

Furthermore, the concept of pre-filtering fits very nicely with our table analogy which can be reflected in the user interface resulting in a very clean user experience and very transparent recommendation approach.

Pre-filtering will help personalise recommendation by allowing users to radically change the pool of content available to our recommendation algorithm from one listening session to the next. This avoids the problem whereby recommenders often produce a single monolithic user profile which doesn't account for non-static taste preferences.

5.7.3.2 Exposed Metadata

To keep our system easy to use, we will initially only expose the sort of metadata people are already familiar with. Specifically our initial metadata attributes will be:

User metadata

- Favourite / Most played artist
- Favourite / Most played album
- Favourite / Most played track
- Playlists

Track Metadata
- Genre
- Release date
- Artist

As the system develops and people become more accustomed to reviewing content and profiles and using other data to help them filter their content pool, this could be built upon. Once a usage precedent and pattern have been established, additional attributes could exposed which were user generated by means of social tagging.

5.7.3.3 Social Tagging

Our system will include social tagging but introduce it strategically to avoid data sparsity issues. Users will immediately be able to tag profiles and content with arbitrary string tags. However, these tags will not be made visible in the HITL interfering interface until a rich tag taxonomy has been developed.

This would have the advantage of always being current and reflective of the userbases' natural means of categorising their music. This could also allow for the emergence of expert profiles. A user profile could be rated and tagged as excellent by other users for generating music recommendations for dinner parties for instance. This reincorporates the personalisation strengths of earlier manual collaborative filtering systems without requiring all the users to actually know one another.

Allowing our system to grow organically and gradually increasing the data that users can make use of in the HITL filtering, helps to minimise user disruption and allows the system to scale naturally thereby avoiding data-sparsity issues. For example, social tagging attributes could be released to users to make use of in their HITL efforts only once a certain percentage of profile or content has been tagged.

5.7.3.4 Profile Creating & Sharing

Users will be able to save their listening session profiles and share their profiles with other users. This is a core aspect of our HITL personalisation approach. Allowing users to capture and share individual listening sessions as listening profiles serves a dual purpose. First, it avoids users becoming statically represented and accounts for their variations in taste. Second, it provides a convenient means of building user generated content sub-pools (by recording the tracks listened to as part of a given profile) which can be drawn upon when producing bespoke recommendations.

5.7.3.5 Collaboration

The final HITL approach our system will include is multi-user collaboration. Users will be able to generate collaborative listening sessions with other users of the system. These collaborative profiles will allow users to easily incorporate multiple people’s listening preferences into a particular profile which might then be saved by them for the purposes of a particular event, e.g. a house party.

Allowing collaborating reduces the rigidity of the system and allows for multiple taste preferences to be accounted for which could be very useful when listening to music in social situations.

5.7.4 Algorithmic Approach

In order to maximise transparency and minimise the time required to deliver the system as well as interpret its results, a collaborative filter nearest-neighbour based recommendation system would be a good initial algorithm to use. It would also be good to modularise the system and divorce the recommendation algorithm from the HITL components and interface design such that multiple algorithms could be contrasted and compared.

5.7.5 UX Considerations

5.7.5.1 System Usage & Application Flow

Upon starting the application, a user would be asked to fill in a brief profile that might approximate their listening history in a music streaming service.

The system would then display a grid layout of thumbnail images representing the userbase of the recommender system. Hovering over a thumbnail will reveal a trading card which will contain profile information about that user such as their top artists and favourite albums.

A given user would be able to manipulate which users are included in the userbase for generating their recommendations by interacting with a control panel which will be displayed below the userbase grid. The control panel will allow participants to filter the userbase based on any of the attributes shown in the user profile cards.

Filtering the userbase will have the effect of changing which users get exposed to the nearest-neighbour collaborative filtering algorithm.

Once a given user has finished filtering the userbase, they would click start and be presented with a set of recommendations consisting of albums, tracks and playlists.

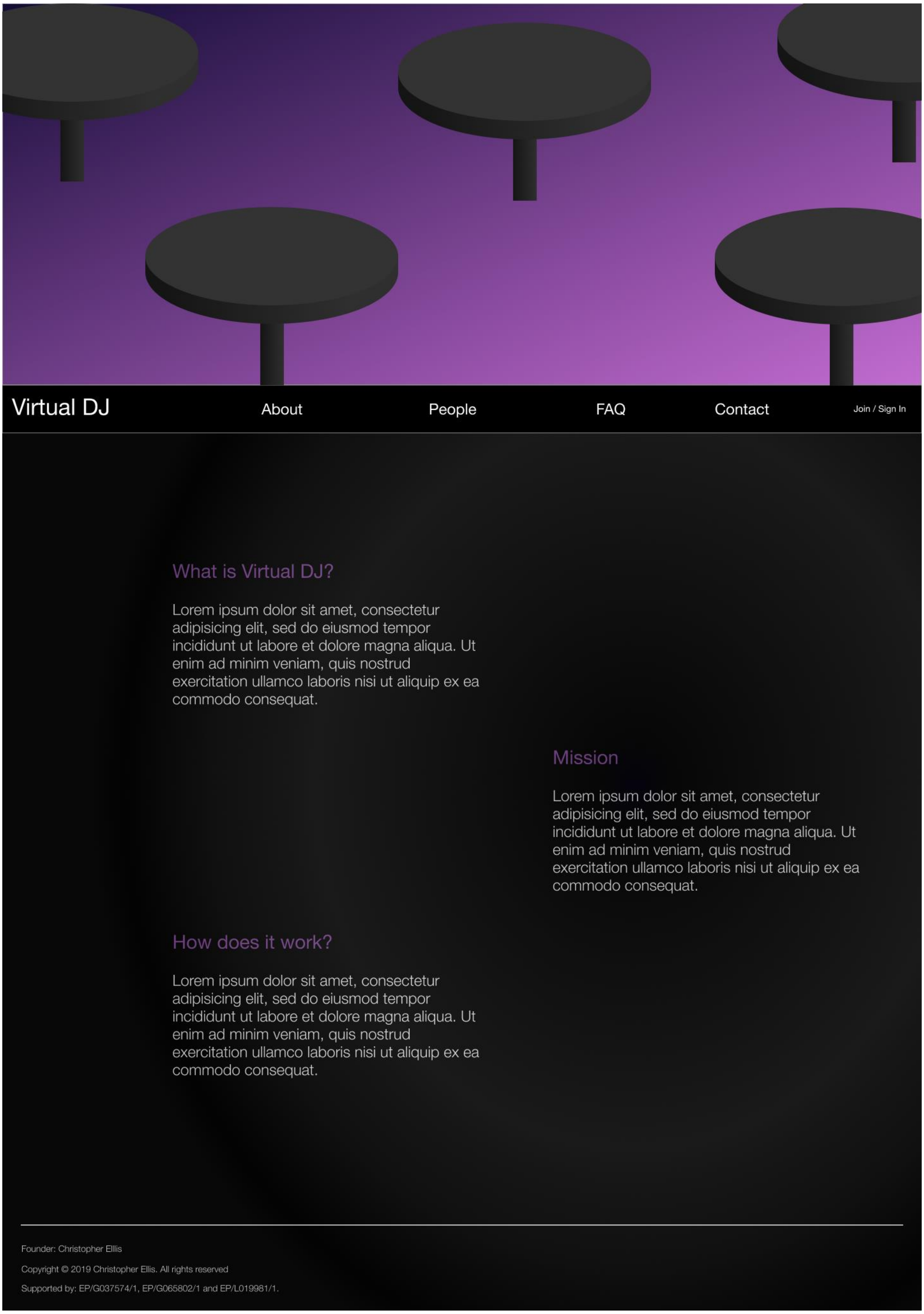
5.7.5.2 User Interface Designs

In the first part of this chapter we stressed the importance of introducing users to a new sort of recommendation system (in this case a HITL music recommender) by means of a familiar analogy or metaphor. For the purposes of this exercise, we will use the restaurant analogy presented at the start of this chapter and previously mentioned in the introductory chapter as a means of explaining the limitations of existing recommender systems.

Functionally speaking from a human-in-the-loop perspective, the user interface has to facilitate the following interactions:

- Allow users to view the userbase and content available
- Allow users to get an impression of other users' and artists' musical tastes
- Allow users to observe a given recommendation and alter the subset of data from which it was produced
- Present the relational mapping of the userbase as viewed by the underlying algorithms in an intuitive manner such that users can act knowledgeably when interfering with it and making changes
- Present information about users and artists in the system without overwhelming users with too much information at any given point in the recommendation process

5.7.5.2.1 Landing



What is Virtual DJ?

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Mission

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

How does it work?

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Founder: Christopher Elllis

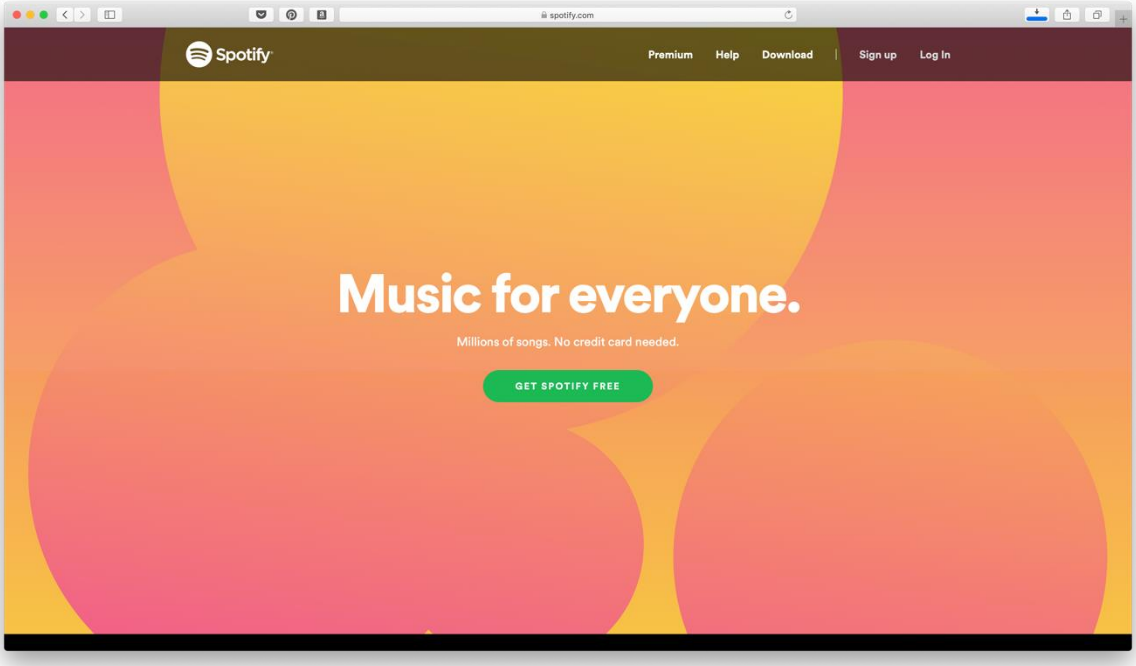
Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.

Figure 37: Landing page

The above screenshot shows the first screen or landing page of a HITL music recommender system based around the restaurant analogy. It achieves several key objectives whilst remaining clean and simple to navigate. First, it provides a space for informing users what the service is. Second, it provides a mission statement area which can be used to inform potential customers what the system offers above and beyond conventional recommender systems (namely personalised and highly tailored music recommendations). It also provides a space for briefly telling users how the system works in layman's terms. This would be a good place to introduce the system's design and usage metaphor, in this case the restaurant analogy. The metaphor is also subtly enforced in the background design by displaying a series of recognisable tables at the top of the website.

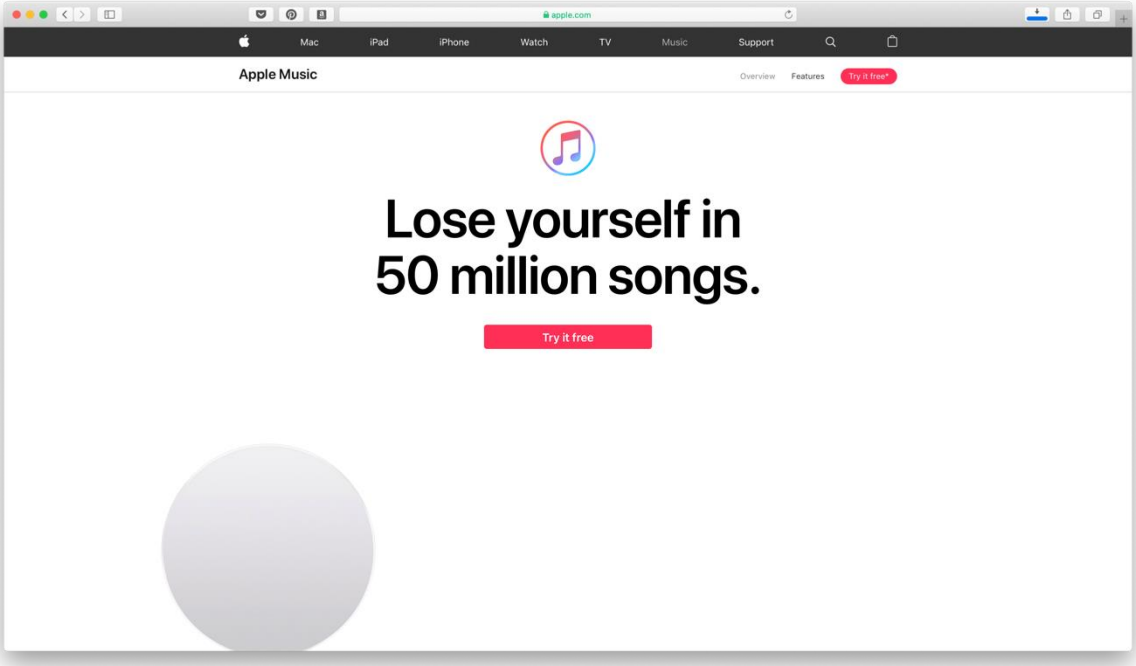
Finally it presents users or potential users with the opportunity to create an account / sign-in and get started using the system by presenting a join / sign-in option on the top right-hand side of the navigation bar. This is a standard practice employed by big recommender companies like Spotify, Apple and Netflix (see screenshots below) and is therefore likely to be familiar to potential customers.



Music for everyone.

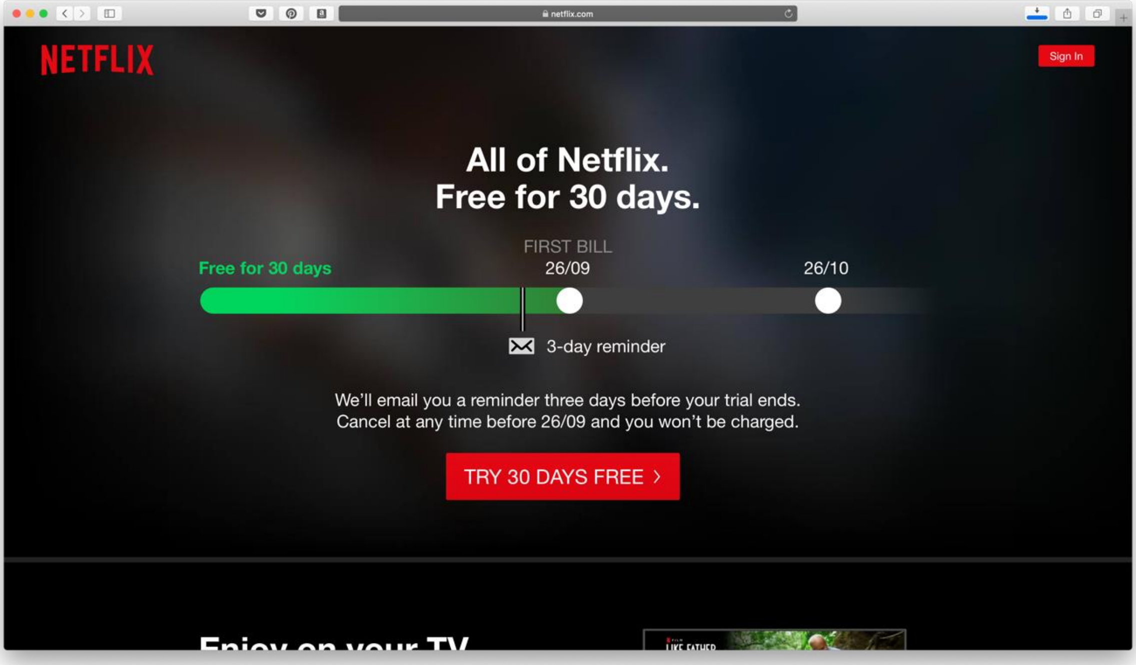
Millions of songs. No credit card needed.

Figure 38: Spotify landing page



Lose yourself in
50 million songs.

Figure 39: Apple music landing page

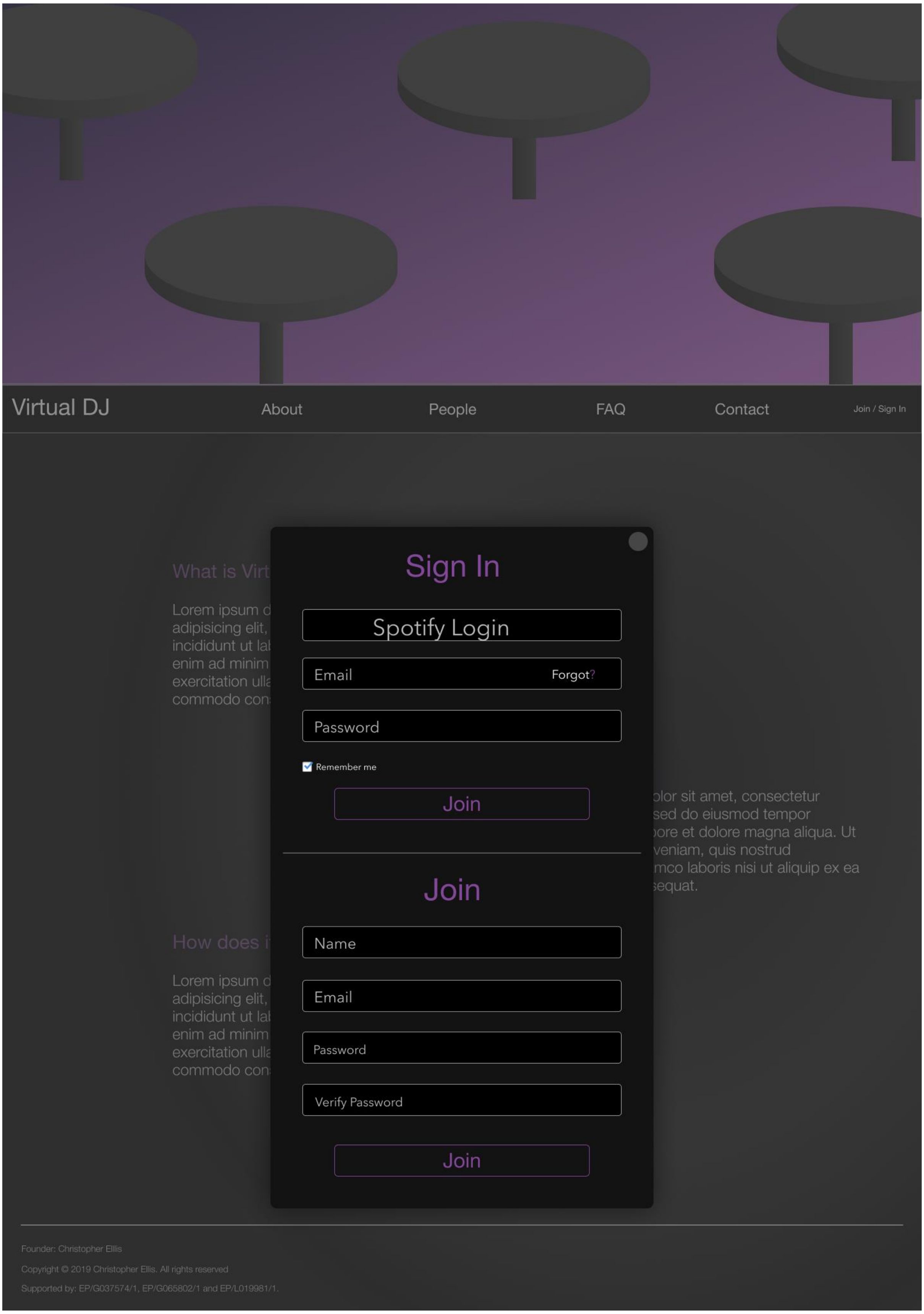


All of Netflix.
Free for 30 days.

We'll email you a reminder three days before your trial ends.
Cancel at any time before 26/09 and you won't be charged.

Figure 40: Netflix landing page

5.7.5.2.2 Join / Sign In



Founder: Christopher Ellis

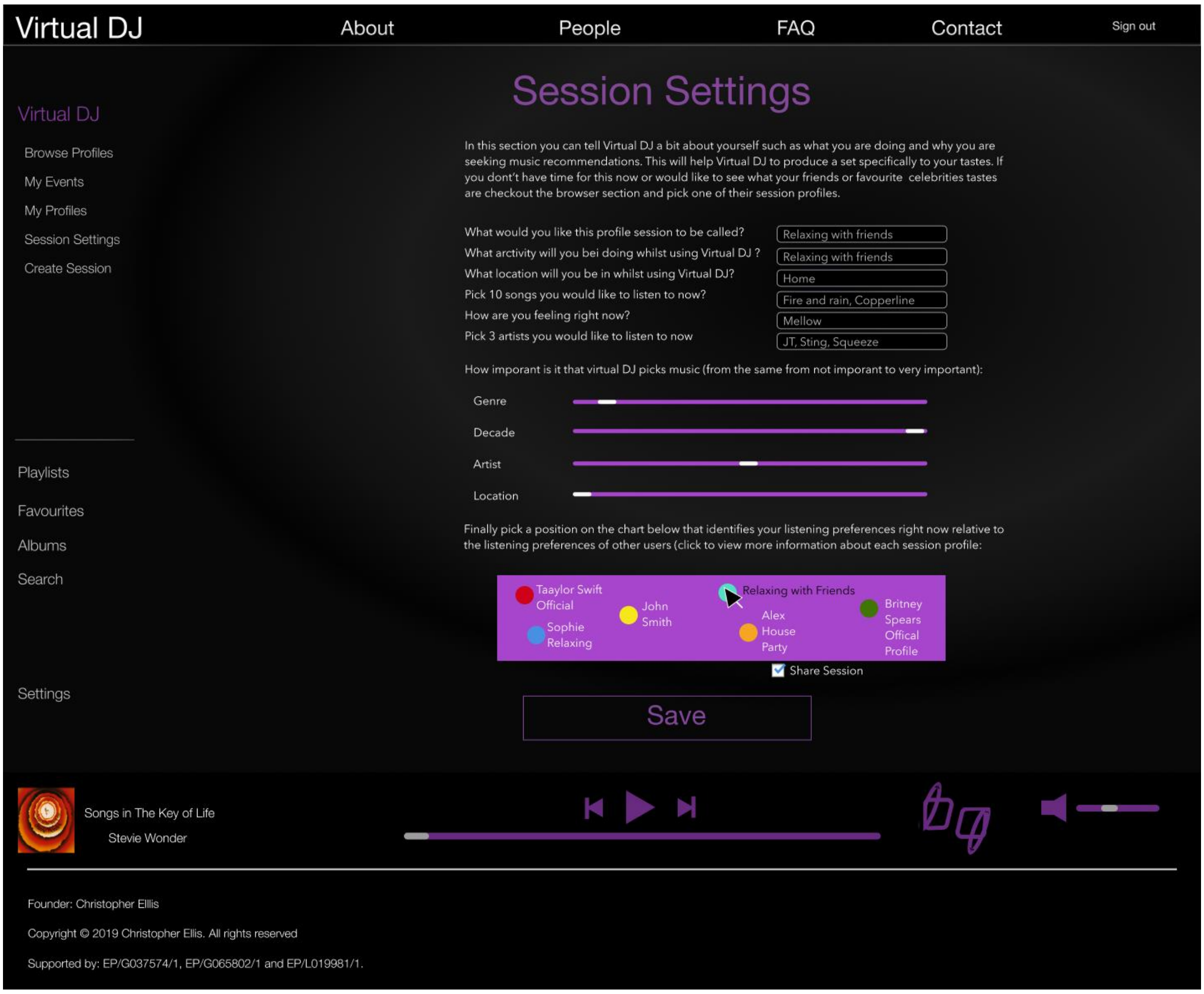
Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.

Figure 41: Login page

Upon clicking the join / sign in link in the top right corner, users are presented with a form which again follows standard practices and additionally allows them to login with existing account details they may already have which aids user conversion rates. Allowing users to join with an existing account also has other advantages. The existing account could be any common web based Oauth log in service like Facebook. If, however, you allow users to log in with an account from another music recommendation services (such as Spotify) you can potentially vastly reduce the likelihood of cold-start issues by often importing their listening history and basic profile information from that other service. This again helps to ease your user into a new system and present them, at least initially, with what is already familiar. In fact, if at this point your user were to skip the following session setting and initialization step and not engage with any of the HITL tools of the system, they would likely get sensible recommendations. This is so even if the system simply ran a conventional basic collaborative filtering algorithm on the entire userbase with the profile information for the active user having been imported from Spotify. The real advantage here is that at the point of joining, the user potentially begins from exactly where they were with their old service without having had to do anything at all. From here by engaging in the HITL tools, they can improve their recommendations from what is hopefully an already satisfying baseline.

5.7.5.2.3 Session Settings & Initialization



Session Settings

In this section you can tell Virtual DJ a bit about yourself such as what you are doing and why you are seeking music recommendations. This will help Virtual DJ to produce a set specifically to your tastes. If you dont't have time for this now or would like to see what your friends or favourite celebrities tastes are checkout the browser section and pick one of their session profiles.

What would you like this profile session to be called? Relaxing with friends
What arctivity will you bei doing whilst using Virtual DJ ? Relaxing with friends
What location will you be in whilst using Virtual DJ? Home
Pick 10 songs you would like to listen to now? Fire and rain, Copperline
How are you feeling right now? Mellow
Pick 3 artists you would like to listen to now JT, Sting, Squeeze

How imporant is it that virtual DJ picks music (from the same from not imporant to very important):

Finally pick a position on the chart below that identifies your listening preferences right now relative to the listening preferences of other users (click to view more information about each session profile:

Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.

Figure 42: System settings page

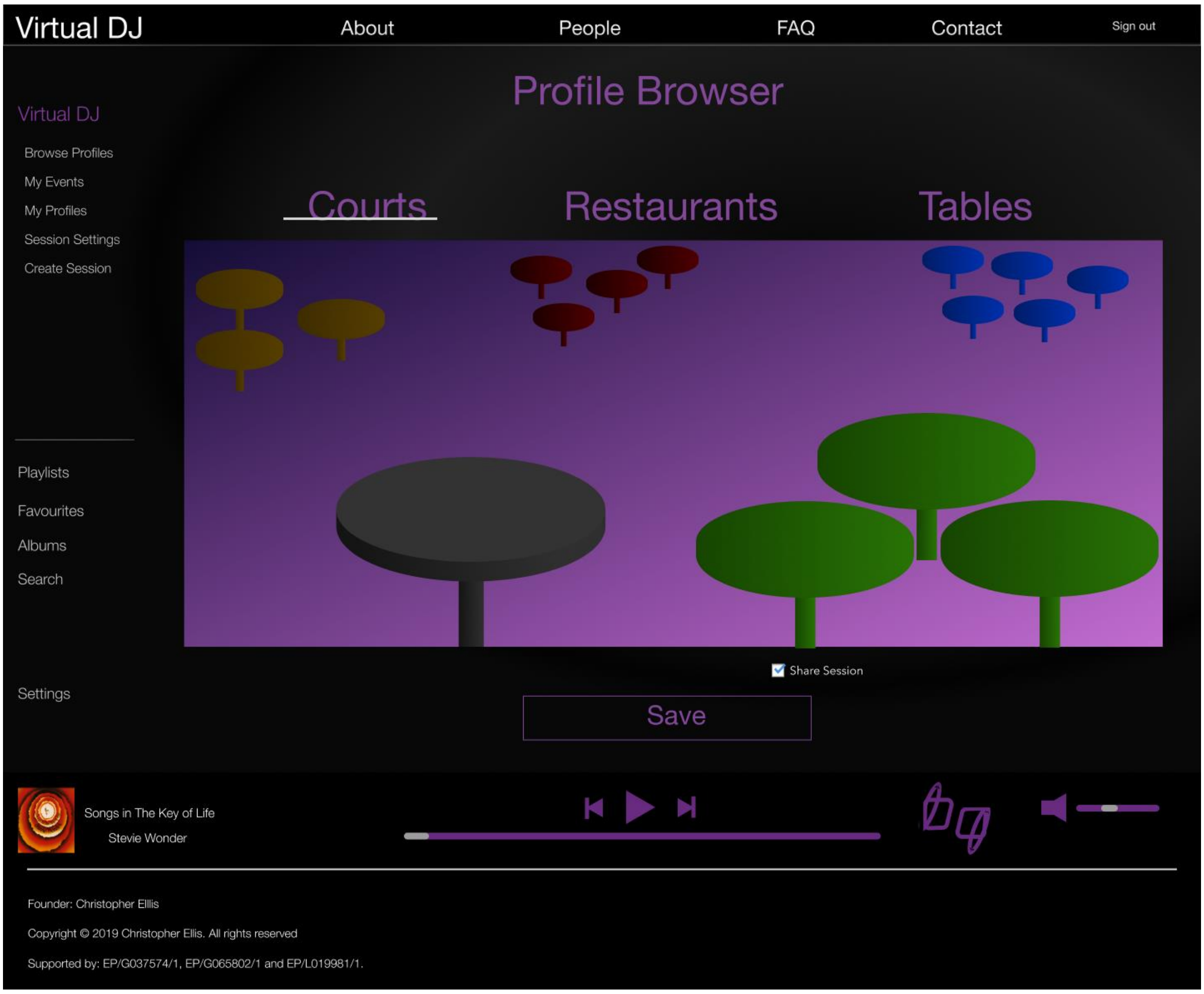
Having logged in, the user is then presented with a session settings screen. If the user has logged in via a pre-existing service, some of the settings can be pre-initialised from imported values. It is important at this stage not to overwhelm the user or present them with technical jargon. As such, the above design possesses a few natural language questions regarding musical taste that you might ask a friend. These are the sort of questions and topics users reported and discussed in the first study presented in the second chapter of this thesis. For instance, what activity are you planning on doing now whilst listening to music? Or, pick 10 tracks which you fancy listening to right now. This last one is important as it helps get an immediate reference point for where the user resides in the pre-existing userbase. In the first study, users often talked about how they would frequently start listening to a series of tracks and let the playlist grow organically or let these seed tracks then be expanded / built up by the inbuilt recommender engine of the service they were using. Many music recommender systems such as Spotify, Pandora and Lastfm have responded to this usage pattern by presenting users with the option to play artist radios which start with seed tracks from a given artist and then using item-item based collaborative filtering to produce additional recommendations based on these initial tracks.

The user is then presented with a series of sliders allowing them to express a personal preference for the key musical categorisations. This again developed out of the findings from our first study which showed that users often had diametrically opposed views of the importance of certain categorisations. Some users really cared that their playlists were thematic or all from a particular genre or artist, whilst others didn't care at all and simply wanted a varied set of tracks which suited their current mood regardless of the artist or genre these tracks were from.

Finally, users are presented with a small localised userbase map which shows where the system currently thinks the user resides in the userbase in terms of taste, based on any imported profile data and their answers in the session's settings. Note that this userbase view is somewhat different to a conventional recommender in that it relates user session profiles (as opposed to single user vector profiles) and artist profiles (which are artificial user profiles generated on the assumption that artists would like to listen to the type of music they produce). An individual user may, and indeed should, have multiple session profiles corresponding to their different listening scenarios and tastes. Representing the userbase this way bridges the divide between content producers (artists) and users. Having session profiles appear as virtual users allows existing collaborative filtering and content-based algorithms to function whilst providing a greater degree of flexibility when representing individual users. This in turn helps facilitate the HITL tools which will be introduced in the subsequent design mock-ups.

This userbase view could notionally be expanded to fill the entire device screen and show many more users by double clicking on it. Users could even notionally pan around the entire userbase map. By clicking on individual session profiles, they can get information about that profile such as 'top tracks' and 'featured albums' which would help them to locate themselves in the userbase for this current listening session. They can position themselves by simply clicking on the point in the userbase at which they would like to appear . This would create a session profile like the others shown in the map which they could then label as they wished. As a privacy consideration, you could then allow them to select whether this session was made public for other users to view when they logged on to the service and browsed the userbase.

5.7.5.2.4 Profile Visualiser / Dining Hall



Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.

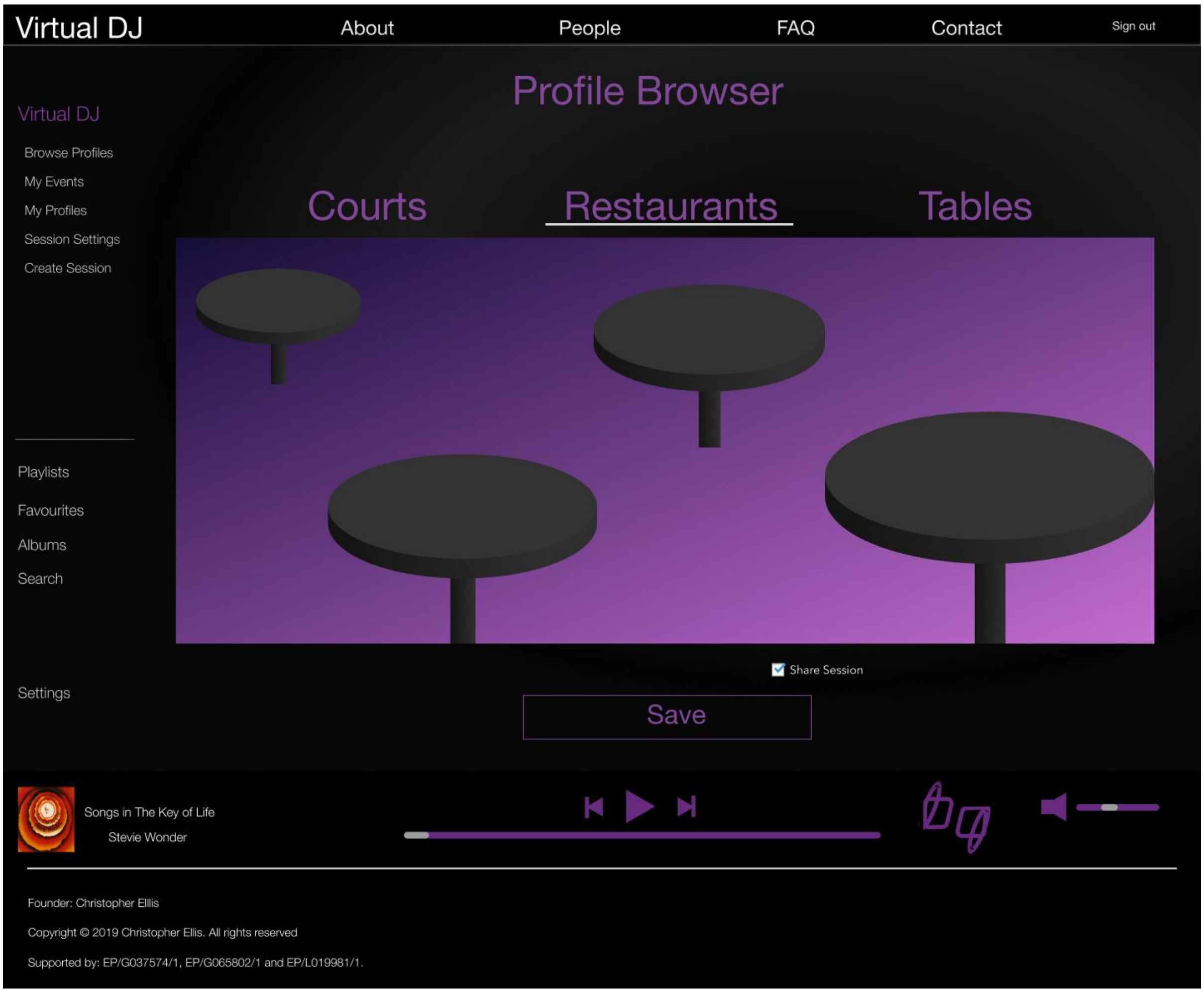
Figure 43: Profile visualiser

After initializing their listening session, users will be taken to the profile browser. Here they will be presented with three different tabulated views, Courts, Restaurants and Tables. The courts view (shown above) shows a collection of restaurants, which are differentiated by colour. Each restaurant represents a collection or grouping of tables which themselves represent a collection of profiles.

As the system developed, users could annotate the restaurants by clicking on them and adding arbitrary descriptive tags which would be displayed when a user hovered over a restaurant to help them decide if they might be interested in it. For instance, a user might tag a restaurant as, “1950’s Jazz” or “Workout Music”. These two tags represent very different ways of grouping music. The first is temporal and genre defined, whilst the latter is purely scenario defined. The useful thing about this is it allows users to navigate the userbase according to the tags they personally find most useful.

The system could even support a popularity extension whereby tags could be voted as useful or not by other users and the most useful ones could be displayed physically larger or retained whilst the less useful ones disappeared after a given amount of time, say 30 days. This would have the interesting consequence of generating a HITL music classification taxonomy that developed and evolved over time, always reflecting the most current way in which people wished to think about and categorise their music.

Suppose, for instance, that a decade from now a new genre emerged called x-beats, people could start labelling appropriate courts, restaurants and tables with the tag x-beats. This would allow users to navigate the system according to a categorisation that didn't exist when the majority of the music and users within the system were added.



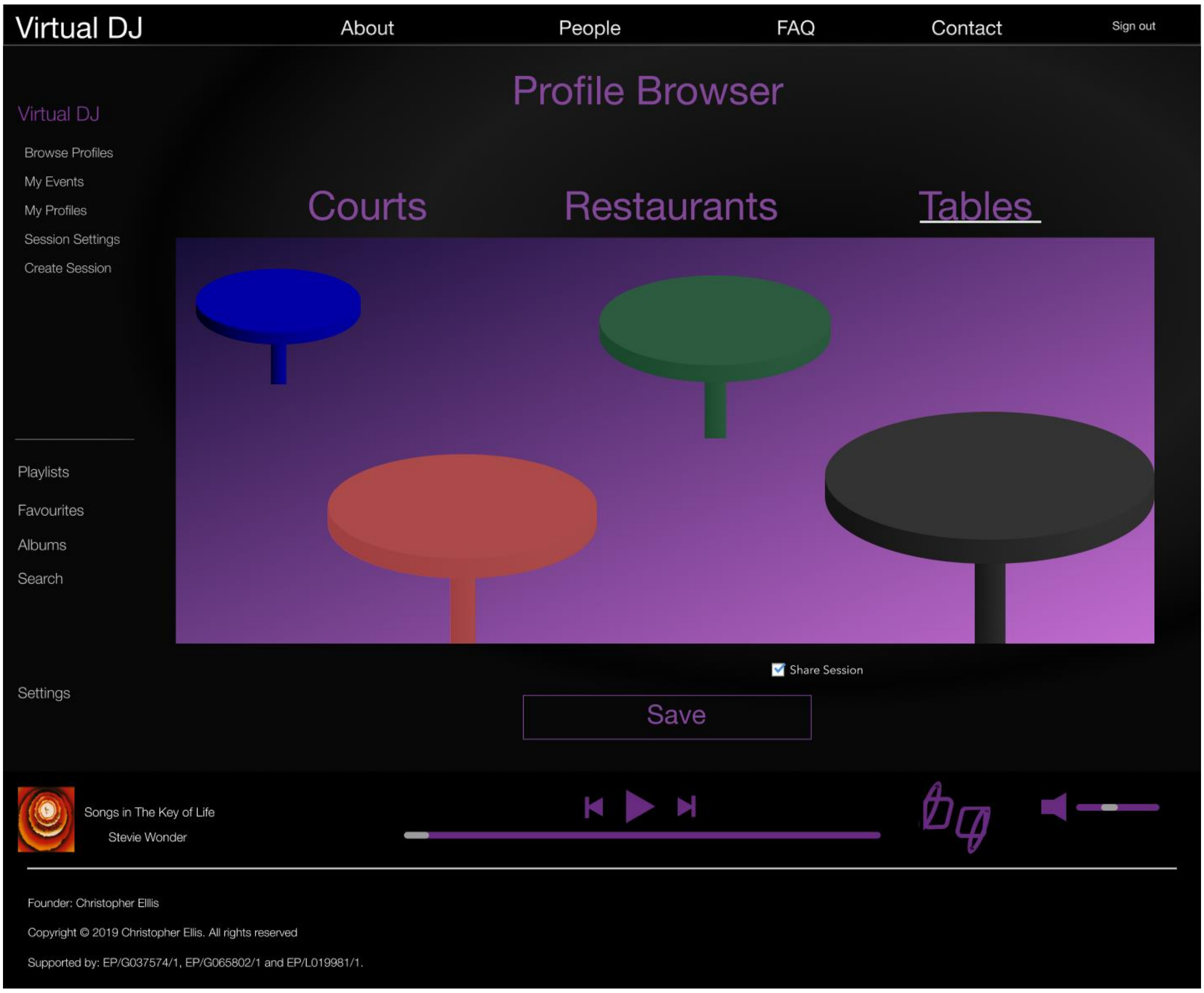
Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.

Figure 44: Restaurants tab

The restaurant tab shows a zoomed in view of the food court focused on one particular restaurant. In the real-world, people are familiar with the idea that food courts contain a wide variety of different types of food in the form of multiple restaurants. This expectation of grouping prepares them to be confronted with a natural grouping, not of culinary choices but of musical choices.



Founder: Christopher Ellis

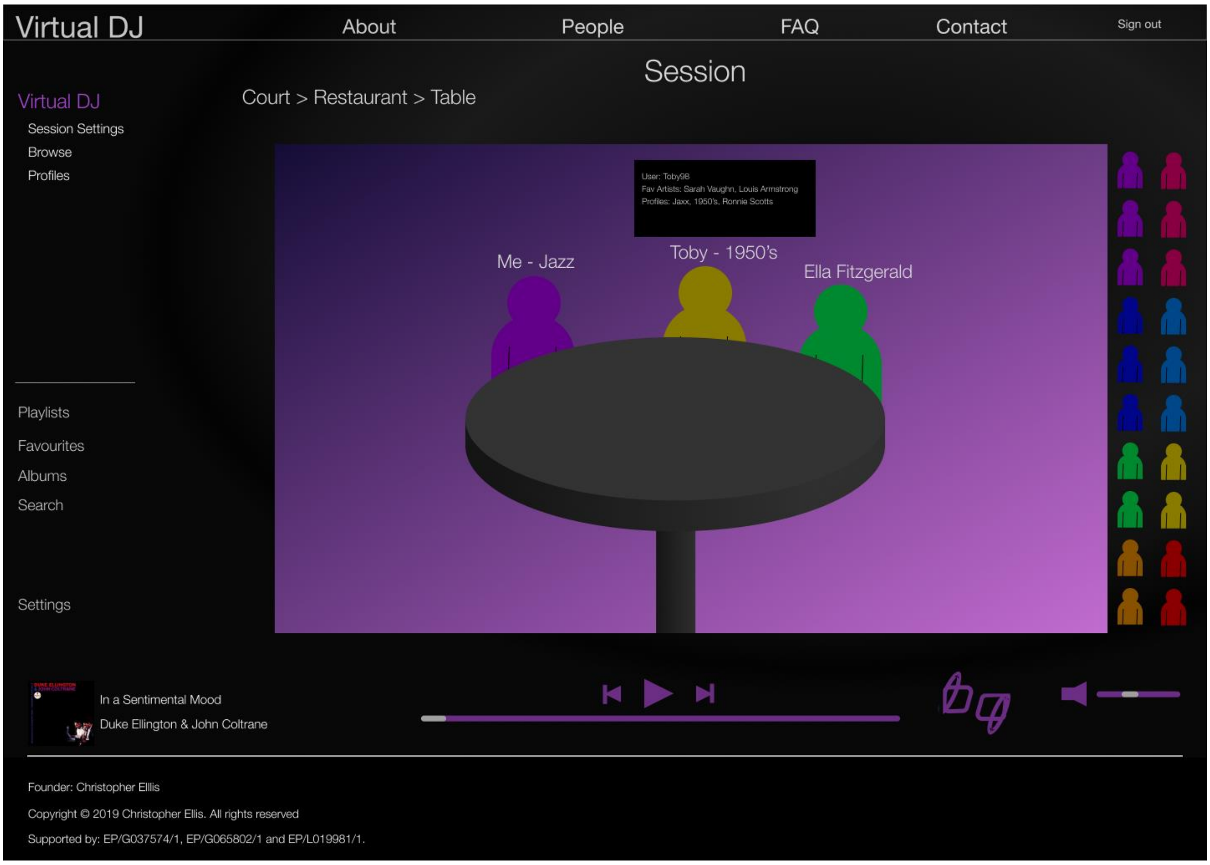
Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.

Figure 45: Tables tab

The tables view shows a collection of profiles which can be made up of real users' listening sessions and virtual artist profile sessions. A given table could contain all virtual artist profiles or all user listening sessions or a mixture of the two. Additionally a table could contain listening sessions from multiple real people. This idea is explored further in the next screenshot representing a collaborative listening session.

5.7.5.2.5 Session view



User: Toby98
Fav Artists: Sarah Vaughn, Louis Armstrong
Profiles: Jaxx, 1950's, Ronnie Scotts

In a Sentimental Mood
Duke Ellington & John Coltrane

Founder: Christopher Elllis
Copyright © 2019 Christopher Ellis. All rights reserved
Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.

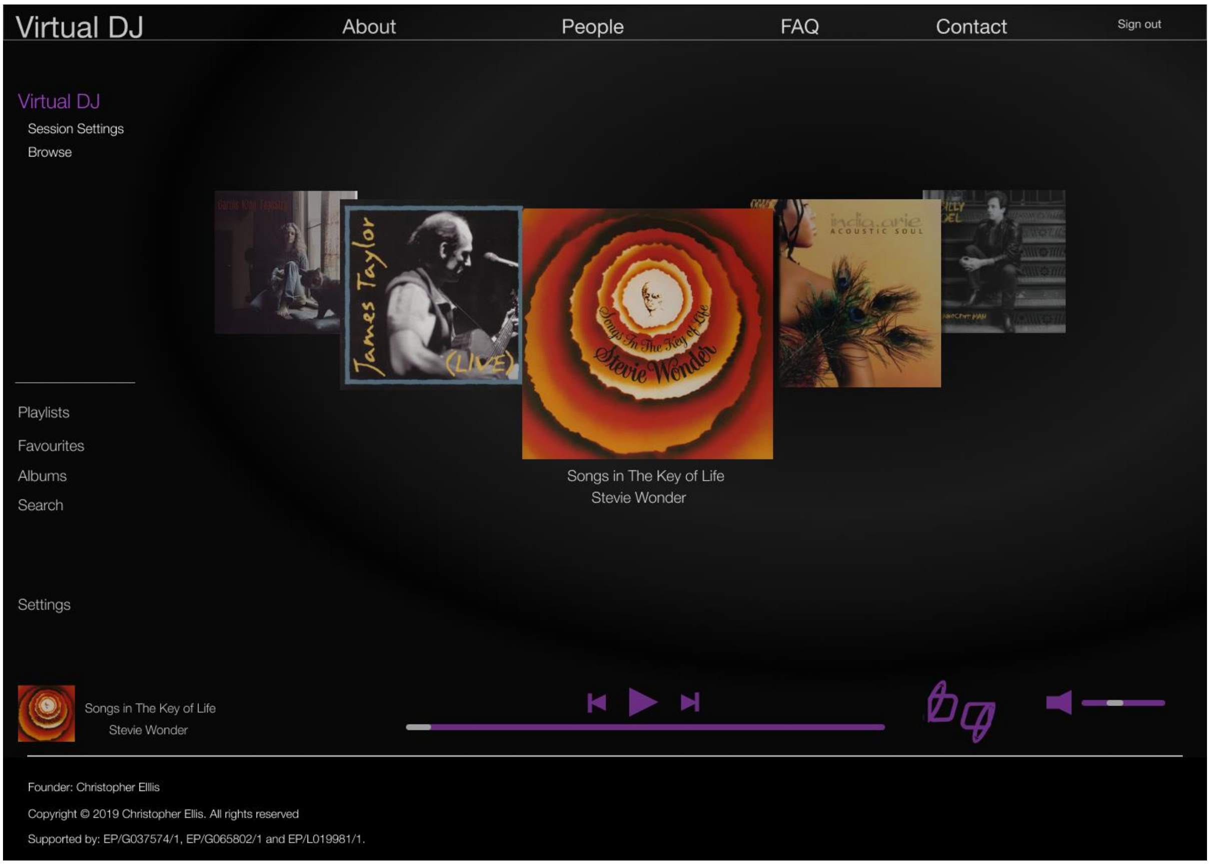
Figure 46: Listening session screen

After having picked a table, the user will be taken to the session view which displays the profiles which are at the active table. In the example above you can see a table which contains all three types of profile, personal (Me – Jazz), other user (Toby – 1950's) and artist (Ella Fitzgerald). The user can remove people from the table for this session by dragging them away from the table. Similarly they can scroll through available profiles in the userbase on the right-hand side and add them to the table by dragging them to it. Hovering over a profile displays information about it such as the user it originated with, their top tracks and favourite albums. Notionally this could be extended with arbitrary tagged data which could be added by users during a listening session or at a later time from within their profiles view .

Upon exiting the application or quitting, a listening session is automatically added to the userbase and allocated to an appropriate court and restaurant. The user can then decide if they wish for it to be kept private so that only they can view it or whether they want other users to be able to use it as well.

This could be built upon further to enable multiple users to compile a table which could be a fun way of creating shared recommender experiences at events like parties. Tables could even support a sort of versioning so you could view what profiles were at a table throughout its history of active listening sessions by multiple users.

5.7.5.2.6 Playback visualiser



Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.

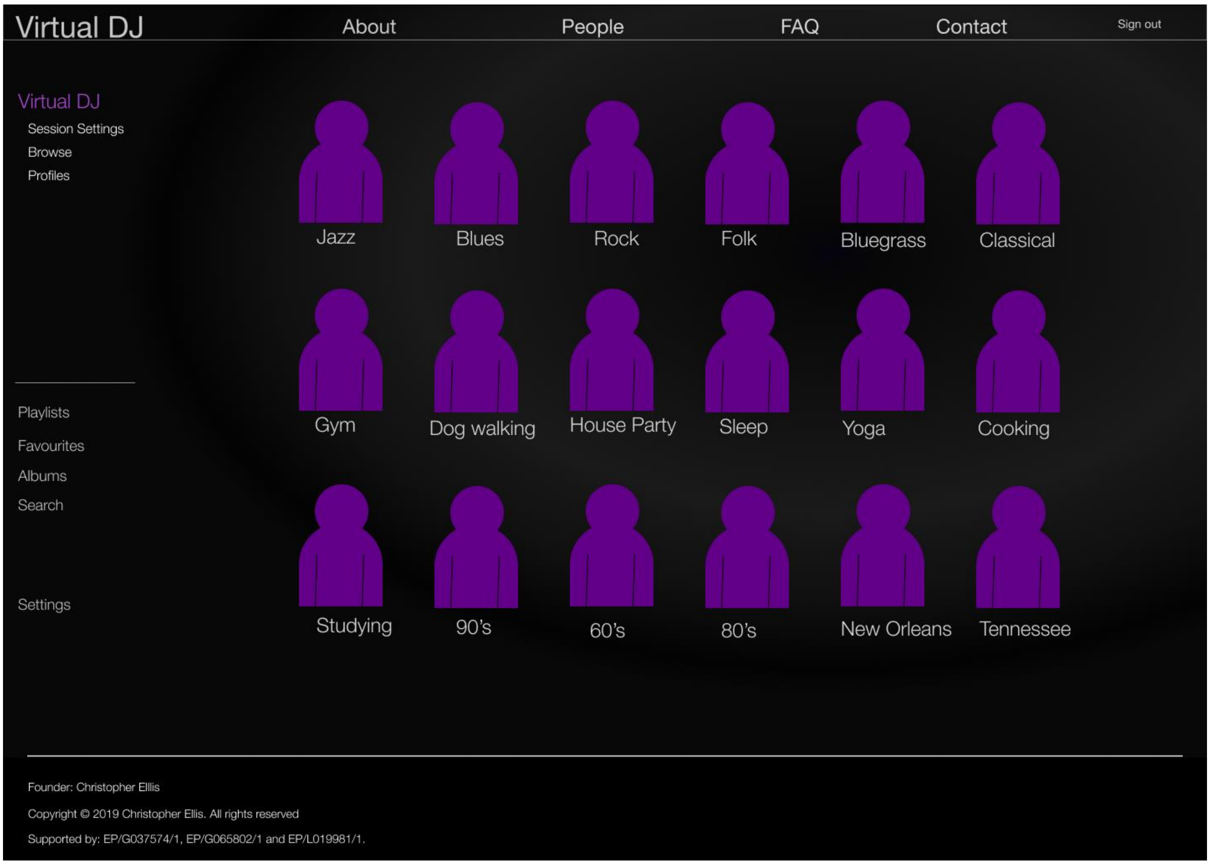
Figure 47: Playback visualiser

The playback visualiser represents the standardised music stream service view adopted by numerous services including Spotify, Apple, Tidal and Deezer. It places users in a familiar setting and presents them with a set of controls they already know how to use. The key thing here is that if they wished to skip everything above after signing in they could navigate to a music streaming and recommendation service that defaulted to something reasonable that they would likely have already come across in another streaming service.

This allows users to engage with the HITL features on a sporadic basis as much or as little as they like whilst still providing a good service that is, at its most basic, no worse than their existing recommendation / streaming service and, at its best, far more customizable and capable of producing superior and tailored personal recommendations.

Allowing the user to selectively engage in the HITL aspects of the system also allows them to learn at their own rate and experience the difference and benefits of the HITL architecture for themselves.

5.7.5.2.7 My Profiles



Founder: Christopher Elllis

Copyright © 2019 Christopher Ellis. All rights reserved

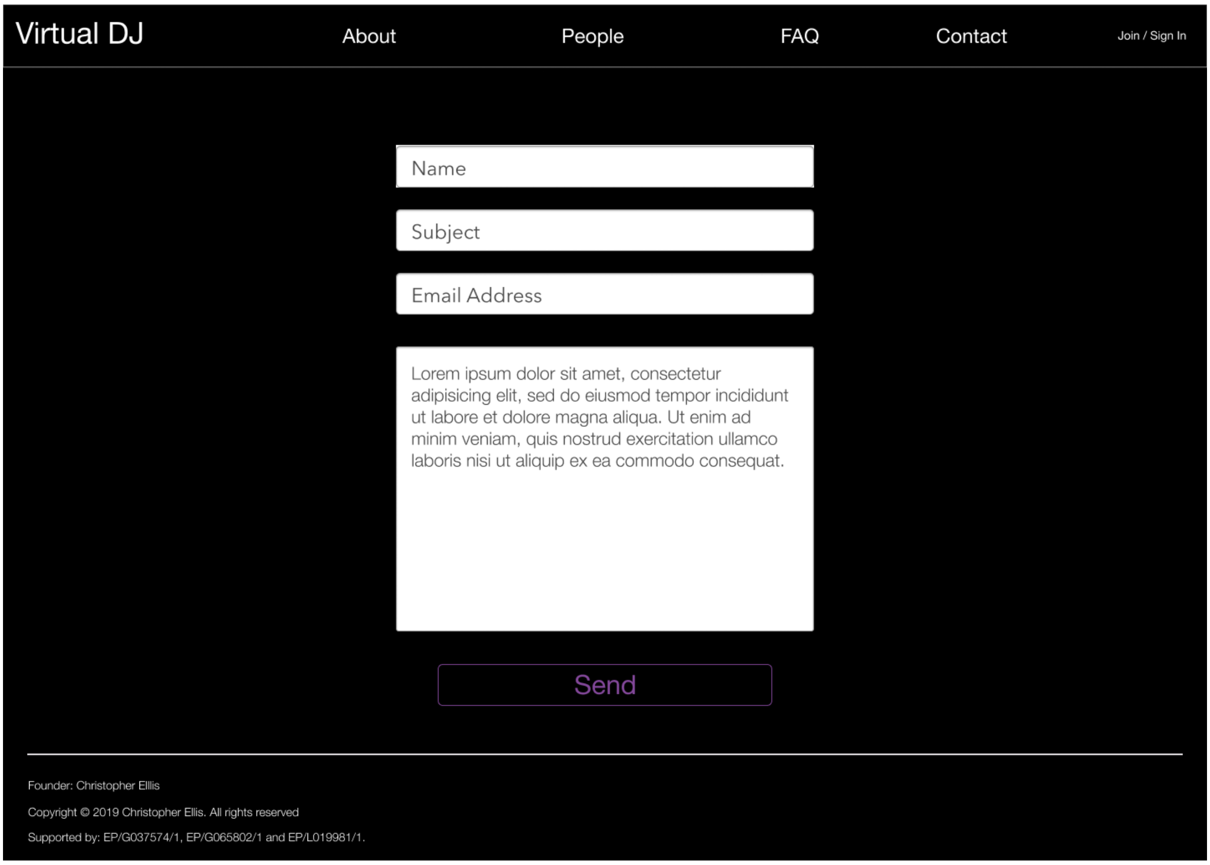
Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.

Figure 48: Profiles screen

The profiles view allows users to browse through their past listening session profiles and label them. There could also notionally be given the option to make specific profiles shared or private so that only the listening sessions people wanted to be included in the userbase would show up for other users.

From here users could select a given profile by double clicking to resume a previous listening session. Alternatively they could select multiple session profiles to create a table which would then show up in the session view screen as the active table to be used when generating recommendations.

5.7.5.2.8 Contact



Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.

Figure 49: Contact form

The contact page is a standard component of most any web-service which allows users to ask for advice and ask questions which are not currently within the frequently asked questions(FAQs) part of the website. This is of particular importance to include, at least initially in a HITL system of any kind, as no standard practice yet exists and users are currently (as of 2019) unfamiliar with how to use such services.

5.8 Validation with Users

Having constructed a series of hypothetical system screenshots in the design exercise it seemed logical to show them to some real people and conclude that chapter with small user experience (UX) study. This provides a nice way to identify any obvious low hanging fruit in design and captures a slightly broader perspective on the various merits and challenges of metaphor driven human-in-the-loop design beyond my thoughts. Crucially it should be acknowledged however that this exercise is not meant as a full-scale study from which any kind of generalisable conclusions is to be drawn. Rather it is a clean way to conclude the chapter and the thesis by returning to ‘the users’ and asking for a few reflections.

5.8.1 UX Feedback Study Design

Before beginning the study each potential participant was given an information sheet and consent form. This told them the nature of the study and informed them of their right to withdraw at any point. A copy of this form is included in the appendix for reference.

Once they had signed the consent form they were given a participant number and scheduled for an hour skype session to complete the study.

Upon starting the skype session they were read a script which provided them with an introduction to the core concepts behind the study, namely recommender systems, HITL design and the personalisation problem (see appendix for the script). At this point, they were asked if they had any questions about any of the concepts covered.

Once any questions had been answered they were provided with a weblink to a slideshow presentation of the recommendation system screens designed earlier in this chapter. They were then talked through sizing the presentation so that it displayed properly on their monitor and took up the entire browser screen. This mimicked how they might experience the real system if they were accessing it from a web browser.

Once they had successfully done this they were informed that they were going to be walked through the screenshots twice. On the first pass, they were introduced to the HITL design elements of the system. On the second pass, they were shown how a standard listening or usage session of the system might look.

Once the second walkthrough was completed they were then left to navigate back and forth through the screenshots and play at mocking up their usage of the system to evaluate it.

After evaluating the system participants were to complete a short 10-15 minute semi-structured interview during which brief written anonymous notes were taken. During the interview, they were asked to provide feedback on things they liked or disliked about the way the system worked and whether or not they had any comments regarding the user interface design and usability. A set of guiding questions can be found in the appendix.

After answering these questions the participants were shown a final screen representing a set of virtual user profiles in the system. These were constructed from the playlists provided by participants in the exploratory study earlier in the thesis. Participants were asked to choose a subset of these profiles from which a recommendation could be generated for them. They were then presented with two recommendations one constructed by running a collaborative filtering algorithm over all of the profiles and the other by just running over the participant selected profiles. Participants were asked if they preferred one or other of the recommendations and whether they had any final thoughts on the use of HITL features to personalise recommendations.

5.8.2 Participant Demographics

Since this study was aimed at being a final UX gathering exercise rather than a full-scale generalisable study an opportunistic recruitment strategy was used. Originally the plan was to recruit participants via both email and posters left around campus. The COVID-19 outbreak and subsequent lockdown situation meant that this strategy had to be altered and participants were subsequently recruited via mailing chains and email exclusively.

The study was conducted over the course of 10 days from 12 May 2020 to 22 May 2020. A table of all 10 participants along with the gathered demographic information is provided below:

Participant ID	Profession	Age	City	Gender
1	Research Assistant	29	Nottingham	Male
2	Information security officer	29	Oxford	Female
3	Legal Assessment Marker	61	Twyford	Female
4	Dog Walker	60	Twyford	Male
5	Nanny	41	Bradford	Female
6	Sales and marketing manager	32	Enschede	Female
7	Assistant prof	34	Enschede	Male
8	Program director IBM Security Division, IBM	-------	Cambridge	Female
9	Senior offering manager	40-50	Cambridge	Male
10	Postgraduate Student	26	Nottingham	Female

Although 10 participants may seem like a small number this is relatively common practise in user-experience testing where the aim is to expose a design to a general audience. This general audience of external reviewers are not caught up in the design of the system and may be capable of identifying any low hanging fruit or design oversights that the developers of the system are too close to spot. The goal is not to produce a robust set of principles or generalisable design guideline but rather to identify any common oversights, strength or weakness in the design of a system or product.

5.8.3 Materials

All materials used in this study can be found unabridged in the appendix. The materials used were:

1. The set of screenshots developed for a HITL system called VirtualDJ presented earlier in this chapter
2. A SharePoint document for taking anonymised interview notes
3. Skype for conducting participant interviews

4. Figma for designing and distributing the screenshot system user interface mockups. Figma is an interface design application that can be run in the web-browser. Figma designs can be shared online via weblinks to facilitate UX testing and preliminary feedback gathering.

5.8.4 Participant Observations

Whilst this study does not represent a full thematic analysis it did reveal some interesting insights into the HITL design and yielded several questions which warrant further research.

5.8.4.1 Flexibility & Control

All 10 participants fundamentally liked the flexibility and control that a HITL system could provide. Here are a few remarks from participants about the flexibility of the system:

> *I like the sessions settings as they would allow me to get different music based on my mood – P4*

> *I like the different levels it isn't the case that I only like one sort of thing like pasta so only want to eat pasta – P5*

> *I like that you haven't lost standard usage but gained more flexibility as and when you want it – P8*

> *I like the ability to navigate a musical landscape and widen or restrict your aperture based on what you are searching for – P9*

It is interesting that in the second quote above the participant explains their appreciation of the flexibility of the system by using the restaurant analogy that the interface is constructed around. This suggests a high level of understanding of the system interface and design as well as the problem that it has been designed to address.

Three participants remarked that HITL features would be of varying use depending on their requirements. Often participants expressed that they would like the HITL features when they were feeling adventurous and wanting to discover new music.

> *I quite like the importance of flexibility in your system. At certain times people are more willing to explore than others. – P1*

> *The adjustable controls are useful because at times I like more similar things but at other times or for other activities I want to try different things – P2*

> *I could imagine using the settings more at different times for instance when I wanted to discover some new music. – P4*

Three participants also suggested that a crucial component in the success and real-world adoption of such tools would be the ability to opt-in and out of using them easily. There was a definite sense with all participants that even if they wanted to use the HITL tools most of the time they would like to be able to rely purely on the automated component of the recommender system on occasion when they didn't have the time or desire to be involved in the recommendation process.

One participant was quoted as saying:

> *The controls are useful …. However, make it clear that you don't have to engage in session settings it is important to know that you can be as involved or not as you want to be – P2*

The notion of profiles and sessions seemed to be universally appealing to all participants. The fundamental aspect of these tools that appealed was the ability to educate the system about their immediate requirements and inform it of their varied and eclectic tastes. Several participants who described themselves as having broad tastes remarked that this feature would benefit them as conventional systems like Spotify often provide them with inappropriate recommendations as it would just try and form one profile from all their listening history.

> *The notion of multiple profiles is really useful. I have an eclectic strange mixed taste in music. Spotify is two narrow when it profiles me so it gives safe recommendations sometimes but then also out of the blue recommendations. – P2*

This participant continued to explain that Spotify didn't understand that although her taste was varied she didn't want to listen to everything she liked at the same time or in some random order. She stated that she had different tastes at different times depending on what she was feeling or doing. She felt encouraged that profiles might provide a way to capture this.

With the addition of multiple profiles and sessions, the above-quoted participant and others felt that they could represent their tastes more faithfully as distinct profiles and then get more enjoyable and appropriate recommendations by selecting only those profiles which matched their immediate listening preferences. This way it was felt they had a chance of avoiding the off-the-wall recommendation that a conventional recommendation would produce as a result of trying to form a single profile of them.

5.8.4.2 Transparency

Another interesting theme to emerge was transparency. Along with liking the flexibility of the HITL system participants commented that they liked the ability to drag profiles into their active listening session. It was felt that the metaphor in tandem with manual control really gave users a sense of transparency as to where their recommendations were coming from and why they might get recommended certain tracks.

> *I like that you directly influence your recommendations and we can see how different profiles are used to make your recommendations. The manipulation is upfront. You have a greater understanding of why you are being recommended certain things -P1*
>
> *I really like the visual means of exploring music and the transparency that brings to your recommendations – P9*

This is interesting as although we had identified transparency as a wider issue for recommender systems earlier in the thesis this was not something we anticipated HITL features would address.

5.8.4.3 Social aspect

Another slightly less surprising consequence of the HITL design was the social aspect. Participants tended to like the ability to have multiple users share and drag their profiles into an active listening session. One participant suggested this could be enhanced further by adding a user forum or chat feature to the system. This way people could engage and ask questions of one another when browsing for profiles to add to their listening sessions. Four participants quoted below claimed to really like the social aspect. In conversation, they remarked that navigating a musical recommender system using the concept of profile browsing could lead to a very social and organic form of music discovery and sharing that would imitate real life.

The best feature was the social aspect -P10

I like the cooperative aspect – P6

I like humans being involved in the discovering from other real people – P3

I like the ability to get my friend's profiles and bring them into my recommendations in a more natural way – P2

As previously mentioned one participant talked about how the social element of the system and organic music discovery could be enhanced with a chat feature. This could open up a range of different novel possibilities. You could imagine artist and album profiles having automated chatbots associated with them. Dialogues could be formed between both real users and automated profile chatbots to aid music discovery. Of course, this would also necessitate a careful privacy and user protection policy being developed.

5.8.4.4 Metaphor

The use of the restaurant analogy by the system to group profiles and facilitate profile search was met with mixed responses. Those participants that were less familiar with conventional recommenders tended to find it useful. Those participants that had made more extensive use of recommender systems tended to find it excessively complex and confusing. The conceptual leap between talking about culinary tastes and musical tastes was not universally liked.

> *The court metaphor is confusing. – P6*

> *The restaurant analogy is complex and confusing. Sometimes you might not know what you fancy so you wouldn't know what profiles to drag to your table – P7*

> *The restaurant analogy is useful for explaining the problem with convention systems but it is confusing as an interface element and might not translate to other cultures – P9*

Those who did like it quickly adapted to it and argued its merits when reporting back that they felt it provided a necessary structure to the application.

> *I like the table analogy dragging people in and out I really like that the metaphor provides a structure for finding and using profiles – P2*

One participant argued that food court analogy might be better than other metaphors precisely because it has nothing to do with music. A closer metaphor like a record store might limit users in their usage of the application as they would attempt to bring their preconceptions of how a record store works into the application.

This sentiment was echoed by another participant that stated that the food court analogy was accessible to them because they were able to embrace it without any preconceptions. For them, it provided a necessary structure and they were able to adapt to it so well as to even express the fact that they liked the flexibility of the VirtualDJ system by referencing the metaphor.

I like the different levels it isn't the case that I only like one sort of thing like pasta so only want to eat pasta – P5

It was interesting that whilst the need to navigate and locate files was universally recognised as important by all participants four participants felt that the specific metaphor for facilitating this was unimportant. One participant stated that so long as the metaphor was clear and reflected in the user interface it didn't matter what the metaphor was so long as it was explained in a usage tutorial or guide.

the analogy doesn't matter so long as usage explained- P4

5.8.4.5 Alternate metaphors

Those who disliked it tended to agree that some form of structure was required but that a metaphor closer to music like the record store analogy mentioned in the previous chapter might be more intuitive. However, this point was debated.

One participant who had liked the restaurant analogy argued that the appeal of the analogy was its universality. They remarked that everyone has a notion of restaurants and food courts and varied culinary tastes.

Furthermore, another participant commented that a record store analogy might not make sense to younger generations who got all their music online and so had no concept of browsing record stores. They also said that the record store analogy might also restrict users who attempted to navigate the system too literally based on their real-world knowledge of record stores.

A record store may be unfamiliar to a younger generation

food court might be better -P3

Whether or not the specific metaphor of a restaurant was appealing participants did seem to agree that some form of metaphor reflected in the UX was necessary as a backbone to the HITL classification and selection process. One participant suggested that a stripped-back more neutral metaphor around social bubbles like google+ circles might be a more appealing metaphor that would allow for arbitrary levels of containment and precision when grouping profiles. Furthermore, they wondered if a looser metaphor of social bubbles might be easier to grasp than either restaurants or record stores as it doesn't bring as many preconceptions about usage into the system.

> *the concept is good but the metaphor is confusing strip it back to be more neutral and flexible containment hierarchy net or bubble – P1*

Another interesting alternative metaphor that one participant suggested was the notion of a map and musical landscape.

> *I really like the idea of exploring a visual landscape of music. I am thinking of maps and map servers where you could zoom in and navigate to coordinates of interest. -P9*

The participant quoted above went on to say that you could also use some kind of aperture metaphor to facilitate focusing in on smaller regions within your musical landscape.

Users of the system could navigate a musical terrain zoom in at coordinates of interest to see a finer level of detail. As with the bubbles analogy above this one holds the advantage of allowing users to zoom in and out to an arbitrary depth rather than constraining them to three levels of granularity as the restaurant analogy did.

5.8.4.6 Tagging vs Hierarchy

One of the ideas discussed with many participants was whether a system of social tagging might be a more natural way to group and/or locate profiles. Surprisingly most felt that this would not be the case. It was argued that social tagging can be error-prone with overlapping labels that quickly get outdated.

> *There is a growing problem of tags vs directories or hierarchies' tags require that you know what you are looking or searching for – P1*
>
> *Tags more chaotic how do you update them over time what about overlapping tags, typos – P2*

It was also pointed out that tagging requires more effort from the user both in labelling other profiles in the system, to begin with, but also when trying to find profiles to add to their listening session. Fundamentally social tags require users to have some idea of what tags they are interested in and looking for which they may not have.

> *There are lots of issues with tagging it requires more input from users and it requires them to know what tags are in the system and what tags might be applied to the sorts of profiles they are looking for. – P2*

By contrast, the restaurant hierarchy gives users a place to start browsing and narrowing down profiles whether they like the specific analogy or not.

However, hierarchical structures also have their faults. In the case of virtual DJ as one participant pointed out you could easily fail to find an artist profile you were interested in simply because you didn't know the era in which the artist performed.

> *I don't like era grouping, might miss out on music you like because you don't know the decade – P10*

If you don't know where something belongs in a hierarchical system you can't find it even if you know what you are looking for. By contrast in a tagging system, you can only find those things you know the tags for.

Perhaps a HITL system might benefit from some means of combining the two methods. This reveals a broader question about social tagging and how it fits alongside conventional hierarchies or directory systems.

5.8.4.7 Usage guide

Six of the participants commented that my walkthrough of the system greatly aided their ability to navigate and evaluate it. There was a strong feeling that a usage guide of some kind would be important in a real system to introduce users to the unfamiliar HITL tools that they had not previously encountered.

> *A usage tutorial could be very helpful - P5*

> *Some kind of forum or online chat could be useful– P8*

Upon first logging, into the system, it could present a series of pop-ups to guide an initial usage session. Alternatively, there could be a first usage guide and tutorial videos on the FAQ page of the system.

5.8.4.8 Barrier to adoption & feature overload

One potential issue with the system that four participants commented on was the possibility of users initially being overwhelmed by the number of features in a HITL system. It was thought that the barrier to adoption could be quite high as the system requests quite a lot from new users both in learning the new HITL tools and then in actively using them to get the best out of the system.

> *It places a lot of demand on the user – P10*

> *Initially it appears very complicated but once users got into it they could use it effectively to find what they like – P3*

One participant liked the idea of being allowed to log in with an existing Spotify log in and thought that this would go some way to reducing the barrier to adoption especially if old Spotify playlists were imported as session profiles.

> *I like that you can sign in via Spotify as it is service people are already making use of it is nice that you don't start blank it shortens setup time – P2*

One participant stated that initially before learning the HITL controls that they might have to "fight against" the system. If they were explicitly informed that the system didn't require them to use the features this initial barrier to adoption would disappear. Then they would be free to use the system exactly as they had been using Spotify. Later on if they felt in an adventurous mood they could explore the setting and try to improve their recommendations. They could explore as many of the HITL elements as they desired at a time which suited them.

5.8.4.9 Other Feedback

A few final interesting ideas to emerge from the study were:

The idea that a social tagging system could potentially be incorporated into a system like VirtualDJ with system-generated tags that labelled tracks and profiles by attributes like 'tempol, 'male vocal' and 'percussion'. This might reduce the initial burden on users to tag profiles. It also allows them to immediately start using tags to search for profiles even before any user has begun adding their own tags to profiles.

Another interesting consideration highlighted by one participant is the notion of privacy. They suggested that the VirtualDJ system could require less effort from users if it was capable of automatically determining the user's location or the time of day. With information such as this, it could make an educated guess at filling in the session settings based on what the user had previously selected at the same time on a previous day. When asked about any privacy concerns they remarked that it didn't concern them but they could see that for others it might. They then suggested some notion of thresholding or opt-in strategy for privacy protection. Certainly, if such a feature was used to initialise the system considerable thought and research would be required to refine an appropriate privacy model.

Furthermore, because HITL by its very nature encourages if not necessitates some level of interpersonal communication between users a carefully designed privacy model would likely be required for any real implementation of a HITL system like virtualDJ. This is likely to be even more important if the system was to incorporate user-generated freeform content like tags as the misuse of such a feature could result in cyberbullying.

5.8.5 Future Research & Final Remarks

The purpose of the exercise above was to identify some initial strengths and weakness of the VirtualDJ HITL recommender system. The results suggest that HITL tools could be useful in helping people to generate more personal music recommendation. Several core aspects of the virtualDJ system seemed to be especially well received by participants. The notion of splitting their musical tastes into multiple profiles appeared to be very desirable to participants. Those with eclectic or varied tastes remarked that it was the lack of most recommenders to provide a means of splitting their tastes that contributed to their getting obvious and/or inappropriate recommendations.

The idea of session setting also appealed as a way of allowing users to voice immediate and dynamic changes in how the recommender profiled them. However perhaps the most well liked aspect of the system was the profile dragging feature which enabled users to directly influence the recommendations in a transparent way whilst also providing the potential for social and collaborative interactions. Several user remarked that it was the use of real humans in generating recommendations and the ability to pull in their friend’s profiles to inform their recommendations that appealed the most. The discussion of facilitating transparency is an interesting one that warrants further research as this was not an area we anticipated HITL tools as being especially useful for. A follow up study that sought to emphasis and study how HITL tools could be maximised for providing transparent recommendations could provide further interesting insight that is broader than the original intention of this study or indeed this thesis.

Whilst the above feedback was useful it is also important to acknowledge that not all aspects of the system were found to be clear or useful by all participants. In conducting the above study it became clear that the food metaphor for enabling HITL interaction was a dividing element of the system that split participants into three camps. Some simply found it far too abstract and complex and disliked that they had to liken musical tastes to the unrelated notion of culinary tastes in order to engage in the HITL aspect of the system. Other participants adapted to it really quickly and liked the structure and degree of control and navigation it facilitated within the application. A third group of participants liked the HITL features that underpinned the restaurant metaphor but argued that the specific metaphor was unimportant so long as they were told how to use the system via some means of usage guide or tutorial.

Finally the use of a strict hierarchy for categorising and retrieving profiles was contrasted by several participants with social tagging and several benefits and limitations of each approach were identified. Fundamentally it was acknowledged that strict hierarchies allowed users to search more generally and browse for profiles when they had no direct idea of what they were necessarily looking for. By contrast social tagging was acknowledged to require more input from users but allow a much more convenient way to locate specific profiles via labelled tags so long as you know the sorts of tags a given profile might have. The possibility of combining the strengths of each of these systems as well as the importance of developing an appropriate privacy and user protection policy would make for interesting future research.

5.8.6 System Unified *Modelling Language* (UML) Diagram

The user validation study of the VirtualDJ interface revealed that HITL tools could be useful in combatting the personalisation problem but it also suggested that the real strengths of the system were not tied to the particular restaurant metaphor in the system. Even users that liked the metaphor acknowledged that other metaphors could have been used. The fundamental strengths of the system appeared to be its ability to allow users to voice their dynamic taste preferences and manually combine human and machine personas in a nuanced manner. The core aspects which users responded to in terms of their potential for generating personal recommendations were profiles, social tagging and the dynamic ability to combine profiles to restrict content that the automated algorithm within the system took as its input.

To investigate these features we chose to use the concept of a food court but this could easily be replaced in a given implementation by another metaphor such as a music festival, sports club or flea market.

By reviewing the study above alongside the previously presented design exercise, it becomes possible to tease out the fundamental structure and general design principles and ideas which a future designer might take note of. A system which follows these is able to treat human and machine personas as first class citizens and flexibly combine them in a dynamic way to produce pleasing recommendations. A unified model of the system is shown below after which the general principles are described.

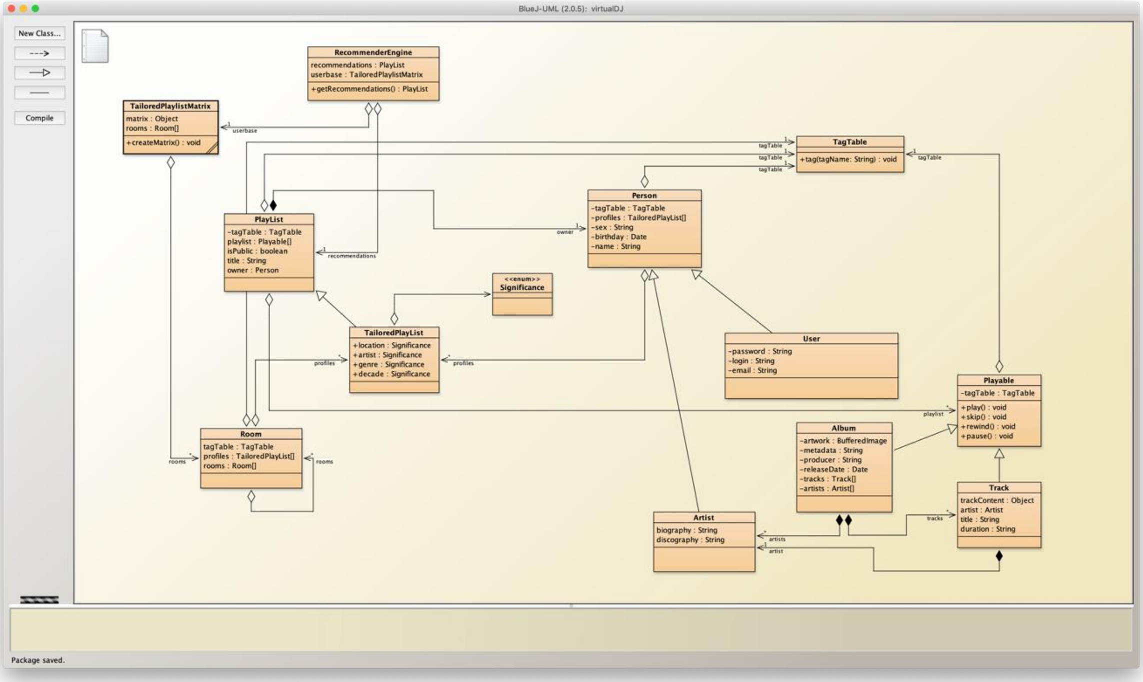


Figure 50: VirtualDJ UML Diagram

5.8.6.1 General Design Principles

Beyond the standard components of conventional recommender systems, the HITL approach proposed by this thesis consists of three core concepts represented in the above UML diagram by the classes TagTable. TailoredPlaylist and Room.

5.8.6.1.1 TagTable

As we saw in chapter two, the earliest manual recommender systems such as Tapestry supported the emergence of expert users who were knowledgeable about certain types of content. Expert users could be followed by other users of the system to receive tailored recommendations about specific sorts of content . Unfortunately this feature required users to have knowledge of the other users in the system in order to identify who their expert users were. As we saw, this rapidly became untenable when systems reached more than a few hundred users.

In our system we reintroduce the idea of expert recommenders to the modern automated recommender through the concept of social tagging. It should be noted from the validation study that further research would be required before implementing a system with social tagging to avoid problems like overlapping tags and duplicate or miss-spelt tags. We outline a tagging framework here that could serve as the basis for this further research.

Most entities in our system are Taggable, that is they have a table of Tags. A tag is a string value with an associated integer usage count. When a new tag is added to an item, the new tag is created with a specified name and an initial count of 1. If a tag is added to an item where a tag of the same string value exists, the existing tag usage count is incremented. Similarly deleting a tag from a tag table decrements the usage count unless the usage count is 1, whereupon the tag is removed from the tag table.

The primary purpose of tags within the system is to aid users in their efforts to interact with the recommendation process and interfere with the content pool which ultimately gets exposed to the recommender engine.

The ability to have tags stored with their usage counts allows for some very interesting implementation options. For instance, a given implementation could opt to only expose tags whose counts are above a certain threshold. This would allow an ever-evolving taxonomy to emerge which continually reflects the most popular way users wish to think about, retrieve and classify their music .

An alternative strategy which could be supported using this framework would be to give individual users the ability to select their own thresholding for determining which tags were visible to them. This would allow user with niche interests to set a lower threshold, thereby allowing them to view less popular tags which may have only been used by a few minority users in the system which share their specific interests.

5.8.6.1.2 TailoredPlaylist

Another core concept to the HITL system is the TailoredPlaylist. TailoredPlaylists consist of a set of playable items such as tracks or albums along with a set of user preferences for common music classification attributes, like location, artist, genre and decade. This allows the user to specify the extent to which they care about these attributes when receiving recommendations.

In the restaurant metaphor system explored earlier in this chapter, profiles are implemented as a TailoredPlaylist. Whenever a user closes the application, their current session preferences and listening history for that active session are saved as a new profile using the TailoredPlaylist class to store this information.

The important characteristic of the TailorPlaylist is that it allows users to apply a weighting to the recommender engine. If they really care about all tracks for a given listening session being from a single genre, they can express this in their session settings. Internally within the system this will get applied as a weighting parameter to playable items whose genre matches the genre of the artists and seed tracks given by the user in their session initialization settings.

5.8.6.1.3 Rooms

The final critical feature of the HITL approach explored in this thesis is the concept of Rooms. Rooms provide the functionality to capture the containment hierarchy of a given metaphor upon which a specific HITL implementation rests. A Room is taggable and comprises a set of TailoredPlaylists and Rooms.

In the restaurant metaphor, the Room class is used to implement the food courts, restaurants and tables features of the system. Using the restaurant metaphor, the system containment hierarchy is three levels deep. Food courts are Rooms which contain a set of small rooms known in the system metaphor as restaurants. Restaurants in turn contain a set of even smaller Rooms referred to as tables. Tables then contain a set of TailoredPlaylists representing profiles.

Whilst the restaurant analogy only has three levels of containment, this is an arbitrary decision and could vary depending upon the metaphor used. For instance, if one were to swap the restaurant metaphor for a record store, you could have a smaller containment hierarchy representing the rows and sections classification system you may find in a physical record store. Alternatively, if the metaphor of a music festival were used, you might have a larger hierarchy of Rooms consisting of fields, quarters, tents and stands.

5.8.6.2 An Alternative Metaphor

By using the core components of TagTable, TailoredPlaylists and Rooms, it becomes possible to design a wide variety of HITL music systems which shape their user experience around a familiar metaphor. The metaphor used then becomes a core part of the user engagement and interaction with the HITL characteristics of the system as demonstrated by the restaurant analogy explored earlier in the chapter.

In essence, the metaphor chosen, be it a restaurant, sports hall or music festival becomes a lens through which users can explore and interact with the system.

By storing both user listening sessions and artist discography settings as TailoredPlaylists it becomes possible to treat human and machine personas (called profiles in the restaurant analogy) as interchangeable entities. This helps break down the divide between the automated machine aspects of the system and the HITL components.

For the sake of clarity, we now briefly summarise how the same UML diagram and class hierarchy could be used to build a system based on the metaphor of a record store rather than the restaurant metaphor.

User listening sessions and artists could be represented as customer avatars (stored internally as TailoredPlaylists) in a music store. You could even envisage the ability to customise your set of avatars by, for instance, having it wear a stereotypical cowboy hat if you were interested in country and western music. This could serve as a visual cue for other users exploring the userbase.

Customers could be grouped into clusters or sections within the store (again represented internally as TailoredPlaylists). You might think of how people in real record stores might cluster around a particular section containing the sort of music they might be interested in.

You could even consider adding an instant messaging chat style form of communication such that users could interact with one another through their customer avatars replicating the sort of conversations that take place in a physical record store. This would again break down the division between automated recommendations and human ones.

Finally, another advantage of this metaphor is that artists in the system would also have customer avatars which could be stylised to resemble their likeness. Users could find themselves in a virtual record store alongside Muddy Waters and Taylor Swift. They could even notionally interact with these avatars via automated chatbots alongside the real chat communications with real online users.

At the top level the store sections could then be grouped into rows representing sets of customers with broadly similar and some overlapping interests.

The brief presentation of an alternative metaphor above demonstrates the flexibility of the general system features identified in this chapter. It also shows how different metaphors could give rise to different system features and advantages whilst all the while serving to break down the divide between machine and human personas and support a natural means of generating personalised HITL music recommendations.

5.9 Chapter Summary

This chapter presented the key findings of this thesis as a list of requirements for a HITL music recommender system. In HCI fashion it then proceeded to explore how these requirements might be met through a detailed exploration of the restaurant analogy used in the introductory chapter of this thesis as a means of explaining some of the limitations of conventions music recommender system.

From the design exploration exercise, the chapter proceeded to generate a general framework and set of design principles which captured the core components of the HITL approach to personalised music recommendation suggested in this thesis. Three core features were identified, including the ability to tag items, the notion of Rooms and the concept of TailoredPlaylists.

After presenting the core design principles and components, a brief description was provided as to how an alternative metaphor of a record store could be used in place of the restaurant metaphor. This served a dual purpose. First, it helped clarify the general principles which had emerged as a guideline for HITL music recommendation from the restaurant exercise. Second, it showed how constructing a different metaphor on top of these general principles can give rise to a different user experience and support different means of interacting with the HITL elements.

Chapter 6: Conclusion

The field of music recommender systems is vast and inherently multi-disciplinary, incorporating elements of information retrieval, machine learning, human-computer interaction, musicology, network theory and datamining. Besides this it is also unusual in having practical application within both academia and industry.

In the literature review chapter of this thesis, we provided a brief overview of the diverse range of problems within the field. Obviously it was beyond the scope of this thesis to address all of these issues in-depth. Within this thesis the focus was restricted to those issues which together comprised the personalisation problem for music recommendation.

Since the personalisation problem is very convoluted and complex, a significant portion of this thesis sought to untangle it from unrelated problems and provide a clean characterisation of the problem. Once this was achieved it became possible to relate the personalisation problem to the field at large and highlight its multiple facets.

Having defined the problem and positioned it within the field, the rest of this thesis was focused on exploring the individual facets of the problem and constructing a methodological framework and approach for addressing them.

While defining the personalisation problem, we observed that early manual recommenders actually performed well in relation to the personalisation aspect but that they failed to scale well. This led us to investigate how people curate and recommend music naturally and what it is about inter-personal recommendation that gives it that personal touch.

Our investigations revealed that music recommendation is highly subjective and that personalisation consisted of balancing predictive accuracy against novelty whilst taking account of the purpose for which an individual or group are seeking recommendations.

After conducting a study into the character and specifics of recommendation purpose we began to draw together the findings of this thesis into a cohesive framework upon which a new style personalised HITL music recommender might be constructed. We explored this by adopting the human-computer interaction practice of scenario design and user validation testing. This allowed us to draw together the central tenets of this thesis and formulate a general set of design principles that might be applied to commercial and academic music recommender systems to generate personalised recommendations.

6.1 Research Questions Revisited

At the beginning of this thesis we introduced the overarching research question below:

How can human-in-the-loop techniques be applied to reincorporate the core tenets of making personalised music recommendations into modern recommender services?

We then proceeded to break this down into two sub-questions to be addressed throughout this thesis. These smaller questions were:

1. **What are the tenets of making personalised music recommendations?**

2. **How can human-in-the-loop practices allow users to inform an automated music recommender of their requirements for personalised recommendations?**

The tenets of personalised music recommendations according to the work conducted within this thesis are:

1. Balancing novelty and predictive accuracy
2. Accounting for the purpose for which the music is intended
3. Accounting for the dynamic nature of people's tastes

Having identified the core tenets of music recommendation that people engage with when making personal recommendations, we then explored how they might be accommodated in a modern automated recommender.

The work done in this thesis gives evidence to suggest that this could be achieved by designing a layered HITL system upon a common metaphor. In chapter 5 we presented VirtualDJ in the design exercise. It was shown how the system could account for people's varying tastes at different times by allowing them to pre-filter the userbase being sent to a standard collaborative filtering recommender. This filtering process was supported by having a layered containment hierarchy structured with the concept of Rooms where Rooms either contain TailoredPlaylists, or further Rooms. The system was also shown to be able to be used dynamically, allowing users to share different facets of themselves and their tastes at different times or when they were seeking recommendations for different purposes.

6.2 Commercial Implications For Designers

This thesis has a multitude of implications for music recommender system designers. Perhaps the biggest implication a commercial entity might wish to take account of is the importance shown in this thesis for constructing a HITL system around a metaphor. This seemingly trivial decision allows for a tightly integrated user interface that guides the user by presenting them with familiar visual elements that provide cues to expected behaviour. This is of particular importance now as no standard practice currently exists for HITL music recommenders.

6.3 Academic Implications For Research

Regarding academic implications this thesis identifies that predictive accuracy has been something of a red herring in recent years in recommender design, especially in the area of music recommendation. A future researcher guided by the findings of this thesis may wish to temper their own system's focus on accuracy and place it within the flexible framework of a HITL system.

Additionally, this system gives reasons to suppose that privacy and transparency will become of increasing importance and a fertile area of research should recommenders move in a more personalised HITL direction.

A final lesser academic contribution can be found in the study presented in chapter 4 which contributed a novel use of crowdsourcing and data analysis to reduce the impact of the WEIRD bias on HCI centred research.

6.4 Conclusions & Future Research

6.4.1 Summary of Contributions

This thesis has produced several novel ideas, frameworks and deliverables which can be made use of by both academic and commercial entities. The main contributions of this are:

1. Defining and positioning the personalisation problem for music recommendation within the wider fields of music recommendation and recommender systems more generally. (Specifically highlighting the role of recommendation purpose and predictive accuracy in shaping the problem)

2. Presenting the VirtualDJ recommender system which incorporates social tagging and a HITL framework for producing personalised music recommendation

3. Outlining a metaphor led framework to constructing HITL music recommender systems which can be used to build familiar user interfaces in the absence of any existing standards or common practices

4. Producing a set of general design principles which can be taken as a road map for building personalised HITL music recommender systems

6.4.2 Limitations & Future Research

As with all research endeavours, there are inevitably certain aspects of this thesis that could be expanded or improved upon. Here we consider the three main areas of this thesis which could be built upon in future research.

6.4.2.1 Testing in The Wild

As mentioned earlier, the subject of this thesis is unusual in its inherently interdisciplinary character and its combined academic and commercial appeal. A regrettable consequence of this, which only became apparent after defining the characteristics of the personalisation problem, is that it must be tested at scale in the wild. There are two reasons for this. The first is that smaller scale systems, whether manual or automated, are not inflicted with the personalisation problem to begin with so you would likely get false positive results in testing. Study participants would likely report satisfactory results and no personalisation issues, but this would be at least in part due to the smaller scale of the system that allows them to navigate the entire user base rather than as a result of engaging with the HITL and social tagging techniques to navigate and restrict the userbase as intended. Second, HITL and social tagging systems only become operational and stable at scale, especially in the case of the system outlined in this thesis where users need to be able to comment on other users with a wide variety of tastes and listening habits. This cannot be imitated in a wizard-of-oz style of research nor by user virtual profiles, as the outlined approach relies on already having virtual profiles in the form of artist profiles.

To mitigate this issue and lay the groundwork for a large organisation to adopt and test a HITL methodology, we used a common technique within HCI involving prototyping and system design as a means of generating a road map for larger organisations to follow in implementing or extending their systems.

An exciting future project exists in implementing and trialling a system such as the virtual DJ system designed in the previous chapter of this thesis. Trialling such a system necessarily involves approaching a large commercial organisation like Spotify which has a large pre-existing user and content base and talking about partnering them to conduct a trial of the system as a joint academic, commercial research and development project. To this end we have entered into discussions with Hugh Rawlinson, a developer at Spotify.

6.4.2.2 Transparency vs Privacy

Although we have touched on the issues of privacy and transparency in this thesis, they have largely remained outside of the scope of investigation. Recommender systems generally have become increasingly opaque as they have grown in complexity to handle scalability issues and in the long pursuit of finding a solution to the cold-start problem.

The approach advocated in this thesis is agnostic in its underlying algorithmic recommender engine, but still vulnerable to any privacy concerns or criticisms of opacity that a chosen recommender engine might face in isolation. Furthermore, the approach in this thesis introduces further nuanced privacy concerns which would need careful consideration before embarking on a real implementation to test in the wild. One of the areas of privacy that the proposed approach needs more research conducted into is the field of social networking and social tagging.

For instance, a full implementation of the proposed virtual DJ system would need to account for what happens to tags when a user opts to leave the platform. Do the tags get removed from the system? Are they anonymized such that they can no longer be traced back to the user profile or even the individual who created them? Conventional systems were protected to some degree from the worst privacy sins by their monolithic approach towards user profiling. They are bound to be inaccurate to some degree due to their inability to model the dynamism of their users. This has the unintended consequence that any privacy breaches are constrained to the particular and incomplete facets and representation of a person in their system. In the approach outlined, people expose multiple facets of themselves via holding multiple profiles. As such, there is the potential to construct a more holistic view of a person from a profiling perspective. This is especially true if you consider the addition of tagging information and instant messaging conversations (a feature which was discussed briefly as an implementation option in the record store metaphor presented in chapter 5).

The very nature of the approach in this thesis incorporating humans and human insight into the system design means that it is collating more data than just that which pertains to a person’s listening history or musical tastes.

In the literary review chapter of this thesis we presented a privacy anecdote whereby the analysis and naïve use of purchasing habits resulted in a teenager’s parents being informed that she was pregnant. In a similar way a naïve or sinister misuse of the potential data accrued in a system like VirtualDJ could also have unintended privacy considerations and consequences. Users may reveal through tagging all manner of personal information which is vulnerable to misuse without realising it. As an artificially obvious example, if a user were to tag a profile as being good for listening to during their Thursday night gym session, it might reveal that their house was empty on Thursday nights.

6.4.2.3 Domain Specific

Finally since this thesis focused specifically on the case of music recommendation, its findings cannot necessarily be directly applied to other recommender domains such as e-commerce or movies. This gives rise to the following future research questions:

1. To what extent does the personalisation problem arise in other recommender domains?

2. How does the personalisation problem as it affects e-commerce or movie recommenders relate to the personalisation problem for music recommendation?

3. What aspects of the HITL and social tagging approach outlined in this thesis might be applicable to solving the related problem (if it exists) for e-commerce or movie recommendation?

6.4.3 Concluding Remarks

Oscar Celma finished his pinnacle thesis on the subject of longtail music recommendation with the following passage:

> *We have an overwhelming number of choices about which music to listen to. As stated in (Schwartz, 2005), we —as consumers— often become paralysed and doubtful when facing the overwhelming number of choices. The main problem, then, is the awareness of content in the tail, not the actual access to the content. Here is where personalised filters and recommender systems enter as part of the solution. Effective recommendation systems should promote novel and relevant material (non–obvious recommendations), taken primarily from the tail of a popularity distribution (Celma, 2010a).*

The findings of this thesis echo his assessment that access to content is not the main problem. Further, this thesis also contends that both discovery and novelty can be an important part of the recommendation task. Indeed a large part of the personalisation problem as characterised in this thesis is caught up in balancing predictive accuracy against recommendation novelty. However this thesis builds on Oscar's work by identifying that neither discovery nor novelty are **necessarily** important in a given recommendation case. The **necessary** feature of a personal music recommender is its ability to cope with multiple **purposes for seeking recommendations** and account for the **dynamic nature of individuals' tastes** (something which Oscar does hint at in his own thesis). This thesis has explored the use of HITL technologies in meeting these requirements and presented an abstracted set of design principles for constructing systems for meeting these requirements.

Oscar Celma's thesis was primarily concerned with novelty and discovery within music recommendation. However, it was also the first to identify that personalisation was lacking from modern music recommender systems. It represents an initial stepping stone in the path towards truly personalised music recommendation.

The goal of this thesis was to define the problem hinted at in Oscar's thesis and explore how HITL technologies might be used to address it. In this way it is hoped that this thesis can serve as another stepping stone moving us ever forward toward the true holy grail of personalised music recommendation. Perhaps the largest task remaining and the next step in the path concerns transparency and privacy, an issue which was hinted at in this thesis but which as yet remains to be explored.

6.4.4 Closing Passage

At the time of writing this, we are beginning to see large scale commercial entities adopt HITL techniques and engage with the personalisation problem with creations like Discover Weekly.

This thesis represents a small step in pursuit of personalised music recommendation by defining the personalisation problem and presenting a preliminary approach to addressing it via social tagging and HITL technologies.

The reality of the truly personalised virtual DJ is getting ever nearer. Soon the personalisation problem like the cold-start problem will be an issue of the past. Between now and then some fascinating research remains to be undertaken in the field of the transparency and privacy of these new personal music recommenders.

Bibliography

0012, J. W., de Vries, A. P., & Reinders, M. J. T. (2006). Unifying user-based and item-based collaborative filtering approaches by similarity fusion. *Sigir*, 501.

Adomavicius, G., & Kwon, Y. O. (2011). Maximizing aggregate recommendation diversity: A graph-theoretic approach. *Proc of the 1st International Workshop on …*.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(6), 734–749. http://doi.org/10.1109/tkde.2005.99

Adomavicius, G., & Tuzhilin, A. (2011). Context-Aware Recommender Systems. In *Recommender Systems Handbook* (pp. 217–253). Springer US. http://doi.org/10.1007/978-0-387-85820-3_7

Al-Shamri, M. Y. H., & Bharadwaj, K. K. (2007). A Compact User Model for Hybrid Movie Recommender System (pp. 519–524). Presented at the Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007) - Volume 01, Washington, DC, USA: IEEE Computer Society. http://doi.org/10.1109/ICCIMA.2007.4

Amazon's algorithm 'suggests bomb-making recipes'. (2017). Amazon's algorithm 'suggests bomb-making recipes'. Retrieved from http://www.independent.co.uk/news/uk/home-news/amazon-algorithm-bomb-making-components-mother-of-satan-channel-4-investigation-a7954461.html

Anna, B. (2016). *Recommender Systems—It's Not All About the Accuracy*.

Baccigalupo, C., Plaza, E., & Donaldson, J. (2008). Uncovering Affinity of Artists to Multiple Genres from Social Behaviour Data. *Ismir*, 275–280.

Balabanović, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, *40*(3), 66–72. http://doi.org/10.1145/245108.245124

Basu, C., Hirsh, H., & Cohen, W. (1998). Recommendation as classification: Using social and content-based information in recommendation. *Aaai/Iaai*.

Ben Schafer, J., Konstan, J. A., & Riedl, J. (2001). E-Commerce Recommendation Applications. In *Applications of Data Mining to Electronic Commerce* (pp. 115–153). Boston, MA: Springer US. http://doi.org/10.1007/978-1-4615-1627-9_6

Bennett, J., & Lanning, S. (2007). The Netflix Prize (pp. 3–6). Presented at the Proceedings of the KDD Cup Workshop 2007, New York: ACM.

Billsus, D., & Pazzani, M. J. (1998). Learning Collaborative Information Filters. *Icml*.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering (pp. 43–52). Morgan Kaufmann Publishers Inc.

Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, *12*(4), 331–370. http://doi.org/10.1023/A:1021240730564

Byrd, D. (n.d.). *Organization and search of musical information*. Retrieved from http://informatics. indiana.edu/donbyrd/Teach/I545Site-Spring08/Syllabus I545.html

Celma, O. (2008). *Music Recommendation and Discovery in the Long Tail*. Barcelona. Retrieved from http://mtg.upf.edu/system/files/publications/PhD_ocelma.pdf

Celma, O. (2010a). Introduction. In *Music Recommendation and Discovery* (pp. 1–13). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-13287-2_1

Celma, O. (2010b). Music recommendation and discovery : the long tail, long fail, and long play in the digital music space. Heidelberg ; New York: Springer.

Celma, O., & Herrera, P. (2008). A new approach to evaluating novel recommendations (pp. 179–186). Presented at the Proceedings of the 2008 ACM Conference on Recommender Systems.

Celma, O., & Lamere, P. (2008). If you like the beatles you might like...: a tutorial on music recommendation. *ACM Multimedia*, 1157–1158. http://doi.org/10.1145/1459359.1459615

Chee, S. H. S., Han, J., & Wang, K. (2001). RecTree: An Efficient Collaborative Filtering Method. In *Data Warehousing and Knowledge Discovery* (Vol. 2114, pp. 141–151). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/3-540-44801-2_15

Condliff, M. K., Lewis, D. D., Madigan, D., & Posse, C. (1999). Bayesian mixed-effects models for recommender systems. *ACM SIGIR'99 Workshop on …*.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., & Furnas, G. W. (1990). Indexing by latent semantic analysis. *JAsIs*.

Degemmis, M., Lops, P., & Semeraro, G. (2007). A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. *User Modeling and User-Adapted Interaction*, *17*(3), 217–255. http://doi.org/10.1007/s11257-006-9023-4

Donnat, O. (2018). Évolution de la diversité consommée sur le marché de la musique enregistrée, 2007-2016. *Culture Etudes*, *n° 4*(4), 1–32.

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative Filtering Recommender Systems. *Foundations and Trends® in Human–Computer Interaction*, *4*(2), 81–173. http://doi.org/10.1561/1100000009

Ghazanfar, M. A., & Prugel-Bennett, A. (2010). A scalable, accurate hybrid recommender system (pp. 94–98). Presented at the Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on, IEEE.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61–70. http://doi.org/10.1145/138859.138867

Goodman, D., & Keene, R. (1997). Man Versus Machine: Kasparov Versus Deep Blue. H3.

Han, B.-J., Rho, S., Jun, S., & Hwang, E. (2009). Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, *47*(3), 433–460. http://doi.org/10.1007/s11042-009-0332-6

Hariri, N., Mobasher, B., & Burke, R. (2012). Context-aware music recommendation based on latenttopic sequential patterns. *the sixth ACM conference* (pp. 131–138). New York, New York, USA: ACM. http://doi.org/10.1145/2365952.2365979

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The Weirdest People in the World? *SSRN Electronic Journal*. http://doi.org/10.2139/ssrn.1601785

Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. *the 22nd annual international ACM SIGIR conference* (pp. 230–237). New York, New York, USA: ACM. http://doi.org/10.1145/312624.312682

Herlocker, J., Konstan, J. A., & Riedl, J. (2002). An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms. *Information Retrieval*, *5*(4), 287–310. http://doi.org/10.1023/A:1020443909834

Hill, K. (2012, February 16). How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did. Retrieved October 11, 2017, from https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#441a3d326668

Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. *the SIGCHI conference* (pp. 194–201). New York, New York, USA: ACM Press/Addison-Wesley Publishing Co. http://doi.org/10.1145/223904.223929

Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 89–115. http://doi.org/10.1145/963770.963774

Horsey, R. (2002). The art of chicken sexing. *UCL Working Papers in Linguistics*, *14*.

Iaquinta, L., Gemmis, M. de, Lops, P., Semeraro, G., Filannino, M., & Molino, P. (2008). Introducing Serendipity in a Content-Based Recommender System (pp. 168–173). Presented at the 2008 8th International Conference on Hybrid Intelligent Systems (HIS), IEEE. http://doi.org/10.1109/HIS.2008.25

Iaquinta, L., Gentile, A. L., Lops, P., de Gemmis, M., & Semeraro, G. (2007). A Hybrid Content-Collaborative Recommender System Integrated into an Electronic Performance Support System (pp. 47–52). Presented at the 7th International Conference on Hybrid Intelligent Systems (HIS 2007), IEEE. http://doi.org/10.1109/HIS.2007.30

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, *16*(3), 261–273. http://doi.org/10.1016/j.eij.2015.06.005

Jalali, M., Gholizadeh, H., & Hashemi Golpayegani, S. A. (2014). An improved hybrid recommender system based on collaborative filtering, content based, and demographic filtering. *International Journal of Academic Research*, *6*(6), 22–28. http://doi.org/10.7813/2075-4124.2014/6-6/A.3

Jia Rongfei, Jin Maozhong, & Wang Xiaobo. (2007). Web Objects Clustering Using Transaction Log (pp. 182–186). Presented at the 2010 3rd International Conference on Knowledge Discovery and Data Mining (WKDD 2010), IEEE. http://doi.org/10.1109/WKDD.2010.69

Kaminskas, M., & Ricci, F. (2012). Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, *6*(2-3), 89–119. http://doi.org/10.1016/j.cosrev.2012.04.002

Karwowski, W. (2006). International Encyclopedia of Ergonomics and Human Factors - 3 Volume Set. (Informa Healthcare & W. Karwowski, Eds.) (0 ed.). CRC Press. http://doi.org/10.1201/9780849375477

Kefalidou, G., & Sharples, S. (2016). Encouraging serendipity in research: Designing technologies to support connection-making. *International Journal of Human-Computer Studies*, *89*, 1–23. http://doi.org/10.1016/j.ijhcs.2016.01.003

Kim, J., Lee, S., & Yoo, W. (2013). Implementation and analysis of mood-based music recommendation system (pp. 740–743). Presented at the Advanced

Communication Technology (ICACT), 2013 15th International Conference on, IEEE.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). GroupLens: applying collaborative filtering to Usenet news. *Communications of the ACM*, *40*(3), 77–87. http://doi.org/10.1145/245108.245126

Lamere, P. (2008). Social Tagging and Music Information Retrieval. *Journal of New Music Research*, *37*(2), 101–114. http://doi.org/10.1080/09298210802479284

Landauer, T. K., Littman, M. L., Bell Communications Research, Inc. (1994). Computerized cross-language document retrieval using latent semantic indexing.

Lekakos, G., & Caravelas, P. (2008). A hybrid approach for movie recommendation. *Multimedia Tools and Applications*, *36*(1-2), 55–70. http://doi.org/10.1007/s11042-006-0082-7

Lemire, D., & Maclachlan, A. (2005). Slope One Predictors for Online Rating-Based Collaborative Filtering. *Sdm*.

Levy, M., & Sandler, M. (2009). Music Information Retrieval Using Social Tags and Audio. *IEEE Transactions on Multimedia*, *11*(3), 383–395. http://doi.org/10.1109/TMM.2009.2012913

Linden, G. D., Jacobi, J. A., & Benson, E. A. (2001). *Collaborative recommendations using item-to-item similarity mappings*. Google Patents.

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*, *7*(1), 76–80. http://doi.org/10.1109/mic.2003.1167344

Maltz, D., & Ehrlich, K. (1995). Pointing the way (pp. 202–209). Presented at the Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '95, New York, New York, USA: ACM Press. http://doi.org/10.1145/223904.223930

Mamunur Rashid, Al, S. K. L., Karypis, G., & Riedl, J. (2006). ClustKNN: a highly scalable hybrid model-\& memory-based CF algorithm.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006a). Being accurate is not always good: How accuracy metrics have hurt recommender systems. ACM Special Interest Group on Computer Human ….

McNee, S. M., Riedl, J., & Konstan, J. A. (2006b). Being accurate is not enough (pp. 1097–1101). Presented at the CHI '06 extended abstracts, New York, New York, USA: ACM Press. http://doi.org/10.1145/1125451.1125659

Miyahara, K., & Pazzani, M. J. (2000). Collaborative Filtering with the Simple Bayesian Classifier. In *PRICAI 2000 Topics in Artificial Intelligence* (Vol. 1886, pp. 679–689). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/3-540-44533-1_68

Montaner, M., López, B., & la Rosa, de, J. L. (2003). A Taxonomy of Recommender Agents on the Internet. *Artificial Intelligence Review*, *19*(4), 285–330. http://doi.org/10.1023/A:1022850703159

Moravec, H. (1988). Mind Children. Harvard University Press.

Nanopoulos, A., Rafailidis, D., Symeonidis, P., & Manolopoulos, Y. (2010). MusicBox: Personalized Music Recommendation Based on Cubic Analysis of Social Tags. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(2), 407–412. http://doi.org/10.1109/TASL.2009.2033973

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, 111–125. http://doi.org/10.1109/SP.2008.33

Nielsen, J. (1989). Usability engineering at a discount, 394–401.

Nielsen, J. (2012). How many test users in a usability study? Retrieved 28 August 2012.

Park, H.-S., Yoo, J.-O., & Cho, S.-B. (2006). A Context-Aware Music Recommendation System Using Fuzzy Bayesian Networks with Utility Theory. In *Fuzzy Systems and Knowledge Discovery* (Vol. 4223, pp. 970–979). Berlin, Heidelberg: Springer, Berlin, Heidelberg. http://doi.org/10.1007/11881599_121

Pennock, D. M., Horvitz, E., & Giles, C. L. (2000a). Social choice theory and recommender systems: Analysis of the axiomatic foundations of collaborative filtering. *Aaai/Iaai*.

Pennock, D. M., Horvitz, E., Lawrence, S., & Giles, C. L. (2000b). Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach (pp. 473–480). Morgan Kaufmann Publishers Inc.

Pontis, S., Kefalidou, G., Blandford, A., Forth, J., Makri, S., Sharples, S., et al. (2015). Academics' responses to encountered information: Context matters. *Journal of the Association for Information Science and Technology*, *67*(8), 1883–1903. http://doi.org/10.1002/asi.23502

Popescul, A., Pennock, D. M., & Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments (pp. 437–444). Morgan Kaufmann Publishers Inc.

qFiasco, F. (2018). Deep Thinking, Where Machine Intelligence Ends and Human Creativity Begins, Garry Kasparov, Mig Greengard. John Murray, London (2017), 262 pages with notes and an index, ISBN - 978-1-47365-350-4. *Artif. Intell.*, *260*, 36–41. http://doi.org/10.1016/j.artint.2018.04.001

Raimond, Y., Abdallah, S. A., Sandler, M. B., & Giasson, F. (2007). The Music Ontology. *Ismir*.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens (pp. 175–186). Presented at the the 1994 ACM conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/192844.192905

Ricci, F. (2012). Context-aware music recommender systems: workshop keynote abstract. *the 21st international conference companion* (pp. 865–866). New York, New York, USA: ACM. http://doi.org/10.1145/2187980.2188215

Robles, V., Larranaga, P., Menasalvas, E., Pérez, M., & Herves, V. (2003). Improvement of naive Bayes collaborative filtering using interval estimation (pp. 168–174).

Saiedian, H., Kumarakulasingam, P., & Anan, M. (2004). Scenario-based requirements analysis techniques for real-time software systems: a comparative evaluation. *Requirements Engineering*, *10*(1), 22–33. http://doi.org/10.1007/s00766-004-0192-6

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *the tenth international conference* (pp. 285–295). New York, New York, USA: ACM. http://doi.org/10.1145/371920.372071

Sasaki, S., Hirai, T., Ohya, H., & Morishima, S. (2013). Affective Music Recommendation System Reflecting the Mood of Input Image (pp. 153–154). Presented at the 2013 International Conference on Culture and Computing (Culture Computing), IEEE. http://doi.org/10.1109/CultureComputing.2013.42

Shani, G., Brafman, R. I., & Heckerman, D. (2002). An MDP-based recommender system (pp. 453–460). Morgan Kaufmann Publishers Inc.

Shardanand, U., & Maes, P. (1995). Social information filtering: algorithms for automating "word of mouth." *the SIGCHI conference* (pp. 210–217). New York, New York, USA: ACM Press/Addison-Wesley Publishing Co. http://doi.org/10.1145/223904.223931

Shirky, C. (2005). Ontology is overrated: categories, links, and tags. Clay Shirky's Writings about the Internet: Economics & Culture, Media & Community.

Slave to the algorithm? How music fans can reclaim their playlists from Spotify. (2016). Slave to the algorithm? How music fans can reclaim their playlists from Spotify. Retrieved from https://www.theguardian.com/books/2016/feb/19/slave-to-the-algorithm-how-music-fans-can-reclaim-their-playlists-from-spotify

Smyth, B. (2007). Case-Based Recommendation. *The Adaptive Web*, *4321*(Chapter 11), 342–376. http://doi.org/10.1007/978-3-540-72079-9_11

Song, S., Kim, M., Rho, S., & Hwang, E. (2009). Music Ontology for Mood and Situation Reasoning to Support Music Retrieval and Recommendation. *2009 Third International Conference on Digital Society (ICDS)* (pp. 304–309). IEEE. http://doi.org/10.1109/ICDS.2009.50

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. http://doi.org/10.1108/00220410410560573

Su, X., & Khoshgoftaar, T. M. (2006). Collaborative Filtering for Multi-class Data Using Belief Nets Algorithms. *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, 497–504. http://doi.org/10.1109/ICTAI.2006.41

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, *2009*(12), 4–19. http://doi.org/10.1155/2009/421425

Sutcliffe, A. G., RE'98, M. R. O. R. E., 1998. (n.d.). Experience with SCRAM, a scenario requirements analysis method. *Ieeexplore.Ieee.org*

.

Sutcliffe, A., Gault, B., & Maiden, N. (2004). ISRE: immersive scenario-based requirements engineering with virtual prototypes. *Requirements Engineering*, *10*(2), 95–111. http://doi.org/10.1007/s00766-004-0198-0

Torres, R., McNee, S. M., Abel, M., Konstan, J. A., & Riedl, J. (2004). Enhancing digital libraries with TechLens. *the 2004 joint ACM/IEEE conference* (pp. 228–236). New York, New York, USA: ACM. http://doi.org/10.1145/996350.996402

Ungar, L., & P Foster, D. (2000). Clustering Methods for Collaborative Filtering.

Wang, X., Rosenblum, D., & Wang, Y. (2012). Context-aware mobile music recommendation for daily activities. *ACM Multimedia*, 99–108. http://doi.org/10.1145/2393347.2393368

Yu-Shian Chiu, Kuei-Hong Lin, & Jia-Sin Chen. (2011). A Social Network-based serendipity recommender system (pp. 1–5). Presented at the 2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS 2011), IEEE. http://doi.org/10.1109/ISPACS.2011.6146073

Zhang, Y. C., Séaghdha, D. Ó., Quercia, D., & Jambor, T. (2012). Auralist: introducing serendipity into music recommendation. *Wsdm*, 13–22. http://doi.org/10.1145/2124295.2124300

Appendix

Study 1 - The Nature of Personalisation

Information Sheet

An Investigation Into Mood Based Music Categorisation

Participant Information Sheet

Importance

The emergence of music streaming services like Pandora and Spotify have enabled people to explore vast libraries of music far bigger than they would previously have had access to via traditional record stores. This can be seen as a great thing for expanding peoples musical horizons but it also presents a problem of content overload. 20 million tracks is too large a collection to be searched through manually. As a consequence people are increasingly making use of recommendation systems to navigate these libraries and find the tracks they want.

Unfortunately these systems are often found to be inaccurate and impersonal as they commonly fail to take account of an individuals preferences of taste (especially initially) when categorising their content and only recommend music based on broad properties like artist and genre.

Throughout my PhD I hope to further the field of recommendation systems by following a user led design approach and creating recommendation systems that emulate and learn from human practises. The aim of this study is to gain an understanding the human practice of categorise music with respect to mood.

Brief

The first part of this study involves creating two playlists to suit two different moods of your choosing. After having created the playlists you will then be asked to participate in a semi-structured interview to explain how you approached this task. The interview will take no longer than 30 minutes and your time will be compensated for with a £10 amazon voucher. Note taking and audio recording will be conducted by the researcher, Christopher Ellis during the interview.

The personal data collected will be stored in accordance with the Data Protection Act 1998, in a password protected folder and will have no names associated with it, just IDs that are mapped with the IDs supplied on the consent forms. It will be accessed and analysed by the researcher, Christopher Ellis and his supervisors' Steve Benford, Genovefa Kefalidou and Max Wilson. Any data used in reports will be fully anonymised following the ID masking procedure.

You may withdraw from the study at any time during or after the research without explanation by contacting the researcher or project supervisor at the addresses listed below and supplying the ID provided on the consent form. In this event all personal data gathered will be erased.

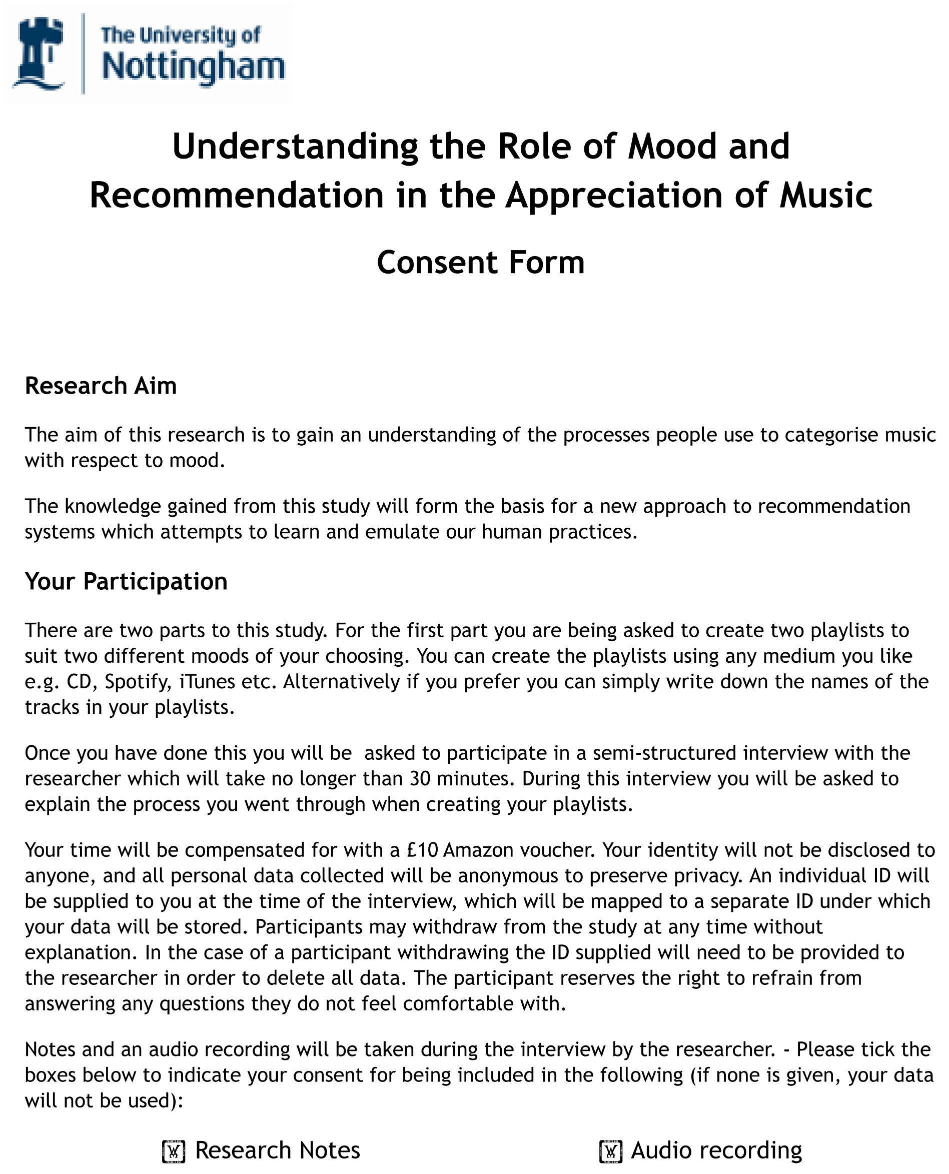
Researcher
Christopher Ellis
e-mail: christopher.ellis@nottingham.ac.uk

Supervisor
Max L. WIlson
e-mail: max.wilson@nottingham.ac.uk

Supervisor
Steve Benford
e-mail: steve.benford@nottingham.ac.uk

Supervisor
Genovefa Kefalidou
e-mail: genovefa.kefalidou@nottingham.ac.uk

Horizon DTC
School of Computer Science,
University of Nottingham,
Jubilee Campus,
Wollaton Road,
Nottingham,
NG8 1BB



The University of Nottingham

Understanding the Role of Mood and Recommendation in the Appreciation of Music

Consent Form

Research Aim

The aim of this research is to gain an understanding of the processes people use to categorise music with respect to mood.

The knowledge gained from this study will form the basis for a new approach to recommendation systems which attempts to learn and emulate our human practices.

Your Participation

There are two parts to this study. For the first part you are being asked to create two playlists to suit two different moods of your choosing. You can create the playlists using any medium you like e.g. CD, Spotify, iTunes etc. Alternatively if you prefer you can simply write down the names of the tracks in your playlists.

Once you have done this you will be asked to participate in a semi-structured interview with the researcher which will take no longer than 30 minutes. During this interview you will be asked to explain the process you went through when creating your playlists.

Your time will be compensated for with a £10 Amazon voucher. Your identity will not be disclosed to anyone, and all personal data collected will be anonymous to preserve privacy. An individual ID will be supplied to you at the time of the interview, which will be mapped to a separate ID under which your data will be stored. Participants may withdraw from the study at any time without explanation. In the case of a participant withdrawing the ID supplied will need to be provided to the researcher in order to delete all data. The participant reserves the right to refrain from answering any questions they do not feel comfortable with.

Notes and an audio recording will be taken during the interview by the researcher. - Please tick the boxes below to indicate your consent for being included in the following (if none is given, your data will not be used):

☒ Research Notes ☒ Audio recording

Data Collected

The data will be used by Christopher Ellis and his supervisors' Steve Benford, Genovefa Kefalidou and Max Wilson to assess how people categorise music with respect to mood. The data will be used for research reports, the PhD thesis and may be used for publications. All data collected will be stored in a password-protected folder on the researchers private server and within The University of Nottingham password and firewall-protected servers and kept in accordance with the Data Protection Act 1998. The data will be kept in its original condition for a minimum of seven years in order to comply with Nottingham University's Code of Research Conduct.

By signing below, you are agreeing that:

(1) You have read and understood the Participant Information Sheet supplied

(2) Questions about your participation in this study have been answered satisfactorily,

(3) You are taking part in this research study voluntarily

(4) You give permission for any anonymised data to be used including any quotes used in subsequent reports, papers or the PhD thesis. Only participants over the age of 18 are eligible to participate.

__________________________ __________________________

Participant's Name (Printed)* Participant's ID

__________________________ __________________________

Participant's signature* Date

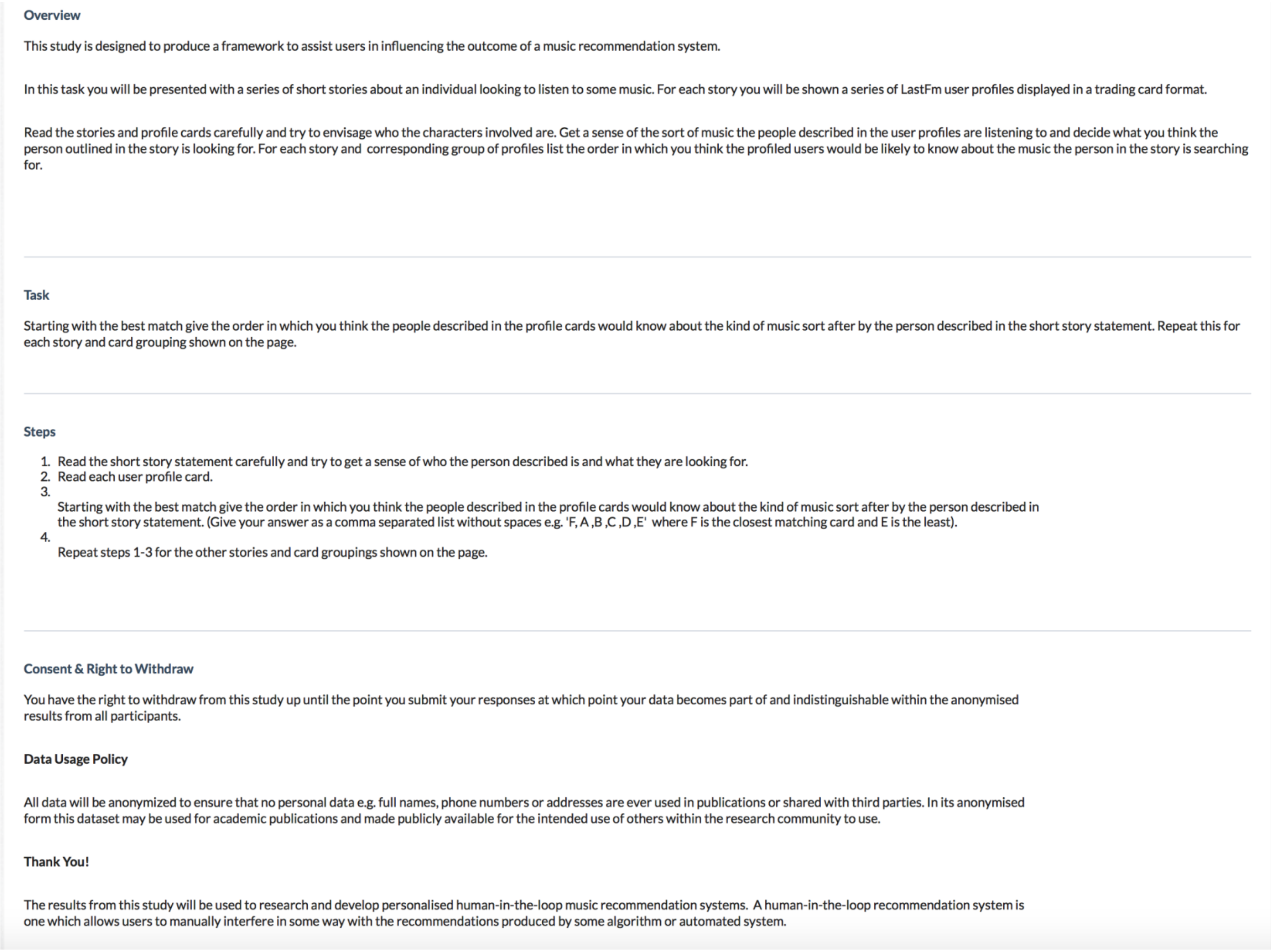
**Participants wishing to preserve some degree of anonymity may use their initials*

Key Interview Questions

1. Describe the process by which you approached the task of creating your playlists?
2. Describe how you felt whilst creating the playlists?
3. Did you create the playlist for just yourself or do you think it would transfer to other people?
    1. If positive answer: What aspects do you think are transferable and why?
4. Describe the environment in which you completed the task. Did you choose this environment for any particular reason?
5. Did you use the playlists for anything?
6. Do you often create playlists and if so why?
7. Was there anything you particularly liked or disliked about the task in this study?
8. Are there any other things you would like to meantion about the task or study in general?
9. Did you use anything to help you create the playlist?
10. Could that help have been improved in anyway?
11. Was order important when creating your playlist?
12. Did you create the playlist collaboratively? Is this something that might interest you?

Study 2 – The Role of Purpose

Information & Consent Form



Overview

This study is designed to produce a framework to assist users in influencing the outcome of a music recommendation system.

In this task you will be presented with a series of short stories about an individual looking to listen to some music. For each story you will be shown a series of LastFm user profiles displayed in a trading card format.

Read the stories and profile cards carefully and try to envisage who the characters involved are. Get a sense of the sort of music the people described in the user profiles are listening to and decide what you think the person outlined in the story is looking for. For each story and corresponding group of profiles list the order in which you think the profiled users would be likely to know about the music the person in the story is searching for.

Task

Starting with the best match give the order in which you think the people described in the profile cards would know about the kind of music sort after by the person described in the short story statement. Repeat this for each story and card grouping shown on the page.

Steps

1. Read the short story statement carefully and try to get a sense of who the person described is and what they are looking for.
2. Read each user profile card.
3. Starting with the best match give the order in which you think the people described in the profile cards would know about the kind of music sort after by the person described in the short story statement. (Give your answer as a comma separated list without spaces e.g. 'F, A ,B ,C ,D ,E' where F is the closest matching card and E is the least).
4. Repeat steps 1-3 for the other stories and card groupings shown on the page.

Consent & Right to Withdraw

You have the right to withdraw from this study up until the point you submit your responses at which point your data becomes part of and indistinguishable within the anonymised results from all participants.

Data Usage Policy

All data will be anonymized to ensure that no personal data e.g. full names, phone numbers or addresses are ever used in publications or shared with third parties. In its anonymised form this dataset may be used for academic publications and made publicly available for the intended use of others within the research community to use.

Thank You!

The results from this study will be used to research and develop personalised human-in-the-loop music recommendation systems. A human-in-the-loop recommendation system is one which allows users to manually interfere in some way with the recommendations produced by some algorithm or automated system.

LastFM Profile extraction script

```
#!/usr/bin/python
# -*- coding: utf-8 -*-


"""
Author: Christopher Ellis
Version: 1.0

This is a simple script to scrape 50 user profiles from the Last.fm website.

"""
from dateutil.parser import parse
from dateutil.rrule import rrule, DAILY
from os import getcwd
from time import sleep
import urllib2
import random
import json
import re

API_KEY = 'ecb1694f85eaee547c35a8c0ac6d0c4f'
PATH = getcwd()
FORMAT = 'json'

# friends in lastfm just means the users you are following

#NAME
#REALNAME
#IMAGES
#URL
#COUNTRY
#AGE
#GENDER
#SUBSCRIBER
#PLAYCOUNT
#PLAYLISTS
#TYPE

MIN_FRIENDS = 10
MIN_LOVEDTRACKS = 3
MIN_RECENTTRACKS = 2
MIN_TOPALBUMS = 2
MIN_TOPARTISTS = 2
MIN_TOPTRACKS = 3
MIN_WEEKLY_ALBUMCHART = 1
MIN_WEEKLY_ARTISTCHART = 1
MIN_WEEKLY_TRACKCHART = 1
MIN_WEEKLY_CHARTLIST = 1


PROFILE_COUNT = 50
USERNAMES = open('/Users/cellis/ownCloud/Horizon DTC/PhD Stuff/Studies/Three PhD
Studies/Factors Useful For Identifying Users Who Likely Listen To The Stuff You Want To Listen
To/NAMES.txt', 'r').read().splitlines() #open('/Users/cellis/ownCloud/Horizon DTC/PhD
Stuff/Studies/Three PhD Studies/Factors Useful For Identifying Users Who Likely Listen To The
Stuff You Want To Listen To/lastfmUserList.txt', 'r').read().splitlines()
PROFILE_ATTRIBUTES = {'Info': 'error', 'Friends': MIN_FRIENDS, 'LovedTracks':
MIN_LOVEDTRACKS, 'RecentTracks': MIN_RECENTTRACKS, 'TopAlbums':
MIN_TOPALBUMS, 'TopArtists': MIN_TOPARTISTS, 'TopTracks': MIN_TOPTRACKS,
'WeeklyAlbumChart': MIN_WEEKLY_ALBUMCHART, 'WeeklyArtistChart':
MIN_WEEKLY_ARTISTCHART, 'WeeklyChartList': MIN_WEEKLY_CHARTLIST,
'WeeklyTrackChart': MIN_WEEKLY_TRACKCHART}

print __doc__

#Search for profiles
count = 1
profiles = [None] * PROFILE_COUNT

while count <= PROFILE_COUNT:

    #Select a random username from the usernames worldlist
    user = filename = random.choice(USERNAMES)

    #Check if user exists
    url = 'http://ws.audioscrobbler.com/2.0/?method=user.getInfo&user='+ user
+'&limit=500&api_key='+ API_KEY + '&format=json'
    if 'error' in json.loads(urllib2.urlopen(url).read()):
#       sleep(0.5)
        continue

    profile = profiles[count] = {}
    attributesFound = 1

    #Get profile attributes for selected user
    for attribute in PROFILE_ATTRIBUTES:

        url = 'http://ws.audioscrobbler.com/2.0/?method=user.get'+attribute+'&user='+
user +'&limit=500&api_key='+ API_KEY + '&format=json'
        profile[attribute] = json.loads(urllib2.urlopen(url).read())
#       sleep(0.5)
        #Check if returned profile has the necessary attributes
        if profile[attribute] == {}:
            break
        elif((profile[attribute].has_key(attribute.lower()) and
profile[attribute][attribute.lower()]['@attr'].has_key('total')) and
(int(profile[attribute][attribute.lower()]['@attr']['total']) >= PROFILE_ATTRIBUTES[attribute])):
            attributesFound += 1
        elif((profile[attribute].has_key(attribute.lower()) and 'Chart' in attribute) and
(len(profile[attribute][attribute.lower()][re.sub( r"([A-Z])", r" \1", attribute).split()[1].lower()]) >=
PROFILE_ATTRIBUTES[attribute])):
            attributesFound += 1


    if attributesFound == len(PROFILE_ATTRIBUTES.keys()):
        print 'New user ' + user + ' found'
        count += 1

        #print profile
        #Create user profile file
        file = open(PATH + '/' + filename + '.' + FORMAT,'w')
```

```
		#Combine attributes
		#data = {}
		#for entry in profile:
		#	data.update(entry)

		file.write(json.dumps(profile, sort_keys=True, indent=4, separators=(',', ': ')))
		file.close()

#Generate summary stats for user profiles
```

Profile card generation script

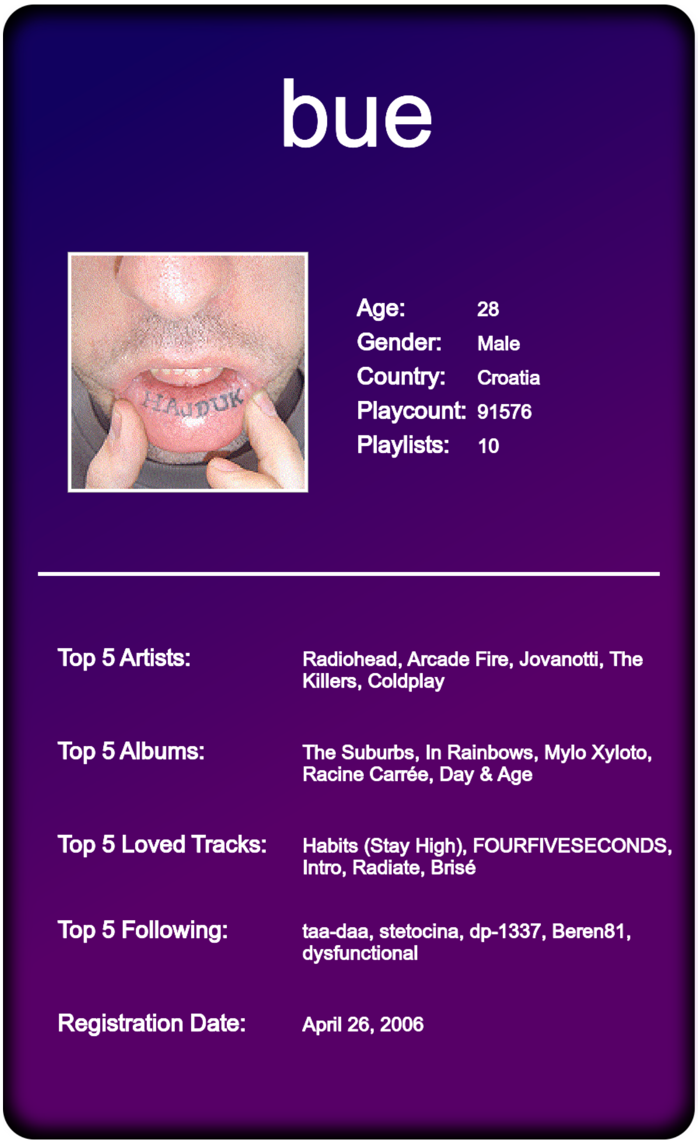
```
#!/usr/bin/python

import glob
import os
import json
import random
import textwrap
import datetime
from wand.image import Image, COMPOSITE_OPERATORS
from wand.drawing import Drawing
from wand.color import Color
from urllib2 import urlopen


PATH = os.getcwd()

PROFILE_ATTRIBUTE_KEYS = ["Info_user_name", "Info_user_gender", "Info_user_age",
"Info_user_country", "Info_user_image_#text3", "Info_user_url", "Info_user_registered_unixtime",
"Info_user_playlists", "Info_user_playcount", "Friends_friends_@attr_total",
"LovedTracks_lovedtracks_@attr_total", "TopArtists_topartists_@attr_total",
"TopAlbums_topalbums_@attr_total", "RecentTracks_recenttracks_@attr_total",
"Friends_friends_user_name0", "Friends_friends_user_url0",
"TopArtists_topartists_artist_name0", "RecentTracks_recenttracks_track_name0",
'RecentTracks_recenttracks_track_artist0_#text', "TopTracks_toptracks_track_name0",
"TopTracks_toptracks_track_artist0_name", "TopAlbums_topalbums_album_name0",
"TopAlbums_topalbums_album_artist0_name",
"LovedTracks_lovedtracks_track_name0","LovedTracks_lovedtracks_track_artist0_name",
"Friends_friends_user_name1", "Friends_friends_user_url1",
"TopArtists_topartists_artist_name1", "RecentTracks_recenttracks_track_name1",
'RecentTracks_recenttracks_track_artist1_#text', "TopTracks_toptracks_track_name1",
"TopTracks_toptracks_track_artist1_name", "TopAlbums_topalbums_album_name1",
"TopAlbums_topalbums_album_artist1_name",
"LovedTracks_lovedtracks_track_name1","LovedTracks_lovedtracks_track_artist1_name",
"Friends_friends_user_name2", "Friends_friends_user_url2",
"TopArtists_topartists_artist_name2", "RecentTracks_recenttracks_track_name2",
'RecentTracks_recenttracks_track_artist2_#text', "TopTracks_toptracks_track_name2",
"TopTracks_toptracks_track_artist2_name", "TopAlbums_topalbums_album_name2",
"TopAlbums_topalbums_album_artist2_name",
"LovedTracks_lovedtracks_track_name2","LovedTracks_lovedtracks_track_artist2_name",
"Friends_friends_user_name3", "Friends_friends_user_url3",
"TopArtists_topartists_artist_name3", "RecentTracks_recenttracks_track_name3",
'RecentTracks_recenttracks_track_artist3_#text', "TopTracks_toptracks_track_name3",
"TopTracks_toptracks_track_artist3_name", "TopAlbums_topalbums_album_name3",
"TopAlbums_topalbums_album_artist3_name",
"LovedTracks_lovedtracks_track_name3","LovedTracks_lovedtracks_track_artist3_name",
"Friends_friends_user_name4", "Friends_friends_user_url4",
"TopArtists_topartists_artist_name4", "RecentTracks_recenttracks_track_name4",
'RecentTracks_recenttracks_track_artist4_#text', "TopTracks_toptracks_track_name4",
"TopTracks_toptracks_track_artist4_name", "TopAlbums_topalbums_album_name4",
"TopAlbums_topalbums_album_artist4_name",
"LovedTracks_lovedtracks_track_name4","LovedTracks_lovedtracks_track_artist4_name"]
```

```
profiles = []

os.chdir(PATH + '/UserProfilesFlatterned/')
for file in glob.glob("*.json"):
   background = Image(filename='/Users/cellis/Desktop/Card Template.png')
   fileName = os.path.splitext(file)[0]
   with open(file) as json_file:
      json_data = json.load(json_file)
      attributeValues = [json_data[key].encode('utf-8') for key in PROFILE_ATTRIBUTE_KEYS if
key in json_data.keys()]
      #print attributeValues

      # Fromat attribute values
      attributeValues = ['N/A' if v is '' else v for v in attributeValues]

      if attributeValues[1] == 'f':
         attributeValues[1] = 'Female'
      elif attributeValues[1] == 'm':
         attributeValues[1] = 'Male'
      else:
         attributeValues[1] = 'N/A'

      try:
         attributeValues[2] = str(random.randint(18, 36)) # Age (assign random ficticious value
between 18-35)
         attributeValues[6] =
datetime.datetime.fromtimestamp(float(attributeValues[6])).strftime("%B %d, %Y") # reg date
         attributeValues[16] += ', ' + attributeValues[27] + ', ' + attributeValues[38] + ', ' +
attributeValues[49] + ', ' + attributeValues[60] # Top artists
         attributeValues[21] += ', ' + attributeValues[32] + ', ' + attributeValues[43] + ', ' +
attributeValues[54] + ', ' + attributeValues[65] # Top albums
         attributeValues[23] += ', ' + attributeValues[34] + ', ' + attributeValues[45] + ', ' +
attributeValues[56] + ', ' + attributeValues[67] # Top loved tracks
         attributeValues[14] += ', ' + attributeValues[25] + ', ' + attributeValues[36] + ', ' +
attributeValues[47] + ', ' + attributeValues[58] # Top following
         attributeValues[7] = str(random.randint(0, 21)) # Playlists (assign random ficticious value
between 0-20)
      except:
         continue

      if attributeValues[4] == 'N/A':
         continue

                  #Create card text
      with Drawing() as draw:
         #print dir(draw)
         #print dir(draw.font_style)
         #draw.font = 'wandtests/assets/AvenirNext.otf'
         draw.font_size = 120
         draw.fill_color = Color('white')
         draw.stroke_color = Color('white')
         draw.text_antialias = True
         x = int( (background.page_width - draw.get_font_metrics(background,
attributeValues[0]).text_width) / 2)
         draw.text(x, 182, attributeValues[0])
```

```
    #Card attributes
    draw.font_size = 30
    draw.text(455,400,'Age: ')
    draw.text(455,444,'Gender: ')
    draw.text(455,488,'Country: ')
    draw.text(455,532,'Playcount: ')
    draw.text(455,576,'Playlists: ')


    draw.text(70,850,'Top 5 Artists: ')
    draw.text(70,970,'Top 5 Albums: ')
    draw.text(70,1090,'Top 5 Loved Tracks: ')
    draw.text(70,1200,'Top 5 Following: ')
    draw.text(70,1320,'Registration Date:')
    draw(background)

    #Card values
    draw.font_size = 25
    draw.text(610,400,textwrap.fill(attributeValues[2], 39))
    draw.text(610,444,textwrap.fill(attributeValues[1], 39))
    draw.text(610,488,textwrap.fill(attributeValues[3], 39))
    draw.text(610,532,textwrap.fill(attributeValues[8], 39))
    draw.text(610,576,textwrap.fill(attributeValues[7], 39))

    draw.text(385,850,textwrap.fill(attributeValues[16], 39))
    draw.text(385,970,textwrap.fill(attributeValues[21], 39))
    draw.text(385,1090,textwrap.fill(attributeValues[23], 39))
    draw.text(385,1200,textwrap.fill(attributeValues[14], 39))
    draw.text(385,1320,textwrap.fill(attributeValues[6], 39))
    #draw.stroke_width = 5
    #draw.line((47, int(background.page_height/2)), (842, int(background.page_height/2)))
    draw(background)

#Add profile picture to card
try:
    response = urlopen(attributeValues[4])
    with Image(file=response) as foreground:
        for o in COMPOSITE_OPERATORS:
            bkground=background.clone()
            frground=foreground.clone()
            with Drawing() as draw:
                draw.composite(operator=o, left=88, top=323,
                            width=frground.width, height=frground.height, image=frground)
                draw(bkground)
                bkground.save(filename= fileName +'Card.png')
finally:
    response.close()
```

Profile cards

bue

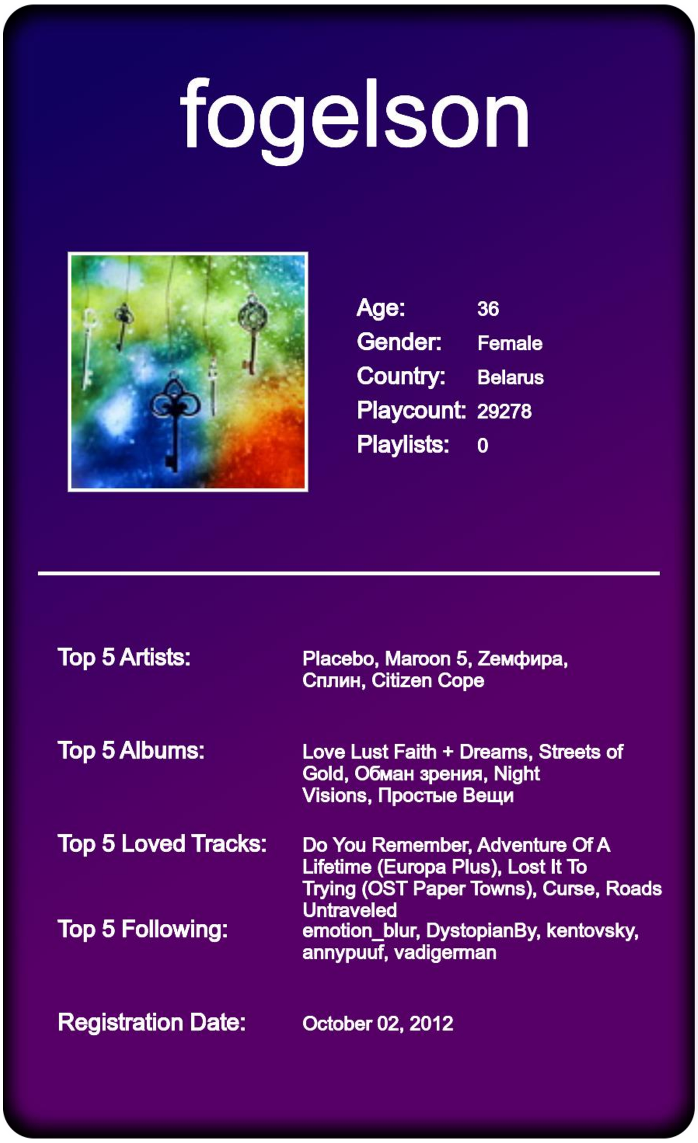


bue

Age:	28
Gender:	Male
Country:	Croatia
Playcount:	91576
Playlists:	10

Top 5 Artists:	Radiohead, Arcade Fire, Jovanotti, The Killers, Coldplay
Top 5 Albums:	The Suburbs, In Rainbows, Mylo Xyloto, Racine Carrée, Day & Age
Top 5 Loved Tracks:	Habits (Stay High), FOURFIVESECONDS, Intro, Radiate, Brisé
Top 5 Following:	taa-daa, stetocina, dp-1337, Beren81, dysfunctional
Registration Date:	April 26, 2006

fogelson

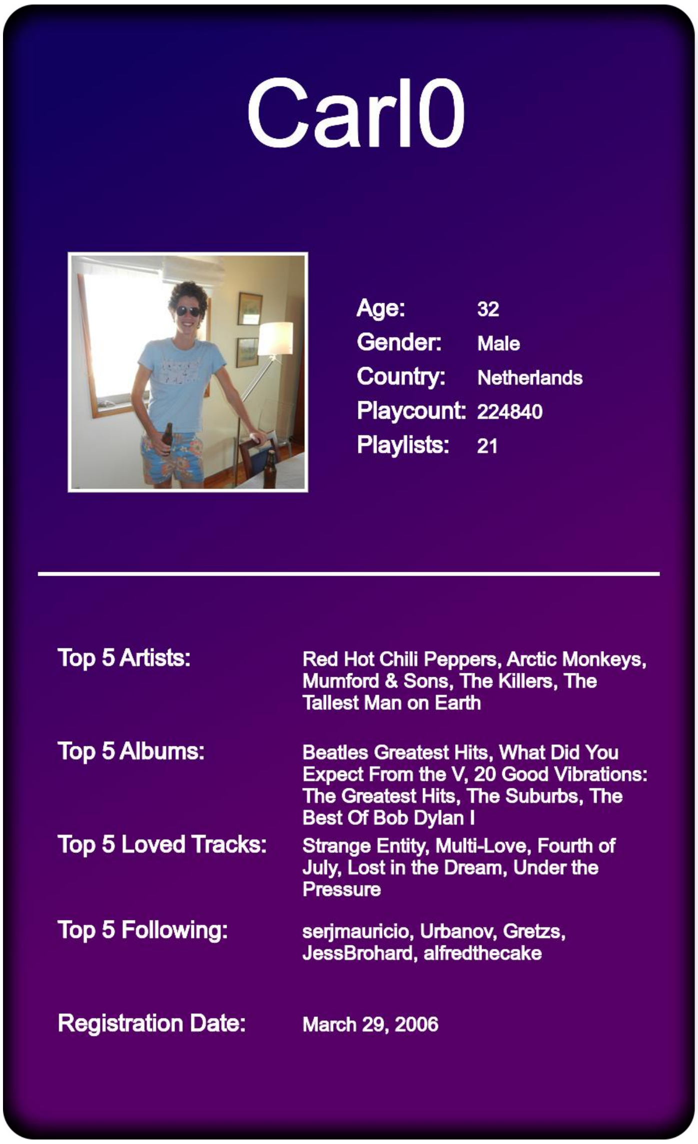


fogelson

Age:	36
Gender:	Female
Country:	Belarus
Playcount:	29278
Playlists:	0

Top 5 Artists:	Placebo, Maroon 5, Земфира, Сплин, Citizen Cope
Top 5 Albums:	Love Lust Faith + Dreams, Streets of Gold, Обман зрения, Night Visions, Простые Вещи
Top 5 Loved Tracks:	Do You Remember, Adventure Of A Lifetime (Europa Plus), Lost It To Trying (OST Paper Towns), Curse, Roads Untraveled
Top 5 Following:	emotion_blur, DystopianBy, kentovsky, annypuuf, vadigerman
Registration Date:	October 02, 2012

Carl0

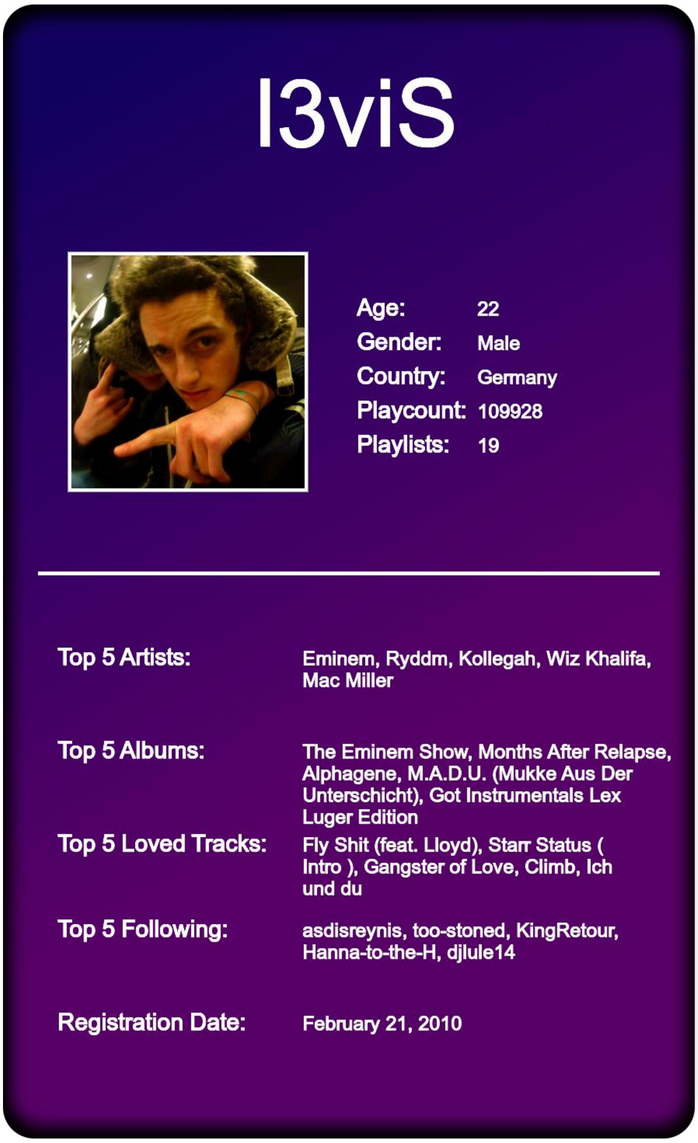


Carl0

Age:	32
Gender:	Male
Country:	Netherlands
Playcount:	224840
Playlists:	21

Top 5 Artists:	Red Hot Chili Peppers, Arctic Monkeys, Mumford & Sons, The Killers, The Tallest Man on Earth
Top 5 Albums:	Beatles Greatest Hits, What Did You Expect From the V, 20 Good Vibrations: The Greatest Hits, The Suburbs, The Best Of Bob Dylan I
Top 5 Loved Tracks:	Strange Entity, Multi-Love, Fourth of July, Lost in the Dream, Under the Pressure
Top 5 Following:	serjmauricio, Urbanov, Gretzs, JessBrohard, alfredthecake
Registration Date:	March 29, 2006

l3viS

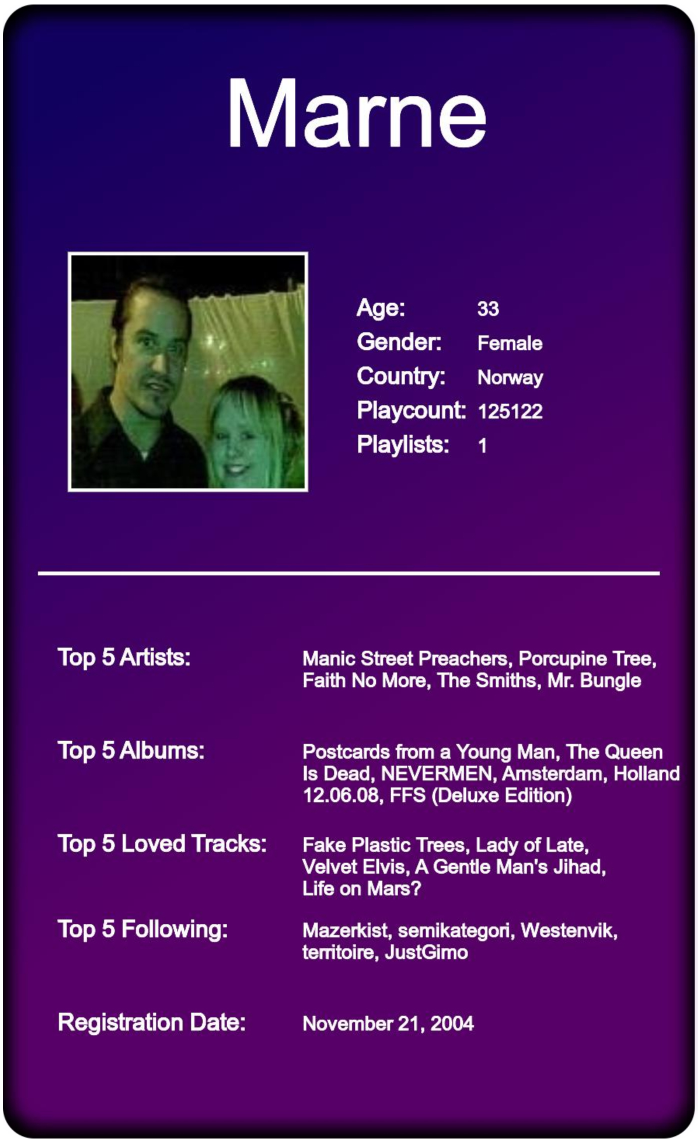


l3viS

Age:	22
Gender:	Male
Country:	Germany
Playcount:	109928
Playlists:	19

Top 5 Artists:	Eminem, Ryddm, Kollegah, Wiz Khalifa, Mac Miller
Top 5 Albums:	The Eminem Show, Months After Relapse, Alphagene, M.A.D.U. (Mukke Aus Der Unterschicht), Got Instrumentals Lex Luger Edition
Top 5 Loved Tracks:	Fly Shit (feat. Lloyd), Starr Status (Intro), Gangster of Love, Climb, Ich und du
Top 5 Following:	asdisreynis, too-stoned, KingRetour, Hanna-to-the-H, djlule14
Registration Date:	February 21, 2010

Marne

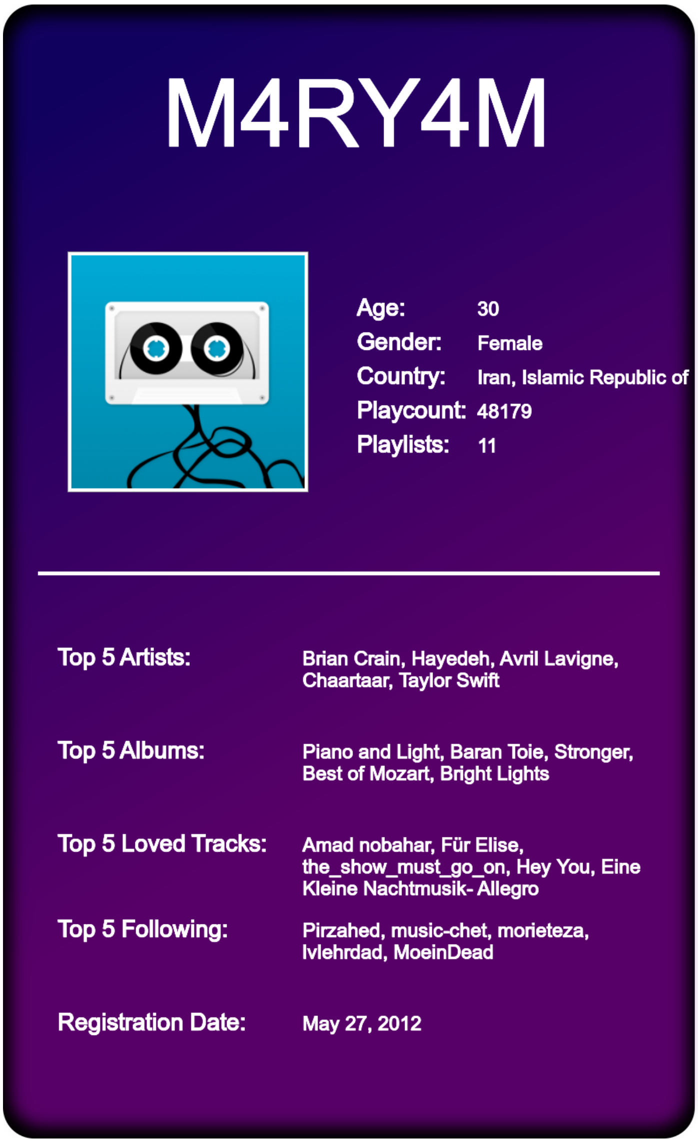


Marne

Age:	33
Gender:	Female
Country:	Norway
Playcount:	125122
Playlists:	1

Top 5 Artists:	Manic Street Preachers, Porcupine Tree, Faith No More, The Smiths, Mr. Bungle
Top 5 Albums:	Postcards from a Young Man, The Queen Is Dead, NEVERMEN, Amsterdam, Holland 12.06.08, FFS (Deluxe Edition)
Top 5 Loved Tracks:	Fake Plastic Trees, Lady of Late, Velvet Elvis, A Gentle Man's Jihad, Life on Mars?
Top 5 Following:	Mazerkist, semikategori, Westenvik, territoire, JustGimo
Registration Date:	November 21, 2004

M4RY4M



M4RY4M

Age:	30
Gender:	Female
Country:	Iran, Islamic Republic of
Playcount:	48179
Playlists:	11

Top 5 Artists:	Brian Crain, Hayedeh, Avril Lavigne, Chaartaar, Taylor Swift
Top 5 Albums:	Piano and Light, Baran Toie, Stronger, Best of Mozart, Bright Lights
Top 5 Loved Tracks:	Amad nobahar, Für Elise, the_show_must_go_on, Hey You, Eine Kleine Nachtmusik- Allegro
Top 5 Following:	Pirzahed, music-chet, morieteza, Ivlehrdad, MoeinDead
Registration Date:	May 27, 2012

Study 3 – UX Human-in-the-loop Validation

Information & Consent Form

The University of Nottingham

A User Evaluation of a Human-in-the-loop music recommender

Information & Consent Form

Research Aim

The aim of this research is to gain some basic feedback and usability comments with regards the a low-level prototype HITL music recommender system which represents the culmination of the findings of my PhD thesis.

The knowledge gained from this study will round out my thesis and provide a final reflection on the application of HITL techniques to the personalisation problem.

Your Participation

There are two parts to this study. For the first part you will be present with a series of user interface screenshots representing the various screens of a human-in-the-loop (HITL) music recommendation system. You will be introduced to the concept of HITL systems and shown how the recommender is intended to be used.

Once you have understood how the system works you will be asked to go through the process of using it in a mock fashion by sliding the screenshots around to mimic navigating the application. Ultimately by engaging in this process you will identify a set of user profiles (shown in the system) which you would like to be used to generate a recommendation for you. You will then be provided with two recommendation one produced using all of the profiles and the other using the profiles you selected using the HITL system.

In the second part of the study you will be asked to complete a short 10-15 minute post study interview during which brief written anonymous notes will be taken. This interview will ask you to describe anything you particularly liked or disliked about the way the system worked and whether or not you have any comments regarding the user interface design and usability. Finally you will be asked what you thought of their recommendations, whether you preferred one over the other and whether you could tell which one was using the profiles you selected.

Your identity will not be disclosed to anyone, and all personal data collected will be anonymous to preserve privacy. An individual ID will be supplied to you at the time of the interview, which will be mapped to a separate ID under which your data will be stored. You may withdraw from the study at any time without explanation.

In the case of a participant withdrawing the ID supplied will need to be provided to the researcher in order to delete all data. The participant reserves the right to refrain from answering any questions they do not feel comfortable with.

Notes will be taken during the interview by the researcher. - Please tick the box below to indicate your consent for being included in the following (if none is given, your data will not be used):

⍰ Research Notes

Data Collected

The data will be used by Christopher Ellis and his supervisors' Steve Benford, Genovefa Kefalidou and Max Wilson to gain some insight into the usability of the system presented. The data will be used for research reports, the PhD thesis and may be used for publications. All data collected will be stored on a secure university shared point server and kept in accordance with the Data Protection Act 1998. The data will be kept in its original condition for a minimum of seven years in order to comply with Nottingham University's Code of Research Conduct.

By signing below, you are agreeing that:

(1) You have read and understood the Participant Information Sheet supplied

(2) Questions about your participation in this study have been answered satisfactorily,

(3) You are taking part in this research study voluntarily

(4) You give permission for any anonymised data to be used including any quotes used in subsequent reports, papers or the PhD thesis. Only participants over the age of 18 are eligible to participate.

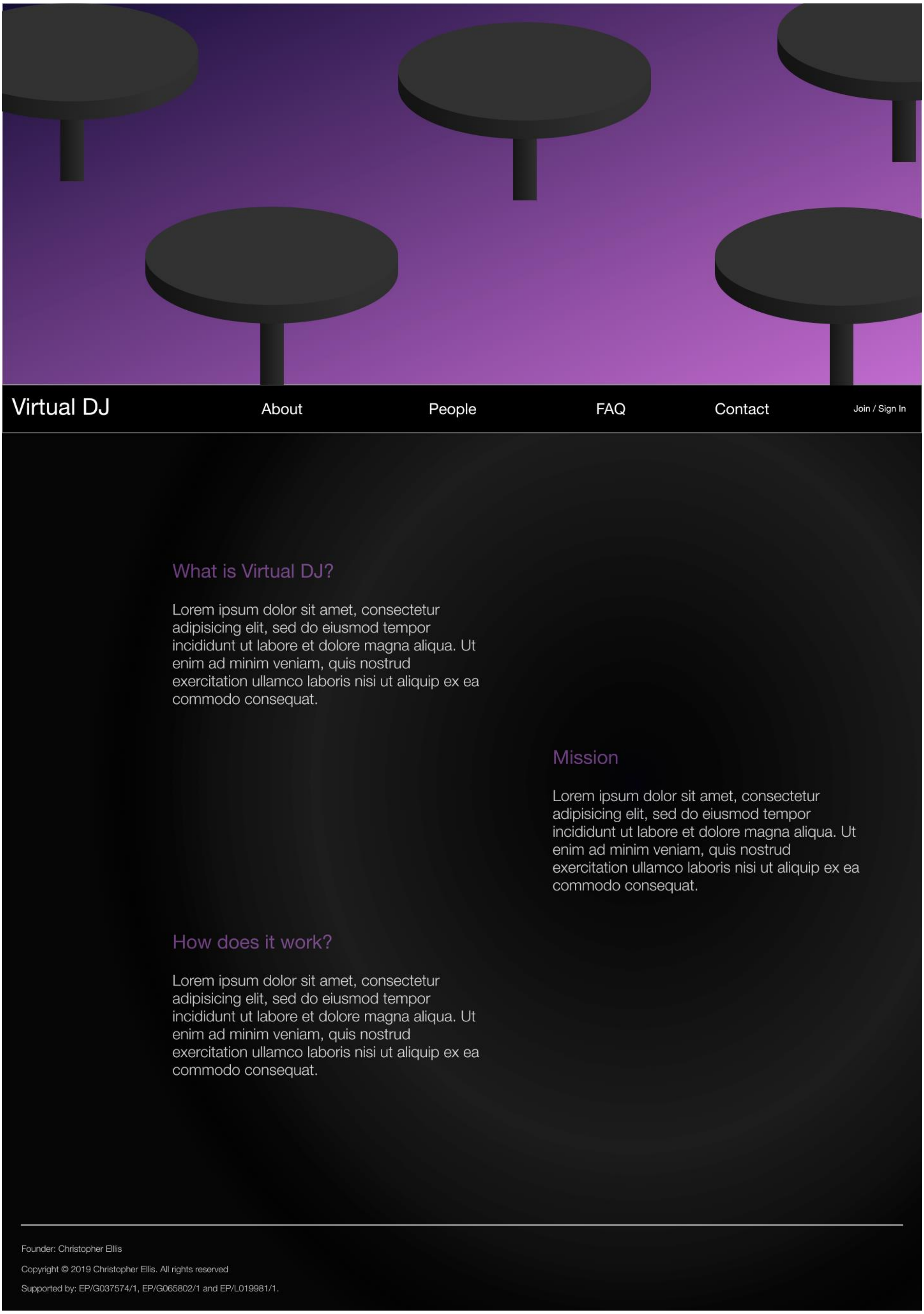
______________________________	______________________________
Participant's Name (Printed)*	Participant's ID
______________________________	______________________________
Participant's signature*	Date

**Participants wishing to preserve some degree of anonymity may use their initials*

Key Interview Questions

0. Best & worst feature of system?
1. anything particular appealed any other comments?
2. Where there any aspects of the system you particularly liked or disliked?
3. Would you consider using a music recommendation system like this?
4. Is there anything that struck you as akward, lacking or missing in the design of the system?
5. Conceptually is there anything you couldn't so that you would have liked to?

VirtualDJ Screens



What is Virtual DJ?

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Mission

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

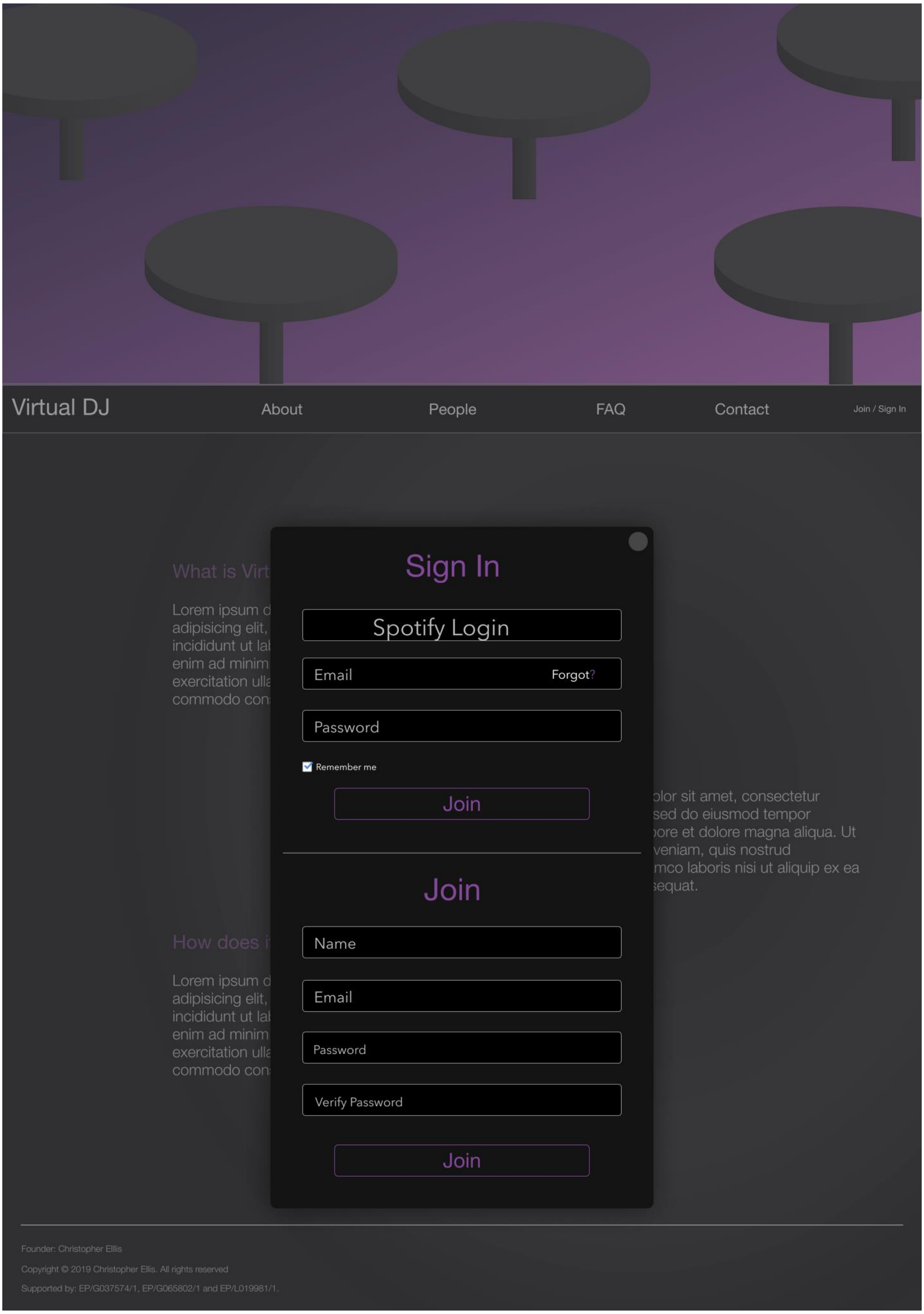
How does it work?

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Founder: Christopher Elllis

Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.



What is Virt

Lorem ipsum d adipisicing elit, incididunt ut lal enim ad minim exercitation ulla commodo con

olor sit amet, consectetur sed do eiusmod tempor oore et dolore magna aliqua. Ut veniam, quis nostrud mco laboris nisi ut aliquip ex ea sequat.

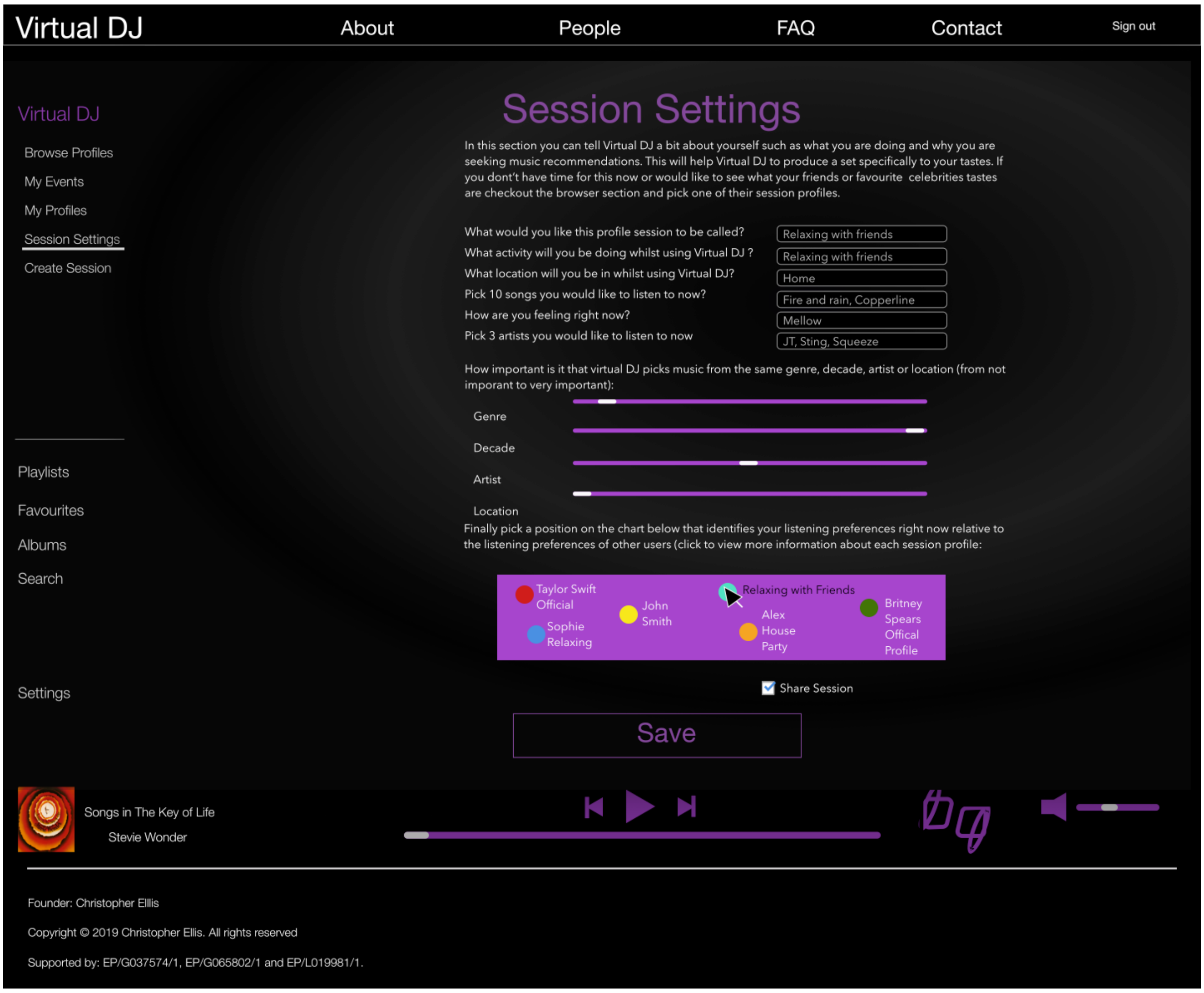
How does i

Lorem ipsum d adipisicing elit, incididunt ut lal enim ad minim exercitation ulla commodo con

Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.



Session Settings

In this section you can tell Virtual DJ a bit about yourself such as what you are doing and why you are seeking music recommendations. This will help Virtual DJ to produce a set specifically to your tastes. If you dont't have time for this now or would like to see what your friends or favourite celebrities tastes are checkout the browser section and pick one of their session profiles.

What would you like this profile session to be called?	Relaxing with friends
What activity will you be doing whilst using Virtual DJ ?	Relaxing with friends
What location will you be in whilst using Virtual DJ?	Home
Pick 10 songs you would like to listen to now?	Fire and rain, Copperline
How are you feeling right now?	Mellow
Pick 3 artists you would like to listen to now	JT, Sting, Squeeze

How important is it that virtual DJ picks music from the same genre, decade, artist or location (from not imporant to very important):

Genre

Decade

Artist

Location

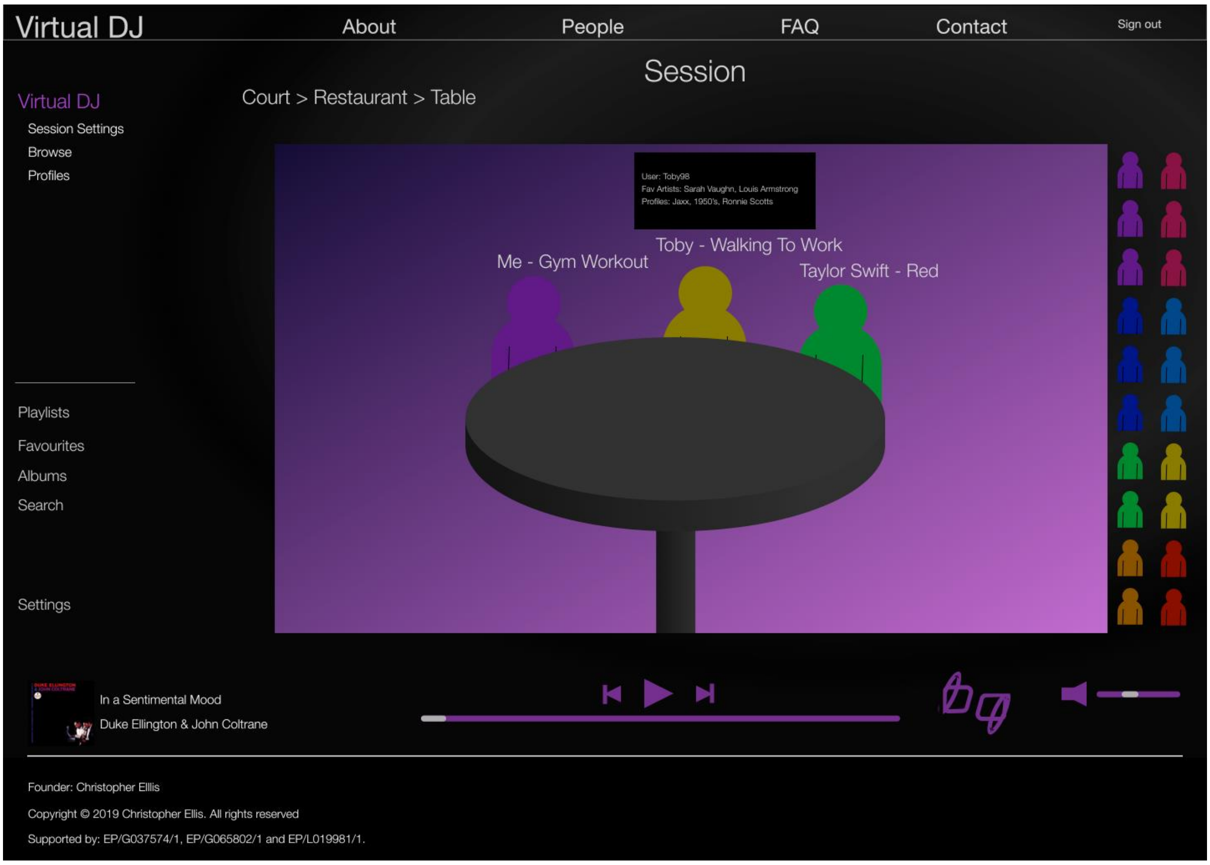
Finally pick a position on the chart below that identifies your listening preferences right now relative to the listening preferences of other users (click to view more information about each session profile:

☑ Share Session

Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.



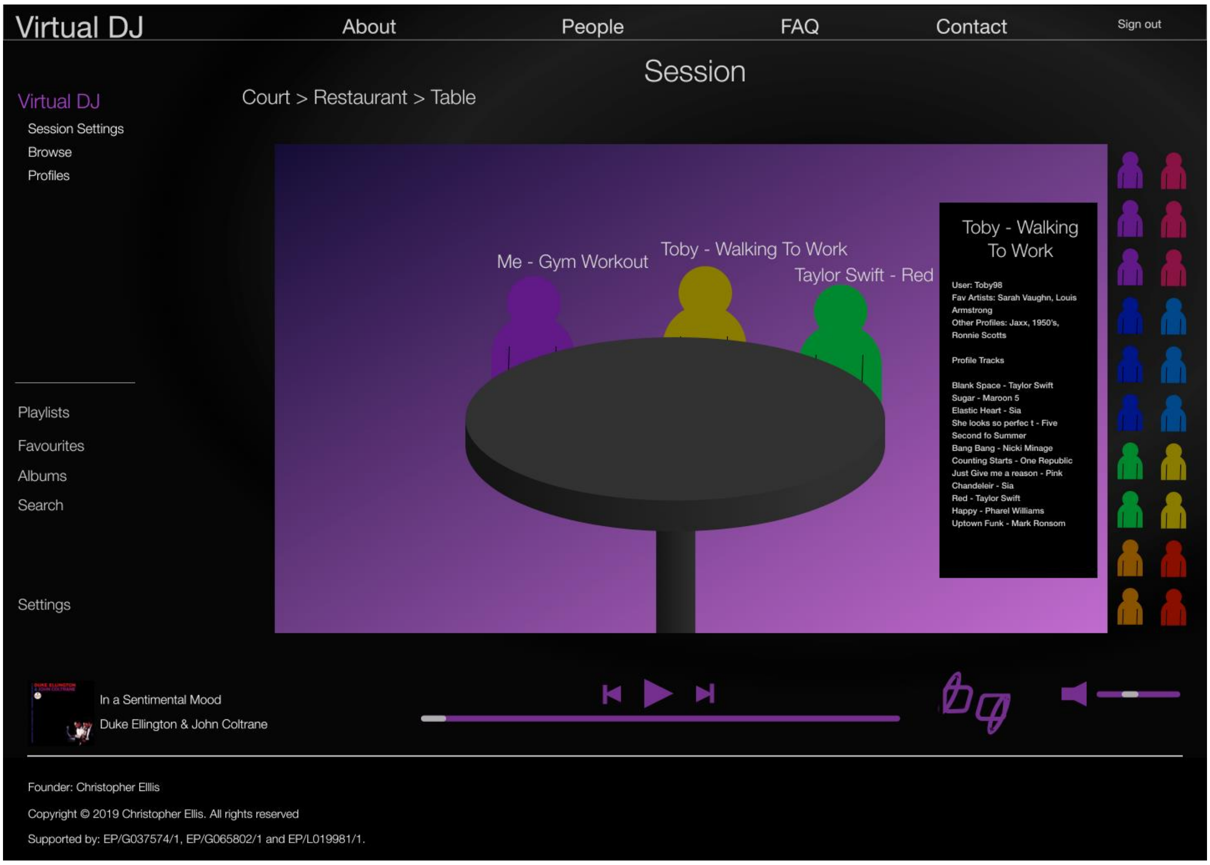
User: Toby98
Fav Artists: Sarah Vaughn, Louis Armstrong
Profiles: Jaxx, 1950's, Ronnie Scotts

In a Sentimental Mood
Duke Ellington & John Coltrane

Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.



Session

Court > Restaurant > Table

Toby - Walking To Work

User: Toby98
Fav Artists: Sarah Vaughn, Louis Armstrong
Other Profiles: Jaxx, 1950's, Ronnie Scotts

Profile Tracks

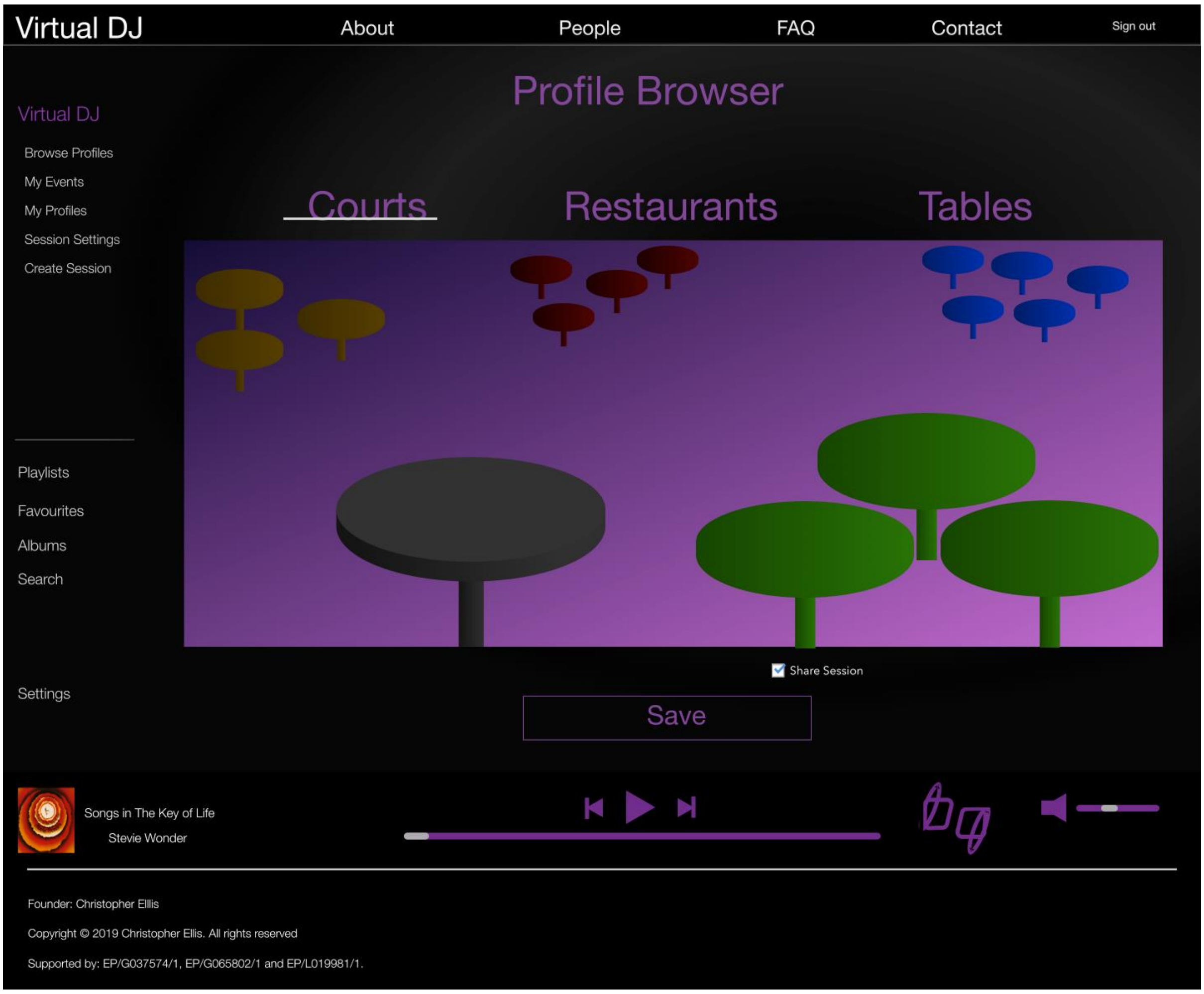
Blank Space - Taylor Swift
Sugar - Maroon 5
Elastic Heart - Sia
She looks so perfec t - Five
Second fo Summer
Bang Bang - Nicki Minage
Counting Starts - One Republic
Just Give me a reason - Pink
Chandeleir - Sia
Red - Taylor Swift
Happy - Pharel Williams
Uptown Funk - Mark Ronsom

In a Sentimental Mood
Duke Ellington & John Coltrane

Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

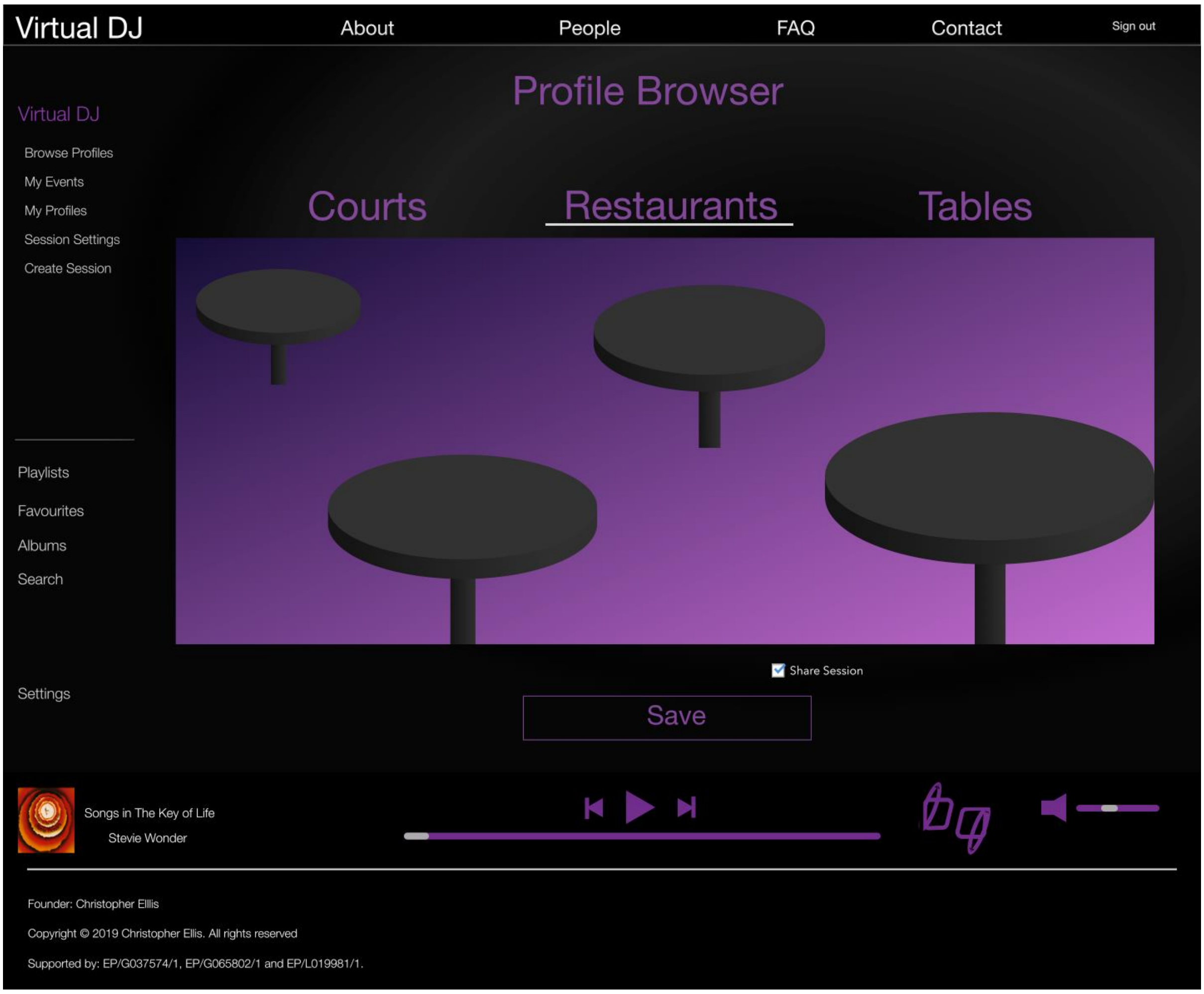
Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.



Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

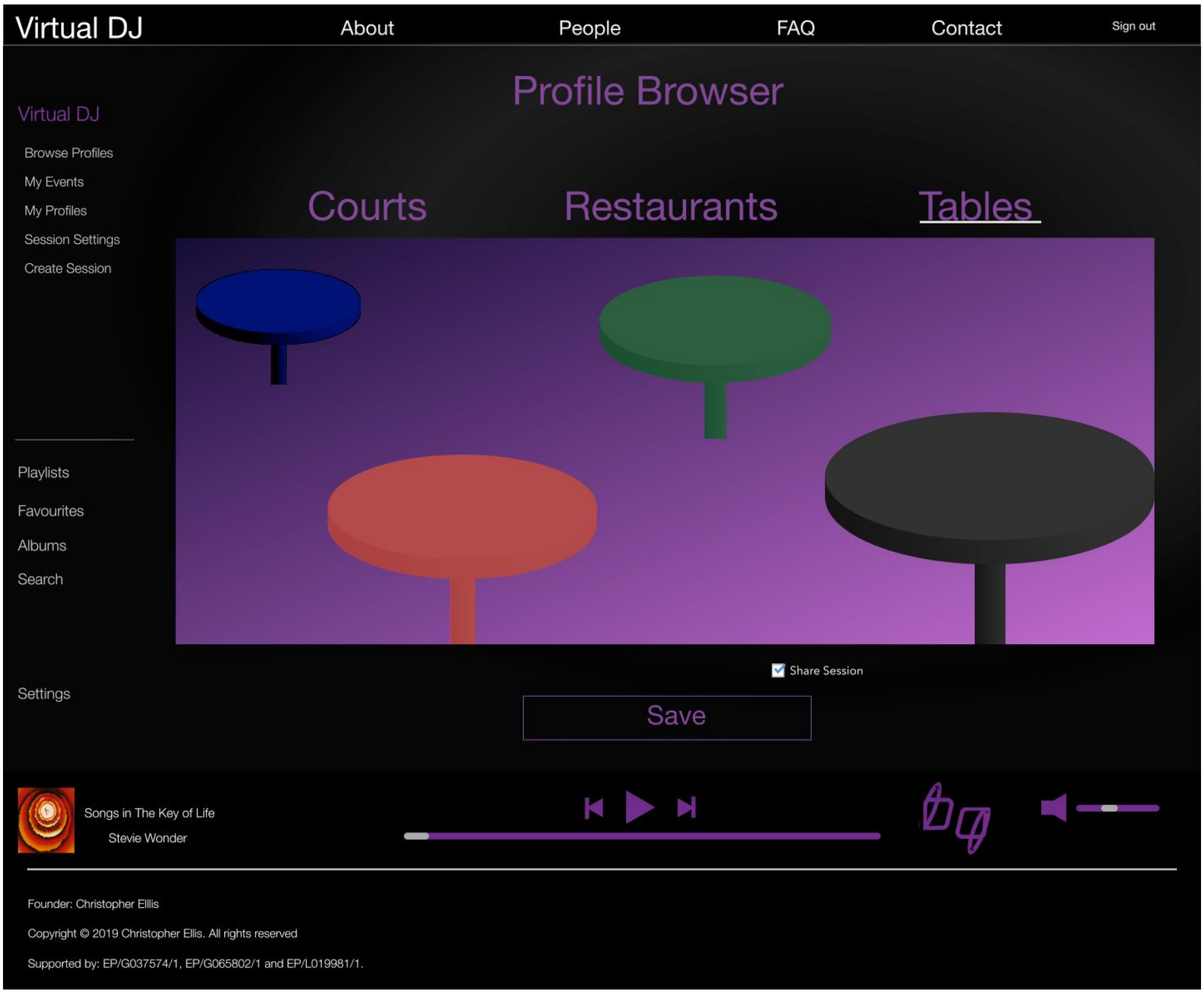
Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.



Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

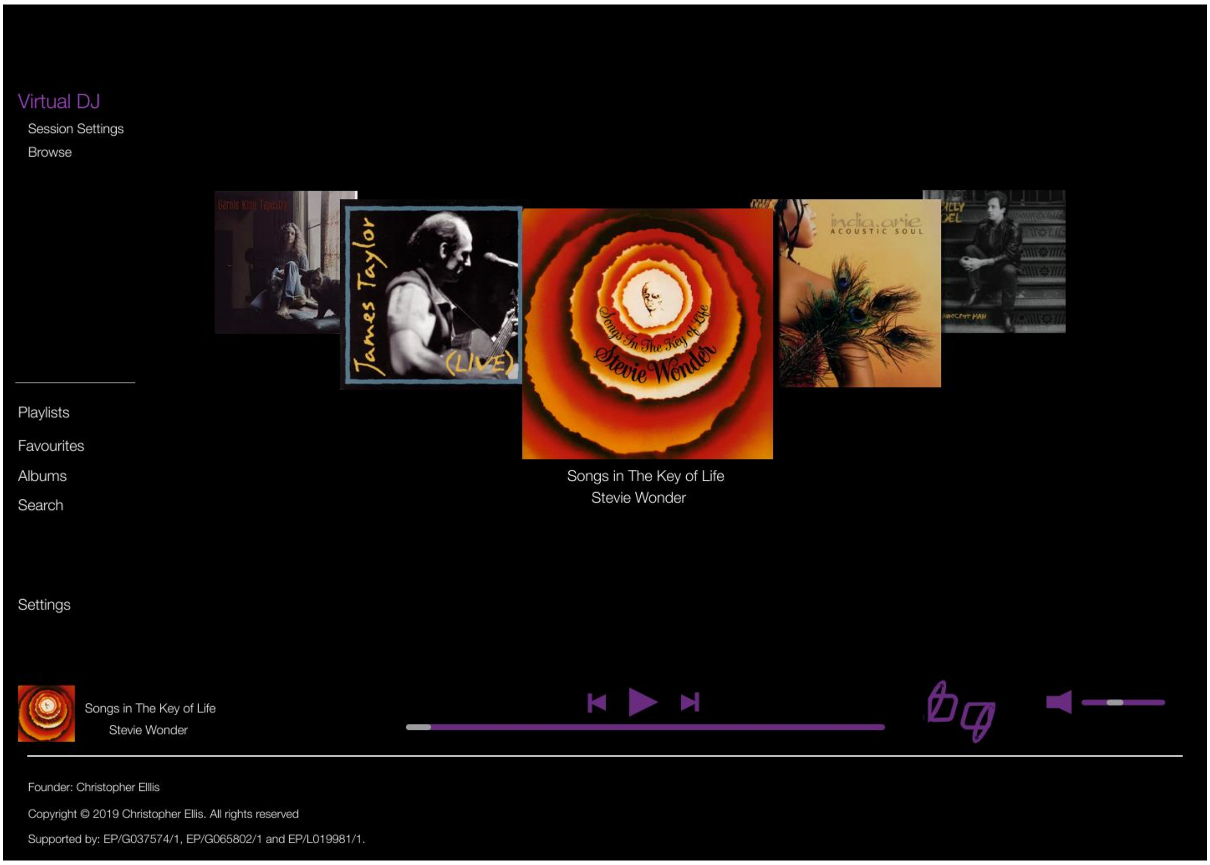
Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.



Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

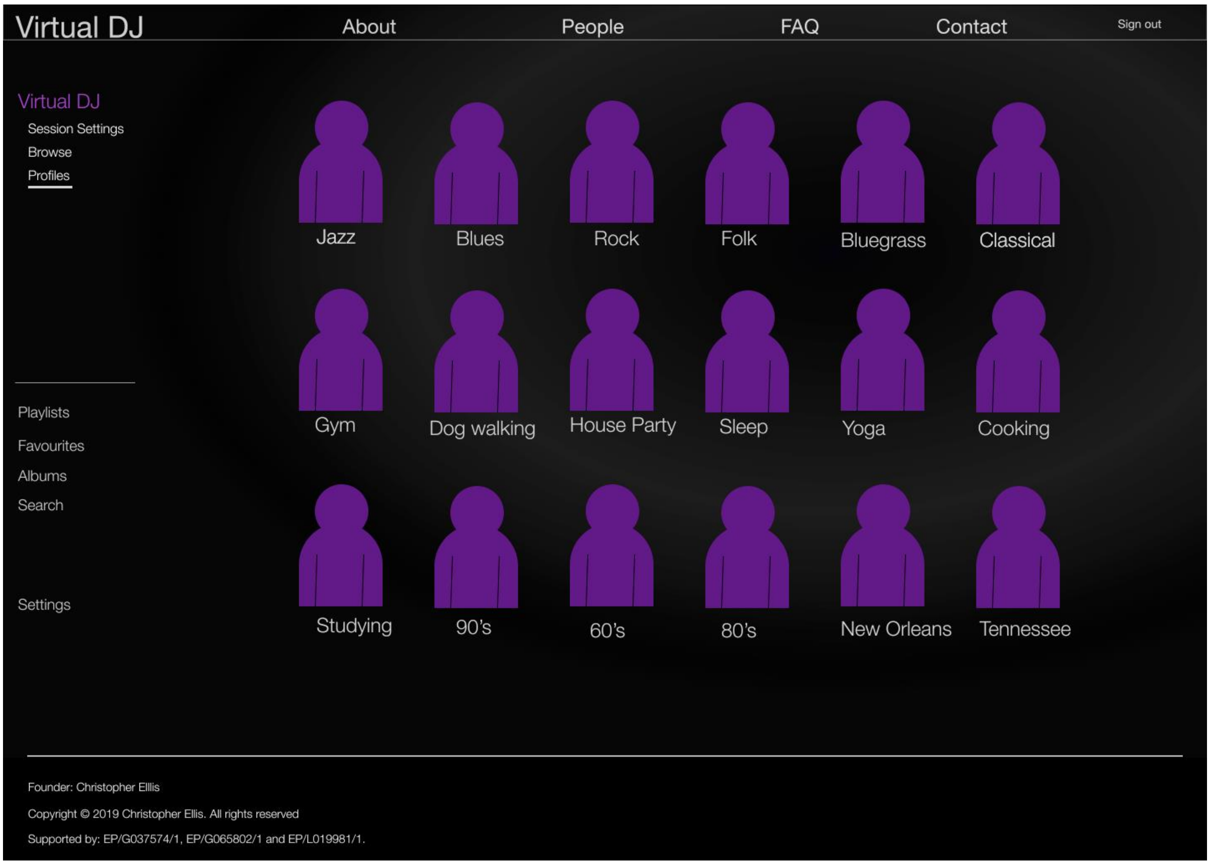
Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.



Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

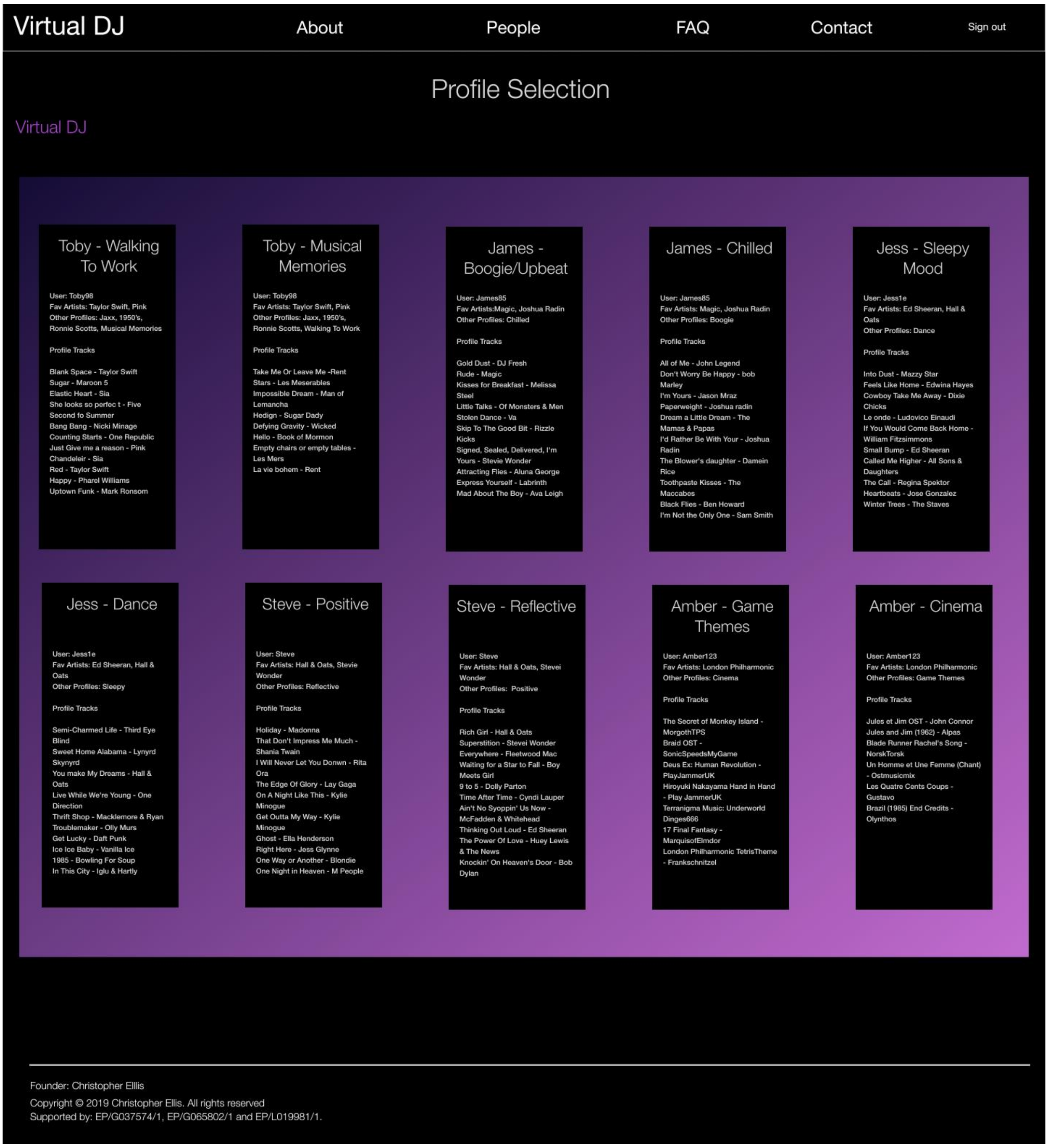
Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.



Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved

Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.



Profile Selection

Virtual DJ

Toby - Walking To Work

User: Toby98
Fav Artists: Taylor Swift, Pink
Other Profiles: Jaxx, 1950's, Ronnie Scotts, Musical Memories

Profile Tracks

Blank Space - Taylor Swift
Sugar - Maroon 5
Elastic Heart - Sia
She looks so perfec t - Five
Second fo Summer
Bang Bang - Nicki Minage
Counting Starts - One Republic
Just Give me a reason - Pink
Chandeleir - Sia
Red - Taylor Swift
Happy - Pharel Williams
Uptown Funk - Mark Ronsom

Toby - Musical Memories

User: Toby98
Fav Artists: Taylor Swift, Pink
Other Profiles: Jaxx, 1950's, Ronnie Scotts, Walking To Work

Profile Tracks

Take Me Or Leave Me -Rent
Stars - Les Meserables
Impossible Dream - Man of Lemancha
Hedign - Sugar Dady
Defying Gravity - Wicked
Hello - Book of Mormon
Empty chairs or empty tables - Les Mers
La vie bohem - Rent

James - Boogie/Upbeat

User: James85
Fav Artists:Magic, Joshua Radin
Other Profiles: Chilled

Profile Tracks

Gold Dust - DJ Fresh
Rude - Magic
Kisses for Breakfast - Melissa Steel
Little Talks - Of Monsters & Men
Stolen Dance - Va
Skip To The Good Bit - Rizzle Kicks
Signed, Sealed, Delivered, I'm Yours - Stevie Wonder
Attracting Flies - Aluna George
Express Yourself - Labrinth
Mad About The Boy - Ava Leigh

James - Chilled

User: James85
Fav Artists: Magic, Joshua Radin
Other Profiles: Boogie

Profile Tracks

All of Me - John Legend
Don't Worry Be Happy - bob Marley
I'm Yours - Jason Mraz
Paperweight - Joshua radin
Dream a Little Dream - The Mamas & Papas
I'd Rather Be With Your - Joshua Radin
The Blower's daughter - Damein Rice
Toothpaste Kisses - The Maccabes
Black Flies - Ben Howard
I'm Not the Only One - Sam Smith

Jess - Sleepy Mood

User: Jess1e
Fav Artists: Ed Sheeran, Hall & Oats
Other Profiles: Dance

Profile Tracks

Into Dust - Mazzy Star
Feels Like Home - Edwina Hayes
Cowboy Take Me Away - Dixie Chicks
Le onde - Ludovico Einaudi
If You Would Come Back Home - William Fitzsimmons
Small Bump - Ed Sheeran
Called Me Higher - All Sons & Daughters
The Call - Regina Spektor
Heartbeats - Jose Gonzalez
Winter Trees - The Staves

Jess - Dance

User: Jess1e
Fav Artists: Ed Sheeran, Hall & Oats
Other Profiles: Sleepy

Profile Tracks

Semi-Charmed Life - Third Eye Blind
Sweet Home Alabama - Lynyrd Skynyrd
You make My Dreams - Hall & Oats
Live While We're Young - One Direction
Thrift Shop - Macklemore & Ryan
Troublemaker - Olly Murs
Get Lucky - Daft Punk
Ice Ice Baby - Vanilla Ice
1985 - Bowling For Soup
In This City - Iglu & Hartly

Steve - Positive

User: Steve
Fav Artists: Hall & Oats, Stevie Wonder
Other Profiles: Reflective

Profile Tracks

Holiday - Madonna
That Don't Impress Me Much - Shania Twain
I Will Never Let You Donwn - Rita Ora
The Edge Of Glory - Lay Gaga
On A Night Like This - Kylie Minogue
Get Outta My Way - Kylie Minogue
Ghost - Ella Henderson
Right Here - Jess Glynne
One Way or Another - Blondie
One Night in Heaven - M People

Steve - Reflective

User: Steve
Fav Artists: Hall & Oats, Stevei Wonder
Other Profiles: Positive

Profile Tracks

Rich Girl - Hall & Oats
Superstition - Stevei Wonder
Everywhere - Fleetwood Mac
Waiting for a Star to Fall - Boy Meets Girl
9 to 5 - Dolly Parton
Time After Time - Cyndi Lauper
Ain't No Syoppin' Us Now - McFadden & Whitehead
Thinking Out Loud - Ed Sheeran
The Power Of Love - Huey Lewis & The News
Knockin' On Heaven's Door - Bob Dylan

Amber - Game Themes

User: Amber123
Fav Artists: London Philharmonic
Other Profiles: Cinema

Profile Tracks

The Secret of Monkey Island - MorgothTPS
Braid OST - SonicSpeedsMyGame
Deus Ex: Human Revolution - PlayJammerUK
Hiroyuki Nakayama Hand in Hand - Play JammerUK
Terranigma Music: Underworld Dinges666
17 Final Fantasy - MarquisofElmdor
London Philharmonic TetrisTheme - Frankschnitzel

Amber - Cinema

User: Amber123
Fav Artists: London Philharmonic
Other Profiles: Game Themes

Profile Tracks

Jules et Jim OST - John Connor
Jules and Jim (1962) - Alpas
Blade Runner Rachel's Song - NorskTorsk
Un Homme et Une Femme (Chant) - Ostmusicmix
Les Quatre Cents Coups - Gustavo
Brazil (1985) End Credits - Olynthos

Founder: Christopher Ellis

Copyright © 2019 Christopher Ellis. All rights reserved
Supported by: EP/G037574/1, EP/G065802/1 and EP/L019981/1.