Exploiting Food Choice Biases for Healthier Recipe Recommendation

David Elsweiler*
University of Regensburg
Germany
david@elsweiler.co.uk

Christoph Trattner*
MODUL University Vienna
Austria
christoph.trattner@modul.ac.at

Morgan Harvey
Northumbria University
The United Kingdom
morgan.harvey@northumbria.ac.uk

ABSTRACT

By incorporating healthiness into the food recommendation / ranking process we have the potential to improve the eating habits of a growing number of people who use the Internet as a source of food inspiration. In this paper, using insights gained from various data sources, we explore the feasibility of substituting meals that would typically be recommended to users with similar, healthier dishes. First, by analysing a recipe collection sourced from Allrecipes.com, we quantify the potential for finding replacement recipes, which are comparable but have different nutritional characteristics and are nevertheless highly rated by users. Building on this, we present two controlled user studies (n=107, n=111) investigating how people perceive and select recipes. We show participants are unable to reliably identify which recipe contains most fat due to their answers being biased by lack of information, misleading cues and limited nutritional knowledge on their part. By applying machine learning techniques to predict the preferred recipes, good performance can be achieved using low-level image features and recipe meta-data as predictors. Despite not being able to consciously determine which of two recipes contains most fat, on average, participants select the recipe with the most fat as their preference. The importance of image features reveals that recipe choices are often visually driven. A final user study (n=138) investigates to what extent the predictive models can be used to select recipe replacements such that users can be "nudged" towards choosing healthier recipes. Our findings have important implications for online food systems.

KEYWORDS

Food RecSys; human decision making; behavioural change; information behaviour

1 INTRODUCTION

Search and recommendation systems play an increasingly important role in the way people choose what they eat: Internet recipe portals are a popular source of food inspiration [8, 20] and often allow users to rate, and receive suggestions of, recipes. People search for recipes in a variety of ways for many different purposes [8, 42] and a relatively large proportion of web search queries are related to food or lead to the visit of a food-related website [39]. As such, systems which provide access to online recipes or make personalised recommendations have been touted as a means to help people nourish themselves more healthily [14, 17]. Nevertheless, despite offering access to healthy content [34], analyses of the systems being used in practice indicate that they tend to promote unhealthy meals [35]. Metrics such as recipe ratings, recipe bookmark frequency and the sentiment scores of recipe comments all tend to correlate positively with recipes that are high in fat, sugar and calorie content [34]. In other words, the recipes consumed most frequently and judged most favourably by users are typically the least healthy. Moreover, when common recommender algorithms are tested on recipe data, it is found that their recommendations are, on average, unhealthier than those rated positively by users themselves [34]. Thus, food access and recommendation systems, by themselves - at least in their current form - are no magic bullet for promoting healthy nutrition and may even serve to increase the likelihood that users will make poor nutritional choices.

Deciding what food one should eat is a complex, multi-faceted process, influenced by many biological, personal and socio-economic factors [5]. Moreover, a large body of evidence demonstrates that the food choices people make can be subtly manipulated with biases and cues, such as the default choice (status quo bias [41]) and the people present when the choice is made (social dependence [38]). Recent work has shown user behaviour with search and recommendation systems to be similarly susceptible to manipulation via psychological and system biases [40]. We bridge these domains by investigating the process of choosing foods via search and recommendation systems. Combining insights gained from analyses of recipes sourced from the large online food portal Allrecipes.com, naturalistic behavioural data detailing how users interact with these recipes, as well as the results of a series of controlled experiments, we seek to understand the processes involved in choosing a recipe online. Furthermore, we use what we learn to establish whether it is possible to algorithmically select recipes to 'nudge' users towards healthier choices.

Our experiments are conceived based on a scenario in which the user has a particular type of dish in mind (e.g. a "stir fry", "cheesy pasta" or "onion soup"), and is searching for a suitable recipe - a scenario naturalistic data show to be commonplace [8] and for which systems have been designed to support [37]. The driving motivation behind our work is to investigate the possibility of replacing meals recommender algorithms predict users will like with healthier versions of similar recipes.

More specifically we address the following research questions:

*Both authors contributed equally to this work.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

DOI: http://dx.doi.org/10.1145/3077136.3080826

- RQ1: To what extent is it possible, using typical online recipe databases, to replace unhealthy versions of recipes with similar recipes with healthier nutritional properties?
- RQ2: To what extent are users able to distinguish between healthy and unhealthy versions of recipes?
- RQ3: How does the information available influence the estimates made?
- RQ4: What biases are involved in the selection of online recipes?
- RQ5: To what extent can these biases be exploited to influence online recipe selections?

Outline. After motivating our contribution in Section 2 by highlighting relevant related work in the areas of food recommendation, food decision-making and biases in human behaviour, Section 3 introduces the data set used as the basis for experiments. The first set of analyses, presented in Section 4 studies the potential for identifying replacement recipes within the data set. In Sections 5.1 and 5.2 we turn to the processes involved in perceiving and choosing recipes, showing by experiment that fat content is poorly estimated by users, but that recipe preference can be predicted. In Section 6 we describe a user study which investigates whether we can use what we have learned to nudge users towards less fatty choices. Finally, in Sections 7 and 8 we discuss the significance of our findings and present our conclusions.

2 RELATED WORK

Food Recommenders. Work in food recommender systems has typically focused on rating prediction, utilising recipe content [14, 33] and contextual information [17] to minimise prediction error. Harvey et al. [17] showed that one important contextual factor for recipe recommendation is the users themselves: a small group explicitly preferred healthier food, while the majority tended to prefer less healthy alternatives. Recent work has tried to incorporate healthiness into the recommendation process by substituting ingredients [33], incorporating calorie counts [16], and generating food plans [12]. Elsweiler et al. and Trattner et al. identified the need for a trade-off between recommending recipes that will be appreciated by users and recipes that can be considered healthy using quantifiable metrics [13, 34]. Experiments show that the trade-off can be improved to some extent using post-filtering [34], but little is yet known regarding how such algorithmic approaches may influence the food decisions people make [35] and whether users will accept the healthier alternatives proffered.

Food Choice. People typically make around 200 food choices every day [38]. Choosing which food to eat is a complex process influenced by a number of context factors at biological, personal, situational, social and socio-economic levels [5]. Choosing food can be cognitively challenging, particularly when the number of options is large [32], leading to decisions often being driven by primal instinct and heavily influenced by simple stimuli, such as colour [7].

Neuroscience research has revealed food choices to be guided by competing behavioural controllers [9]. *Pavlovian control* induces pre-programmed responses when exposed to specific stimuli; *Habitual control* offers more flexible responses based on the previous history of rewards; and *Goal-directed control* allows decisions to reflect goals, such as weight loss. The evidence suggests that environmental factors, such as time constraints and marketing campaigns, which induce cognitive load, lead to dominance of the Pavlovian controller [28]. Thus, modern busy lifestyles, where people have limited time and attention resources and are bombarded with advertising designed to appeal to sensory instincts, mean that making healthy food choices is naturally difficult for many, a consequence that has been linked to problems such as obesity [28]. Compounding this, research shows [10] that foods that are high in calories and fat are the most palatable but are also the least satiating, meaning that we often choose such foods due to their appetising nature but must eat an excessive quantity of them to feel full.

Biases in Decision Making. It is generally accepted that, precisely because of limited cognitive resources, people often base their decisions on heuristics rather than a rational differentiation between avaiable options [19]. While heuristics can work quite well, the choices people make of what to eat can be biased in countless ways. For example, people make poor decisions when stimulated (e.g. when hungry and surrounded by the sights and smells of calorie rich food) [38] or when emotional [24] or stressed [26]. People adapt their behaviour to their social context: obese individuals are more likely to be friends with other obese individuals [6] and people consume more when they eat in groups, rather than alone [38].

Debate exists as to whether it is more effective or ethically appropriate to 'nudge', where biases are exploited to change behaviour, or 'boost', where people are supplied with information so that they can take more informed and, hopefully, better choices. With food this debate resolves around efforts such food-labelling [11] versus, for example, the language used to describe food products, where positive adjectives, for instance, make people more likely to accept a recommendation [15]. Both approaches have been applied to search and recommenders. Educational approaches, such as in [3] or [25] can be considered to be examples of boosting, whereas query suggestions and manipulating the search box size [4] are nudges. What is lacking in the literature, however, is an understanding of how boosts and nudges can be applied to food choices from search and recommender systems and whether such approaches can lead to any real behavioural change.

In summary, previous work has highlighted a trade-off between recommending users meals they will find appealing and those that are healthy. One approach to optimising this trade-off would be to substitute meals that would typically be recommended to users (as in [14, 17]) with similar but healthier dishes. For this strategy to be successful, however, a number of prerequisites need to be fulfilled: 1) recipes need to exist that are sufficiently similar in style and content, but different in health properties. 2) there needs to be potential in the human decision-making process to allow for the selection of healthier dishes, if available. That is, people need to be unable to tell the difference between the healthy and unhealthy versions of meals, if their preferences are so influenced, or other factors need to be identified that can outweigh healthiness in the decision process. 3) It must be possible to select replacement dishes that are more attractive to users than the original suggestions. The remainder of this paper addresses exactly these points.

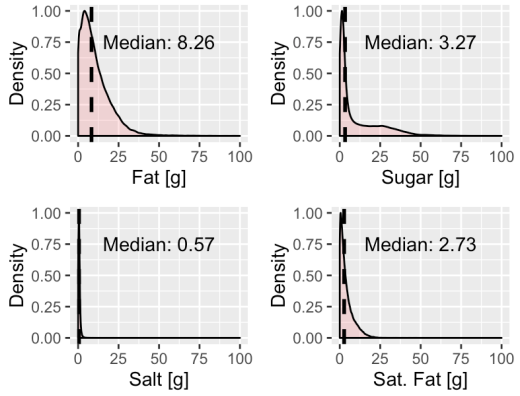

Figure 1: Density plots and medians for fat, sugar, salt and sat. fat content per 100g in the recipes of Allrecipe.com.

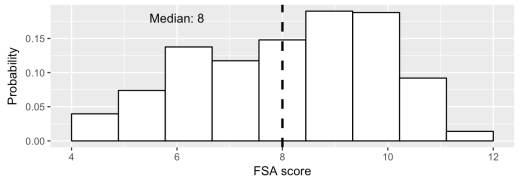

Figure 2: FSA score distribution (Probability: 1=100%).

3 DATASET

To address our research aims we obtained recipe and nutritional data from the Web by implementing a standard Web crawler. Between the 20^{th} and 24^{th} of July 2015, the crawler collected 242,113 recipes published between the years 2000 and 2015 on the Allrecipes.com website. We focus only on recipes published on the main site and ignore personal recipes, which are often incomplete and do not provide nutrition information. The primary reason for choosing Allrecipes.com was that, at the time of writing, it claims to be the world's largest food-focused social network and mantains a community of 40 million users from 24 countries accessing 3 billion recipes annually [2]. For our analysis, we relied only on those recipes for which all ingredients were present in the Allrecipes.com food database (out of 242,113 originally crawled recipes 58,263 had nutrition information available). We chose to focus on "fat", "saturated fat", "sugar" and "sodium" (measured in 100g per recipe) because they allow us to determine the healthiness of a recipe according to international standards introduced in 2007 by The Food Standards Agency (FSA) [1]. Following the procedure described in [18], for each meal we calculated the nutritional content per portion by dividing the total content by the number of portions in the meal. This allowed a so-called FSA health score to be calculated which measures, on a discrete scale, the extent to which a recipe is healthy or unhealthy [1]. The FSA front of package labelling system [1] relates to 4 macro-nutrients (sugar, sodium, fat and saturated fat). The scale is green (healthy), amber and red (unhealthy) and seeks to provide a clear and understandable indication of how healthful the product is. As in [30] we first assign an integer value to each colour (green=1, amber=2 and red=3) then sum the scores for each macro-nutrient resulting in a final range from 4 (very healthy) to 12 (very unhealthy). This metric, referred to as the "FSA score", provides a proxy for the healthiness of recipes. Figures 1 and 2, which plot the distributions of FSA scores, fat, saturated fat, sugar and salt for the dataset, reveal that the collection does contain recipes considered healthy according to these dimensions, but that the majority of recipes are indeed unhealthy.

Table 1: Overall probability of finding similar recipe pairs with 4 different thresholds (Probability: 1=100%).

Sim*	≥ 0.2	≥ 0.4	≥ 0.6	≥ 0.8
Probability	.051	.021	.007	.002

Note: * Cosine Similarity.

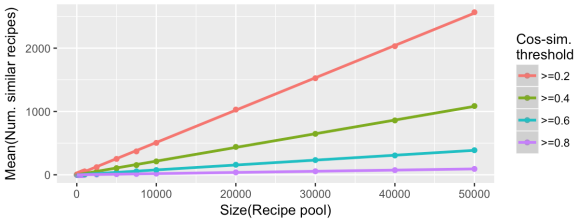

Figure 3: Recipe pool size vs mean number of similar recipes available for 4 different cosine similarity cut-off values.

4 RQ1: FINDING SUITABLE REPLACEMENTS

To realise our goal of replacing recipes with healthier alternatives, we must first establish if and when it is possible to find suitable replacements in the collection. We address this in three stages: First, following a method similar to [33], we establish recipe pairs based on their pairwise similarities; second, after such pairs have been established, we look at the distribution of various health properties across pairs to determine to what extent healthier replacements can be found. Finally, since replacements are unlikely to be accepted if the overall ratings are poorer than the original, we consider the rating distributions within pairs.

Table 1 shows the results of our first analysis, which examines the availability of similar recipes in the entire Allrecipes.com collection. We examined several similarity metrics, but report the cosine similarity as it was the metric used by [33], a standard reference from the literature[1][2]. The table shows how the probability of finding a similar recipe changes when the cosine similarity threshold is varied. For a given recipe, there are typically far fewer similar recipes than dissimilar ones, however some similar recipes do tend to exist. In particular, for a cosine similarity threshold of ≥ 0.2[3], there is a 5.1% probability of finding a partner recipe, while for a stricter threshold of ≥ 0.8 this reduces to 0.2%. In other words, with the recipe pool of 58,263 we have, out of 3.4 billion possible pairs approximately 6.8 million of these will have a similarity of 0.8 or higher. This means that each recipe will have an average of 116 potential replacements.

To understand the relationship between collection size and the probability of finding suitable replacements, we repeated the above

[1]There is a high correlation between cosine similarity and Jaccard coefficient (ρ=0.87) suggesting 1) ingredient quantities are not very important and 2) similar results would be expected regardless of the distance metric applied.
[2]Vector elements (ingredients) were weighted by the proportion of the recipe they represent. We experimented with various other weighting schemes but these produced pairs poorer results.
[3]The value used by [33], which we feel is not strong enough for our aims and thus report other threshold values.

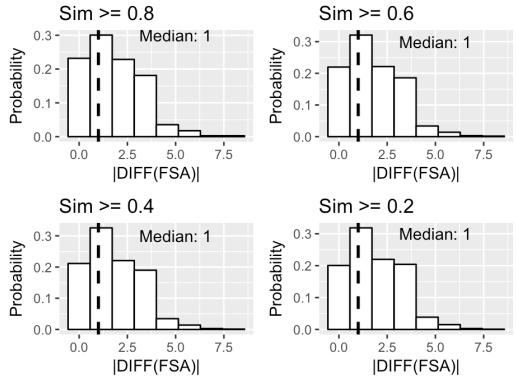

Figure 4: Distributions of recipe pairs for difference in FSA healthiness score and different sim. thresholds (Probability: 1=100%).

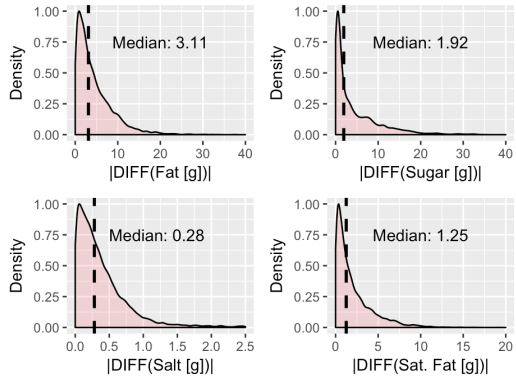

Figure 5: Density plots for recipe pairs for difference in fat, sat. fat, sugar and salt and with $sim \geq 0.8$.

procedure using recipe pools of various sizes drawn randomly from the full collection. To obtain statistically valid results, for each recipe pool we ran a boot-strap procedure with 1000 iterations.

The results of this experiment reveal that the relationship between pool size and number of similar recipes is linear, i.e., the larger the pool of recipes, the larger the number of similar recipes from which we can choose. The slope of the model is determined by the similarity threshold - the lower the threshold, the more similar recipes are available. For example, with a pool containing 100 recipes, we can expect to find on average 0.18 recipes with $sim \geq 0.8$, 0.78 for $sim \geq 0.6$, 2.2 for $sim \geq 0.4$ and 5.2 for $sim \geq 0.2$. A larger pool of 1000 recipes yields an average of 1.8 recipes with $sim \geq 0.8$ and so on. These results indicate that at least 555 recipes are required to find one similar recipe using a threshold 0.8 or higher and thus, even for relatively small recipe pools, it is typically possible to find very similar replacement recipes.

Next, we consider the relative healthiness of recipes and the extent to which we can find similar recipes that also exhibit very different nutritional properties. Figure 4 shows the results of experiments that explore the probability of finding similar recipes with different thresholds and with minimal differences in their FSA healthiness scores. Regardless of the similarity threshold chosen, the median difference in the FSA scores between the recipe pairs is 1, meaning that half of all the pairs differ in terms of FSA score by more than this. The 75th percentile is 2 for similarity thresholds

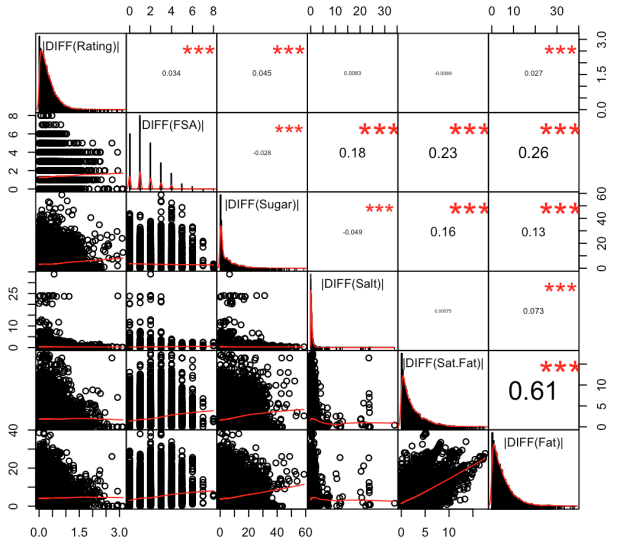

Figure 6: Correlation (spearman) matrix incl. distributions for recipe pairs with $sim \geq 0.8$ comparing ratings to macro-nutritional facts and the FSA health score (*p < 0.05, **p < 0.01, ***p < 0.001).

($\geq 0.4, \geq 0.6$, *and* ≥ 0.8) and 3 for $sim \geq 0.2$. As such, given a target recipe, it is feasible to find a number of alternative recipes that are similar in terms of content, but quite different in terms of nutrition. Figure 5 explores this relationship further on a macro-nutrient level, demonstrating that it is also possible to find recipe pairs that have quite different fat, sugar, salt and saturated fat levels. The plots also demonstrate that fat is the macro-nutrient with the greatest average difference in grams (Md=3.11)[4]. The tail of the distribution is also thick, suggesting it is easier to find similar pairs with very great difference in fat (g) than in the case of, for example, sugar.

Finally, we turn our attention to ratings. Figure 6 depicts a correlation matrix computed using only recipe pairs with a cosine similarity ≥ 0.8. The plot reveals a very slight correlation between user feedback in the form of ratings and the FSA score-based estimates of a recipe's healthiness (ρ=0.03), fat (ρ=0.03) etc. This contrasts with much stronger correlations between ratings and nutritional properties we reported in [34], where analysed the same data set, but did not restrict the analyses to similar pairs. We interpret this to mean that it is possible to find pairs of similar recipes where rating is not determined by healthiness. A further encouraging discovery with respect to our goals is provided by the distribution of rating divergence for recipe pairs with sim ≥ 0.8 (see top row in Figure 6), which shows that similar recipes have similar ratings (density peaks at zero).

In this section we have considered three properties of recipe pairs to establish whether it is possible to find appropriate healthy equivalents and have shown that, given a large enough pool, healthier recipes can typically be found. We have shown it is possible to

[4]In terms of Reference Intakes (RIs) based on a 2,000kCal daily diet, it is advised not to exceed 70g of fat, 20g sat. fat, 90g of sugar and 6g of salt [1]. As such, for the avg. portion size of 135.6g in our collection, 3.11g of fat difference amounts to $(3.11/70) * 135.6 = 6\%$ of daily recommended fat intake. For sugar this is $(1.92/90) * 135.6 = 2.9\%$, for salt $(0.28/6) * 135.6 = 6.32\%$ and for sat. fat. $(1.25/20) * 135.6 = 8.5\%$. Hence in respect to RIs the largest change can be made regarding sat. fats.

Table 2: A selection of recipe pairs.

French Crepes	Basic Crepes
Asparagus Soup in Seconds	Cream of Fresh Asparagus Soup II
Florentine Stuffed Chicken	Mom's Mozzarella Chicken for Drew
Ranch Crispy Chicken	Marinated Ranch Broiled Chicken
Buttermilk Coleslaw	Restaurant-Style Coleslaw I
French Toast I	Peanut Butter French Toast
French Onion Soup II	Lance's French Onion Soup

find similar recipes with different health characteristics and that these health characteristics are, in turn, only loosely correlated with rating. In sum, our analyses suggest replacing recipes with similar, healthy and comparably or better rated recipes is feasible.

5 INVESTIGATING PERCEPTION OF RECIPES

Now that we are satisfied that suitable replacement candidates can be found, we turn our attention to the decision processes involved in accepting recommendations. We wish to understand which informational cues are used to make these decisions and how the cues used relate, not only to a person's ability to correctly identify the healthiest recipe, but also to the choice of which recipe they select to eat. To this end we perform two experiments, both of which follow the same basic experimental design:

Basic Design. Participants were presented with a series of 10 recipe pairs, chosen randomly from a pool of 50 pairs in total. These were selected algorithmically such that recipes in the pair were very similar (similarity ≥ 0.8), but were different in fat content (one of the recipes contained at least twice the fat content of the other). A selection of example pairs is shown in Table 2.

Participants were questioned about each pair in a random order and, in each case, the two recipes were presented side-by-side. For each pair participants were asked 1) which recipe they found most appealing, 2) which they believed to contain the most fat, and 3) which piece of information most informed their opinion of relative fat content.

We questioned participants on a specific macro-nutritional property rather than asking them which recipe was "healthiest" because "health" is an ambiguous, subjective and multi-dimensional concept and, therefore, open to interpretation. We chose fat because our analyses above indicate that replacement recipes (i.e. similar recipes with lower fat content) would be plentiful. Future work will repeat our experiments with other macro-nutrient components.

The information presented for each recipe varied over the two studies: In the first study participants were shown only the recipe title as presented on Allrecipes.com and the first image available for the recipe on the site. Many recipes have several user-contributed images available, but we chose the first because this is the main one used for recommendation and search presentation on the website. In the second study, in addition to the title and image, participants were also provided ingredient lists for both recipes. The idea of having two studies with varying informational cues was to establish how much information was required and if extra cues can change the outcome.

Participants. 107 undergraduate information science students (64.5% male) took part in the first experiment. The students reported eating home cooked meals regularly (median= 5 days per week, IQR =2). The frequency with which they reported using online recipe websites varied. The distribution was spread uniformly over the categories "on a weekly or daily basis","on a monthly basis", "roughly every 3 months" and "hardly ever". The sample included 13 vegetarians, 3 vegans, and 6 pescatarians. On a 5-point Likert scale from "cooking is torture" to "I love cooking", the median value was 4 (IQR=1). When choosing a meal, the majority of participants perceived taste to be the most important criteria. However, some participants reported that the healthiness of the meal or the tastes of fellow diners were also important criteria.

A second group of 111 undergraduate information science students (59.5% male) participated in the follow-up experiment. The second group reported cooking at home in a very similar distribution to the first, perceived the cooking experience similarly (median=4) and also generally thought that taste was most important when choosing food, however they were less likely to use recipe websites on a daily basis. Among the participants, perception of the cooking experience was a significant predictor of preference towards healthy food (R^2=0.05, p=0.013,). There were 12 vegetarians, 2 vegans and 8 pescatarians.

Overall these groups represent convenience samples from a relatively homogeneous population of well-educated individuals primarily aged between 18-28. That being said the sample is diverse in terms of food preferences, priorities regarding the food they choose, the enjoyment of the cooking process, as well as the frequency with which online recipe sites are used, which we argue makes a good starting point to investigate our research questions.

5.1 RQs 2 & 3: Judging fat content

Participants in the first study were only able to correctly identify the recipe containing the most fat in 51.1% of cases, which is not significantly better than random (χ^2 = 0.61, df = 1, p = 0.43). A Krippendorff's alpha value of 0.101 moreover suggests little agreement among the participants. Unsurprisingly, in the second study, where participants were also provided with the ingredients list to make their judgements, the fattiest dish was correctly identified more often (in 56.7% of cases), which, despite being significantly better than random (χ^2 = 21.4, df = 1, df = 1, p < 0.001), is still hardly reliable. The Krippendorff's alpha value 0.116 is also slightly higher than without the ingredients list, but the agreement across participants is still only considered slight according to [23] .

In the first study, the participants reported mostly relying on the images to make their decision (71.9% of the time the image was the decisive cue) rather than the title. However, when one examines the success rate when different cues are used it seems that the cue used relates to the performance. Using the image resulted in a 50.4% success rate compared to 53.2% with the title. In the second study, decisions were often taken with the ingredients list as the decisive cue (in 52.7% of cases) compared to the image (37.4%) and title (10.0%). A Krippendorff's alpha value of 0.181 again indicates only slight agreement across participants. The ingredients list was associated with the highest accuracy (62.5% of estimates were correct) compared to the image (49.5%) and title (53.0%). A chi-square test confirms a relationship between the cue used and accuracy rate (χ^2 = 18.509, df = 2, p<0.01).

Figure 7: Example of a misleading image cue.

To investigate the extent to which cues can inform or mislead we divide pairs into groups along two dimensions: *agreement / little agreement* (with respect to the best cue on which to base judgments) and *high / low success-rate* (regarding the extent to which judgements were correct). Cue agreement across users was established at the level of recipe pair by calculating Shannon entropy for each pair. To avoid zeros we smoothed the probabilities by allocating 0.01% of the probability evenly across cues. Across the pairs there was considerable difference in the agreement (median = 1.24, IQR=0.34). The pair with the highest agreement had an entropy score of 0.81 and for this pair ingredients list was chosen 9 times, the image 3 times and the title was unused. The pair with the max entropy score (1.56) was rated 8 times, with 2 participants naming image as the primary cue, 3 the ingredients and 3 the title.

When the participants agreed on the best cue (i.e. entropy < median), the success rate was 60.8%, significantly better than 52.2% when there was poor agreement (χ^2 = 8.693, df = 1, p< 0.01). This means that typically, when participants agreed on the best cue, it tended to be a reliable indicator. This was not always the case however. Manually examining the cases where agreement was high and accuracy low allowed us, in some cases, to understand why the participants misjudged the fat content. Poor quality and misleading photographs were one source of bias. Pair 47 (shown in Figure 7) is a good example of how images can lead to incorrect judgements. The image for the dish with most fat (left) is misleading because it contains vegetables, which are not actually present in the recipe. The recipe on the right is similarly biased but in the opposite direction. The photographed accompaniment is cheesy pasta (again, not in the recipe). In the image the chicken looks as if it were deep-fried, whereas in the recipe it is actually baked, which reduces the fat content. Moreover, the plate in the recipe image (right) looks to feature what look like traces of fat, which may have influenced judgements. Examining the other user uploaded images for the same recipes suggests to us that if these were shown the participant estimates may have been very different.

Recipe titles can be similarly deceptive. For example, in the case of pair 38 ("Simple Red Sauce with Pasta","Holy Smoked Bacon and Mushroom Penne"), one third of the participants who incorrectly judged the first recipe to contain least fat, did so on the basis of the title, presumably associating "smoked bacon" with high fat content.

The title was not always informative of the fat content, which explains the infrequency with which it is cited as the best cue. However, in some cases (e.g. pair 8: "Banana Nut Bread III","Lower Fat Banana Bread II"), the title contained an obvious clue. In this case, only 57.9% of participants cited the title as the determining factor and 31.6% actually answered incorrectly. Amongst the participants who answered incorrectly was 1 participant who cited the title as his cue, but obviously doubted the veracity of its content.

In cases where the ingredients list was used as a cue and often incorrectly estimated revealed limited nutritional knowledge amongst some participants. It seemed, for example, that some participants were unaware that the fat content of red meats is higher than in white meats. In other words it was not only misleading cues, such as image or title biases, which led to incorrect judgements. Limited knowledge appears to be another factor.

In the demographic questionnaire, some participants cited the healthiness of the meals as being an important factor in their food choices. We split the sample into two groups. Those who rated the importance of healthiness below the median (n=27) and those who rated the importance >median (n=36). The healthy group seemed to be better at judging the fattier meal (58.1% vs 52.5%), but the difference is not significant (χ^2 = 1.7862, df = 1, p = 0.1814).

Summary. In summary, judging the fat content of online recipes is challenging: we observed poor accuracy in judgements and little agreement across participants. Even people who described health as being a priority when they choose a meal, were not significantly better than those who do not at judging which recipe contained most fat. Our analyses reveal different explanations for this: *lack of information* (the titles and images did not always provide the information necessary to judge), *lack of knowledge* to interpret the ingredient list correctly, and *misleading cues* - in many cases certain cues led users to falsely estimate the fat content of recipes.

5.2 RQ 4: Biases influencing selections

In the previous section, we showed that users find it hard to determine the fat content of a recipe and that certain cues (image of a recipe, recipe title or ingredient lists) can bias their interpretation. At first glance, this is a puzzling discovery as it does not fit well with the evidence from the literature suggesting that people, in the main, prefer fatty, calorie-rich recipes. In this section we focus on user preferences; concretely we investigate cues and biases influencing recipe selection using the data collected in the studies described above, as well as the naturalistic data set from Allrecipes.com. The main empirical contribution in this section is formulated as a prediction task, whereby we attempt to algorithmically estimate which of two recipes a user will prefer.

We take a machine learning approach to understand how various factors influenced the decision to choose one recipe over another. The prediction task is set up as follows: given a recipe pair (a, b), where recipes have a similarity $\geq$ 0.8, predict whether recipe(a) will be selected over recipe(b). Hence, in the prediction data set, each observation consists of a set of predictor variables or features that represent information about two recipes, and the response variable is a binary indicator with value "true" in the case when a was selected over b and the value "false" when b was selected over a. This is the setup previously employed in [33].

Feature Engineering. We selected 76 features relating to the three types of cues (recipe title, recipe image and recipe ingredients)

investigated in the previous sections. Furthermore, we selected 20 additional features relating to the nutritional content of a recipe as well as its popularity and the extent to which it is appreciated on Allrecipes.com. Below we briefly summarise these features and their corresponding sets:

- **Title:** For the title feature set, we derived 27 features. Four were simple text metrics, e.g. length in words and characters and text entropy [27]. We also measured the sentiment of the title and counted the words appearing in the Oxford English Dictionary. The remaining title features refer to POS-tags[5], e.g. number of Adjectives, Nouns, etc..
- **Image:** For the image feature set we derived 5 features capturing image sharpness, brightness, colorfulness, contrast and entropy. These features have been successfully used in the past to determine the attractiveness of Flickr images [31].
- **Ingredients:** We also created a set of simple features based on the ingredients used in a recipe, e.g. number of ingredients, number of words and chars, equivalent to those for the title [6].
- **Popularity & Appreciation:** We used popularity indicators such as number of ratings and bookmarks as well as appreciation measures, e.g. average rating and sentiment (via comments) provided by users in Allrecipes.com as a predictor.
- **Nutrition:** Finally, we derived features based on nutritional facts of the recipes as features. These include: the number of calories, fat, saturated fat, sodium and sugar per 100g contained in a recipe and the FSA health score.

Feature Selection & Classification Setup. The classification experiment was conducted with the help of the Weka[7] machine learning suite and R. Classifiers employed for the experiment were Random Forest, Logistic Regression and Naive Bayes. The evaluation protocol employed was 10-fold cross-validation. The order of recipe combinations (i.e. which is A and which is B) was rotated to ensure balanced classes. Throughout our experiments, we use feature selection methods to reduce the feature dimensionality and to ensure the models estimated were as robust and interpretable as possible. The discriminative power of features was measured using Information Gain (IG), which weights features according to their correlation with class attribute (=user preference) based on entropy. For the purpose of our study, we used IG to determine the top-10 features in each of the prediction experiments conducted (see Table 3).

Prediction Results. Table 3 presents the main results of the prediction experiments. The results are organised in 4 sections, reflecting different training and testing data sets. The first two sections report the results using data from studies 1 and 2, respectively. The third section presents the results of the same experiments using pairs generated from a sample of 10,000 recipes drawn randomly from the subset of the Allrecipes.com pool, which had been rated by at least 10 users. We use the highest average rating as provided by the users in Allrecipes.com, as a proxy for user choice [8]. A fourth section reports the performance of models trained on the Allrecipes.com sample, while predicting the choices made in studies 1 and 2.

Table 3: Results of the prediction experiment.

Accuracy				
Feature Set	Rand.For.	Logistic	Naive Bay.	Num. Feat.
Study 1 (Instances = 1102)				
Title	49.18%	48.63%	49.36%	54
Image	**64.25%**	**58.43%**	**60.16%**	10
Ingredients	62.25%	57.89%	55.71%	12
Nutr.	**64.25%**	58.25%	54.99%	12
Pop. & Appr	64.15%	55.53%	57.89%	8
Best (Top-10)	64.24%	60.61%	60.79%	10
All	64.33%	63.06%	63.52%	96
Study 2 (Instances = 1181)				
Title	48.43%	48.09%	49.87%	54
Image	**66.21%**	**61.64%**	**59.61%**	10
Ingredients	64.35%	60.96%	53.51%	12
Nutr.	65.96%	58.59%	54.19%	12
Pop. & Appr	65.96%	59.52%	58.59%	8
Best (Top-10)	66.04%	64.86%	61.05%	10
All	66.04%	64.86%	61.05%	96
Random Sample - Avg. Rating (Instances = 14,568)				
Title	62.95%	56.80%	54.60%	96
Image	**77.12%**	53.13%	52.83%	10
Ingredients	57.83%	52.42%	52.24%	12
Nutr.	75.41%	53.93%	52.79%	12
Pop. & Appr*	77.05%	**71.84%**	**69.54%**	6
Best (Top-10)	79.79%	71.29%	67.90%	10
All	84.78%	72.57%	65.95%	136
Random Sample (train) / Study 1 & 2 (test)				
Study1 (top-10)	**56.98%**	55.35%	52.54%	10
Study2 (top-10)	58.34%	**59.94%**	57.15%	10
Study1 (images)	**54.08%**	53.90%	49.63%	10
Study2 (images)	55.54%	**57.15%**	56.56%	10

Note: * Ratings were excluded.

Examining the results for studies 1 and 2 shows that comparable performance was achieved in the studies with the best performance (approx. 64-66%) being achieved by the random forest classifier. The strongest feature set was the image set, which outperformed the popularity / appreciation feature set for all three classifiers employed, although the differences are small. Unsurprisingly the title features, which we know from the user studies are not always informative, perform rather poorly.

Better performance was achieved all round when the same experiments were performed using the data from the Allrecipes.com random sample. This may simply be because more training data was available. Using all of the features available resulted in 84.78% being achieved; with the top 10 features this reduces to 79.79%. Again the image features provide solid predictive power by themselves and, as in the first two studies, when applying a random forest with these features alone, the best performance of all individual sets is achieved (77.12% accuracy). The title features seem to

[5]POS-tags were calculated with the popular Stanford NLP tagger see: http://nlp.stanford.edu/software/tagger.shtml. As title strings are short, we employed the GATE english pos-tagger model, see: https://gate.ac.uk/wiki/twitter-postagger.html

[6]We also experimented with words as features for both title and ingredients. However, since the number of words for recipe titles and ingredients in study 1 and 2 only cover a very small fraction of total words in the corpus, we decided to not train our model on these as we felt it would limit our chances of estimating a generalisable model

[7]http://www.cs.waikato.ac.nz/ml/weka/

[8]We also ran the same experiments using both number of ratings and bookmarks, as well as average sentiment as a proxy. Due to space limitations, we only present the results of the rating indicator. However, all experiments showed the similar performance and useful features.

Table 4: Top-10 features in each of the the 3 studies according to Information Gain (IG).

	Study 1		Study 2		Rand. Sample (rating)	
Rank	IG	Feature	IG	Feature	IG	Feature
1	.0933	IMG:contrast1	.0743	NUT:fat1	.1018	POP:sent2
2	.0829	IMG:brightness1	.0634	IMG:contrast2	.1016	POP:sent1
3	.0719	IMG:entropy1	.0573	IMG:colorfullness1	.0679	IMG:colorfullness1
4	.0707	POP:rating2	.0568	NUT:cal1	.0609	NUT:fat2
5	.0703	IMG:entropy2	.0542	NUT:satfat1	.0605	NUT:cal1
6	.065	POP:sent2	.0512	NUT:fat2	.0562	POP:book1
7	.0612	POP:book2	.0484	NUT:salt2	.0549	POP:book2
8	.0568	NUT:cal2	.0454	IMG:entropy1	.0430	IMG:sharpness1
9	.0551	IMG:colorfullness2	.0417	ING:charCount2	.0361	POP:ratings2
10	.055	POP:ratings1	.0390	IMG:entropy2	.0344	NUT:satfat2

do better in naturalistic environments, with the models trained on these features consistently outperforming the ingredient models on this data set. Table 4 lists the top 10 features for each data set estimated with IG. This shows that the image features are amongst the most important regardless of data set; the nutritional features help most in the second study, whereas for the Allrecipes.com sample the most discriminative features are spread across the popularity, nutritional and image sets.

To illustrate why the image features work so well, in Figure 8 we present the images associated with a series of recipe pairs. A model trained only on image features judged one recipe (top) from the pair to be particularly likely to be chosen while the other (bottom) was judged to be particularly unlikely to be chosen. In our subjective opinion, the top images are more attractive, particularly in the case of the 4 left-most examples. The 3 right-most examples are, in our opinion, less clear. We test the persuasive power of images selected by this model more thoroughly in Section 6.

As a final experiment we trained models using: 1) only the top-10 features; and 2) only image-related features) on the Allrecipes.com sample and tested how effective these models are at predicting the choices made by participants in the two user studies. The results (shown in the bottom section of Table 3), demonstrate that a maximum performance of 56.98% and 59.94% accuracy can be achieved with the top-10 features model for studies 1 and 2, respectively. Slightly poorer performance (54.08% and 57.15%) was achieved by the image feature models. In other words, significantly better than random[9] prediction performance can be achieved using only features, such as low-level image properties and general popularity indicators, trained on a data set with completely different users, collected in a different way. This despite knowing nothing about the individual preferences of the users. We view this as a strong indicator of the predictive power of the features and the robustness of the models.

In this section we have shown that when selecting recipes, user decisions are influenced by numerous cues. Despite not being consciously able to differentiate the fat content of recipes (see Section 5.1), users tended to, on average, select the recipe with the most fat content from the recipe pairs. Other good indicators included popularity metrics - it seems users in the main prefer recipes popular with other users - and low-level image properties, indicating that recipe choices are often visually driven.

[9] χ^2 tests show all the results to be significant, p<0.01.

6 RQ 5: NUDGING HEALTHIER CHOICES

The results presented above suggest the prerequisites for nudging we set out at the end of Section 2 can be met: replacement pairs exist (see Section 4), as does doubt in estimating fat content (see Section 5.1). We have also identified strong cues regarding the recipes people prefer (see Section 5.2). This section describes a final experiment, which determines if we can utilise what we have learned to realise the nudging of healthier recipes in practice.

The final study repeats the basic design reported above with participants choosing recipes from displayed pairs. In this case, however, pairs were selected using the models reported in the previous section to test to what extent it is possible to nudge people towards meals with significantly lower fat content. Thus, in this study participants are only required to indicate their recipe preference and were not required to make any explicit judgement with respect to the nutritional content of the meals. As in the previous experiments, a pool of 50 recipe pairs were chosen. We wanted to test to what extent it is possible to nudge people towards meals with significantly lower fat content, therefore we first restricted pairs to those in the top 30% in terms of difference in fat. From this subset we selected 25 pairs for which the random forest top-10 model trained on the Allrecipes.com sample predicted that the recipe with lowest fat content would be selected. Similarly, we choose 25 pairs where the random forest image-based model, trained on the Allrecipes.com sample, predicted the least fatty recipe would be selected. Further criteria for the selections were that 1) the pairs had to be comparable i.e. human users would consider them to be replacements for each other and 2) the same recipes did not feature repeatedly in the pairs.

138 participants, this time a more heterogeneous sample recruited via email lists and social media marketing, selected from 16 pairs. Based on a coin flip it was decided whether the next pair would be drawn randomly from top-10-model pairs or image-model pairs, thus a similar number of pairs were judged for each model.

The difference in fat content was similar for pairs selected by different models (top10: median Δfat =8.38g/100g, IQR=2.26; image: median Δfat = 8.34g/100g, IQR= 3.72). This represents a median nudge of 16.1% of the daily recommended fat intake for the avg. 2000kcal diet (see footnote 3). However, the certainty in prediction was significantly higher for the top-10 model (top10: median Δprediction = 0.82, IQR =0.08; image: median Δprediction = 0.70, IQR= 0.22).

Of the 134 participants who took part, 56% (n=75) were male and 78 reported their occupation. The most commonly stated occupation (n=40) was student, but others included historians, bar managers, lawyers and educators. Most participants (n=79) were between the ages of 18 and 24 and only 14 stated that they were older than 44. Similarly to the previous groups, participants reported eating home cooked meals regularly (median=5 days per week, IQR=3) and there were 14 vegetarians, 3 vegans and 10 pescatarians. Most rated taste as the most important factor when choosing what to eat, although many also stated that the healthiness of a recipes and social factors are important to them. As with previous groups, the median response to the 5-point Likert scale from "cooking is torture" to "I love cooking" was a 4 (IQR=1), indicating that most enjoyed cooking.

Figure 8: Example images from recipes a model trained on image-features predicts would be selected (top row) vs would not be selected (bottom row). The recipes are considered comparable (i.e. sim≥0.8).

Overall, in 62.2% of cases, the participants chose the recipe in the pair with the least fat, that is, the model predicted correctly. This is significantly better than random (χ^2 = 129.9, df = 1, p < 0.01) and the opposite of typical trends - people, as we know, generally choose recipes with most fat, demonstrating that we are indeed able to algorithmically "nudge" people and influence their food choices. In terms of the two competing models, the image model was able to predict the choice 65.2% of the time, outperforming the top-10 features model, which was correct in 59.3% of cases. Although both models individually were significantly better than random, the image-based model significantly outperforms the top-10 model (z=2.72, p < 0.01). This contrasts with the performance achieved in Section 5.2, where the top-10 model performed best.

It seems that the vegetarians were harder to sway - considering only the results from vegetarians, the percentage of correct predictions lowers to 56.4% over all pairs. With vegetarians removed, the accuracy of our methods increases to 62.8%, which is significantly better (p<0.05). The image-based model still works well for vegetarians (correct in 65.1% of cases), but the top-10 model performs very poorly and is correct only 48.7 % of the time. One plausible explanation is that for non-vegetarian recipes, these participants simply chose the best quality image. Meat-eaters, on the other hand it seems, can be nudged for such dishes using other cues.

In this section, we have demonstrated empirically that by selecting replacement recipes based on the predictive models trained in Section 5.2, we can tempt users into selecting the recipe containing the least fat.

7 SUMMARY & DISCUSSION

The main findings with respect to our RQs are summarised as follows:

- RQ1: The analyses in Section 4 show that, at least in the case of one extremely popular online recipe collection, it is possible to replace recipes with similar, healthy and comparably or better-rated alternatives.
- RQ2: Preference for fatty foods seems to be an implicit one as participants cannot tell the difference when asked, but typically select the fattier one as their preference.
- RQ3: Perception of fat content can be influenced by the information available and, in some cases, misleading cues (image or title) can bias and result in a false impression.
- RQ4: User preferences are predictable: several features can be useful predictors, however, the utility of low-level image features was consistent. We initially found this surprising, but perhaps we should not have - this is, after all, the pavlovian control in action.
- RQ5: We can exploit the biases to nudge people towards choosing the option with least fat. The high performance of the model trained only on image features shows how visually driven online food choices can be. Indeed, our approach shows that we can manipulate recommendations such that the pavlovian controller - the source of many unhealthy food decisions - can actually lead to choosing recipes containing less fat.

Taken together our results show that when a user is given a selection of two comparable recipes, we can select a pair such that the user is "nudged" towards the least fatty of the two. This is an extremely powerful finding and could potentially have far-reaching consequences. It does not mean, however, that the user is happy with the choice made nor that the recipe would actually be cooked and eaten in practice. Future research is needed to complement this work by experimenting in different settings, for example, in the context of SERPs. We are currently planning such studies and intend to measure additional outcomes, including user satisfaction with end choices.

Another limitation of this work is that the recipe pairs as we derived them are ignorant of individual user preferences. We are able to make accurate predictions in most cases, but any predictions may be undone because a user is allergic to eggs, does not particularly like broccoli or even icing on a cake. We believe that significant performance improvements could be achieved if we account for user preferences and future work will explore this in greater depth.

The lack of personalisation in our study also means that in some cases neither of the two recipes in a pair will have appealed to the participant. As discussed above, for instance, in some cases, vegetarians were required to choose from two meat-based dishes, which would be unlikely under normal circumstances. The fact that the image-based model performed better for non-meat eating participants suggests vegetarians' choices in these cases were perhaps even more strongly biased by the image than for two dishes they might actually consider eating. For example, they may have chosen one recipe because the image contained vegetables or salad or perhaps the image simply showed good lighting or presentation.

Our work, despite offering a new way of incorporating healthiness into the food recommendation problem, has only scratched the surface in terms of understanding how people make food choices online and how these choices can be influenced by search or recommendation systems. We are currently planning a series of studies to research this further, including eye-tracking studies to investigate how user behaviour changes when different information (e.g. nutritional information, food labels, recipe descriptions) are shown in different ways.

Our investigations of user perception and selection were restricted to the influence of fat content. While not all fats are unhealthy and low-fat does not necessarily mean healthy, current guidelines advise cutting down on all fats and replacing saturated fat with some unsaturated fat[10]. We plan on repeating our studies for other nutritional properties, such as sugar, carbohydrates and calories, to determine if similar effects can be achieved. We will also test if we can nudge to increase a nutritional element: can we nudge people to increase fibre or protein, for instance?

There are many other ways in which healthier recommendations could be achieved. The nutritional properties of recipes can be changed either by substituting individual ingredients [33] or simply by reducing the portion sizes. It would be interesting to study what kind of effects can be achieved with this approach. Finally, it is well known that cultural differences exist with respect to food choice and indeed the role food has in everyday life [29, 36]. We have begun to investigate whether or not the trends reported here are repeated in data collected from recipe websites from different countries. Our preliminary investigations with the German-based food portal Kochbar.de [21, 22] seem to indicate that many are.

8 CONCLUSIONS

This work combines insights from a broad range of empirical tools - analyses of online recipes, analysis of naturalistic behavioural data regarding how users interact with these recipes, as well as a series of controlled online experiments - to determine the feasibility of replacing online recipes with healthier equivalents. The results show that, despite (or perhaps even due to) the complexity of human food choices, the recommendation process can be manipulated through nudging such that a particular, "healthier" recipe will be chosen more often than would be expected by chance alone. This research provides the groundwork for the development of more sophisticated nudging techniques to build systems that help people to choose healthier meals whilst enjoying those choices even more.

REFERENCES

[1] 2007. FSA nutrient and food based guidelines for UK institutions. available at http://www.food.gov.uk/sites/default/files/multimedia/pdfs/nutrientinstitution.pdf. Last accessed on 20.6.2016. (2007).

[2] 2016. Allrecipe.com Press report. available at http://press.allrecipes.com/. Last accessed on 20.6.2016. (2016).

[3] Scott Bateman, Jaime Teevan, and Ryen W White. 2012. The search dashboard: how reflection and comparison impact search behavior. In *Proc. of SIGCHI'12*. ACM, 1785–1794.

[4] Nicholas J Belkin, Diane Kelly, G Kim, J-Y Kim, H-J Lee, Gheorghe Muresan, M-C Tang, X-J Yuan, and Colleen Cool. 2003. Query length in interactive information retrieval. In *Proc. of SIGIR'03*. ACM, 205–212.

[5] F Bellisle. 2005. The determinants of food choice. *EUFIC Review* 17, April (2005), 1–8.

[6] Nicholas A Christakis and James H Fowler. 2007. The spread of obesity in a large social network over 32 years. *New England journal of medicine* 357, 4 (2007), 370–379.

[7] Fergus M Clydesdale. 1993. Color as a factor in food choice. *Critical Reviews in Food Science & Nutrition* 33, 1 (1993), 83–101.

[8] Sally Jo Cunningham and David Bainbridge. 2013. An analysis of cooking queries: Implications for supporting leisure cooking. (2013).

[9] Nathaniel D Daw and John P O'Doherty. 2013. Multiple systems for value learning. *Neuroeconomics: Decision Making, and the Brain*, (2013).

[10] Adam Drewnowski. 1998. Energy density, palatability, and satiety: implications for weight control. *Nutrition reviews* 56, 12 (1998), 347–353.

[11] Andreas C Drichoutis, Panagiotis Lazaridis, and Rodolfo M Nayga Jr. 2006. Consumers' use of nutritional labels: a review of research studies and issues. *Academy of marketing science review* 2006 (2006), 1.

[12] David Elsweiler and Morgan Harvey. Towards automatic meal plan recommendations for balanced nutrition. In *Proc. of RecSys'15*. ACM, 313–316.

[13] David Elsweiler, Morgan Harvey, Bernd Ludwig, and Alan Said. Bringing the "healthy" into Food Recommenders. In *Proc. of 2nd International Workshop on Decision Making and Recommender Systems, Bolzano, Italy, October 22-23*. 33–36.

[14] Jill Freyne and Shlomo Berkovsky. 2010. Intelligent food planning: personalized recipe recommendation. In *Proc. of IUI'10*. ACM, 321–324.

[15] Lorenzo Gatti, Marco Guerini, Oliviero Stock, and Carlo Strapparava. 2014. Sentiment variations in text for persuasion technology. In *International Conference on Persuasive Technology*. Springer, 106–117.

[16] Mouzhi Ge, Francesco Ricci, and David Massimo. Health-aware Food Recommender System. In *Proc. of RecSys '15*. 333–334.

[17] Morgan Harvey, Bernd Ludwig, and David Elsweiler. 2013. You are what you eat: Learning user tastes for rating prediction. In *Proc. of SPIRE'13*. Springer, 153–164.

[18] Simon Howard, Jean Adams, Martin White, and others. 2012. Nutritional content of supermarket ready meals and recipes by television chefs in the United Kingdom: cross sectional study. *BMJ* 345 (2012).

[19] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[20] Marian Knopp. 2011. Information needs, preferences, and behaviors of home cooks. *Library and Information Research* 35, 109 (2011), 40–54.

[21] Tomasz Kusmierczyk, Christoph Trattner, and Kjetil Nørvåg. Temporal Patterns in Online Food Innovation. In *Proc. of WWW'15 Companion*.

[22] Tomasz Kusmierczyk, Christoph Trattner, and Kjetil Nørvåg. Temporality in online food recipe consumption and production. In *Proc. of WWW'15*.

[23] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[24] Michael Macht. 2008. How emotions affect eating: a five-way model. *Appetite* 50, 1 (2008), 1–11.

[25] Neema Moraveji, Daniel Russell, Jacob Bien, and David Mease. 2011. Measuring improvement in user search performance resulting from optimal search tips. In *Proc. of SIGIR'11*. ACM, 355–364.

[26] Georgina Oliver, Jane Wardle, and E Leigh Gibson. 2000. Stress and Food Choice: A Laboratory Study. *Psychosomatic Medicine* 62 (2000), 853–865.

[27] Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proc. of EMNLP'08*. Association for Computational Linguistics, Stroudsburg, PA, USA, 186–195.

[28] Antonio Rangel. 2013. Regulation of dietary choice by the decision-making circuitry. *Nature neuroscience* 16, 12 (2013), 1717–1724.

[29] Paul Rozin, Claude Fischler, Sumio Imada, Allison Sarubin, and Amy Wrzesniewski. 1999. Attitudes to food and the role of food in life in the USA, Japan, Flemish Belgium and France: Possible implications for the diet–health debate. *Appetite* 33, 2 (1999), 163–180.

[30] Gary Sacks, Mike Rayner, and Boyd Swinburn. 2009. Impact of front-of-pack traffic-light,nutrition labelling on consumer food purchases in the UK. *Health promotion international* 24, 4 (2009), 344–352.

[31] Jose San Pedro and Stefan Siersdorfer. 2009. Ranking and Classifying Attractiveness of Photos in Folksonomies. In *Proc. of WWW'09*. ACM, New York, NY, USA, 771–780.

[32] Benjamin Scheibehenne, Rainer Greifeneder, and Peter M Todd. 2010. Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research* 37, 3 (2010), 409–425.

[33] Chun-Yuen Teng, Yu-Ru Lin, and Lada A. Adamic. Recipe Recommendation Using Ingredient Networks. In *Proc. of WebSci'12*. 298–307.

[34] Christoph Trattner and David Elsweiler. 2017. Investigating the Healthiness of Internet-Sourced Recipes: Implications for Meal Planning and Recommender Systems. In *Proc. of WWW'17*. ACM.

[35] Christoph Trattner, David Elsweiler, and Simon Howard. 2017. Estimating the Healthiness of Internet Recipes: A Cross-sectional Study. *Frontiers in Public Health* 5 (2017), 16. DOI: http://dx.doi.org/10.3389/fpubh.2017.00016

[36] Viet Phu Tu, Dominique Valentin, Florence Husson, and Catherine Dacremont. 2010. Cultural differences in food description and preference: Contrasting Vietnamese and French panellists on soy yogurts. *Food Quality and Preference* 21, 6 (2010), 602–610.

[37] Youri van Pinxteren, Gijs Geleijnse, and Paul Kamsteeg. 2011. Deriving a recipe similarity measure for recommending healthful meals. In *Proc. of IUI'11*. ACM, 105–114.

[38] Brian Wansink. 2006. *Mindless eating*. Bantam Books.

[39] Robert West, Ryen W White, and Eric Horvitz. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proc. of WWW'13*. 1399–1410.

[40] Ryen White. 2013. Beliefs and biases in web search. In *Proc. of SIGIR'13*. ACM, 3–12.

[41] Jessica Wisdom, Julie S Downs, and George Loewenstein. 2010. Promoting healthy choices: Information versus convenience. *American Economic Journal: Applied Economics* 2, 2 (2010), 164–178.

[42] Michiko Yasukawa, Fernando Diaz, Gregory Druck, and Nobu Tsukada. 2014. Overview of the NTCIR-11 Cooking Recipe Search Task.. In *NTCIR*.

[10] http://www.nhs.uk/Livewell/Goodfood/Pages/Fat.aspx