PlateMate: Crowdsourcing Nutrition Analysis from Food Photographs

Jon Noronha, Eric Hysen, Haoqi Zhang, Krzysztof Z. Gajos
Harvard School of Engineering and Applied Sciences
33 Oxford St., Cambridge, MA 02138, USA
{noronha,hysen,hqz,kgajos}@seas.harvard.edu

ABSTRACT

We introduce PlateMate, a system that allows users to take photos of their meals and receive estimates of food intake and composition. Accurate awareness of this information can help people monitor their progress towards dieting goals, but current methods for food logging via self-reporting, expert observation, or algorithmic analysis are time-consuming, expensive, or inaccurate. PlateMate crowdsources nutritional analysis from photographs using Amazon Mechanical Turk, automatically coordinating untrained workers to estimate a meal's calories, fat, carbohydrates, and protein. We present the Management framework for crowdsourcing complex tasks, which supports PlateMate's nutrition analysis workflow. Results of our evaluations show that PlateMate is nearly as accurate as a trained dietitian and easier to use for most users than traditional self-reporting.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

General terms: Design, Human Factors

Keywords: Human computation, Crowdsourcing, Mechanical Turk, Nutrition, Remote Food Photography

INTRODUCTION

The majority of Americans perceive healthy eating as complicated [5]. Seeking comprehensible and actionable advice, Americans spend over $40 billion each year on diets and self-help books [18], but achieve little success: the majority eventually regain any lost weight and more [13].

There are many factors that may impact successful long-term change in eating habits. Our work is based on the observation that food intake monitoring is a popular component of many diets. For people who make a commitment to changing their eating habits, accurate logs of what they eat may help in monitoring progress toward set goals [11]. Currently, food logging is typically done by hand using paper diaries, spreadsheets, or a growing number of specialized applications. This process is both time-consuming and error-prone [17, 6]. Nutritionists have explored alternative methods such as daily interviews with trained experts. While these methods improve accuracy, they are costly and still require substantial time investment.

Our work is inspired by the *Remote Food Photography Method* (RFPM) [16], a novel approach from the nutrition literature. Rather than remembering foods or writing down records, users take two photographs of each meal: one at the beginning of the meal and one at the end documenting the leftovers. These images are analyzed by a third party, making logging easier and discouraging self-deception. The challenge is in finding a qualified third party without prohibitive costs. Expert nutritionists are scarce and costly, limiting the system to wealthy users or patients with particular conditions.

To make accurate food logging easier and more affordable, we introduce *PlateMate*, a system for crowdsourcing nutritional analysis (calories, fat, carbohydrates, and protein) from photographs of meals using Amazon Mechanical Turk. Complex tasks like this are hard problems for crowdsourcing, as workers may vary drastically in experience and reliability. To achieve accurate estimates, we propose a workflow in which the overall problem is decomposed into small, manageable, and verifiable steps. PlateMate uses this workflow to assign tasks to contributors, to validate and combine results, and to appropriately route tasks for further processing.

This paper makes three main contributions:

1. We present PlateMate, an end-to-end system for crowdsourced nutrition analysis from food photographs.
2. We discuss the results of a two-part evaluation, which suggests PlateMate can be as accurate as experts and self-report methods, and more usable than manual logging for everyday use.
3. We introduce the Management framework—inspired by the structure of human organizations, it provides effective support for managing crowdsourcing of complex heterogeneous tasks.

PlateMate implements the first step in the Remote Food Photography Method. In the last section we suggest how it can be extended to also support the second step: the analysis of photographs of food waste.

UIST'11, October 16-19, 2011, Santa Barbara, CA, USA.
Copyright 2011 ACM 978-1-4503-0716-1/11/10...$10.00.

In the next section we review relevant prior work. We then describe the design and implementation of the PlateMate system and its components. Next, we discuss our Management framework. We then present an evaluation of the accuracy and usability of PlateMate and discuss the results. Finally, we consider future extensions to PlateMate.

RELATED WORK

Nutritionists have established several methods for measuring food intake. One prominent approach is 24-hour recall, in which a trained dietitian interviews a subject about her consumption over the previous day [16]. Accuracy depends on the subject's memory and honesty, and the technique requires a costly expert to conduct analysis. The main alternative is food journals, in which subjects record meals and estimate portions themselves, usually with pen and paper.

Both methods require significant time and self-reports also suffer from limited accuracy. A review of nine studies found error rates from -76% (underestimates) to $+24\%$ (overestimates) [19]. Prior work also suggests a dangerous bias in self-report methods. Most subjects selectively underreport fat intake, and obese people underestimate portions much more than leaner ones [17, 6]. These errors imply a larger problem of self-deception, especially in vulnerable groups.

A number of online interfaces exist to simplify the process of food logging. Smartphone applications and online calorie databases improve on earlier methods by performing calculations automatically. However, they still require tedious logging that discourages recording. Self-reports using these interfaces are no more accurate than pen and paper [2, 21].

The Computer Science community has explored additional alternatives, such as automatic analysis of chewing sounds [1] and scanned grocery receipts [12]. These methods, while potentially more scalable and less time-consuming than current approaches, remain inaccurate.

Martin et al. recently suggested an alternative approach called the Remote Food Photography Method (RFPM) [16]. Rather than typing names of foods and estimating portions, users take photographs of their plates both at the beginning of the meal and at the end to accurately capture how much food was actually eaten. Trained dietitians identify the pictured foods remotely and estimate portions. The results of laboratory studies showed that dietitians using RFPM underestimated calories by 5-7% compared to the ground truth obtained by directly weighing the foods [16].

RFPM thus combines the accuracy of direct observation by experts with the convenience of free-living conditions. Users of the method found it extremely satisfying and easy to use [16]. The problem is cost. RFPM relies on experts to analyze each photograph, limiting the system's accessibility and potential scale.

Kitamura et al. attempted to use computer vision to cheaply implement RFPM [7]. They were successful in algorithmically detecting if a photograph contained food and in estimating amounts of general categories of food, such as meats, grains, and fruit. They did not attempt to identify the specific foods in a photo or provide actual intake totals.

The cost of experts and limitations of computer vision suggest an opportunity for crowdsourced nutritional analysis. Prior research indicates that the most difficult part of nutritional analysis is estimating portion size [16], and that trained amateurs have low bias but high variance [15]. The "wisdom of crowds" is ideally suited to these situations, since the average of amateur estimates often beats a single expert [20].

A recent iPhone application demonstrates, however, that naive approaches to crowdsourcing for nutritional analysis are not sufficient. In April, 2011, the fitness website Daily Burn released Meal Snap, which allows users to photograph foods and receive calorie estimates by so-called "pure magic."[1] Meal Snap creates a single Mechanical Turk task for each image. Workers provide a free text description of food, and the application appears to match this description with a database of average consumption to estimate a range of possible calories. This approach is appealing, but critics have accused it of failing to provide useful data[2] and our evaluation showed that Meal Snap's results do not correlate with the meal's actual caloric content.

PhotoCalorie[3] is a recent on-line tool that encourages users to upload photographs of their meals, but it uses them just to illustrate the user's personal photo journal. The apparent similarity to PlateMate is superficial because to obtain calorie estimates, users have to enter short descriptions of the contents of the meals and manually estimate the amounts eaten.

PLATEMATE

PlateMate allows users to upload food photographs and receive nutrition estimates within a few hours. The estimates consist of a list of foods in the photograph, with associated measurements of serving size, calories, fat, carbohydrates, and protein for each food item. The information is displayed to the user via the user interface shown in Figure 1.

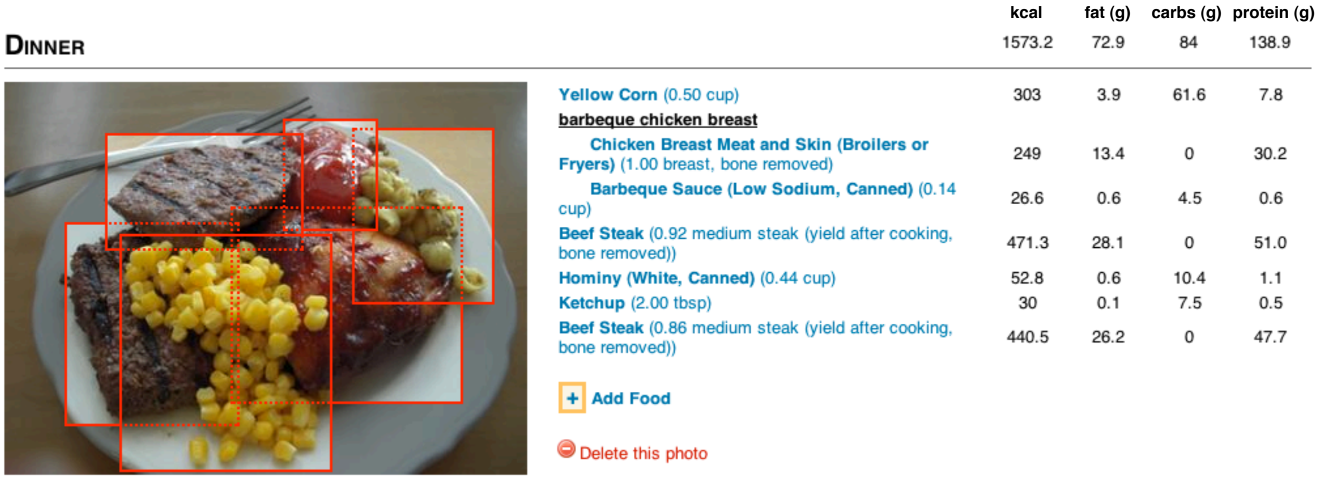
Estimates are generated from a series of tasks on Amazon Mechanical Turk. Crowdsourcing nutritional analysis presents several challenges in interface and workflow design. First, Turkers are inexperienced, and may thus produce unreliable estimates. Second, most Mechanical Turk tasks are simple, and Turkers may be unaccustomed to performing complex operations like nutritional analysis if presented as a single, complex task. Finally, any individual Turker may be biased in their estimates or have trouble recognizing certain foods contained in a photograph, making it necessary to select from or combine the outputs of multiple workers.

To best design a workflow for crowdsourcing nutritional analysis, we started by observing a dietitian as she determined nutritional data from several photographs. Her process consisted of three distinct steps: identifying foods in each image, estimating their portions, and then calculating the corresponding nutrition data. The final step can be fully computerized, but PlateMate implements the first two with

[1] http://mealsnap.com/, accessed July 5, 2011

[2] http://www.mobilecrunch.com/2011/04/05/too-lazy-to-count-calories-now-you-can-just-take-a-picture-of-your-meal/

[3] http://photocalorie.com/, accessed on July 5, 2011



Dinner	kcal	fat (g)	carbs (g)	protein (g)
	1573.2	72.9	84	138.9
Yellow Corn (0.50 cup)	303	3.9	61.6	7.8
barbeque chicken breast				
Chicken Breast Meat and Skin (Broilers or Fryers) (1.00 breast, bone removed)	249	13.4	0	30.2
Barbeque Sauce (Low Sodium, Canned) (0.14 cup)	26.6	0.6	4.5	0.6
Beef Steak (0.92 medium steak (yield after cooking, bone removed))	471.3	28.1	0	51.0
Hominy (White, Canned) (0.44 cup)	52.8	0.6	10.4	1.1
Ketchup (2.00 tbsp)	30	0.1	7.5	0.5
Beef Steak (0.86 medium steak (yield after cooking, bone removed))	440.5	26.2	0	47.7

Figure 1: The PlateMate user interface. Users upload photographs of their meals, which are processed through Mechanical Turk to produce a list of foods, serving sizes, and nutrition information.

crowdsourcing. Following Soylent [3], we also add an input decomposition stage at the start to parallelize work.

The result is a workflow with three major stages, shown in Figure 2. *Tag* takes photos and labels them with boxes drawn around distinct foods on a plate. *Identify* matches each box to one or more foods in a commercial nutrition database. *Measure* returns portion estimates for each identified food.

Step 1: Tag

The goal of the Tag stage is to find every food item in a photograph. One picture may depict several plates, and each plate might contain several distinct foods. Tag discovers these foods and distinguishes them by drawing a rectangle around each. The result is a group of boxes overlaid on the picture. Each box corresponds to a single food item, like a sandwich.

This step has the same benefits as the Find stage in Soylent's Find-Fix-Verify pattern [3]. Results can be surfaced more naturally in the user interface, and this makes estimates easier to understand and correct. Parallel work can also be combined more carefully, since we know which identifications describe each pictured food. Finally, the Tag step encourages completeness, preventing "Lazy Turkers" from ignoring or forgetting to match certain foods.

Drawing Boxes Tag's first Human Interactive Task (HIT) asks workers to draw boxes around each food in the picture. Workers need cultural background knowledge to understand how foods on a plate fit together. Pure computer vision can detect edges and boundaries, but it cannot recognize that an open-faced hamburger with half of the bun off to the side is in fact one item. The HIT relies on Turkers' general intuition about food items, and provides examples of sandwiches, salads, and pasta with vegetables as appropriate items.

Similarity Comparison and Voting Two Turkers are asked to tag each photo, and a combination of machine and human computation is used to select the better box group. Once both assignments are completed, they are algorithmically compared in the number, size, and position of boxes. If the two groups are sufficiently similar, one is picked at random as the final answer.

If the box groups differ significantly, three additional Turkers are shown each set overlaid on the photo and asked to select the better option, using similar guidelines. The box group receiving more votes is returned as the final result.

Step 2: Identify

The Identify step matches a tagged box to one or more food entries in a commercial nutrition database. While each box output from Tag should only contain one food, some composite items do not exist in the database. For example, if "ham and cheese sandwich" is missing, Identify should choose "wheat bread," "sliced ham," and "American cheese."

There are two main challenges in this stage. Identifications must be correct, and when several correct identifications exist, the most compact one should be used in order to simplify measurement and eventual presentation of data to end users.

In an initial pilot study, Identify was performed in a single HIT. Workers used an autocomplete text input to list each food in the box. Their answers were frequently incorrect or incomplete. Workers appeared to type a one-word description of the picture, like "chicken," and then select the first option regardless of how closely it fit. Like the "Lazy Turkers" in [3], they performed the minimal work necessary to get paid and nothing more.

These problems also occurred because the interface asked Turkers to perform two conceptually different tasks sequentially but only produce one final output. Turkers first had to identify the food in their own minds, and then locate the corresponding entries in the database. To correct for this, we developed a workflow that contained two simpler HITs. The first asks workers to describe the food in their own words. The second asks them to match this description to items in the database.

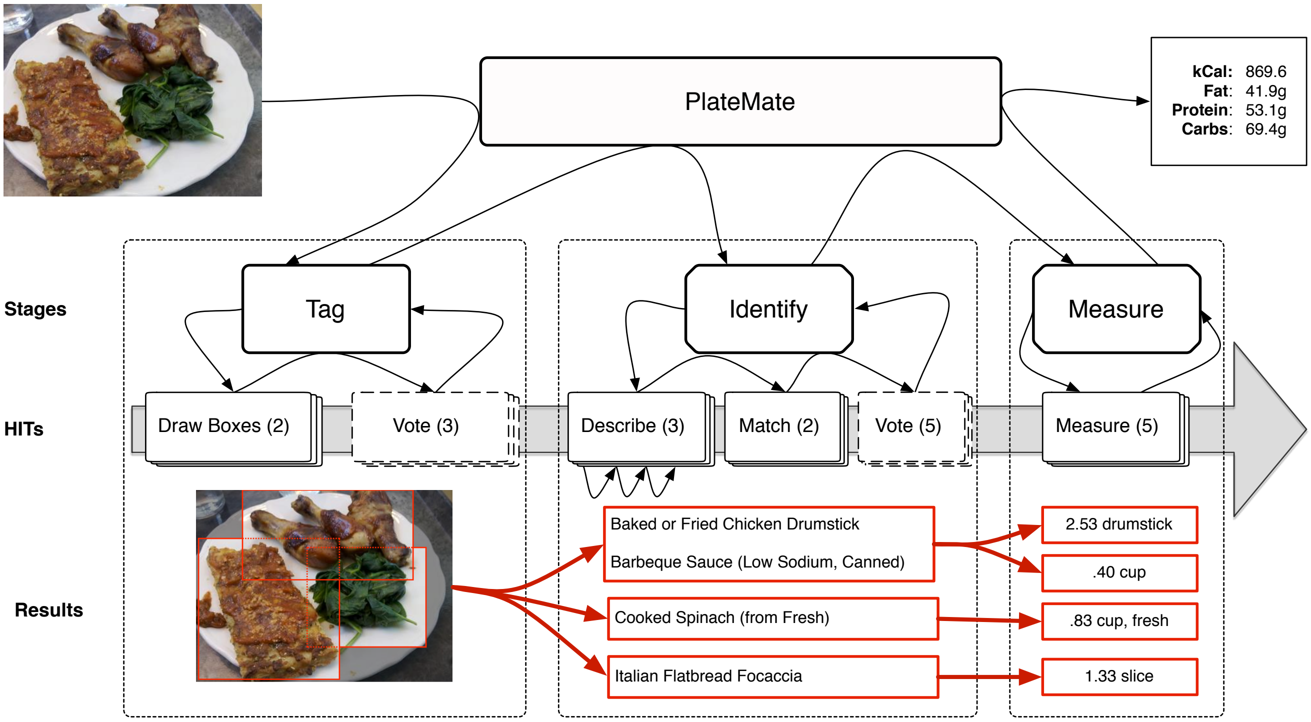


Figure 2: The PlateMate system. Work travels between stages and Human Intelligence Tasks (HITs) along the black arrows, starting from the input on the left and concluding with the output on the right. The system takes submitted photos and creates Tag tasks to annotate these photos with boxes. Each box becomes the input to a series of Identify tasks which end with a list of foods from a commercial food database. Each individual food is then input to a Measure task, which produces a unit and amount. Dashed boxes represent optional stages, which may be skipped during routing.

Describing Items In this HIT, Turkers see a box on a photo. One question asks "What is this food?", requesting one-line descriptions like "pepperoni pizza" or "salad with chicken." Another asks "What is it made of?", providing a free-form text field where workers can list component parts. For simple foods like broccoli these fields will be identical, but for composite foods the fields should have different answers that are each useful.

Following successful prior experiments in describing images [9], we made this step *iterative*. One worker starts from blank boxes. Her answer becomes input to another HIT, where the next Turker is asked to improve on it by correcting mistakes and adding detail. This process is well-suited to the "Eager Beavers" of [3], who provide minute details and list many possibilities. It also handles "Lazy Turkers" well, since terse descriptions are progressively expanded.

Matching Foods After three iterations, the output of the Describe task is fed into a Match HIT. Here, workers see the photo and the final descriptions. They are asked to select the best entry or set of entries in the database to match the box, with the descriptions as a suggestion for what to search. Workers first attempt to locate the description of the box as a whole in the database. If they find no good match, they search for each part. For example, workers should first search for "salad with chicken and tomatoes." If this fails, they should look for "chicken breast", "romaine lettuce", and "cherry tomatoes."

The search interface is modified from a standard autocomplete. Search results display below the input box, but the keyboard cannot be used for quick selection. Turkers must use the mouse to click the correct items to add. The interface also makes it clearer that multiple items can be selected through several searches. These changes negate the instinct of "Lazy Turkers" from the pilot study to select the first item they see.

This decomposition makes each step manageable for Turkers moving rapidly through HITs. The results of the Describe step are not necessary for the end goal of calculating nutrition information, but the generated descriptions reduce the mental work required for the Match step. We can then ask Turkers working on Match HITs to find the simplest representation in the database, using the Describe results as a guide.

Agreement Detection and Voting Two workers are asked to complete each Match HIT. If each returns a list pointing to the exact same item or items in the food database, then that list is used. Otherwise, five workers complete a Vote HIT to decide between them.

Step 3: Measure

The Measure step produces an estimated portion size for each food matched in Identify. Following this stage, the nutrition data for a photo can be calculated by multiplying the per-unit nutrition breakdown from the food database with the estimated measurement for each identified food.

Measure uses only one HIT, which shows Turkers a photo with a box highlighted along with the name of one food in that box. They are asked to first select a measurement unit and then provide a numeric estimate in terms of that unit. The units provided by the food database are specific to each food. "Pepperoni pizza" includes options like "slice, large" or "whole pie, medium," while "white rice, cooked" uses cups or ounces.

Measurement is considered the most difficult step of this process for amateurs [16], so the Measure stage uses a number of techniques to produce accurate results. Presenting multiple measurement options is helpful, since many of these only require counting rather than estimating a weight or volume. For example, it is much easier to count florets than to estimate grams of broccoli.

Not every food can be measured by counting. To help in cases where weight or volume estimates are necessary, HITs include a portion guide which provides common approximations for different measurements. For example, 3oz of meat looks like a deck of cards, and a quarter cup is roughly the size of a golf ball. These approximations are more error-prone than simple counting, but they allow workers to estimate portions without any training.

The interface also warns Turkers who appear to be making common errors. Pilot testing revealed that measurements in weight were significantly less accurate than those using volume or counting, so a warning is presented when Turkers choose grams, ounces, or pounds. Testing also indicated that some workers misunderstood the serving types. For example, for "chicken nuggets," one worker selected "serving, 6 nuggets" and then entered 6 as the value. This indicated 6 servings of 6 nuggets each for 36 total.

To reduce these errors, the interface generates a calorie estimate on the fly and asks workers to eyeball their answer. They are given common calorie ranges for different meals and shown warnings if the count becomes unusually low or high. These warnings cannot prevent all errors, but they encourage Turkers to double-check their answers.

Aggregating Measurements Five Turkers are presented with Measure HITs. The results from these HITs can be compared in the common units of calories. This means estimates can be aggregated without any additional human computation like voting. Drawing on the principle that averaging many high variance by low bias estimates can lead to accurate results [20], we remove outliers and then return the mean of the remaining estimates.

Turker Qualifications

After several iterations during pilot testing, we decided to accept only Turkers located in the United States who had previously completed at least 200 HITs and had a 98% HIT acceptance rate. We chose to require American Turkers due to the unique cultural context required for most elements of the process. Pilot tasks with foreign workers showed common mistakes like identifying the ketchup on a hamburger bun as strawberry jam, showing the necessity of cultural context.

THE MANAGEMENT FRAMEWORK

In this section, we introduce a programming framework for solving problems with crowds based on a human organizational hierarchy. This approach differs conceptually from prior work, which has focused on creating "crowd programming languages" that combine human and machine computation. For example, TurKit [9] lets requesters program crowds in JavaScript, Qurk [14] integrated crowds into SQL, and CrowdForge [8] parallelized work with MapReduce scripts. In each case, these toolkits have attempted to make working with crowds more like working with computers. This approach emphasizes computation as the natural glue for combining individual worker contributions and the resulting artifact is a computer program with some of the primitive operations implemented as "functional calls" to human workers [9].

Because PlateMate relies primarily on human work, divided into a number of heterogenous and interacting tasks, and because the issues of worker skill and motivation were central to our design process, we found it conceptually helpful to use human organizational hierarchies as the metaphor for designing our system. Specifically, we observe that in the real world, expert-level work (e.g., building a table) can be reproduced by less skilled workers—each working on a specific part of the process—supervised by managers who are not necessarily skilled craftsmen themselves, but who know how to assign tasks, route work among workers, and verify the quality of the work.

Thus, to implement division of labor for crowdsourcing, we created a new framework organized around objects called *managers.* Managers communicate with their supervisors and their *employees* using asynchronous message passing: managers assign tasks by placing them in inboxes of lower level managers and communicate with their superiors by placing results of completed tasks in their own outboxes. This hierarchical message-passing approach allows programmers to implement workflows by decomposing problems into progressively smaller steps.

As illustrated earlier in Figure 2, the root of this tree is a *chief* manager, which gathers new inputs and produces completed outputs. In PlateMate, the chief has three employees: Tag, Identify, and Measure. Each of these are in turn managers and have their own employees, corresponding to the individual HITs described above.

This hierarchical structure creates a flexible workflow consisting of modules connected by higher-level managers. Managers can route work intelligently among their employees, and may dynamically alter the sequence of steps in the process depending on a situation. For example, PlateMate's Tag manager compares the outputs from its DrawBoxes employee. If they are sufficiently different, they are sent to the VoteBoxes manager to decide between them. Otherwise, one answer is chosen randomly and sent up the hierarchy as Tag's completed output. All managers work in parallel, each processing its own stream of work.

When multiple tasks are submitted, processing is done just-in-time: for example, as soon as one photograph is tagged, the Identify manager begins the process of finding out what foods are present in each of the boxes without waiting for the remaining photographs to be tagged.

At the lowest level of the hierarchy are managers whose employees are the crowd workers. Managers at this level create jobs (such as asking for the food in one tagged box on a photo to be identified) and receive responses. Programmers create HIT templates and validation functions which are used by the framework to create HITs and approve work. Managers simply assign work to the crowd and receive validated outputs that can be passed up the tree.

Of course, the Management Framework *is* a computational framework, and it naturally supports a number of the recently introduced design patterns for programming the crowds. For example, the Tag step is an analog of the map step in MapReduce and the Describe step (part of Identify, see Figure 2) relies on iterative refinement [10] to improve the level of detail of the descriptions.

Management is implemented as an extension of Django, a web application framework for Python. It builds on several useful features from Django, including an HTML template language for defining HIT instructions, examples, and interfaces. It also uses Django's object-relational mapper, which automatically stores Python objects in a MySQL database. This means that the precise state of the system is always stored, including managers' inboxes and outboxes, active HITs and completed assignments, and intermediate inputs and outputs. This simplifies later analysis, since requesters can go back and query responses from each stage in the workflow. It also protects completed work from program errors or service outages; after crashes, execution simply resumes from the last good state.

EVALUATION

Our evaluation focused on PlateMate's feasibility as a replacement for traditional food logging. We considered three broad criteria:

1. **Accuracy.** How accurate were crowdsourced estimates compared to current alternatives? Could users trust them?
2. **Usability.** How much effort or discomfort would users experience in photographing food, uploading the photos, and correcting errors in PlateMate's estimates?
3. **Robustness.** How well does the PlateMate system fare with "real world" photographs?

We designed two experiments to answer these questions. In the first, nutrition data returned from PlateMate was compared with ground truth, expert dietitian estimates, and a recent commercial application. In the second study, ten participants used PlateMate and a manual food-logging system for four days.

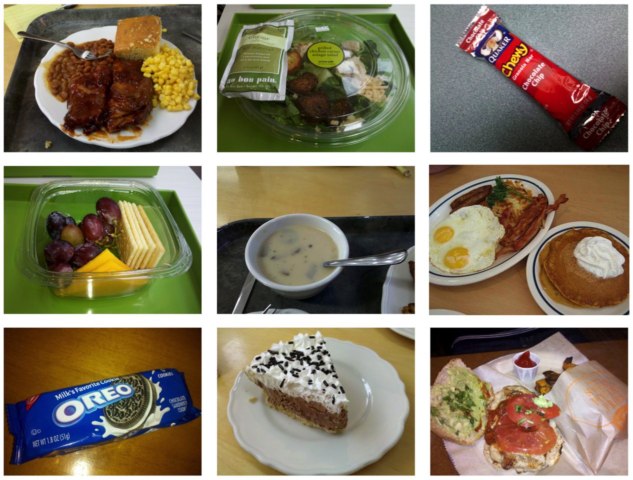


Figure 3: Examples of photos from the study of PlateMate's accuracy.

Evaluation of Accuracy

Our first study had two goals. The first was to determine the accuracy of PlateMate with ground truth data obtained from manufacturers or preparers. The second was to compare PlateMate's performance with two alternative approaches to remote food photography: analysis by experts and results from Meal Snap. Because Meal Snap only returns calorie information and to make the task manageable for our expert participants, we limited our comparison to estimated calories even though PlateMate generates reports that also include fat, protein, and carbohydrates.

Method We conducted the experiment with a sample of 18 photographs showing 36 distinct foods. Some depicted individual foods or packages, while others showed complex plates containing many items, as shown in Figure 3. Each pictured food had nutritional data available through the manufacturer or preparer, and foods were weighed when necessary to ensure accuracy. These foods were selected to span a variety of meals and sources, including restaurants, cafeterias, and grocery items. We also included a mix of simple foods and composite items like salads and sandwiches.

We recruited three professional dietitians to provide expert estimates: one was a private nutrition counselor, and the other two were employed by a hospital. They received compensation for their time and provided estimates from their own offices. They were encouraged to use any aids, like books and calorie databases, that they would typically use for a similar task.

Our third set of estimates came from Meal Snap, a recent commercial application. Meal Snap returns a range of calories rather than a definitive answer, so we used the mean of its high and low values.

Results In terms of mean absolute error on calorie estimates, PlateMate was not significantly different from the human experts or the Meal Snap application. Figure 4 illustrates the results in detail. As expected, trained dietitians

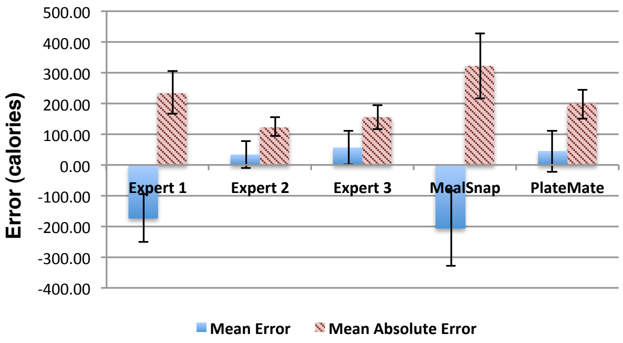


Figure 4: Mean errors (i.e., overall bias) and mean absolute errors (average magnitude of an error) for estimates made by the human experts, the Meal Snap application, and PlateMate compared to data provided by manufacturer or preparer. Error bars correspond to standard error.

were the most accurate on average. Their mean absolute error rates were 39.4%, 20.8%, and 26.1%, for an average of 172.0 calories or 28.7% per photograph. The best expert was off by just 124.5 calories, on average. PlateMate was close behind with a mean absolute error rate of 198 calories, or 33.2%. MealSnap was farther behind, with an average error rate of 322.8 calories or 53.9%.

Absolute error rates reflect the average magnitude of the error, but not the biases in each method. To understand how estimates from each source would add up over time, we also measured mean error without taking absolute values. The best expert overestimated by just 32.75 calories on average, for a mean error rate of +5.5%. The other two experts had error rates of +9.2% and −27.5%.

In comparison, PlateMate had a mean error rate of +44.1 calories, or +7.4%, which was much closer than Meal Snap's −34.4%. Expert and PlateMate results are significantly correlated with the ground truth data ($r^2 = .8626$, .9062, and .9378 for the experts, and $r^2 = .8622$ for PlateMate, all with $p < .0001$), while Meal Snap results were not correlated with the actual nutritional content of the meals ($r^2 = .2352$, $p = .3475$).

PlateMate's error rate compares favorably to amateur self-reports, where error rates can be greater than 400 calories/day and range from −76% to +24% [19, 4]. It also lacks the systematic bias towards underestimation in self-reports, especially among vulnerable users. These results indicate that PlateMate's answers, while imperfect, can be a useful nutritional guide.

Error Analysis Most errors in the study corresponded to single failures in specific parts of the pipeline. In the Tag stage, boxes were sometimes drawn improperly, leading to missing or duplicate identifications. In one photo of a brownie and banana on a small plate, only one box was drawn covering the entire banana and most of the brownie. As a result, the workers at the Identify stage omitted the brownie. On a photo of a hamburger with mushrooms, overlapping boxes were drawn over the burger and topping. In this case, the mushrooms were identified in both boxes.

Most errors occurred in the Identify stage. Turkers had trouble distinguishing similar types of a food, which sometimes had large nutrition differences. A plate of vegetarian baked beans was identified as regular baked beans, tripling the calorie count. Branded foods also caused problems: a relatively low-calorie chicken sandwich was identified as a sandwich from the restaurant Chili's, which had over twice as many calories. Another common situation involved duplication with both a composite item and one or more foods included in that composite both being selected. A slice of pizza with pepperoni and olives was identified as "Pizza with Meat and Vegetables," "Pepperoni," and "Black Olives," duplicating the toppings.

During measurement, many very small quantities were overestimated, especially when a small amount of a food was spread over a large area. A dash of parsley on a sandwich was overestimated as .27 cups, for example. Other errors occurred when one food appeared in several boxes. This led to a hamburger bun being counted as two buns when each half of the bun was seen in its own box.

User Study

Our second study looked at the subjective experience of using PlateMate as an end-to-end system for nutritional monitoring, compared to manual logging. We looked for insights about the system's usability, in terms of the inconvenience of taking photographs and the effort required to correct errors. Finally, we wanted to observe how robustly PlateMate functioned in the "real world," without any constraints on the types of photographs submitted to the system.

Method We recruited 10 participants (4 male, 6 female) via advertisements posted on several university email lists. Seven of the participants were undergraduates, two were graduate students, and one was a faculty member.

To help us evaluate the quality of the nutritional estimates generated by PlateMate and by the participants in this study, we recruited four dietitians employed at a local hospital. Two of them had also participated in the experiment evaluating the accuracy of PlateMate, where they produced the most accurate results. Participants and dietitians were compensated for their time.

Users were interviewed before and after the experiment. In the first interview, we discussed prior experiences tracking meals and trained participants on using the system. In the exit interview, we discussed their experiences using both logging methods.

During the study, we asked the participants to take photographs of their meals for four days and upload them to PlateMate once a day. For two of the days, participants received estimates generated by PlateMate and could correct those estimates. For the other two days, participants were not shown estimates and manually logged their food. Half of the participants used the manual system first and half used PlateMate first. We designed the interface for manual logging and

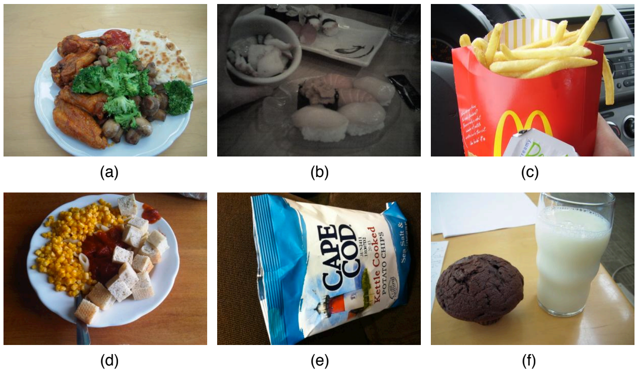

(a)
(b)
(c)
(d)
(e)
(f)

Figure 5: Example user photos. PlateMate handled (a) - (c) well, while (d) - (f) caused problems. In (d) the pasta with sauce in the middle was hard to see; in (e) there is no sense of scale to determine the size of the bag; and in (f) the type of milk is unclear.

correcting estimates to resemble existing commercial tools for manual food logging.

In order to assess the results produced by PlateMate compared to current consumer best practices, we used PlateMate to generate hidden estimates for participants' photos from the two days of manual recording. Participants only saw this data during the exit interviews, when they were asked to compare their own logging with the automatic estimates.

Findings from the Initial Interviews Our pre-interviews with participants confirmed that existing methods of logging food are cumbersome. All but one of the 10 had tried logging their food at some point, but most gave up after a few days or weeks. Only two participants still tried, and neither reported doing so consistently or reliably. They recalled their attempts to keep track of food as "annoying," "inconvenient," and "tedious." One subject recalled six separate failed attempts, each lasting just a few days. Some reported success when they were required to log for athletics or class projects, but recording lapsed when those commitments ended.

Despite these challenges, participants found nutritional information valuable. Eight participants reported looking at nutrition labels on packaged foods. Several reported looking up new foods out of curiosity. One participant reported revelations like "Oh wow, that's a lot of calories in dressing," and another now dilutes her juice with water after discovering how much sugar it contains.

Exit Interview Reactions We asked all ten participants their preference between using PlateMate for automatic estimates and logging manually (by any method). Seven participants said they would prefer using PlateMate in the future, citing its ease of use and the convenience of "having someone else do that for me rather than guess myself." One subject explained, "My answers were closer to guesswork; this felt more like science." The three subjects who did not prefer PlateMate felt that they could not trust it or that the process of taking photos and correcting the estimates was too cumbersome.

Subjects were divided in their perceptions of PlateMate's accuracy. Seven of 10 found the answers at least as good as their own and of these four found PlateMate's estimates to be more accurate than self reports. After seeing his own estimates and PlateMate's for the same meals, one subject said the exercise "confirmed my suspicions that you guys were more accurate than I was. The tendency is always to say 'oh, I didn't have that much."' Three others found their own estimates and PlateMate's basically equivalent.

The other three subjects all found PlateMate less accurate than their own estimates. One said that PlateMate's answers were close, "like 80-90%, but not perfect. I want to be sure." Another still preferred PlateMate even though she could not fully trust its results. She explained, "For some people if it's not perfect they'll never use it. But for me it was great...Even if it is only half of them correct, that is fewer I have to enter manually, and the happier I am." Another user disagreed, feeling that it took more effort to correct PlateMate estimates than "do it right myself the first time."

In total, seven users said that PlateMate required less effort than manual logging, which most of these users considered unpleasant and tedious. They called it "annoying", "boring," and "not excruciating but not insignificant either." These participants said PlateMate was "definitely easier" and "much simpler," concluding that it "definitely saves me time and effort." They also found receiving the results exciting. One user explained, "it was more fun...I got the email, and my friend was like, 'Oh! Do we get to see it now?"' Another was discouraged by the difficulty of manually estimating portions, so she found it "really helpful to have someone else do that for me rather than guess myself."

PlateMate Performance Next, we analyze PlateMate's performance on photographs collected by our users. We wanted to investigate the system's robustness given a broad variety of meals and realistic variations in photograph quality. Ideally, the PlateMate estimates and manual logging data from the user study could be compared to ground truth to determine accuracy, but such data were clearly not available.

Instead, we first looked at the differences between participant and PlateMate estimates. Comparing results from 112 photos for which we had both participant and PlateMate estimates, we found the two sets of results to be positively correlated ($r^2 = 0.62$, $p < .0001$). PlateMate's estimates were slightly higher than participants', with a mean difference of $+41.0$ calories (median $+18.8$) or $+11.5\%$ that was not statistically significant (Wilcoxon $z = 686$, *n.s.*).

To gain further insight into relative accuracies of PlateMate and our participants, we presented 50 of these photographs together with both sets of nutritional estimates to 4 professional nutritionists. The nutritionists worked in pairs. Each pair was presented with a photo and two sets of numbers representing total calories, protein, fat, and carbohydrates. One of these sets came from a participant and one from PlateMate, and the experts were blind to the source of the data. They were then asked to pick the more accurate set, taking as much time as necessary and using any necessary tools and references. The dietitians in each pair were allowed to talk to each other and could choose to agree on one data set as more accurate, disagree, or say they were unable to pick one data set as more accurate.

Of the 50 user study photos, the first pair could not decide which set of nutritional estimates was more accurate in 5 cases and the second pair could not in 12 cases. Out of the decisive photos, PlateMate data was selected as more accurate 44.4% and 47.4% of the time by the two pairs. These results suggest that neither method was obviously more accurate, especially since nearly half (49.2%) of photos had estimates within 100 calories of each other.

When disagreements did happen, PlateMate's estimates were larger 63.5% of the time. This is consistent with our finding in the first study that PlateMate slightly overestimates and prior research that suggests a strong bias in manual recording towards underestimation. [17, 6]. PlateMate's estimates for daily energy intake were +229.8 calories higher than self-reports on average, a difference equivalent to four Oreo cookies every day.

Error Analysis Many of the errors seen from the user study results were similar to those already discussed from the ground truth study, but some new issues emerged. In measurement, we saw difficulty estimating portions when extreme close-up photos were taken with no sense of scale. Turkers could not agree if a bag of potato chips (Figure 5) was a portion or large bag. Scale was also a problem in identification: a small tangerine was identified as a larger orange. Other identification errors occurred when foods with nearly correct names but vastly different nutrition were selected, like "grapefruit juice" and "juice concentrate," which has eight times the calories. One Subway chicken sandwich was identified as "Subway Roasted Chicken Patty," which could be interpreted as the whole sandwich but in fact just contained the chicken.

Human errors during manual logging mostly occurred when participants forgot to log a certain food. In photo (f) of Figure 5 the participant only recorded the milk and forgot to log the muffin, which represented most of the photo's calories. In a photo of french fries, a participant forgot to record the dipping sauce next to the fries. Similar errors occurred when participants sought to simplify their recordings to save time. One subject ate a bowl of several types of fruit but recorded the entire bowl as raspberries, while PlateMate correctly identified each fruit.

Cost and Wait Times During the course of both evaluations we analyzed 262 photos using PlateMate, generating 1,553 HITs that were assigned to 199 total Turkers 4,332 times. The average cost of a single photo was $1.40. The mean time to complete analysis was 94.14 minutes, with 73% of photos completing in less than 2 hours and all photos completing in less than 6 hours.

DISCUSSION AND FUTURE DIRECTIONS

The results from our evaluations of PlateMate suggest that through careful coordination, untrained workers can approach experts in their accuracy in estimating nutrition information from photographs of meals. These estimates are close to those logged by the people who actually ate the meals. However, several issues which became apparent during the course of our evaluations could be addressed through future work.

PlateMate consistently struggled to produce good results on liquids like beverages and salad dressing. One participant drinks a low-fat latte each morning, but PlateMate consistently identified it as coffee with cream. Another only used low-fat salad dressings, which were identified as their full-fat versions. These issues could be addressed by introducing personalization mechanisms. For example, the interface could give users access to images of foods they eat frequently—instead of taking a picture of today's latte, a user would simply select a picture of last week's, ensuring correct logging and obviating the need for engaging the crowd. Statistical methods could also be used to adapt the Turker interface to emphasize the foods most common in a user's diet and thus most likely to appear in their photos. These approaches could result in improvements to both reliability and cost.

Geolocation capabilities available in many mobile devices could be used to further improve accuracy of the crowd-sourced analysis of restaurant meals. Photos could be annotated with the cuisine of the restaurant in which they were taken, providing Turkers with helpful context while maintaining the privacy of user's actual location. Integrating with existing local "check-in" applications like Foursquare[4] would make it even simpler to associate meals with their places of origin.

Permitting optional textual annotations by users (e.g., "skim latte", "mango curry") would naturally further improve accuracy and reduce cost. So would employing computer vision and machine learning for parts of the process: over time and continued use, PlateMate could build a large database mapping tagged sections of photographs to specific foods and portions. An algorithmic approach could be taken to analyze new photos for similarity with previously processed images. This could result in fully computerized analysis based on the prior crowdsourced work, extending the vision approach in [7], or these potential similar items could be surfaced in alternate HIT interfaces to Turkers as a way of skipping unnecessary stages of the PlateMate process.

This work was done on the assumption that lowering the barrier to monitoring one's food intake may result in a larger number of people persisting in their attempts to alter their eating habits. We are aware, however, that making the process too easy may reduce the opportunities for reflection. Ultimately, PlateMate's success depends on the users' willingness to engage with the information it provides. But if they do, PlateMate can help its users correct misconceptions about nutritional content of the foods they consume and to improve their ability to estimate portion sizes.

The Remote Food Photography Method relies on two images of each meal: a photograph of the original portion and a photograph of any food that was left uneaten. We have explored the first part of the process and we expect that the second can

[4] https://foursquare.com/

be performed in a similar manner. A major difficulty in analyzing images of leftovers is likely to be in identifying the foods in the photo. But as such foods are already identified in the first part of the process, a reasonable approach to extend PlateMate may be to display an annotated image of the original plate next to the photograph of the leftovers, and ask Turkers to identify portions of the second image where the original foods are present and, in the subsequent step, to estimate the amounts of the leftover foods. Our future work will aim to test the efficacy of such an approach and to thus fully implement the Remote Food Photography Method.

CONCLUSION

This paper presents PlateMate, which allows users to take photos of their meals and receive estimates of the meals' nutrition content. PlateMate builds on a concept of remote food photography developed recently by the nutrition community. While the original method relies on expert dietitians providing the estimates, PlateMate uses Amazon Mechanical Turk to make this approach more affordable and scalable.

Through careful decomposition of the process into small and verifiable steps, PlateMate achieves accuracy comparable to trained dietitians: the results of our evaluation demonstrate that PlateMate overestimated caloric content by $+7.4\%$ on average, while the best of three trained dietitians overestimated by $+5.5\%$. In our user study, which compared PlateMate to the currently most common practice of manual self-logging of meals, most participants found PlateMate easier and faster to use and at least as accurate. Four dietitians were unable to differentiate between nutrition estimates produced by PlateMate and those manually logged by our study participants, further suggesting parity with the current methods.

Overall, PlateMate is an attractive alternative to existing solutions because it reduces user effort compared to manual logging, achieves good accuracy, is affordable, and can be conceivably deployed to support a large number of users.

We suggest ways in which the accuracy can be further improved and cost reduced by combining crowdsourcing with machine learning, computer vision, personalization and location information.

PlateMate is one of the first complex crowdsourcing systems to combine—in a real world application—several of the recently introduced design patterns for programming the crowds. In the process of building PlateMate, we have developed the Management framework, a modular software framework inspired by the structure of human organizations. The Manager abstraction conveniently supported hierarchical problem decomposition as well as modular development and debugging. The choice of message passing as the main communication mechanism cleanly supports asynchronous just-in-time processing of sub-tasks. PlateMate may serve as a useful case study for future developers of complex crowd-based applications.

REFERENCES

1. Amft, O., Stäger, M., Lukowicz, P., and Tröster, G. Analysis of chewing sounds for dietary monitoring. In *Ubicomp*. Springer, 2005, volume 3660 of *Lecture Notes in Computer Science*, 56–72.
2. Ann Yon, B., Johnson, R., Harvey-Berino, J., and Gold, B. The use of a personal digital assistant for dietary self-monitoring does not improve the validity of self-reports of energy intake. *Journal of the American Dietetic Association*, *106*, 8 (2006), 1256–1259.
3. Bernstein, M., Little, G., Miller, R., Hartmann, B., Ackerman, M., Karger, D., Crowell, D., and Panovich, K. Soylent: a word processor with a crowd inside. In *Proc. UIST'10*. ACM, 2010, 313–322.
4. Champagne, C., Bray, G., Kurtz, A., Monteiro, J., Tucker, E., Volaufova, J., and Delany, J. Energy intake and energy expenditure:: A controlled study comparing dietitians and non-dietitians. *Journal of the American Dietetic Association*, *102*, 10 (2002), 1428–1432.
5. Dinkins, J. Beliefs and attitudes of americans towards their diet. US Department of Agriculture Center for Nutrition Policy and Promotion (2000).
6. Goris, A., Westerterp-Plantenga, M., and Westerterp, K. Undereating and underrecording of habitual food intake in obese men: selective underreporting of fat intake. *The American journal of clinical nutrition*, *71*, 1 (2000), 130.
7. Kitamura, K., de Silva, C., Yamasaki, T., and Aizawa, K. Image processing based approach to food balance analysis for personal food logging. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2010, 625–630.
8. Kittur, A., Smus, B., and Kraut, R. *CrowdForge: Crowdsourcing Complex Work*. Technical Report CMU-HCII-11-100, CMU (2011).
9. Little, G., Chilton, L., Goldman, M., and Miller, R. Turkit: human computation algorithms on mechanical turk. In *Proc. UIST'10*. ACM, 2010, 57–66.
10. Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. Exploring iterative and parallel human computation processes. In *Proc. ACM SIGKDD Workshop on Human Computation*. ACM, New York, NY, USA, 2010, HCOMP '10, 68–76.
11. Locke, E. and Latham, G. Building a practically useful theory of goal setting and task motivation. *American Psychologist*, *57*, 9 (2002), 705–717.
12. Mankoff, J., Hsieh, G., Hung, H. C., Lee, S., and Nitao, E. Using low-cost sensing to support nutritional awareness. In *Proc. UbiComp '02*. Springer-Verlag, London, UK, 2002, 371–376.
13. Mann, T., Tomiyama, A., Westling, E., Lew, A., Samuels, B., and Chatman, J. Medicare's search for effective obesity treatments: Diets are not the answer. *American Psychologist*, *62*, 3 (2007), 220.

14. Marcus, A., Wu, E., Karger, D., Madden, S., and Miller, R. Crowdsourced databases: Query processing with people. In *CIDR*. 2011.

15. Martin, C., Anton, S., York-Crowe, E., Heilbronn, L., VanSkiver, C., Redman, L., Greenway, F., Ravussin, E., and Williamson, D. Empirical evaluation of the ability to learn a calorie counting system and estimate portion size and food intake. *British Journal of Nutrition*, *98*, 02 (2007), 439–444.

16. Martin, C., Han, H., Coulon, S., Allen, H., Champagne, C., and Anton, S. A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method. *British Journal of Nutrition*, *101*, 03 (2009), 446–456.

17. Pikholz, C., Swinburn, B., and Metcalf, P. Under-reporting of energy intake in the 1997 national nutrition survey. *The New Zealand Medical Journal*, *117*, 1202 (2004).

18. Reisner, R. and Thompson, D. The diet industry: A big fat lie. *BuisinessWeek* (2008).

19. Schoeller, D., Bandini, L., and Dietz, W. Inaccuracies in self-reported intake identified by comparison with the doubly labelled water method. *Canadian journal of physiology and pharmacology*, *68*, 7 (1990), 941.

20. Surowiecki, J. *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday, 2004.

21. Yon, B., Johnson, R., Harvey-Berino, J., Gold, B., and Howard, A. Personal digital assistants are comparable to traditional diaries for dietary self-monitoring during a weight loss program. *Journal of behavioral medicine*, *30*, 2 (2007), 165–175.