Biased Bytes: On the Validity of Estimating Food Consumption from Digital Traces

KRISTINA GLIGORIĆ, EPFL, Lausanne, Switzerland
IRENA ĐORĐEVIĆ*, University of Niš, Niš, Serbia
ROBERT WEST, EPFL, Lausanne, Switzerland

Given that measuring food consumption at a population scale is a challenging task, researchers have begun to explore digital traces (e.g., from social media or from food-tracking applications) as potential proxies. However, it remains unclear to what extent digital traces reflect real food consumption. The present study aims to bridge this gap by quantifying the link between dietary behaviors as captured via social media (Twitter) vs. a food-tracking application (MyFoodRepo). We focus on the case of Switzerland and contrast images of foods collected through the two platforms, by designing and deploying a novel crowdsourcing framework for estimating biases with respect to nutritional properties and appearance. We find that the food type distributions in social media vs. food tracking diverge; e.g., bread is 2.5 times more frequent among consumed and tracked foods than on Twitter, whereas cake is 12 times more frequent on Twitter. Controlling for the different food type distributions, we contrast consumed and tracked foods of a given type with foods shared on Twitter. Across food types, food posted on Twitter is perceived as tastier, more caloric, less healthy, less likely to have been consumed at home, more complex, and larger-portioned, compared to consumed and tracked foods. The fact that there is a divergence between food consumption as measured via the two platforms implies that at least one of the two is not a faithful representation of the true food consumption in the general Swiss population. Thus, researchers should be attentive and aim to establish evidence of validity before using digital traces as a proxy for the true food consumption of a general population. We conclude by discussing the potential sources of these biases and their implications, outlining pitfalls and threats to validity, and proposing actionable ways for overcoming them.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; **Social media**; • **Applied computing → Health informatics**; • **Information systems → Data mining**; **Social networking sites**; • **General and reference → Empirical studies**; **Validation**.

Additional Key Words and Phrases: biases, validity, social media, food, images, Twitter, crowdsourcing

ACM Reference Format:
Kristina Gligorić, Irena Đorđević, and Robert West. 2022. Biased Bytes: On the Validity of Estimating Food Consumption from Digital Traces. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 497 (November 2022), 27 pages. https://doi.org/10.1145/3555660

1 INTRODUCTION

Diets determine health. Eating healthy helps prevent malnutrition as well as a range of diseases and conditions, including diabetes, heart disease, stroke, and cancer [46]. In order to be able to improve diets, researchers and stakeholders need to know what foods people consume, but monitoring diets

*Work done during an internship at EPFL.

Authors' addresses: Kristina Gligorić, EPFL, Lausanne, Switzerland, kristina.gligoric@epfl.ch; Irena Đorđević, University of Niš, Niš, Serbia, irenadj@elfak.rs; Robert West, EPFL, Lausanne, Switzerland, robert.west@epfl.ch.

2573-0142/2022/11-ART497 $15.00
https://doi.org/10.1145/3555660

at a population scale is challenging. Traditionally, nutritional studies rely on survey-based methods [24] employing questionnaires [105] and personal food journals [29, 30, 90], which are prone to biases, most notably social and cognitive biases, such as false recall and social desirability bias [15]. Traditional methods are also costly to organize. Researchers and practitioners might not have access to extensive surveying, and it might be hard to collect reliable statistics—even though a large and ever-growing [79] portion of the population has access to advanced technology including smartphones with Internet access [13].

In light of the challenges of traditional methods on the one hand, and the opportunities afforded by widespread Internet access on the other hand, there is great promise in using passively collected digital data to estimate food consumption. Digital datasets are unmatched in terms of scale and immediacy [66, 87], do not rely on self-reports, and do not suffer from biases typical of traditional methods. Given this potential, researchers have been developing and applying their expertise to studying diets via passively collected digital data, whose tremendous potential for providing insights into food consumption has been showcased numerous times [2, 26, 89].

The promises of Web and social media data notwithstanding, important methodological questions remain: Are researchers measuring what they aim to measure? Do digital traces reflect actual food consumption? Do effects estimated from online signals hold in the offline world? Are predictive models trained on online signals accurate in the offline world? In other words, the validity of studying diets with digital data remains opaque.

Online data is not primarily collected with scientific studies in mind and is therefore sometimes referred to as "found data" [87]. Found data overcomes several of the biases typical of traditional methods, but may introduce new biases that threaten validity in their own ways [57, 74, 100]. Despite their potential, error-prone and unreliable data and methods may do more harm than good if handled without the required caution [28]. As our community increasingly relies on large-scale digital data sources, methods that offer insights into the validity of new measures thus become increasingly necessary [57].

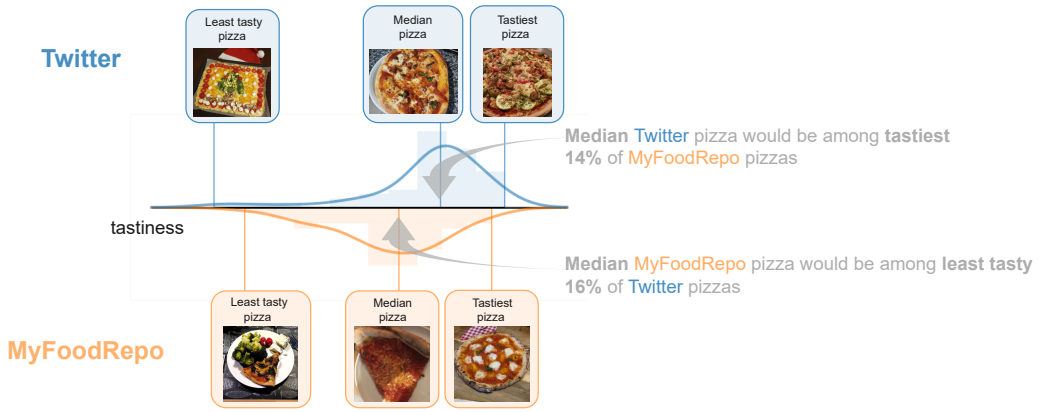


Fig. 1. **Illustration of bias in perceived tastiness.** Perceived tastiness of tweeted food (*top*) vs. actually consumed and tracked food (*bottom*) of type "pizza". Histograms summarize tastiness scores estimated in our crowdsourcing framework. As illustrated, tweeted pizzas are perceived as considerably tastier than actually consumed and tracked pizzas.

1.1 Research questions

In this spirit, our overall research question asks: *Is social media a biased or a truthful mirror of actual food consumption, as measured via food tracking?* We focus on those dietary aspects in particular that researchers frequently study with social media: food type [34], nutritional properties [2], and appearance and subjective perception [65]. In order to establish a link between online and offline dietary behaviors at a population scale, we study images of food that people consumed and tracked via a food-tracking application, and contrast them with images of food posted on Twitter, addressing the following specific research questions:

RQ1 *Bias of food-type distribution:* To what extent do food images posted on Twitter reflect the types (beef, bread, burger, etc.) of actually consumed food, as measured via food tracking?

RQ2 *Biases within food types:* For a given food type, to what extent do food images posted on Twitter reflect the nutritional properties, perceived tastiness, and appearance of actually consumed and tracked food of that type?

In order to address RQ1, we investigate whether food images posted on Twitter are a faithful reflection of the types of actually consumed and tracked food or not. *A priori,* one might envision the following potential outcomes:

a) "Food images posted on Twitter are a faithful reflection of the types of actually consumed and tracked food, consistent with the demonstrated potential of Twitter to provide insight into dietary choices [2]."

b) "Food images posted on Twitter are not a faithful reflection of the types of actually consumed and tracked food, given a variety of challenges in the practices of social media use for research [74]."

In order to address RQ2, we investigate whether or not food images posted on Twitter are a faithful reflection of actually consumed and tracked food in terms of how healthy, caloric, and tasty-looking the food is. We investigate whether the two sources diverge, and if so, in what direction. *A priori,* one might envision the following potential outcomes:

a) "Food images posted on Twitter are a faithful reflection of actually consumed and tracked food in terms of how healthy, caloric, and tasty the food is, consistent with the demonstrated potential of Twitter to provide insight into dietary choices [2]."

b) "Tweeted food is healthier, less caloric, and less tasty than consumed and tracked food. Social media is increasingly used to promote trendy ingredients and recipes, and clean and healthy eating [25]. Social media inspires and connects people interested in healthy eating [67]."

c) "Tweeted food is less healthy, more caloric, and tastier than consumed and tracked food, consistent with a documented fetishization of food online. Users share appetizing pictures of culinary experiences where exaggerated foods such as sugary desserts dominate over more standard local cuisines [65]."

1.2 Contributions

To the best of our knowledge, ours is the first attempt to investigate the link between online and offline dietary behaviors by studying food images as measured via two platforms, in our case Twitter and the MyFoodRepo[1] food-tracking app. We design and apply a novel crowdsourcing framework for estimating biases (Sec. 3), and we perform a case study of food consumption in Switzerland (Sec. 4). Controlling for location, period, and food types, we contrast an extensive set of tweeted food images with images of consumed and tracked food.

[1]https://www.myfoodrepo.org/

We find that food type distributions among social media foods vs. among consumed and tracked foods diverge (RQ1, Sec. 4.1). Controlling for the discrepant food-type distributions by studying food types individually (RQ2, Sec. 4.2), we find that Twitter still provides a biased view of food consumption as measured via food tracking. Tweeted food is, on average across food types, perceived as more caloric, less healthy, less likely to have been consumed at home, and tastier (example in Fig. 1), compared to actually consumed and tracked food. For example, on average across food types, a median-tasty Twitter dish is among the top 26% tastiest MyFoodRepo dishes, and a median-caloric Twitter dish is among the top 34% most caloric MyFoodRepo dishes. While social media traces can be a reasonable proxy of tracked consumption for certain foods types (Fig. 5), we find that, overall, food shared on social media and consumed and tracked food significantly diverge from each other (Fig. 4a and 5, Table 1).

We discuss the relationship between three distributions: all foods consumed by the general population, food consumption estimated via MyFoodRepo, and food consumption estimated via Twitter. The fact that there is a divergence between food consumption measured via the two platforms—food tracking and social media—implies that at least one of the two is not a faithful representation of true food consumption in the general Swiss population. We argue that it is less likely that food tracking is the main source of bias, and we conclude that researchers should be attentive and try to establish evidence of validity before using digital traces as a proxy for the true food consumption in the general population.

Measuring biases in digital traces is the first step towards correcting them and drawing valid conclusions despite their presence [101]. Through a case study of the Twitter and MyFoodRepo platforms in Switzerland, contrasting tweeted food images with consumed and tracked foods, we provide grounding and first insights by controlling for location, period, and food types. Our findings cannot, however, be assumed to generalize globally, and future work should apply our framework to other populations, other social media platforms and Web traces, and other food tracking apps. Our study may serve the purpose of a "proof by counterexample": we have identified one common setting where there is a bias between two types of digital trace data. Hence, we should assume that there can be bias in other populations and platforms, too.

We conclude the paper with a discussion (Sec. 5) of how the methods and findings reported here can inform researchers in their efforts to leverage digital traces for various applications, in the context of food and beyond.

2 BACKGROUND AND RELATED WORK

2.1 Estimating food consumption from digital traces

We start by reviewing related work leveraging social media to study nutrition and dietary behaviors. Studying diets through social media posts has been an active area of CSCW research. Instagram [41, 72, 78, 93] and Twitter [2, 34, 63, 64, 66, 104] have emerged as particularly promising platforms. Researchers have studied specific dietary issues and harmful behaviors. In particular, in work with important implications for the health and well-being of vulnerable populations, researchers studied reports of eating disorders [19, 75], dietary choices, nutritional challenges in food deserts (places with poor access to healthy and affordable food) [34], and obesity patterns in online behaviors [65]. Related work has also studied eating disorder support online communities, quantifying and predicting disease severity and recovery [20, 32].

Although social media has emerged as a rich data source, food shared or discussed on Instagram and Twitter might not be representative of food that people actually consume. Researchers have compared, at a population scale, statistics extracted from tweet text with public health statistics

regarding the prevalence of obesity and diabetes [2, 65, 86]. However, the content of posted images and the foods themselves have not been contrasted with actually consumed foods to date.

Beyond social media, researchers have long been applying their expertise to analyze health and nutrition behaviors using other kinds of digital trace data. First, researchers monitor food consumption with smartphone tracking applications and wearables [3, 12]. Researchers analyzed compliance and contextualization of such platforms by investigating perceived and true snacking and meal consumption [11] and the potential for inferring population-level eating routines [42].

Second, more distant proxies were previously used to analyze nutrition behaviors, including search engine logs [43, 97, 103], purchase logs [6, 7, 17, 35, 44, 53, 54], online recipes [81, 86, 95, 96, 98, 99], reviewing platforms and websites [23, 31, 48, 80, 102], crowdsourcing platforms [36, 52], and geolocation signals [83]. While food shared on social media might not be representative of consumed food, the above-listed, more distant proxies make it even harder to determine validity. For example, do recipe searches on search engines correspond to eating the food? Does reading an online recipe imply that the food was prepared and consumed? It is unknown to what extent such proxies imply food consumption, and it is not clear whether studies of food consumption via such digital traces truly measure the quantities intended to be measured.

In more distantly related research, researchers have been utilizing user-generated food content to train and develop machine learning models. Current AI applications that use online food images include mining food photos to perform segmentation [73], recognize food [9, 14, 85, 106], learn food and recipe embeddings [88], and perform calorie [71] and nutrient [39] estimation. However, if the food that people consume is systematically different from food shared online, models trained and evaluated on online datasets might not generalize to real-world scenarios.

2.2 Biases of studying digital traces

Next, we review related work studying biases of digital traces. The goal of measurements using behavioral trace data is to extract meaning from raw data that most often was not collected with the extraction of scientific insight in mind. Data-driven research has thus been criticized for asking questions that appear to be opportunistically answerable with the data at hand, overlooking different types of biases [45].

Lazer et al. [57] argue that the digital traces need to be linked to known constructs before we can use the data to answer scientific questions. Thus, the key challenge of studying digital data is determining whether measurements accurately capture the construct that one would ideally want to examine. For example, if one is measuring physical activity based on mobile phone location traces, how consequential is the omission of stationary activities such as treadmill or yoga [57]? If one is tracking influenza with Web search logs of symptoms, how consequential are searches from persons not experiencing any symptoms [58]?

The mismatch between the theoretical understanding of a concept and its operationalization, known as the issue of construct validity, can have harmful consequences [100]. In particular, when data that allows for measurement (e.g., arrest records) does not properly match the actual social construct that the measurement is intended to capture (e.g., a criminal act), measurements can replicate, mask, or exacerbate existing social issues [28].

Related work has thus aimed to establish the validity of studying human behaviors with Web and social media traces. Example studies include studying the validity of screening depression [55], location traces [51], inferring political approval [91], sentiment analysis [76], or using Twitter's APIs [70]. De Choudhury et al. [33] have studied seeking and sharing health information online by comparing search engines and social media. Researches have also studied decisions around whether to post content online [1], political, racial and gender biases in Web systems [47, 56], and how Web systems influence offline user behavior [8].

Further related work includes studies that issue calls to carefully scrutinize the use of social media data against biases and provide practical advice to aid researchers in performing their data-driven studies. Sen et al. [92] proposed a total error framework for digital traces of human behavior on online platforms, Olteanu et al. [74] identified a variety of challenges in the practices of social media use for research, and Hofman et al. [50] advocated for measuring the extent to which causal estimates made in one domain transfer to another domain. Whereas related work [74, 82, 92] aims to put the biases into a unified framework cutting through different domains, we aim to specifically establish the validity of estimating food consumption from digital traces.

3 DATA AND METHODS

3.1 Food tracked via MyFoodRepo

To get as close as possible to capturing true food consumption, we use a novel dataset of food images collected via the MyFoodRepo mobile app [69]. By design, the food present in these images was actually consumed, for the purpose of the application is to track users' personal food consumption. Through the app, volunteer users from Switzerland are asked to provide images of their complete daily food intake, mainly in the context of being enrolled in a digital cohort called Food & You [37].[2] MyFoodRepo thus captures all foods that compliant individuals consume, in any context.

The images are publicly available as part of the Food Recognition Challenge.[3] The dataset has been annotated such that the individual foods are mapped onto an ontology of food types. Images were logged between 2017 and 2020. In our analyses, we study the training-set portion of the dataset, comprising 24,120 images, along with their corresponding 39,328 food-type annotations.

3.2 Food shared on Twitter

To answer the question of whether images shared on social media diverge from food consumption as measured via food tracking, we aim to contrast images of consumed and tracked food with images of food posted on social media. To this end, we curate a dataset of food images shared on Twitter in Switzerland during the same period spanned by the images collected via the food tracking app, this way controlling for location and time.

Since our goal is to investigate the validity of studying diets with social media, in our Twitter data collection strategy, we, first, aim to follow data collection methods present in the existing literature closely, to be able to make conclusions that can be relevant for researchers working in this area, as opposed to inventing novel strategies that would be less relevant. Our data collection pipeline is therefore similar to pipelines described in related work, extracting nutritional information from social media posts with keywords (Sec. 2). Note that, since we follow existing work, specific data collection decisions are not limitations per se. Instead, the impact of data collection based on user-specified keywords is intended to be measured, since this is how researchers usually collect Twitter posts to study food consumption.

Second, we aim to gather a complete dataset, i.e., to collect all food images posted on Twitter by the relevant population in the relevant time frame. To this end, we use the full-archive search endpoint, available to researchers via Twitter's Academic Research product track,[4] which allows searching Twitter's complete archive going back until March 2006.

Third, we aim to find images posted on Twitter that actually contain food, as we are interested in studying the posted food itself, rather than only how it is described. To this end, we apply

[2]https://www.digitalepidemiologylab.org/projects/food-and-you
[3]https://www.aicrowd.com/challenges/food-recognition-challenge
[4]https://developer.twitter.com/en/docs/twitter-api/tweets/search/

automated and manual annotation of the collected images. With these three goals in mind, we employ the following data collection pipeline.

Step 1: Twitter data collection.. We start from the set of MyFoodRepo image annotations (e.g., "bread", "banana"). We remove drinks and merge small types that are similar, obtaining 155 food types. We map each type to suitable handcrafted high-precision keywords, translate the keywords from English to German, French, and Italian (the large Swiss national languages) via Google Translate, and use the disjunction ("OR") of keywords (separately per language) to query Twitter's full archive search API for the respective food type. We thus obtain all posts that in the text contain at least one of the keywords related to the food, in one of the four languages, in either singular or plural form (if relevant). For example, for the type "bread", we retrieve all English tweets containing "bread" or "breads", French tweets containing "pain" or "pains", Italian tweets containing "pane" or "pani", and German tweets containing "Brot" or "Brote". Additional restrictions ensure that tweets were posted between 2017 and 2020 (the period when images of MyFoodRepo food were logged) from a location in Switzerland and contain at least one image. This step yields 33,425 unique images.

Step 2: Automated annotation.. We are interested in studying the foods themselves, so we make sure that images indeed contain food. To that end, we perform detection of food in images with the ResNet50 model trained on ImageNet [49], which we finetuned for food-vs.-not-food classification on the publicly available Food-5K food image dataset [94], with 98% recall and 96% precision on the task of detecting food images on the held out 20% test set (using a threshold of $p = 0.5$). Inspection of the images revealed that the images that do not contain food most frequently occur in food types where keywords have homonymous meanings. Two food types with the largest fraction of images that do not contain food are "date" (which can signify a fruit or "day of a year" or "social appointment") and "apple" (which can signify a fruit or Apple Inc. and its products). After this step, we keep 7,723 tweets with images that contain food.

Step 3: Manual annotation.. We manually inspect the images to verify that an image contains the food that the user mentions in the tweet text, even if a small quantity. The visible food item needs to be edible, e.g., a silver pendant of lemon shape or a carved and decorated Halloween pumpkin does not qualify. Additionally, the image needs to contain a prepared dish, and not all the ingredients laid out separately, nor an uncooked caught fish. Finally, no explicit content can be present in the background for the image to be safe for crowd workers.

Due to the completeness of Twitter's full archive search and the manual inspection of collected images, at the end of the above process, we obtain all tweets posted from Switzerland between 2017 and 2020 with images that contain a food that is mentioned in the tweet text, for a total of 3,692 images of food along with their corresponding 4,481 food-type annotations.

In summary, the two datasets we analyze contain 24,120 images of *consumed and tracked food* and 3,692 images of *tweeted food*. Images are mapped on the food-type level and contain foods that we can compare in order to address our research questions. See Fig. 2 for examples of images of type "pizza".

Having described the data, we continue by outlining our crowdsourcing framework for measuring biases. We then describe how we implement this framework on Amazon Mechanical Turk.

3.3 Crowdsourcing framework for estimating biases

Beyond food types, previous work (Sec. 2) has most notably used social media to estimate nutritional properties of food [2], as well as its appearance and perception [65]. Based on these themes, we operationalize four pertinent dimensions along which we contrast tweeted and consumed and

tracked food, capturing how (1) healthy, (2) tasty, (3) caloric, and (4) likely to have been consumed at home the food is.

For each dimension, we aim to estimate a score for each image. Contrasting the scores of tweeted vs. consumed and tracked food then allows us to assess biases. In principle, we could obtain scores directly via human annotation by asking, e.g., "How tasty does this dish look, on a scale from 1 to 10?" It is, however, challenging for humans to place items on an interval scale that is consistent across individuals [5]. Based on the fact that judging between two alternatives is generally easier and more intuitive for humans [21], we instead adopt a pairwise paradigm, where we confront human raters with pairwise choices (e.g., "Which of these two dishes looks tastier?") and later infer latent scores from the pairwise preferences.

Consider a given dimension (we use tastiness for concreteness in the following exposition) and a given food type. Then, for two images a and b showing food of the same type, we use the notation "$a \succ b$" to express that a is preferred over b by a human rater. Note that human preference is a random variable: different raters may have different preferences with respect to a given pair. We assume, however, that certain images show inherently tastier dishes and are thus more likely to be preferred. More formally, following the Bradley–Terry (BT) model [16], we assume that each image i has a latent tastiness score $s(i)$ and that the probability that a rater will prefer image a over image b [image b over image a] in a pairwise comparison is proportional to the score of a [score of b]:

$$\Pr(a \succ b) = \frac{s(a)}{s(a) + s(b)}. \tag{1}$$

Given this setup, maximum likelihood estimation [62] can be used in order to infer the latent scores that best explain the empirically observed pairwise preferences. Thus, although only pairwise choices are made by humans, we can rank all images in a total order based on their latent scores s. The BT model is appropriate for our purposes, as it has a well-understood interpretation and is well-suited to model human preferences [21]. In practice, we fit a so-called Plackett–Luce model [62], a generalization of BT that does not require comparisons for all image pairs.

3.4 Implementation on Amazon Mechanical Turk

In the remainder of this section, we describe how we implemented the above-described framework on Amazon Mechanical Turk. To estimate the latent scores of images of MyFoodRepo and of Twitter food in terms of the four dimensions, we selected 24 well-represented food types (Fig. 5a). The types are selected such that each type has at least 50 tweeted and at least 50 consumed and tracked food images. The different food types are considered as independent "tournaments", so we obtained a separate ranking per type.

We first sampled the same number of tweeted food images and consumed and tracked food images per type, to account for potentially different food-type frequencies. Recall that location and period are already controlled for in the data collection. We randomly sampled 100 images from each type, 50 tweeted images and 50 images of consumed and tracked foods, resulting in 2,400 competing images in total.

We then performed random sampling of comparison pairs. From each set of images for a given food type, we sampled N pairwise comparisons, each time randomly sampling one image of consumed and tracked food and one image of tweeted food, constrained such that each image participates in the same number of comparisons. We chose the number of "duels" per food type based on rank inference simulations, with the goal of ensuring that we can infer true ranks accurately, as follows. We assumed 100 items divided into two groups with item quality sampled from the standard normal distribution. We ranked the items and then randomly sampled N comparisons between items from the two groups. We sampled the outcome of a duel based on the items' quality scores

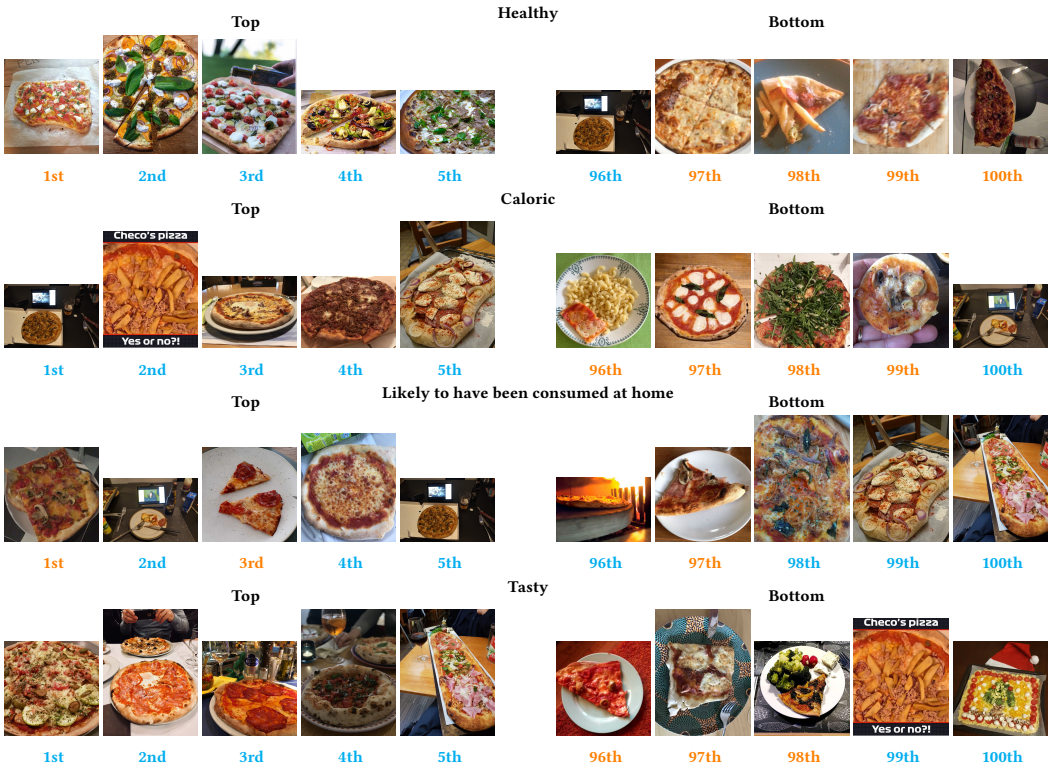


Fig. 2. **Example food images of type "pizza".** The overall top five (rank 1st–5th, *left*) and overall bottom five (rank 96th–100th, *right*) images with respect to estimated rank according to four criteria (one criterion per row). Twitter foods are marked in blue and MyFoodRepo foods, in orange.

and estimated quality and rank with a BT model (Eq. 1). Fig. 3 depicts how well the true ranking can be recovered for different numbers N of comparisons. As more comparisons are performed, the estimated rank (y-axis) correlates more strongly with the ground-truth rank (x-axis). Based on these results, we chose to perform 10 comparisons per image ($N = 500$, Kendall's $\tau = 0.80$). That is, at $N = 500$, each of the 50 images is compared to 10 competitors, and rank can be accurately inferred with Kendall's $\tau = 0.80$.

In every rating task, a participant was shown a random pair of images containing food of the same type (images in a pair were scaled to the same size and shown in randomized order), and asked to give a preference label for each of the four dimensions (healthiness, tastiness, caloric content, likelihood to be consumed at home). The pairwise comparison task had no neutral option; participants were required to choose one image. As the order within pairs was randomized, this is a valid way of breaking ties, and recommended practice [77]. Additionally, we asked participants to explain how they perceived both images by providing between one and three free-form tags (e.g., "dull", "greasy"). (Prior to data collection, we did not make hypotheses about specific biases as revealed by the tags, but rather explore them post-hoc in order to gain insights about how people describe the appearance of tweeted vs. consumed and tracked food.) In total, we collected 12,000 pairwise comparisons (500 duels for each of 24 types) for each of the four dimensions.

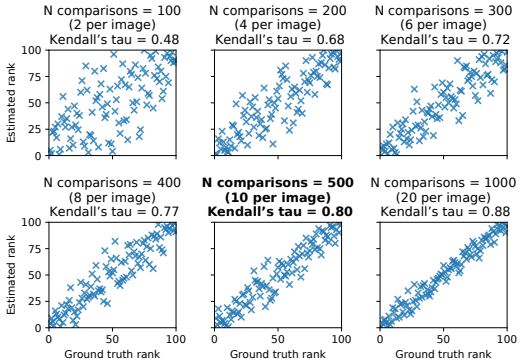


Fig. 3. **Determining a sufficient number of comparison pairs via rank reconstruction simulations.** Ground truth rank (x-axis) and estimated rank (y-axis), for varying number of comparisons (N). In our subsequent analysis, we chose to perform 10 comparisons per image ($N = 500$).

3.4.1 Participants. Since the tasks require reading and writing text in English, participants were restricted to those residing in the United States, Canada, or the United Kingdom. To ensure high-quality answers, we admitted only workers with approval rates greater than 99% and with more than 1,000 previously approved tasks. We collected the 12,000 pairwise preferences through 24 batches with 500 assignments each, over the course of five days. The task was performed by 595 distinct workers, who performed 20.2 pairwise comparisons each, on average.

3.4.2 Compensation. We targeted a pay rate of $9 per hour. Participants were paid $0.15 per pairwise comparison. The mode of the time taken per comparison was 57 seconds, which corresponds to an estimated hourly rate of $9.50 (the U.S. federal minimum hourly wage in 2021 was $7.25 per hour, for reference). Note that this is likely an underestimate of the hourly rate since crowd workers often use scripts that make it possible to automatically accept a task they are interested in, and hold it assigned while not actively working on it.

3.4.3 Instructions. To ensure reproducibility of our experiment, below we quote the instructions as they were displayed to participants:

> *Please take a look at the two images displayed below. Please focus on the food itself, and not the other contents of the image. Answer the questions about the pizza shown in the images by entering either* ***1*** *for Image 1, or* ***2*** *for Image 2. Additionally, please explain your preferences by adding* ***at least one word or short phrase to describe the foods*** *appearing in each image. Write a word or a short phrase, and not full sentences.*
>
> 1: Which image contains pizza that appears more **tasty**?
> 2: Which image contains pizza that appears more **healthy**?
> 3: Which image contains pizza that appears more **caloric**?
> 4: Which image contains pizza that appears more likely to have been **consumed at home**?
> 5: Pizza shown in Image 1 is …
> (Add a word or a short phrase to describe food in Image 1)
> 6: Pizza shown in Image 2 is …
> (Add a word or a short phrase to describe food in Image 2)

4 RESULTS

4.1 RQ1: Bias of food-type distribution

To address RQ1, we start by comparing MyFoodRepo food images with images of foods posted on Twitter. We compare the prevalence of the 155 food types across the two sets (Fig. 4a).

First, we observe a significant positive correlation (Spearman's rank correlation coefficient $\rho = 0.49, p = 8.5 \times 10^{-11}$). The more frequent a food type is among consumed and tracked foods, the more frequent it tends to be among tweeted foods. That said, although food type frequencies are correlated, important deviations can be observed. For instance, bread and butter are more likely to be observed in MyFoodRepo foods compared to Twitter foods, i.e., these foods are underrepresented on Twitter. Bread is 2.5 times more frequent among MyFoodRepo foods, while butter is 5.5 times more frequent among MyFoodRepo foods. On the other hand, cake, soup, chocolate, raclette, burgers, etc., are more likely to be observed among Twitter foods, compared to MyFoodRepo foods, i.e., these foods are overrepresented on Twitter. Soup is 11.5 times, cake 12.0 times, burger 10.0 times, and raclette 9.5 times more frequent among Twitter foods.

4.2 RQ2: Biases within food types

4.2.1 Duel outcomes. Controlling for the different food-type distributions, we address RQ2, which is concerned with biases within fixed food types. As an initial look into the duel outcomes, across all duels, we first consider the fraction where the Twitter image won. Together with this fraction, we report p-values from two-sided binomial tests, where the null hypothesis is that the outcome of comparisons is random, i.e., that the Twitter image wins in 50% of duels. Across all duels, in 58.46% of duels the Twitter image is chosen as more caloric ($p < 10^{-70}$), in 45.96% as more healthy ($p < 10^{-10}$), in 38.08% as more likely to have been consumed at home ($p < 10^{-140}$), and in 61.73% as more tasty ($p < 10^{-100}$).[5]

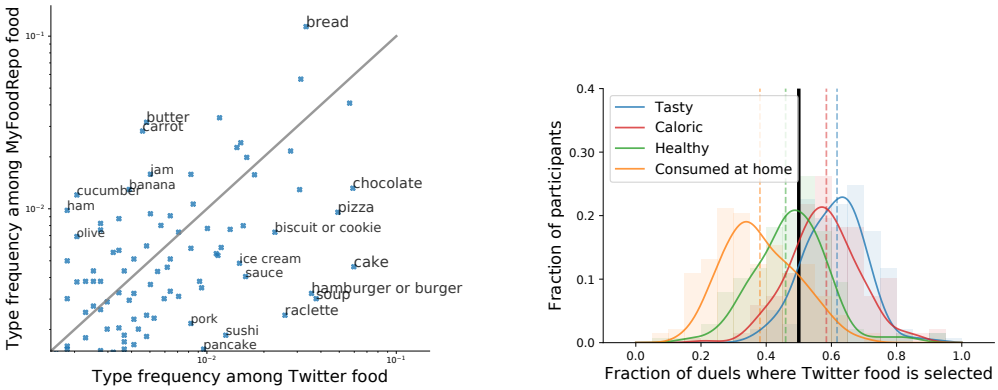
4.2.2 Bias measurement: score estimations. Next, in order to compare images, as opposed to the outcomes of duels, we fit the BT model (Eq. 1) on the collected preferences and estimate a score that represents the latent quality of each competing image concerning how healthy, tasty, caloric, and likely to have been consumed at home the food appears.

Consider a given dimension (we again use tastiness for concreteness). Let each MyFoodRepo image $i \in \{1, ..., N_M\}$ have a an estimated tastiness score $\hat{s}(i)$, and each Twitter image $j \in \{1, ..., N_T\}$ have an estimated tastiness score $\hat{s}(j)$. The tastiness bias $b(T, M)$ between food consumption measured with Twitter and food consumption measured with MyFoodRepo can be expressed as the difference in the average estimated tastiness scores measured via the respective data sources, T and M:

$$b(T, M) = T - M = \frac{1}{N_T} \sum_{i=1}^{N_T} \hat{s}(i) - \frac{1}{N_M} \sum_{j=1}^{N_M} \hat{s}(j). \tag{2}$$

The measured bias $b(T, M)$ (with 95% confidence intervals obtained via bootstrap resampling of images) is 0.52 $[0.46, 0.56]$ for "tasty", 0.39 $[0.35, 0.45]$ for "caloric", -0.18 $[-0.23, -0.11]$ for "healthy", and -0.58 $[-0.63, -0.52]$ for "likely to have been consumed at home". These bias measurements indicate that, on average, food posted on Twitter is perceived as significantly tastier, more caloric, less healthy, and less likely to have been consumed at home, compared to consumed

[5]We also examined the macro-average duel outcomes, where we consider the preference of each crowd worker and then average over workers. There appears to be no rater bias where some workers overwhelmingly prefer one or the other, as the estimates are consistent, and no notable outlier crowd workers emerge (Fig. 4b).



(a) (b)

Fig. 4. **Bias of food-type distribution, and within food-type preferences.** (a) Comparison of food type frequency among Twitter food (x-axis) vs. MyFoodRepo food (y-axis). Categories where the larger frequency is at least two times greater than the smaller frequency are annotated. Gray diagonal line marks identity, where two frequencies are equal. (b) Histogram of crowd workers' preferences. On x-axis, fraction of duels where Twitter food image is selected when compared to a MyFoodRepo food image. On y-axis, fraction of workers with such preferences. Dashed vertical lines mark average fractions of wins, across participants. Solid vertical line marks 0.5.

and tracked food. As estimated scores are not located on an interpretable scale, we shift our focus to ranks instead of raw scores next.

4.2.3 Bias measurement: rank estimations. Once latent scores are estimated, images can also be ranked, either jointly, or separately, among Twitter and MyFoodRepo food. Given its intuitive interpretation, our main method of analysis is quantifying the shifts in distributions via ranks, as depicted in Fig. 1. For concreteness, it is helpful to consider an example before studying images on a more aggregate level across food types. Fig. 2 contains the top and bottom portions of the joint rankings (one ranking for each of the four dimensions) for food type "pizza".

Now, for each food type, we rank the two sets (MyFoodRepo food vs. Twitter food) separately and determine to which percentile of MyFoodRepo food each percentile of Twitter food corresponds. We focus on the median Twitter food image (also referred to as the "typical" Twitter image), and compute the percentile rank of its score, relative to scores of MyFoodRepo food images. The rank of the median Twitter image among MyFoodRepo food images is presented in Fig. 5a, across the 24 food types.

Bias in nutritional properties. We find that Twitter foods are perceived as more caloric, less healthy, and less likely to have been consumed at home. On average across food types, the median-caloric Twitter food is among the top 34% most caloric MyFoodRepo foods. The median-healthy Twitter food is among the bottom 42% most healthy MyFoodRepo foods. And finally, the median Twitter food is among the bottom 27% most likely home-consumed MyFoodRepo foods.

Examining food types separately, regarding how healthy the foods are estimated to be, we find no significant differences in 15 out of the 24 types. In eight food types, tweeted food is perceived as *less healthy*, and one food type (vegetables) is found to be more healthy on Twitter. With respect to perceived caloric content, we find no significant differences in 12 out of the 24 types. For 11 types, tweeted food is perceived as more caloric. Vegetables are again an exception, found to be

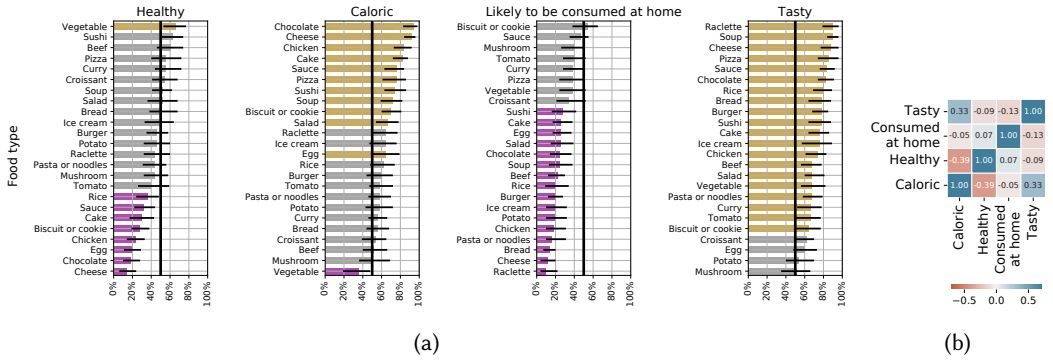


Fig. 5. **Rank bias estimations and correlation between estimated qualities.** (a) Relative rank (with respect to estimated latent scores) of median tweeted food among consumed and tracked foods, where 100% corresponds to top score. Colored bars mark estimates that significantly differ from median (i.e., 50% relative rank). Yellow marks ranks higher, purple marks ranks lower than median, while gray marks non-significant differences. Error bars mark 95% confidence intervals obtained via bootstrap resampling of duels. Example: relative rank of median tweeted raclette with respect to tastiness is 90% (95% CI [79%, 96%]); i.e., median-tasty tweeted raclette is ranked among top 10% [4%, 21%] tastiest consumed and tracked raclette images. (b) Correlation matrix between estimated qualities among 2,400 competing images.

less caloric on Twitter compared to MyFoodRepo foods. We find that in most of the types (16 out of the 24), tweeted foods are less likely to have been consumed at home. For eight food types, there are no significant differences in likelihood of having been consumed at home.

Although there are food types with no biases in nutritional properties, we observe large biases for certain foods, where a median-healthy Twitter food is among the bottom 20% of MyFoodRepo food images with respect to healthiness. For instance, median-healthy Twitter cheese is among the bottom 14% (95% confidence interval [6%, 24%]) on MyFoodRepo; median-healthy Twitter chocolate, among the bottom 18% [10%, 28%] on MyFoodRepo; and median-healthy eggs, among the bottom 19% [11%, 29%] on MyFoodRepo.

Moreover, the median-caloric Twitter food is among the top 20% caloric MyFoodRepo food. For example, median-caloric chocolate is among the top 6% [2%, 17%] on MyFoodRepo; median-caloric cheese, among the top 8% [4%, 16%] on MyFoodRepo; median-caloric chicken, among the top 16% [8%, 27%] on MyFoodRepo; and median-caloric cake, among the top 18% [12%, 28%] on MyFoodRepo.

Bias in tastiness. Next, we find substantial bias in how tasty the foods are perceived to be. There are more discrepancies in perceived tastiness compared to the above nutritional properties (Fig. 5a). On average across food types, the median-tasty tweeted food is among top 26% tastiest consumed and tracked foods. A median-tasty tweeted food is ranked significantly higher than the median-tasty consumed and tracked food image in 20 out of the 24 types. In only four types there are no significant differences regarding tastiness (mushrooms, potato, egg, croissant).

We note that, for a number of foods, a median-tasty tweeted food is ranked as high as among the top 20% of consumed and tracked food. The median-tasty Twitter raclette is among the top 10% [4%, 21%] on MyFoodRepo; the median-tasty soup, among top 10% [4%, 16%] on MyFoodRepo; the median-tasty cheese, among top 12% [4%, 23%] on MyFoodRepo; the median-tasty pizza, among top 14% [4%, 26%] on MyFoodRepo; the median-tasty sauce, among top 16% [18%, 24%] on MyFoodRepo; and the median-tasty Twitter chocolate, among top 18% [9%, 26%] on MyFoodRepo.

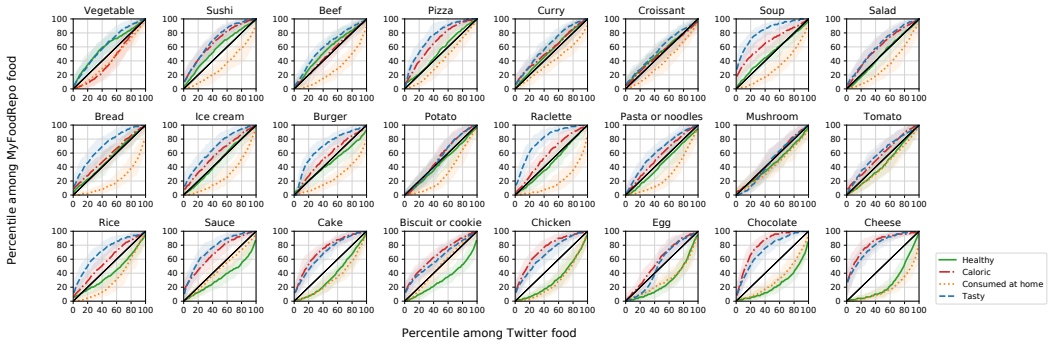


Fig. 6. **Complete rank comparisons across food types.** For each percentile among Twitter food (on the x-axis), the corresponding percentile among MyFoodRepo food (on the y-axis). That is, a Twitter image at percentile rank x among Twitter images would place at percentile rank y if it were ranked against MyFoodRepo images instead. Error bars mark 95% confidence intervals obtained via bootstrap resampling. Rank comparison is displayed for 24 types of food, across four estimated qualities: how healthy, caloric, likely to have been consumed at home, and tasty the food is. Food types are sorted according to the rank of the median-healthy tweeted food among consumed and tracked foods. Black diagonal line marks identity, where percentile distributions of Twitter food and MyFoodRepo food are equal.

4.2.4 Correlations. Next, we inspect the correlation between tastiness and the other nutritional properties. Recall that we obtained 2,400 images (100 images for each of the 24 types), with four estimated quality scores. Computing Pearson's correlation between the estimated scores (Fig. 5b), we observe a moderate positive correlation between how caloric and tasty ($\rho = 0.33$, $p < 10^{-60}$) foods are, and a negative correlation between how caloric and how healthy ($\rho = -0.39$, $p < 10^{-80}$) they are. Whereas the negative correlation between how caloric and how healthy foods are is expected, the correlation between how tasty and how caloric they are might indicate that tweeted food might be perceived as overwhelmingly tastier because it is more caloric and exaggerated.

4.2.5 Complete rank comparisons. Whereas so far we focused on the median Twitter image and studied its relative rank among MyFoodRepo images, in Fig. 6 we present the full percentile rank comparisons, for completeness. The black diagonal line marks identity, where the percentile distributions of Twitter food and MyFoodRepo food are equal. In the example of cheese (bottom right), percentiles among tweeted food correspond to higher percentiles among consumed and tracked food regarding how caloric and tasty the cheese is (tweeted food ranks are above the diagonal line). Percentiles among tweeted food correspond to lower percentiles among consumed and tracked food regarding how healthy and home-consumed the cheese is (tweeted food ranks are below the diagonal line).

4.2.6 Bias in appearance. Finally, we analyze the tags provided by crowd workers, in order to understand further how MyFoodRepo food and Twitter food differ in their appearance.[6] Recall that crowd workers were asked to enter up to three tags per image (on average, participants entered 2.4 tags per image). Analyzing the 58,645 collected tags, we ask: *How do people perceive social media foods compared to consumed and tracked foods?*

We first performed normalization of the tags provided by crowd workers. We split commas, convert to lowercase, remove stop-words at the beginning of the text (e.g., "looks", "seems"), and

[6]Note that this is an exploratory post-hoc analysis to gain deeper insights into potential mechanisms that drive the observed biases. Our main analyses are related to perceived nutritional properties and tastiness, as estimated via pairwise comparisons.

Table 1. **Bias in appearance.** Tags most distinctive of MyFoodRepo food (*left*) or of Twitter food (*right*), as determined by pointwise KL divergence (cf. Sec. 4.2.6); $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$).

Food	Top tags typical for MyFoodRepo food			Top tags typical for Twitter food		
Overall	plain****	bland****	simple****	fancy****	fresh****	colorful****
	boring****	small****	thin****	delicious	tasty***	raw****
	dry****	healthy***	homemade****	gourmet****	flavorful****	large****
Beef	grilled**	bland**	simple*	raw***	fancy*	flavorful
Biscuit or cookie	shortbread**	chocolate*	plain*	decorated***	festive**	chocolate chip**
Bread	white****	plain**	dry**	delicious**	fresh**	fancy*
Burger	simple***	small size*	fast food*	interesting**	attractive*	filling*
Cake	small****	chocolate****	chocolatey****	decorated****	fruity***	fancy**
Cheese	boring***	plain***	cold***	melted****	warm****	hot****
Chicken	plain****	bland****	dry****	fried***	tasty***	spicy**
Chocolate	dark****	bitter***	broken***	delicious***	flavoured**	fancy**
Croissant	delicious*	buttery*	stuffed*	warm*	fresh	healthy
Curry	rice**	white*	homestyle*	yummy*	healthy*	flavorful*
Egg	plain***	simple***	bland**	decorated**	raw**	colorful**
Ice cream	chocolate***	nice*	simple*	delicious	decadent	fancy
Mushroom	colorful	simple	pizza	raw**	creamy	large
Pasta or noodles	plain**	bland**	simple**	fancy***	delicious**	yellow*
Pizza	small****	pepperoni**	saucy*	large***	fresh**	delicious*
Potato	boiled****	peeled****	plain*	raw***	unpeeled**	whole*
Raclette	healthy***	greasy***	unappetizing*	sliced**	hot**	tasty*
Rice	bland***	plain**	dry*	flavorful**	mixed*	fried*
Salad	simple**	plain**	homemade**	filling*	great*	fresh
Sauce	thin****	watery***	light****	thick****	creamy****	red****
Soup	bland****	simple****	watery***	hearty***	colorful***	delicious**
Sushi	boring***	simple**	plain**	appetizing*	variety*	fresh
Vegetable	chopped*	overcooked*	mixed	fresh**	raw*	delicate
Tomato	small*	healthy	salad	mouth-watering*	juicy*	flavorful

we map versions of words with a dash to a single form (e.g., mapping "mouth watering" and "mouthwatering" to "mouth-watering").

A tag is typical for one set of images if used frequently within the set, but at the same time unlikely to be used in the other set. Additionally, it is not only the discrepancy between the two probabilities that matters; a tag should also appear frequently in a set to be considered typical of the set. This intuition is captured by the pointwise Kullback–Leibler (KL) divergence between the distributions of tags for Twitter food images and for MyFoodRepo food images, respectively. Specifically, the distinctiveness of a tag t with respect to MyFoodRepo food compared to Twitter food images is calculated as

$$D_{\mathrm{KL}}(p_{\mathrm{M}}(t) \| p_{\mathrm{T}}(t)) = p_{\mathrm{M}}(t) \log \frac{p_{\mathrm{M}}(t)}{p_{\mathrm{T}}(t)}, \tag{3}$$

where $p_{\mathrm{M}}(t)$ is the probability of observing tag t among MyFoodRepo food images, and $p_{\mathrm{T}}(t)$ the probability of observing t among Twitter food images. On the other hand, since the KL divergence is not symmetric, the distinctiveness of Twitter food compared to MyFoodRepo food is calculated as

$$D_{\mathrm{KL}}(p_{\mathrm{T}}(t) \| p_{\mathrm{M}}(t)) = p_{\mathrm{T}}(t) \log \frac{p_{\mathrm{T}}(t)}{p_{\mathrm{M}}(t)}. \tag{4}$$

In Table 1, we present the tags with the largest pointwise KL divergence, separately for tags distinctive of consumed and tracked food (left) and of tweeted food (right). For each tag, a χ^2 test on

the two frequencies is used to measure significance, under the null hypothesis that the two groups do not differ in frequency. We now examine—first overall, then separately by type—how foods differ in their appearance. We see that, overall, the tags most indicative of consumed and tracked food are "plain", "bland", "simple", "boring", "small", "thin", "dry", "healthy", and "homemade". On the contrary, tweeted food is more likely to be described as "fancy", "fresh", "colorful", "delicious", "tasty", "raw", "gourmet", "flavorful", and "large". Zooming into specific food types, the exact differences for specific food type become apparent. For example, tweeted pizzas are more likely to be described as "large", "fresh", and "delicious", whereas consumed and tracked pizzas are seen as "small", "pepperoni", or "saucy".

Inspecting the most discriminative tags overall and separately across types of food, we identified the following four prominent themes in tags that are discriminative of consumed and tracked vs. tweeted food:

(1) *Complexity.* Consumed and tracked food is described as simple and homemade ("plain", "bland", "simple" and "homemade"); tweeted food, as more elaborate ("fancy", "gourmet").
(2) *Portion size.* Consumed and tracked food comes in small portions ("small", "thin"), whereas tweeted food is exaggerated in portion size ("large"). Portion size differences are particularly evident for specific types of food, e.g., consumed and tracked burgers are described as "small size"; pizzas are "small" when consumed and tracked, and "large" when tweeted.
(3) *Ways of preparing.* Tweeted food is perceived as "raw" and "fresh", while consumed and tracked food is described as "dry". The differences are evident when it comes to specific foods. Consumed and tracked beef is more likely to be "grilled", whereas tweeted beef is more likely "raw". Consumed and tracked vegetables are "chopped" and "overcooked", whereas on Twitter, they are "fresh" and "raw". Rice and chicken are "fried" on Twitter, and "dry" when consumed and tracked.
(4) *Presentation.* Tweeted food is visually appealing, whereas consumed and tracked food is more likely to look repulsive. Tweeted food is said to be "colorful", "delicious", "flavorful", and "tasty", whereas consumed and tracked food is usually less appealing, with "watery" soup, "greasy" and "unappetizing" raclette, and "watery" and "thin" sauce.

5 DISCUSSION

Our goal has been to determine the validity of estimating food consumption from digital traces: from social media posts vs. from images of consumed and tracked foods. To this end, we designed a crowdsourcing framework for measuring biases, contrasting tweeted food images with consumed and tracked foods images, and deployed it for the case of Switzerland.

5.1 Summary of main findings

We find that social media does not provide a faithful representation of food types of consumed and tracked food. Measuring biases in food-type distributions, we observe that cake, soup, chocolate, raclette, and burgers are among the most overrepresented foods on Twitter in Switzerland. Cake, soup, and burger are visually appealing food types suitable for sharing on social media, while chocolate and raclette are foods typical of Switzerland. Winter social and sport activities among residents might make them more likely to be shared online. On the other hand, bread and butter—among the most underrepresented on Twitter—are simple everyday foods that tend not to have a lot of potential to look particularly visually appealing.

Controlling for the discrepant food-type distributions, we find that tweeted foods are perceived as less healthy, more caloric, and less likely to have been consumed at home, compared to consumed and tracked food of the same type (Fig. 5a). We also find substantial bias in perceived tastiness. A

median-tasty tweeted food is, on average across food types, ranked among the top 26% of consumed and tracked foods. Exploring free-form tags provided by crowd workers reveals that these biases are likely mediated by differences in portion size, complexity, presentation, and different ways of preparing food. For example, tweeted food is 3.5 times more likely to be described as "large", and 4 times more likely to be described as "fancy".

These results provide evidence that food shared online tends to be exaggerated compared to tracked food. The most biased foods in terms of nutritional properties are foods that can be very caloric and high in fat and carbohydrates: chocolate, cheese, chicken, cake, and egg. On the other hand, we find that some of the foods are not skewed in terms of nutritional properties. For instance, for mushrooms and croissants, no significant difference is observed in any of the four dimensions (healthiness, tastiness, caloric content, likelihood to be consumed at home). This suggests that some foods can still be validly studied via social media as a proxy for consumed and tracked foods.

5.2 Consumed food vs. tracked food vs. tweeted food

Our study attempts to establish a link between online and offline dietary behaviors by studying food images as measured via two platforms: Twitter and the MyFoodRepo food-tracking app. In what follows, we consider the relationship between three distributions: all foods consumed by the general population (the actual phenomenon of interest), food consumption estimated via MyFoodRepo, and food consumption estimated via Twitter.

For concreteness, consider tastiness (but the following argument equally applies to all other dimensions studied here). Let T denote the average tastiness score estimated via Twitter, M the average tastiness score estimated via MyFoodRepo, and G the true (unobserved) average tastiness score of food actually consumed by the general Swiss population. As before (cf. Eq. 2 and Sec. 4.2), let the bias $b(T, M) = T - M$, and analogously, $b(T, G) = T - G$ and $b(G, M) = G - M$. Although G, $b(T, G)$, and $b(G, M)$ are not observed, we have:

$$\begin{aligned} b(T,G) + b(G,M) &= (T - G) + (G - M) \\ &= T - M \\ &= b(T,M), \end{aligned} \tag{5}$$

as illustrated in Fig. 7 for the case $T > M$ (without loss of generality; if $T < M$, we may simply make the argument about $b(M, T)$ instead of $b(T, M)$).

The established substantial and significant bias $b(T, M)$ along the four studied dimensions (Sec. 4) therefore implies a lower bound on the unobserved biases, since at least one of $b(T, G)$ and $b(G, M)$ must be at least $b(T, M)/2$. In other words, along all four studied dimensions, either Twitter or MyFoodRepo differs by at least $b(T, M)/2$ from the general population. Dividing the measured biases $b(T, M)$ by two, we find that the lower bound on the bias still corresponds to significant gaps in how tasty, caloric, healthy, and likely to have been consumed at home the food is. Therefore, at least one of Twitter and MyFoodRepo foods significantly differs from the foods consumed by the general population. For example, the measured bias $b(T, M)$ is 0.52 $[0.46, 0.56]$ for how tasty the food is (Sec. 4.2). The lower bound with the shifted corresponding 95% confidence intervals, $b(T, M)/2 = 0.26$ $[0.20, 0.30]$, still corresponds to significant gaps in how tasty the food is.

Consequently, the fact that there is a divergence between food consumption as measured via food tracking and as measured via social media implies that at least one of the two is not a faithful representation of the true food consumption in the general population concerning how healthy, tasty, caloric, and likely to have been consumed at home the food is. Fig. 7 illustrates possible scenarios in more detail. On the one hand, it might be that the true food consumption in the general population is somewhere between the food consumption as measured with MyFoodRepo, and the

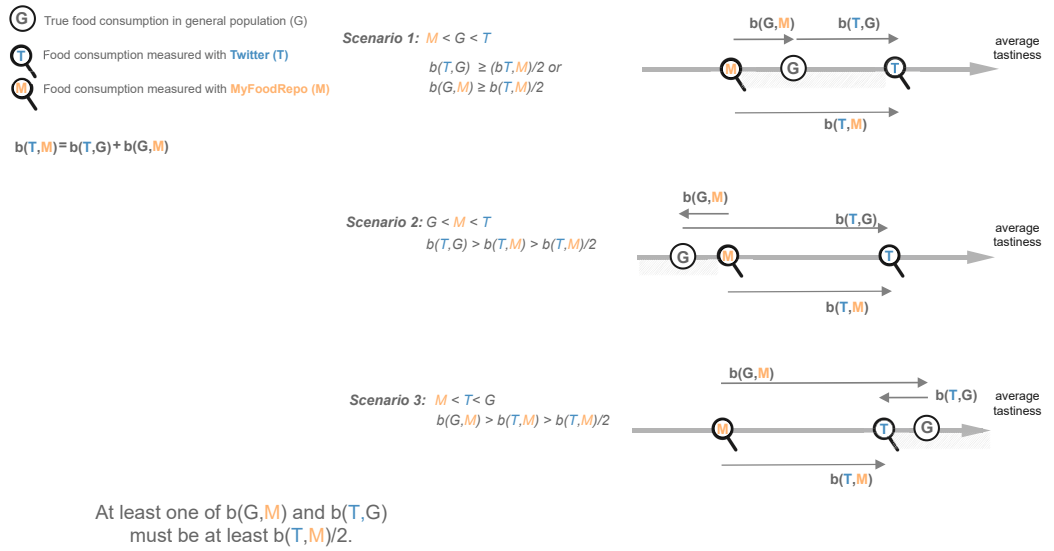


Fig. 7. **Illustration of (tastiness) biases between true food consumption G in general population, tracked food M, and tweeted food T.** The bias between tweeted food and tracked food, $b(T, M)$, is characterized in our study, whereas $b(G, M)$ and $b(T, G)$ are unobserved. Although $b(T, G)$ and $b(G, M)$ are unobserved, at least one of $b(T, G)$ and $b(G, M)$ must be at least $b(T, M)/2$. The three illustrated possible scenarios depict situations where (1) $M < G < T$, (2) $G < M < T$, or (3) $M < T < G$.

food consumption as measured with Twitter (i.e., $M < G < T$, Scenario 1). On the other hand, true food consumption in the general population might be skewed, and even more extreme than either MyFoodRepo (i.e., $G < M < T$, Scenario 2) or Twitter (i.e., $M < T < G$, Scenario 3).

In principle, MyFoodRepo might be the main source of bias (i.e., G might be closer to T than to M) since individuals might not track all consumed foods, and food trackers are known to be situated and contextualized [27, 59]. Similarly, consumed foods logged with the MyFoodRepo app are likely not representative of the full Swiss population. People who log food with the tracking app have access to a smartphone with an Internet connection and care about their diets. However, we argue that it is less likely that MyFoodRepo is the main source of bias since the majority of MyFoodRepo food images are collected from volunteers enrolled in a digital cohort called Food & You[7] who are instructed and reminded to provide images of their complete daily food intake. By design, the food present in these images was actually consumed, and omissions were discouraged. Therefore, it appears less likely that MyFoodRepo could misrepresent the true food consumption in the general population to such an extent that it would fully explain the measured biases $b(T, M)$.

Nonetheless, given the absence of data about the general population and the fact that all the considered scenarios (cf. Fig. 7) are not strictly impossible, we remain agnostic about the true source of bias. We argue that researchers should be attentive and aim to establish evidence of validity before using either social media or tracking apps as a proxy for true food consumption in the general population, since at least one of them differs by at least $b(T, M)/2$ from the general population. Future work should apply our framework for bias estimation to a representative sample of the overall population, with all food consumption recorded. At the present time, doing so remains

[7]https://www.digitalepidemiologylab.org/projects/food-and-you

challenging as the images logged with the tracking app by the volunteers are as good a peek onto actual plates as we can currently get.

5.3 Implications

5.3.1 Implications for research studying food tracking as a proxy for offline behaviors. Based on our considerations of the two platforms (i.e., Twitter and MyFoodRepo), we argue that it is less likely that food tracking is the main source of bias when estimating food consumption in the general population. Nonetheless, researchers relying on food tracking should be attentive before implicitly assuming that the tracked consumption perfectly reflects the true consumption. Whenever possible, further contextualization of the tracked consumption data and investigation of alternative digital traces of the studied persons can be beneficial for examining the validity and establishing robustness. For instance, if there is a concern that users consume food systematically different from the logged food, future deployments should consider designing logging reminders and nudges within the tracking applications, targeted towards and specifically encouraging logging the true behaviors. Future research can also encourage users to assess the accuracy of logging through the tracking applications, for instance, by self-reporting the overall perceived truthfulness.

5.3.2 Implications for social media research studying online traces as a proxy for offline behaviors. Studies using passively collected digital traces as a proxy for real behaviors need to be valid in order to support public health research and have implications for the design of policies and interventions that can impact health outcomes. Based on our findings, we now highlight major potential pitfalls that can threaten the validity of such applications and provide actionable implications for overcoming them.

Actionable implication 1: Addressing over- and underrepresentation of food types. If researchers were to estimate what foods the general population consumes based on the number of tweets containing these foods, the estimates could be biased. We suggest triangulating social media behaviors with known government statistics whenever these are publicly available. For example, although bread is underrepresented on Twitter, while burgers are overrepresented, compared with consumed and tracked foods (Fig. 4a), researchers could—even without access to logs of actual food consumption—identify the implausibly high prevalence of burgers on social media compared to bread by examining publicly available statistics [68]. For comparison, in 2019, the average Swiss consumed 89 kg of products based on grains, compared to roughly half as many kilograms of meat (48 kg). When aggregate country-wide statistics are available for calibration, social media can still be used as a sensor for spatially and temporally fine-grained analyses (e.g., by neighborhood or during holidays). When no such statistics are available, researchers might be able to calibrate their methods on populations where statistics are available and adjust the final estimates. Domain knowledge about the populations being studied can also help in alleviating some of these disproportions. One could consider knowledge about foods that studied populations consume in a social context, foods frequently consumed by visitors and tourists, or only during special periods or occasions, and avoid studies being impacted by such idiosyncrasies.

Actionable implication 2: Foods with very biased nutritional properties. If researchers were to estimate nutritional properties of foods that the general population consumes based on the tweets containing images of foods, the estimates could be biased. Researchers should be careful about bias that stems from certain foods that appear particularly less healthy and more caloric compared to consumed and tracked food, such as chocolate, cheese, chicken, cake, and egg (although these tags might not necessarily represent exactly the same food types in populations beyond Switzerland). Researchers should also be aware of differences in portion sizes that likely mediate the difference in calories. We suggest detecting and examining images with an implausible amount of calories.

For example, a single image might not be taken into account if the estimated amount of calories is not within reasonable bounds around the recommended daily 2,000–2,500 calories for an adult.

Similarly, when training machine learning models with datasets obtained from social media and aiming to generalize to the general population, samples should be adjusted such that the amount of calories more closely mirrors the amount of calories and portion sizes of real food. Otherwise, models trained on social media data to estimate the amount of calories will not make valid estimations outside of the context of exaggerated social media foods.

Actionable implication 3: Addressing systematic discrepancies in appearance. If researchers were to estimate appearance of foods that the general population consumes based on the tweets containing images of foods, the estimates could be biased. Foods that people consume and track tend to appear significantly less tasty, simpler and less elaborate, prepared in different ways, and smaller in portion size, compared to tweeted food. These are challenging biases to overcome, as there is a need to use human annotation or computer vision models. Note, however, that the foods that are most biased in terms of nutritional properties and appearance are also precisely those that are overrepresented on Twitter (Fig. 4a). Therefore, adjusting the bias in the distribution of foods is likely to alleviate the bias in nutritional properties and appearance.

5.3.3 Implications for social media research studying online traces per se. Research studying online communities and online content not as a proxy for real behaviors but as a *phenomenon per se* need not necessarily worry about validity issues and potential pitfalls. Such typical applications include studies characterizing online communities and specific users, such as users self-reporting eating disorders online [19, 75] or online eating disorder support communities [20, 32]. The behaviors of interest in these cases are precisely the online behaviors (i.e., the information that users choose to post). Similarly, studies developing machine learning models leveraging social media data that are *not* concerned with performance generalization beyond the platform and to the general populations are not necessarily impacted by these biases. Such applications might include social media food recognition [9, 14, 85, 106] or learning online food image embeddings [88].

5.3.4 Implications for food representation and users' well-being. Beyond the above implications for social media research, our results have implications for understanding the complex relationship between technology use and the well-being of social media users. In the case of food, Twitter users are exposed to unrealistic mirrors of reality [10], since foods that people actually consume and track are smaller, less "fancy", and less visually appealing (Fig. 5a, Table 1). Such distortions might contribute to the high prevalence of social comparison [40], where, e.g., as much as one-fifth of Facebook users can recall recently seeing a post that made them feel worse about themselves [18]. A user exposed to the social media portrayal of food might therefore believe that other people consume food that is tastier than the food they consume themselves. However, this would likely not be the case, due to the discrepancy between social media foods and consumed and tracked foods. Social media might, in that case, promote an unhealthy relationship with food. Our findings have implications for research about the mechanisms of such social comparison.

5.4 Limitations

Next, we outline key limitations to be kept in mind when interpreting our results. In our main analyses, we study how foods are perceived by non-expert crowd workers. The extent to which expert nutritionists would agree with such non-experts is unknown. Furthermore, while the case study is focused on Switzerland, the crowdsourced workers are located in English-speaking countries, which might influence the food perception due to cultural factors. Although the tags provided by the participants (Table 1) provide insights about factors that guide their ratings, future

work should more deeply investigate the nutritional properties of the studied foods in collaboration with expert annotators.

We note that the number of studied food images posted on Twitter—despite being the results of a best effort for completeness—is relatively small (around 3,700 Twitter photos of food, 2,400 of which were annotated). This number is small mostly due to the fact that we consider geolocated tweets only, and Switzerland is a relatively small country compared to the U.S., which has been studied in most related work [34, 66, 103].

We performed a case study of Twitter in Switzerland. Our findings cannot be assumed to generalize globally, and future work should apply our framework to other populations, other social media platforms and Web traces, and other food tracking apps. That said, this study may serve the purpose of a "proof by counterexample": we have identified one common setting where there is a divergence between food consumption as measured via food tracking and as measured via social media, implying that at least one of the two is not a faithful representation of the true food consumption in the general population. Hence, we should assume that there can be bias elsewhere, too. Researchers studying other populations should thus be attentive and aim to establish evidence of validity before using either social media or tracking apps as a proxy for the true food consumption in the general population.

As a final limitation, we note that we did not include Swiss German dialect forms of keywords, since there is no written standard for Swiss German.

5.5 Potential sources of bias of social media vs. true food consumption

We provide grounding and first insights about the validity of estimating food consumption from digital traces by contrasting consumed and tracked food with tweeted food, controlling for location, period, and food types. Revealing exact mechanisms that can lead to the biases of social media traces as a proxy for true food consumption in the general population is out of the scope of this work. However, in what follows, we consider potential sources of bias. We postulate that biases are driven by both measurement error (related to the operationalization of consumption via the concept of posting on Twitter) and population error (stemming from biases in subpopulations sharing food on Twitter) [92]. Sources of *measurement errors* might include these:

(1) *Construct validity.* On the one hand, many foods that the Swiss consume are not posted on Twitter. Appealing food consumed in certain contexts is more likely to be shared, as positive and anticipated events are more likely to be disclosed on social media in general [84]. Furthermore, photos published on Twitter may be self-selected for higher quality, thus influencing how food is perceived by the annotators. On the other hand, not all foods shared on Twitter are necessarily consumed by the posting individual (especially not in their entirety, given the portion-size bias, Table 1). Conceivably, certain tweets may originate from promotions, restaurants, or recipe sharing, all of which do not necessarily mirror actual consumption. In general, food images do not necessarily need to be related to consumption at all. They can mean something else entirely (e.g., a food can be a meme or a symbol of a political movement), although we did not find evidence of such biases in the studied data.
(2) *Platform effects.* Numerous applications and platforms support improving image quality and editing with filters, which can all contribute to the food image being more visually appealing and appearing tastier [61].
(3) *Community feedback.* Feedback received from other platform members influences the type of dietary content which a social media user posts [4], whereas negative feedback can lead to behavioral changes [22]. Biases in how food is represented online are implied by the design of online platforms.

Biases are also likely in part driven by *population error*. Users of social media platforms do not mirror the general population, neither demographically nor regarding other attributes such as behaviors and interest [57]. Users of public geotagged tweets are not randomly distributed over the general population [38, 60]. In the future, performing individual-level studies, as opposed to the population-level study reported here, will make it possible to disentangle measurement error from population error.

5.6 Future work

Beyond the already outlined future directions, the collected data can be used to further study patterns of sharing food online. Future work should further understand who shares food on Twitter (consumers, skilled individuals, but also non-individual agents, such as restaurants or caterers). We expect that further characterizing user types would not change our main conclusions, but would reveal how biases vary between different strata of Twitter users. Future work should further study where they share it from (residential vs. commercial areas), when, and in what context, as well as what are the predictors of engagement with food on Twitter.

5.7 Implications beyond food

Researching human behaviors beyond food, our crowdsourcing framework can be used to measure many types of biases, including, but not limited to, politics and activism or behaviors important for health and well-being, such as fitness and time spent in nature, travel, fashion and aesthetics, socialization, or pet ownership. Resolving the questions of truthfulness and validity of digital traces beyond food is an important direction for future research. Moving forward, it is increasingly necessary for the our research community to invest in the collection of datasets closely mirroring real behaviors.

5.8 Code and data

Code and data necessary to reproduce our results at are publicly available at https://github.com/epfl-dlab/biased-bytes.

6 CONCLUSION

As we conduct more research analyzing digital trace data, methods that offer insights into the validity of the new measures become increasingly necessary. Controlling for location, period and food types, we find that, overall, foods shared on social media significantly diverge from consumed and tracked foods. The fact that there is a divergence between food consumption as measured via food tracking and as measured via social media implies that at least one of the two is not a faithful representation of the true food consumption in the general population. Our study design lets us identify a lower bound: at least one of Twitter and MyFoodRepo diverges from the general Swiss population by at least half of the measured bias between the two platforms. We envision that the findings reported here will inform researchers in their efforts to study dietary behaviors. We also hope that the crowdsourcing framework for bias estimation and the initial quantifications will be useful broadly for research that leverages digital traces in the context of diets and beyond.

ACKNOWLEDGMENTS

We thank Maxime Peyrard for helpful feedback, and the Digital Epidemiology Lab for making the MyFoodRepo dataset publicly available via the AIcrowd Food Recognition Challenge. We acknowledge support from Microsoft (Swiss Joint Research Center), Swiss National Science Foundation (grant 200021_185043), Collaborative Research on Science and Society (CROSS), European Union (TAILOR, grant 952215), Facebook, and Google. Finally, a hat tip to Marcel Salathé for having baked the top-rated pizza in the MyFoodRepo dataset (see Fig. 1).

REFERENCES

[1] Sofiane Abbar, Carlos Castillo, and Antonio Sanfilippo. 2018. To Post or Not to Post: Using Online Trends to Predict Popularity of Offline Content. In *Proceedings of the 29th Conference on Hypertext and Social Media*.

[2] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You tweet what you eat: Studying food consumption through Twitter. In *Proc. of the 33rd Conference on Human Factors in Computing Systems (CHI)*.

[3] Palakorn Achananuparp and Ingmar Weber. 2016. Extracting food substitutes from food diary via distributional similarity. (2016).

[4] David Ifeoluwa Adelani, Ryota Kobayashi, Ingmar Weber, and Przemyslaw A Grabowicz. 2020. Estimating community feedback effect on topic choice in social media with predictive modeling. *EPJ Data Science* 9, 1 (2020).

[5] Alan Agresti. 2003. *Categorical data analysis*. Vol. 482.

[6] Luca Maria Aiello, Daniele Quercia, Rossano Schifanella, and Lucia Del Prete. 2020. Tesco Grocery 1.0, a large-scale dataset of grocery purchases in London. *Scientific Data* 7, 1 (2020).

[7] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Lucia Del Prete. 2019. Large-scale and high-resolution analysis of food purchases and health outcomes. *EPJ Data Science* 8, 1 (2019).

[8] Tim Althoff, Pranav Jindal, and Jure Leskovec. 2017. Online actions with offline impact: How online social networks influence online and offline user behavior. In *Proc. of the 10th ACM International Conference on Web Search and Data Mining (WSDM)*.

[9] Giuseppe Amato, Paolo Bolettieri, Vinicius Monteiro de Lira, Cristina Ioana Muntean, Raffaele Perego, and Chiara Renso. 2017. Social media image recognition for food trend analysis. In *Proc. of the 40th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*.

[10] Chris Bail. 2021. *Breaking the Social Media Prism*.

[11] Joan-Isaac Biel, Nathalie Martin, David Labbe, and Daniel Gatica-Perez. 2018. Bites 'n'bits: Inferring eating behavior from contextual mobile data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018).

[12] Mehrab Bin Morshed, Samruddhi Shreeram Kulkarni, Richard Li, Koustuv Saha, Leah Galante Roper, Lama Nachman, Hong Lu, Lucia Mirabella, Sanjeev Srivastava, Munmun De Choudhury, Kaya de Barbaro, Thomas Ploetz, and Gregory D Abowd. 2020. A Real-Time eating detection system for capturing eating moments and triggering ecological momentary assessments to obtain further context: system development and validation study. *JMIR mHealth and uHealth* 8, 12 (2020).

[13] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350, 6264 (2015).

[14] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*. Springer.

[15] Ann Bowling. 2005. Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health* 27, 3 (2005).

[16] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952).

[17] David L Buckeridge, Katia Charland, Alice Labban, and Yu Ma. 2014. A method for neighborhood level surveillance of food purchasing. *Annals of the New York Academy of Sciences* 1331, 1 (2014).

[18] Moira Burke, Justin Cheng, and Bethany de Gant. 2020. Social comparison and Facebook: Feedback, positivity, and opportunities for comparison. In *Proceedings of the 2020 Conference on Human Factors in Computing Systems (CHI)*.

[19] Stevie Chancellor, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proc. of the 2016 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*.

[20] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proc. of the 2016 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*.

[21] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proc. of the 6th ACM International Conference on Web Search and Data Mining (WSDM)*.

[22] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How community feedback shapes user behavior. In *Proc. of the eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[23] Martin J Chorley, Luca Rossi, Gareth Tyson, and Matthew J Williams. 2016. Pub crawling at scale: tapping untappd to explore social drinking. In *Proc. of the 10th International AAAI Conference on Web and Social Media (ICWSM)*.

[24] 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 (NEJM)* 357, 4 (2007).

[25] Chia-Fang Chung, Elena Agapie, Jessica Schroeder, Sonali Mishra, James Fogarty, and Sean A Munson. 2017. When personal tracking becomes social: Examining the use of Instagram for healthy eating. In *Proc. of the 2017 Conference*

on Human Factors in Computing Systems (CHI).

[26] Chia-Fang Chung, Jonathan Cook, Elizabeth Bales, Jasmine Zia, and Sean A Munson. 2015. More than telemonitoring: health provider use and nonuse of life-log data in irritable bowel syndrome and weight management. *Journal of Medical Internet Research* 17, 8 (2015).

[27] Chia-Fang Chung, Qiaosi Wang, Jessica Schroeder, Allison Cole, Jasmine Zia, James Fogarty, and Sean A Munson. 2019. Identifying and planning for individualized change: Patient-provider collaboration using lightweight food diaries in healthy eating and irritable bowel syndrome. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 1 (2019).

[28] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2019. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM International Conference on Knowledge Discovery and Data Mining (KDD).*

[29] Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty. 2015. Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture. In *Proc. of the 2015 Conference on Human Factors in Computing Systems (CHI).*

[30] Felicia Cordeiro, Daniel A. Epstein, Edison Thomaz, Elizabeth Bales, Arvind K. Jagannathan, Gregory D. Abowd, and James Fogarty. 2015. Barriers and negative nudges: Exploring challenges in food journaling. In *Proc. of the 2015 Conference on Human Factors in Computing Systems (CHI).*

[31] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proc. of the 22nd International Conference on World Wide Web (TheWebConf).*

[32] Munmun De Choudhury. 2015. Anorexia on tumblr: A characterization study. In *Proc. of the 5th International Conference on Digital Health.*

[33] Munmun De Choudhury, Meredith Ringel Morris, and Ryen W White. 2014. Seeking and sharing health information online: Comparing search engines and social media. In *Proc. of the 2014 Conference on Human Factors in Computing Systems (CHI).*

[34] Munmun De Choudhury, Sanket Sharma, and Emre Kıcıman. 2016. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proc. of the 2016 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW).*

[35] Neele Dijkstra. 2022. *Cross-modal recipe analysis for fine-grained geographical mapping of food consumption from images and supermarket sales.* Master's thesis.

[36] Elizabeth Dunford, Helen Trevena, Chester Goodsell, Ka Hung Ng, Jacqui Webster, Audra Millis, Stan Goldstein, Orla Hugueniot, and Bruce Neal. 2014. FoodSwitch: a mobile phone app to enable consumers to make healthier food choices and crowdsourcing of national food composition data. *JMIR mHealth and uHealth* (2014).

[37] Douae El Fatouhi, Harris Héritier, Chloé Allémann, Laurent Malisoux, Nasser Laouali, Jean-Pierre Riveline, Marcel Salathé, and Guy Fagherazzi. 2022. Associations Between Device-Measured Physical Activity and Glycemic Control and Variability Indices Under Free-Living Conditions. *Diabetes Technology & Therapeutics* 24, 3 (2022).

[38] Casey Fiesler, Michaelanne Dye, Jessica L Feuston, Chaya Hiruncharoenvate, Clayton J Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S Bruckman, Munmun De Choudhury, et al. 2017. What (or who) is public? Privacy settings and social media content sharing. In *Proc. of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW).*

[39] Charles NC Freitas, Filipe R Cordeiro, and Valmir Macario. 2020. Myfood: A food segmentation and classification system to aid nutritional monitoring. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI).*

[40] Eline Frison and Steven Eggermont. 2017. Browsing, posting, and liking on Instagram: The reciprocal relationships between different types of Instagram use and adolescents' depressed mood. *Cyberpsychology, Behavior, and Social Networking* 20, 10 (2017).

[41] Venkata Rama Kiran Garimella, Abdulrahman Alfayad, and Ingmar Weber. 2016. Social media image analysis for public health. In *Proc. of the 2016 Conference on Human Factors in Computing Systems (CHI).*

[42] Daniel Gatica-Perez, Joan-Isaac Biel, David Labbe, and Nathalie Martin. 2019. Discovering eating routines in context with a smartphone app. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers.*

[43] Kristina Gligorić, Arnaud Chiolero, Emre Kıcıman, Ryen W White, and Robert West. 2022. Population-scale dietary interests during the COVID-19 pandemic. *Nature Communications* 13, 1 (2022).

[44] Kristina Gligorić, Ryen W. White, Emre Kiciman, Eric Horvitz, Arnaud Chiolero, and Robert West. 2021. Formation of Social Ties Influences Food Choice: A Campus-Wide Longitudinal Study. *Proc. of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)* (2021).

[45] Venkat N Gudivada, Ricardo Baeza-Yates, and Vijay V Raghavan. 2015. Big data: Promises and problems. *Computer* 48, 03 (2015).

[46] Samira Humaira Habib and Soma Saha. 2010. Burden of non-communicable disease: global overview. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 4, 1 (2010).
[47] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proc. of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*.
[48] Jennifer L Harris, Sarah E Speers, Marlene B Schwartz, and Kelly D Brownell. 2012. US food company branded advergames on the Internet: Children's exposure and effects on snack consumption. *Journal of Children and Media* 6, 1 (2012).
[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2014. Deep residual learning for image recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[50] Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, et al. 2021. Integrating explanation and prediction in computational social science. *Nature* 595, 7866 (2021).
[51] Abigail L Horn, Brooke M Bell, Bernardo Garcia Bulle Bueno, Mohsen Bahrami, Burcin Bozkaya, Yan Cui, John P Wilson, Alex Pentland, Esteban Moro Egido, and Kayla de la Haye. 2021. Investigating mobility-based fast food outlet visits as indicators of dietary intake and diet-related disease. *medRxiv* (2021).
[52] Patrick D. Howell, Layla D. Martin, Hesamoddin Salehian, Chul Lee, Kyler M. Eastman, and Joohyun Kim. 2016. Analyzing taste preferences from crowdsourced food entries. In *Proc. of the 6th International Conference on Digital Health Conference (DH)*.
[53] Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, and Naonori Ueda. 2009. Topic tracking model for analyzing consumer purchase behavior. In *Proc. Twenty-First International Joint Conference on Artificial Intelligence (IJCAI)*.
[54] Noriaki Kawamae. 2010. Serendipitous recommendations via innovators. In *Proc. of the 33rd International ACM Conference on Research and Development in Information Retrieval (SIGIR)*.
[55] Jiin Kim, Zara Ahmad, Yena Lee, Flora Nasri, Hartej Gill, Roger Mclntyre, Lee Phan, and Leanna Lui. 2021. Systematic Review of the Validity of Screening Depression through Facebook, Twitter, and Instagram. *Journal of Affective Disorders* (2021).
[56] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2019. Search bias quantification: Investigating political bias in social media and web search. *Information Retrieval Journal* 22, 1 (2019).
[57] David Lazer, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. 2021. Meaningful measures of human society in the twenty-first century. *Nature* 595, 7866 (2021).
[58] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014).
[59] Yuhan Luo, Peiyi Liu, and Eun Kyoung Choe. 2019. Co-Designing food trackers with dietitians: Identifying design opportunities for food tracker customization. In *Proceedings of the 2019 Conference on Human Factors in Computing Systems (CHI)*.
[60] Momin M Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population bias in geotagged tweets. In *Proc. of the 9th International AAAI Conference on Web and Social Media (ICWSM)*.
[61] Momin M Malik and Jürgen Pfeffer. 2016. Identifying platform effects in social media data. In *Proc. of the 10th International AAAI Conference on Web and Social Media (ICWSM)*.
[62] Lucas Maystre and Matthias Grossglauser. 2015. Fast and accurate inference of Plackett–Luce models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[63] Yelena Mejova, Sofiane Abbar, and Hamed Haddadi. 2016. Fetishizing food in digital age:# foodporn around the world. In *Proc. of the 10th International AAAI Conference on Web and Social Media (ICWSM)*.
[64] Yelena Mejova, Hamed Haddadi, Sofiane Abbar, Azadeh Ghahghaei, and Ingmar Weber. 2015. Dietary habits of an expat nation: Case of Qatar. In *2015 International Conference on Healthcare Informatics*.
[65] Yelena Mejova, Hamed Haddadi, Anastasios Noulas, and Ingmar Weber. 2015. # foodporn: Obesity patterns in culinary interactions. In *Proc. of the 5th International Conference on Digital Health 2015*.
[66] Yelena Mejova, Ingmar Weber, and Michael W Macy. 2015. *Twitter: A digital socioscope*.
[67] Rebecca Mete, Alison Shield, Kristen Murray, Rachel Bacon, and Jane Kellett. 2019. What is healthy eating? A qualitative exploration. *Public Health Nutrition* 22, 13 (2019).
[68] S Meyre. 2017. Agriculture et alimentation. Statistique de poche 2017.
[69] Sharada Prasanna Mohanty, Gaurav Singhal, Eric Antoine Scuccimarra, Djilani Kebaili, Harris Héritier, Victor Boulanger, and Marcel Salathé. 2022. The Food Recognition Benchmark: Using Deep Learning to Recognize Food in Images. *Frontiers in Nutrition* 9 (2022).
[70] Fred Morstatter, Jürgen Pfeffer, and Huan Liu. 2014. When is it biased? Assessing the representativeness of Twitter's streaming API. In *Proc. of the 23rd International Conference on World Wide Web (TheWebConf)*.

[71] Shu Naritomi and Keiji Yanai. 2020. CalorieCaptorGlass: Food calorie estimation based on actual size using hololens and deep learning. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*.

[72] Ferda Ofli, Yusuf Aytar, Ingmar Weber, Raggi Al Hammouri, and Antonio Torralba. 2017. Is saki# delicious?: The food perception gap on Instagram and its relation to health. In *Proc. of the 26th International Conference on World Wide Web (TheWebConf)*.

[73] Kaimu Okamoto and Keiji Yanai. 2021. UEC-FoodPIX Complete: A Large-scale Food Image Segmentation Dataset. In *International Conference on Pattern Recognition*.

[74] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019).

[75] Jessica A. Pater, Oliver L. Haimson, Nazanin Andalibi, and Elizabeth D. Mynatt. 2016. "Hunger hurts but starving works": Characterizing the presentation of eating disorders online. In *Proc. of the 2016 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*.

[76] Max Pellert, Hannah Metzler, Michael Matzenberger, and David Garcia. 2022. Validating daily social media macroscopes of emotions. *Scientific Reports* 12, 1 (2022).

[77] Maria Perez-Ortiz and Rafal K Mantiuk. 2017. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686* (2017).

[78] Thanh-Trung Phan and Daniel Gatica-Perez. 2017. Healthy# fondue# dinner: analysis and inference of food and drink consumption patterns on instagram. In *Proc. of the 16th International Conference on Mobile and Ubiquitous Multimedia*.

[79] Jacob Poushter et al. 2016. Smartphone ownership and internet usage continues to climb in emerging economies. *Pew research center* 22, 1 (2016).

[80] Manoel Horta Ribeiro, Kristina Gligorić, Maxime Peyrard, Florian Lemmerich, Markus Strohmaier, and Robert West. 2021. Sudden attention shifts on Wikipedia during the COVID-19 crisis. In *Proc. of the 15th International AAAI Conference on Web and Social Media (ICWSM)*.

[81] Markus Rokicki, Christoph Trattner, and Eelco Herder. 2018. The impact of recipe features, social cues and demographics on estimating the healthiness of online recipes. In *Proc. of the 12th International AAAI Conference on Web and Social Media (ICWSM)*.

[82] Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213 (2014).

[83] Adam Sadilek, Stephanie Caty, Lauren DiPrete, Raed Mansour, Tom Schenk, Mark Bergtholdt, Ashish Jha, Prem Ramaswami, and Evgeniy Gabrilovich. 2018. Machine-learned epidemiology: real-time detection of foodborne illness at scale. *npj Digital Medicine* 1, 1 (2018).

[84] Koustuv Saha, Jordyn Seybolt, Stephen M Mattingly, Talayeh Aledavood, Chaitanya Konjeti, Gonzalo J Martinez, Ted Grover, Gloria Mark, and Munmun De Choudhury. 2021. What Life Events are Disclosed on Social Media, How, When, and By Whom?. In *Proc. of the 2021 Conference on Human Factors in Computing Systems (CHI)*.

[85] Doyen Sahoo, Wang Hao, Shu Ke, Wu Xiongwei, Hung Le, Palakorn Achananuparp, Ee-Peng Lim, and Steven CH Hoi. 2019. FoodAI: food image recognition via deep learning for smart food logging. In *Proc. of the 25th ACM International Conference on Knowledge Discovery & Data Mining (KDD)*.

[86] Sina Sajadmanesh, Sina Jafarzadeh, Seyed Ali Ossia, Hamid R Rabiee, Hamed Haddadi, Yelena Mejova, Mirco Musolesi, Emiliano De Cristofaro, and Gianluca Stringhini. 2017. Kissing cuisines: Exploring worldwide culinary habits on the web. In *Proc. of the 26th International Conference on World Wide Web (TheWebConf)*.

[87] Matthew J Salganik. 2019. *Bit by bit: Social research in the digital age*.

[88] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[89] Jose Ramon Saura, Ana Reyes-Menendez, and Stephen B Thomas. 2020. Gaining a deeper understanding of nutrition using social networks and user-generated content. *Internet Interventions* 20 (2020).

[90] Jessica Schroeder, Jane Hoffswell, Chia-Fang Chung, James Fogarty, Sean Munson, and Jasmine Zia. 2017. Supporting patient-provider collaboration to identify individual triggers using food and symptom journals. In *Proc. of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*.

[91] Indira Sen, Fabian Flöck, and Claudia Wagner. 2020. On the reliability and validity of detecting approval of political actors in tweets. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[92] Indira Sen, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. 2021. A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly* (2021).

[93] Sanket S Sharma and Munmun De Choudhury. 2015. Measuring and characterizing nutritional information of food and ingestion content in Instagram. In *Proc. of the 24th International Conference on World Wide Web (TheWebConf)*.

[94] Ashutosh Singla, Lin Yuan, and Touradj Ebrahimi. 2016. Food/non-food image classification and food categorization using pre-trained googlenet model. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*.

[95] Lauren N Tobey, Christine Mouzong, Joyce Senior Angulo, Sally Bowman, and Melinda M Manore. 2019. How low-income mothers select and adapt recipes and implications for promoting healthy recipes online. *Nutrients* 11, 2 (2019).
[96] Christoph Trattner, Dominik Moesslang, and David Elsweiler. 2018. On the predictability of the popularity of online recipes. *EPJ Data Science* 7, 1 (2018).
[97] Simeon Vosen and Torsten Schmidt. 2011. Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting* 30, 6 (2011).
[98] Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2014. The nature and evolution of online food preferences. *EPJ Data Science* 3 (2014).
[99] Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2014. Spatial and temporal patterns of online food preferences. In *Proc. of the 23rd International Conference on World Wide Web (TheWebConf)*.
[100] Claudia Wagner, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. 2021. Measuring algorithmically infused societies. *Nature* 595, 7866 (2021).
[101] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. Forecasting elections with non-representative polls. *International Journal of Forecasting* 31, 3 (2015).
[102] Kristi Weber, Mary Story, and Lisa Harnack. 2006. Internet food marketing strategies aimed at children and adolescents: A content analysis of food and beverage brand web sites. *Journal of the American Dietetic Association* 9 (2006).
[103] Robert West, Ryen W. White, and Eric Horvitz. 2013. From Cookies to Cooks: Insights on Dietary Patterns via Analysis of Web Usage Logs. In *Proc. of the 22nd International Conference on World Wide Web (TheWebConf)*.
[104] Michael J Widener and Wenwen Li. 2014. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography* 54 (2014).
[105] Eveline J Wouters, Junilla K Larsen, Stef P Kremers, Pieter C Dagnelie, and Rinie Geenen. 2010. Peer influence on snacking behavior in adolescence. *Appetite* 55, 1 (2010).
[106] Keiji Yanai and Yoshiyuki Kawano. 2014. Twitter food photo mining and analysis for one hundred kinds of foods. In *Pacific Rim Conference on Multimedia*.

Received January 2022; revised April 2022; accepted August 2022