The Lexicocalorimeter:
Gauging public health through caloric input and output on social media

Sharon E. Alajajian[1,2,3,4,*], Jake Ryland Williams[1,2,3,4,5], Andrew J. Reagan[1,2,3,4], Stephen C. Alajajian[6], Morgan R. Frank[7], Lewis Mitchell[8], Jacob Lahne[9], Christopher M. Danforth[1,2,3,4], Peter Sheridan Dodds[1,2,3,4,*],

1 Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05401.
2 Vermont Center for Complex Systems, University of Vermont, Burlington, VT 05401.
3 Computational Story Lab, University of Vermont, Burlington, VT 05401.
4 Vermont Advanced Computing Core, University of Vermont, Burlington, VT 05401.
5 School of Information, University of California Berkeley, 102 South Hall #4600, Berkeley, CA 94720.
6 Women, Infants and Children, East Boston, MA 02128.
7 Media Lab, Massachusetts Institute of Technology, Cambridge, MA, 02139.
8 School of Mathematical Sciences, North Terrace Campus, The University of Adelaide, SA 5005, Australia.
9 Culinary Arts and Food Science, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104.
∗ E-mail: Corresponding authors—sharon.alajajian@gmail.com (SEA) and peter.dodds@uvm.edu (PSD)

Abstract

We propose and develop a Lexicocalorimeter: an online, interactive instrument for measuring the "caloric content" of social media and other large-scale texts. We do so by constructing extensive yet improvable tables of food and activity related phrases, and respectively assigning them with sourced estimates of caloric intake and expenditure. We show that for Twitter, our naive measures of "caloric input", "caloric output", and the ratio of these measures—"caloric balance"—are all strong correlates with health and well-being demographics for the contiguous United States. Our caloric balance measure outperforms both its constituent quantities; is tunable to specific demographic measures such as diabetes rates; provides a real-time signal reflecting a population's health; and has the potential to be used alongside traditional survey data in the development of public policy and collective self-awareness. Because our Lexicocalorimeter is a linear superposition of principled phrase scores, we also show we can move beyond correlations to explore what people talk about in collective detail, and assist in the understanding and explanation of how population-scale conditions vary, a capacity unavailable to black-box type methods.

Introduction

Online instruments designed to measure social, psychological, and physical well-being at a population level are becoming essential for public policy purposes and public health monitoring [1, 2]. These data-centric gauges both empower the general public with information to allow comparisons of communities at all scales, and naturally complement the broad, established set of more readily measurable socioeconomic indicators such as wage growth, crime rates, and housing prices.

Overall well-being, or quality of life, depends on many factors and is complex to measure [3]. Existing techniques for estimating population well-being range from traditional surveys [1,4] to estimates of smile-to-frown ratios captured automatically on camera in public spaces [5], and vary widely in the types of data they amass, collection methods, cost, time scales involved, and degree of intrusion. Many measures are composite in nature with two examples being the Gallup Well-Being Index, which is based on factors

such as life evaluation, emotional health, physical health, healthy behavior, work environment, and basic access to necessary resources [4]; and the Living Conditions measure developed by the United States Census Bureau, which is derived from housing conditions, neighborhood conditions, basic needs met, a "full set" of appliances, and access to help if needed [6].

With the explosive growth of online activity and social media around the world, the massive amount of real-time data created directly by populations of interest has become an increasingly attractive and fruitful source for analysis. Despite the limitation that social media users in the United States are not a random sample of the US population [7], there is a wealth of information in these data sets and uneven sampling can often be accommodated.

Indeed, online activity is now considered by many to be a promising data source for detecting health conditions [8, 9] and gathering public-health information [10, 11], and within the last decade, researchers have constructed a range of online public-health instruments with varying degrees of success.

We cover a few relevant examples in the following section, and then turn to measuring "caloric content" of text, the focus of our present work.

Previous work

In the difficult realm of predicting pandemics [12], Google Flu Trends [13]—initially based very simply on search terms—enjoyed early success and acclaim but proved (unsurprisingly) to be imperfect in need of a more sophisticated approach [14].

In work by several of the current authors and colleagues, Mitchell *et al.* measured the happiness of tweets across the US and found strong correlations with other indices of well-being at city and state level, such as the Gallup Well-being Index; the Peace Index; the America's Health Ranking composite index of Behavior, Community and Environment, Policy and Clinical Care metrics; and gun violence (negative correlation) [15]. Using the same instrument in 10 languages, the Hedonometer, we have also shown that the emotional content of tweets tracks major world events [2, 16].

Paul and Dredze found that states with higher obesity rates have more tweets about obesity, and states with higher smoking rates have more tweets about cancer [11]. They also found a negative correlation between exercise and frequency of tweeting about ailments, suggesting "Twitter users are less likely to become sick in states where people exercise." They further found health care coverage rates to be negatively correlated with likelihood of posting tweets about diseases.

Chunara *et al.* recently found that activity-related interests on Facebook are negatively correlated with being overweight and obese, while interest in television is positively correlated with the same [17].

In an analysis of online recipe queries, West *et al.* found that the number of patients admitted to the emergency room of a major urban hospital in Washington, DC for congestive heart failure (CHF) each month was significantly correlated with average sodium per recipe searched for on the Web in the same month [18].

Eichstaedt and colleagues [19] have demonstrated that psychological language on Twitter outperforms certain composite socioeconomic indices in predicting heart disease at the county level. They were able to show in particular that the expression of negative emotions such as anger on Twitter could be taken as a kind of risk factor at the population scale.

Finally, in work directly related to our present study, Abbar *et al.* [20] have recently performed a similar analysis of translating food terms used on Twitter into calories. They found a correlation between Twitter calories and obesity and diabetes rates for the US, and explored how food-themed interactions over social networks vary with connectedness, finding suggestions of social contagion. While our approaches and results are largely sympathetic, our work incorporates estimates of physical activity which we will show provides essential extra information regarding health; introduces a phrase extraction method we call serial partitioning; and leads to an online implementation of a real-time instrument as part of our proposed 'panometer.'

Lexicocalometrics

It has thus become clear that we can estimate population-scale levels of health and well-being through social media.

Here, we examine the words and phrases people post publicly about food and physical activity on Twitter on a statewide level for the contiguous United States (48 states along with the District of Columbia). As we explain fully below in Sec. and Methods and Materials, Sec. , we group categorically similar words and phrases into lemmas, and we then assign caloric values to these lemmas using the terms and notation "caloric input" for food, C_{in}, and "caloric output" for activity, C_{out}. We define the ratio of caloric output to caloric input to be a third quantity, "caloric balance":

$$C_{\text{bal}} = \frac{C_{\text{out}}}{C_{\text{in}}}. \tag{1}$$

We use "phrase shifts" [2] to show how specific lemmas—e.g., "apples", "cake with frosting", "white water rafting", "knitting", and "watching tv or movie" contribute to the caloric texture of states across the contiguous US. We then correlate all three values with 37 demographic measures, and we find statistically strong correlations with quantities such as high blood pressure, inactivity, diabetes levels, and obesity rates. For ease of language, we will generally speak of phrases rather than lemmas.

We have also generated an accompanying online, interactive instrument for exploring health patterns through the lens of "Twitter calories": the Lexicalorimeter. An initial version of the instrument may be accessed at this paper's Online Appendices, `http://compstorylab.org/share/papers/alajajian2015a/`, and will be housed within our larger measurement platform `http://panometer.org` at `http://panometer.org/instruments/lexicalorimeter`. We note that while our online instrument is based on Twitter, it may in principle be used on any sufficiently large text source, social media or otherwise, such as Facebook.

From this point, we structure the core of our paper as follows. In Sec. , we establish and discuss our findings in depth. Specifically, we: (1) Outline our text analysis of a Twitter corpus from 2011–2012 (Sec.), reserving full details for Methods and Materials in Sec. ; (2) Present caloric maps of the contiguous US contrasting the 48 states and DC through histograms and phrase shifts (Sec.); and (3) Examine how C_{in}, C_{in}, and C_{bal} correlate with a suite of demographic measures. In the Supporting Information, we provide a sample of confirmatory figures as well as all shareable data sets. We offer concluding thoughts in Sec. .

Analysis and Results

Estimating calories from phrases

From a sample of around 50 million geotagged tweets made during 2011 and 2012 in each of the 48 continental states and the District of Columbia, we counted the total number of times each food and physical activity phrase in our database was tweeted about (see Methods and Materials, Sec. , and Supporting Information). We then used these counts to determine the average caloric input C_{in} from food phrase tweets and the average caloric output C_{out} from physical activity phrase tweets as follows.

First, we equate each food phrase s with the calories per 100 grams of that food, using the notation $C_{\text{in}}(s)$. (We also explored serving sizes but the databases available proved far from complete.) We then compute the caloric input for a given text T as:

$$C_{\text{in}}(T) = \frac{\sum_{s \in S_{\text{in}}} C_{\text{in}}(s) f(s \mid T)}{\sum_s f(s \mid T)} = \sum_{s \in S_{\text{in}}} C_{\text{in}}(s) p(s \mid T), \tag{2}$$

where $f(s \mid T)$ is the frequency of phrase s in text T, $p(s \mid T)$ is the normalized version, and S_{in} is the set of all food phrases in our database.

Second, for each tweeted physical activity phrase, we use an estimate of the Metabolic Equivalent of Tasks, or METs, which we then converted to calories expended per hour, assuming a weight of 80.7 kilograms, the average weight of a North American adult [21]. Analogous to $C_{in}(T)$ above, we then have

$$C_{out}(T) = \sum_{s \in S_{out}} C_{out}(s)p(s|T), \tag{3}$$

where now S_{out} is the set of all phrases in our activity database.

We emphasize that both our food and exercise phrase data sets and Twitter databases are necessarily incomplete in nature. The values of C_{in} and C_{out} are thus not meaningful as absolute numbers but rather have power for comparisons. We also acknowledge that our equivalences are crude—e.g., each mention of a specific food is naively turned into the calories associated with 100 grams of that food—and later on we address our choices in more depth. Nevertheless, our method is pragmatic yet—as we will show—effective, and offers clear directions for future improvement.

For simplicity and ultimately because the results are sufficiently strong, we did not filter tweets beyond their geographic location. Tweets may thus come from individuals, restaurants, sports stores, resorts, news outlets, marketers, fitness apps, tourists, and so on, and further improvements and refinements may be achieved by appropriately constraining the Twitter corpus.

Finally, we take the ratio of $C_{out}(T)$ to $C_{in}(T)$ to obtain the text's caloric balance $C_{bal}(T)$. In general, we observe that a higher value of $C_{bal}(T)$ at the population scale would appear to be intuitively better, up to some limit indicating negative energy balance. We note that $C_{bal} = 1$ is not salient and should not be taken to mean a population is 'balanced calorically'. Using the difference $C_{out} - C_{in}$ generates similar results but, from a framing perspective, we have reservations in creating a scale with a 0 point given the approximate nature of our measures.

Caloric maps of the contiguous US

We now move to our central analysis and exploration of how our lexicocalorimetric measure varies geographically. We start with visual representations and then continue on to more detailed comparisons. In Fig. 1, we show three choropleth maps of our overall 2011–2012 measures of Twitter's caloric input C_{in}, caloric output C_{out} and caloric balance C_{bal}, for the contiguous US. For all three maps, quantities increase as colors move from light to dark green.

These maps immediately allow for some basic observations which we will delve into and harden up as our analysis proceeds. For the food calories map, we see C_{in} is generally largest in the Midwest and the south while Colorado and Maine stand out as states with the lowest calories.

We see a different texture in the activity calories map with the highest calories appearing in the three-state block of Wyoming, Colorado, and Utah, as well as Vermont. Tweet-based caloric output drops to a low in Mississippi and the surrounding states, while Michigan also appears to have a low value of C_{out}.

Visually, the third map shows the highest values of C_{bal} are found for Colorado, Wyoming, and Vermont, and secondarily for Maine, Minnesota, Oregon, and Utah. Low values of C_{bal} appear in the region of Mississippi, Louisiana, Alabama, and Arkansas, as well as West Virginia.

Visually, we see that C_{out} is clearly more well aligned with C_{bal} than C_{in}. The reason is that for the present version of the Lexicocalorimeter, C_{out} has a larger dynamic range than C_{in}, roughly 250 to 285 versus 160 to 210 giving ratios of $\frac{210}{160} \simeq 1.31$ and $\frac{285}{250} \simeq 1.14$. We could assert that C_{in} is fundamentally less informative but:

1. In Sec. , we will find that some demographic measures correlate more strongly with C_{in} and C_{out};

2. We may adjust the dynamic range of either measure by rescaling, introducing a kind of tunability [2] to the instrument (a feature we will reserve for future iterations); and

3. Because our food phrase database is a factor of 10 smaller than our activity phrase one, revisions of our instrument may elevate the power of C_{in}.

To provide some support for point 1, we compare C_{out} and C_{in} in Fig. 2 (see also Fig. S1). Importantly, we see that the two measures are indeed not well correlated, indicating they contain different kinds of information (Pearson correlation coefficient $\hat{\rho}_{\text{p}} \simeq 0.13$, p-value $= 0.37$). This demonstrates why we might expect C_{in} or C_{out} to separately correlate more strongly with other population-level measures, and justifies the incorporation of both C_{in} and C_{out} into some kind of composite measure.

Regarding point 2 above, we have evidently made a number of choices in computing C_{in} and C_{out} that mean we have already introduced an arbitrary tuning of the ratio C_{bal} (e.g., assuming 100 grams of a food and an hour's worth of activity). Having no principled way of rescaling (i.e., one that is not a function of the data set being studied), we have chosen to leave the measures as computed, but we envisage introducing tunability of the dynamic ranges of C_{in} and C_{out}—altering the bias of the measure toward food or activity—will allow the Lexicocalorimeter to be refined for a range of purposes such as estimating correlates of diabetes levels versus cancer rates (see Sec.).

For the food and activities maps, we also overlay the specific phrase whose increased prevalence most contributes to moving each population's Twitter calorie scores away from the overall average for the contiguous US. We will return to how we determine these phrases later with phrase shifts (Sec.), but for now we see a diverse spread of terms.

We find a number of phrases make for apparently reasonable representations:

- "lobster" in Maine and Massachusetts;

- "grits" in Georgia;

- "skiing" in Vermont, New Hampshire, and Utah;

- and "running" in Colorado and a number of other locations.

Prototypical unhealthy foods rise to the top in various states:

- "donuts" in Texas;

- "cake" in Mississippi;

- "chocolate candy" in Louisiana;

- and "cookies" in Indiana.

By contrast, a few "virtuous" foodstuffs appear such as "green beans" in Oregon and "tomato" in California.

Our activity list also includes some rather low intensity ones and we see:

- "eating" rising to the top in Texas, the south, and a number other states;

- "watching tv or movie" in Pennsylvania and elsewhere;

- "sitting" in Tennessee;

- "talking on the phone" in Delaware;

- "getting my nails done" in New Jersey;

- and simply "lying down" in Michigan.

Now, we do not pretend that these phrases all come from individuals diligently recording their present meals or activities. And some phrases are problematic in their generality of meaning, most especially "running" (the word "run" currently has the most meanings in the Oxford English Dictionary). Nevertheless, as we dig deeper into all the phrases found for a particular state, we will continue to find commonsensical lexical patterns.

Rankings for the contiguous US

Having taken in the maps of our three measures C_{in}, C_{out}, and C_{bal}, we now explore the rankings quantitatively, first through the histograms shown in Fig. 3. We order the 48 states and DC by C_{bal} (rightmost plot) and all bars are relative to the overall average of the specific measure. Numeric rankings for each measure are given next to each bar. In Figs. S2 and S3, we present the same histograms re-sorted respectively by C_{in} and C_{out}.

As was the case for the maps, we again see that C_{bal} is more strongly driven by C_{out} than C_{in} due to the former's larger dynamic range. The states with the highest values of C_{bal} achieve their scores through high levels of C_{out} but more variable levels of C_{in}. Wyoming (26), Vermont (21), and Utah (24) are all middling in C_{in} while Colorado (48) and Maine (49) have the lowest ranks for caloric intake. At the trailing end, we see by contrast that low activity ranks are coupled with high ranks for caloric intake.

A few of the more anomalous states are both evident in the C_{in} and C_{out} histograms and as those appearing furthest away from the best line of fit in the scatter plot of Fig. 2. South Dakota has both high values of C_{in} and C_{out} (ranks of 1 and 7) that arrange to give it a ranking of 25 for C_{bal}. Maryland ranking 43rd and 45th in C_{in} and C_{out}, is the only state in the 'bottom' 10 of both measures.

Phrase shifts

In any study involving word- or phrase-based measures of texts, it is plainly essential that we demonstrate that the measures involved behave reasonably by exploring the effects of changes in word frequencies (and yet many papers fail to do so, e.g., [22]). Such a requirement limits the scientific value of potentially powerful "black box" type instruments, but as our measure is linear in nature, the lexicocalometric contributions of individual terms can be laid bare.

In our work on measuring happiness, we have developed and extensively used "word shifts" to show which words make a given text appear more positive than another text in aggregate (see [2] and [16]). Such visualizations not only provide our necessary test, but also allow us to draw insight from the lexical tapestry of texts. Here, we will explain and use analogously constructed phrase shifts for both C_{in} and C_{out} to examine the states at the extremes of our C_{bal} rankings, Colorado and Mississippi. Interactive food and activity phrase shifts for the 49 regions of the contiguous US form a central part of our online Lexicocalorimeter: `http://panometer.org/instruments/lexicocalorimeter`.

We start with two texts: a base "reference text" T_{ref}, and a "comparison text" T_{comp} which we wish to compare to T_{ref}. In this paper, we will use the Contiguous US as the reference text (weighting the phrase distributions of each state equally), but in principle any text can be used (e.g., in comparing two states, one would be selected as a reference). Our interest is in determining which words or phrases most contribute to or go against the difference in estimated calories. $C_{\mathrm{i/o}}(T_{\mathrm{comp}}) - C_{\mathrm{i/o}}(T_{\mathrm{ref}})$ where i/o stands for in or out. Following [2] and using Eq. (2), we can express the difference as

$$
\begin{aligned}
& C_{\mathrm{i/o}}(T_{\mathrm{comp}}) - C_{\mathrm{i/o}}(T_{\mathrm{ref}}) \\
& = \sum_{s \in S_{\mathrm{i/o}}} C_{\mathrm{i/o}}(s) \Big[p(s \,|\, T_{\mathrm{comp}}) - p(s \,|\, T_{\mathrm{ref}}) \Big] \\
& = \sum_{s \in S_{\mathrm{i/o}}} \Big[C_{\mathrm{i/o}}(s) - C_{\mathrm{i/o}}^{(\mathrm{ref})} \Big] \Big[p(s \,|\, T_{\mathrm{comp}}) - p(s \,|\, T_{\mathrm{ref}}) \Big].
\end{aligned}
$$

$$(4)$$

We now have a sum contributions due to all phrases. We normalize these contributions as percentages

and annotate their structure as follows:

$$\delta C_{\mathrm{i/o}}(s) =$$
$$\frac{100}{\left|C_{\mathrm{i/o}}^{(\mathrm{comp})} - C_{\mathrm{i/o}}^{(\mathrm{ref})}\right|} \underbrace{\left[C_{\mathrm{i/o}}(s) - C_{\mathrm{i/o}}^{(\mathrm{ref})}\right]}_{+/-} \underbrace{\left[p_s^{(\mathrm{comp})} - p_s^{(\mathrm{ref})}\right]}_{\uparrow/\downarrow},$$

$$(5)$$

where $\sum_{s \in S_{\mathrm{i/o}}} \delta C_{\mathrm{i/o}}(s) = \pm 100$. We use the symbols $+/-$ and $\uparrow/\downarrow$ to respectively encode whether the calories of a phrase exceed the average of the reference text, and whether a phrase is being used more or less in the comparison text. We call $\delta C_{\mathrm{i/o}}(s)$ the "per food/activity phrase caloric expenditure shift". Finally, we sort phrases by the absolute value of $\delta C_{\mathrm{i/o}}(s)$ to create each phrase shift.

In Fig. 4, we present food phrase shifts which help to illustrate why:

- Colorado ranks 48/49 for caloric input C_{in} (Fig. 4A),

- Mississippi ranks 16/49 for caloric input C_{in} (Fig. 4B),

- Colorado ranks 2/49 for caloric output C_{out} (Fig. 4C),

- and Mississippi ranks 49/49 for caloric output C_{out} (Fig. 4D).

These shifts display phrases that fall into four categories:

$+\uparrow$, yellow: Phrases representing above average quantities (here calories) being used more often. Examples: "cookies" for Mississippi in Fig. 4B and "rock climbing" for Colorado in Fig. 4C.

$-\downarrow$, pale blue: Phrases representing below average quantities being used less often. Examples: "watching tv or movie" for Mississippi in Fig. 4B and "laying down" for Colorado in Fig. 4C.

$+\downarrow$, pale yellow: Phrases representing above average quantities being used less often. Examples: "chocolate candy" for Colorado in Fig. 4A and "running" for Mississippi in Fig. 4D.

$-\uparrow$, blue: Phrases representing below average quantities being used more often. Examples: "reading" for Colorado in Fig. 4A and "catfish" for Mississippi in Fig. 4B.

Note that depending on the quantity, higher or lower may be "better" and the four categories flip signs in their support. For example, C_{in} and C_{out} increase with $+\uparrow$ phrases; after we examine correlations with demographics in Sec. , we will be able to interpret this as "bad" for C_{in} and "good" for C_{out}.

At the top of each phrase shift, the bars indicate the total contribution of each of the four types of phrases, and the black bar the net change. We see that the four net changes arise in different ways.

- Fig. 4A: Colorado is lower than average for C_{in} largely due to tweeting more about relatively low calorie (per 100 grams) foods: "noodles", "egg", "pasta", and "turkey". We also find less tweets about high calorie foods such as "candy", "cake", and "cookies." Going against these phrases, we see Colorado does tweet relatively more about "bacon" and "olive oil", and less about some relatively lower calorie foods "chicken", "ice cream", "shrimp", and "corn". We note that this does not mean these foods are low calorie in absolute terms ("ice cream" is a good example), just that 100 grams of them are low calorie in comparison to the US baseline.

- Fig. 4B: Mississippi almost equally tweets less about a variety of low calorie foods, e.g., "pasta", "banana", and "crab" (pale blue bar) while also tweeting more about the complementary range of such foods including "shrimp", "peaches", and "pineapple" (dark blue bar). The modest net gain is mostly due to a small increase in tweeting about high calorie foods such as "cake", "cookies", and "sausage".

- Fig. 4C: For physical activity, tweets from Colorado show a preponderance of relatively high caloric expenditure phrases ($+\uparrow$, yellow) including "running", "skiing", "hiking", "snowboarding" and so on. Tweeting less about low effort activities is the only other contribution of any substance— Colorado tweets less about "eating", "laying down", and "watching tv or movie".

- Fig. 4D: Mississippi's low ranking in activity is largely due to tweeting less about high output activities ($+\downarrow$, pale yellow): less "running", "dancing", "walking", and "biking". The second most important category is an increase in low output activity phrases such as "eating", "attending church", and "talking on the phone."

In Figs. S4, S5, S6, and S7, we complement the four phrase shifts of Fig. 4 by showing the top 23 phrases for each of four ways phrases may contribute. Interactive phrase shifts for all of the contiguous US are housed at `http://panometer.org/instruments/lexicocalorimeter`.

Overall, we find the lexical texture afforded by our phrase shifts is generally convincing, but we expect future improvements in our food and activity data sets will iron out some oddities (we again use the example of ice cream). We also note that phrase shifts are very sensitive and that terms that seem to be being evaluated incorrectly may easily be removed from the phrase set, and that doing so will minimally change the overall score for sufficiently large texts.

Correlations with other health and well-being measures

We now turn to a suite of statistical comparisons between our three measures—caloric input, caloric output, and caloric balance—and a collection of demographic quantities.

We use Spearman's correlation coefficient $\hat{\rho}_s$ to examine relationships between C_{in}, C_{out}, and C_{bal} and 37 variables variously relating to food and physical activity, "Big Five" personality traits, and health and well-being rankings (a total of 111 comparisons). To correct for multiple comparisons, we calculate the q-value for each correlation coefficient using the Benjamini-Hochberg step-up procedure [23] (the q-value is to be interpreted in the same way as a p-value). We then consider correlations in reference to the standard significance levels of 0.01 and 0.05.

We must first acknowledge that many of the variables we test against our measures are highly correlated with each other. The food and physical activity-related variables are in the areas of physical activity levels, produce intake and availability rates (including trends in public schools), chronic disease rates, and rates of unhealthy habits. Many of these variables are well known to be influenced by diet and physical activity (e.g., obesity rates [24]), and others may be less directly related (e.g., percent of cropland in each state harvested for fruits and vegetables [25]).

To give some grounding for the full set of comparisons, we show in Fig. 5 how six demographic quantities vary with caloric balance C_{bal}. We see strong correlations with $|\hat{\rho}_s| \geq 0.68$, and the highest value for Benjamini-Hochberg q-value is 5.7×10^{-7}.

We present a summary of all results in Tab. 1 where we have ordered and numbered demographic quantities in terms of ascending Benjamini-Hochberg q-values for C_{bal}.

Surveying the health-based demographics, we found C_{bal} was significantly correlated with all chronic disease-related rates we tested against (high blood pressure (#1), adult diabetes (#4), heart disease deaths (#6), adult overweight and obesity (#7), adult obesity (#8), childhood overweight and obesity (#13), high cholesterol (#19), and colorectal cancer (#22)). All of these but colorectal cancer rate were also significantly correlated with C_{out}.

Caloric input C_{in} results were more mixed. Rates of high blood pressure, adult diabetes, childhood overweight and obesity, and high cholesterol were not significantly correlated with C_{in} after correcting for multiple comparisons.

The variables relating to unhealthy habits (smoking (#16) and binge drinking rates (#26)) both correlated significantly with all three of our measures with the one exception of binge drinking and

caloric input. The direction of correlations for these two habits are opposite each other (e.g., negative for smoking and C_{bal}, positive for binge drinking and C_{bal}), consistent with recent work on alcohol consumption [26].

The two variables relating to physical activity rates (percent of population that has been physically active in past 30 days (#2), and percent of population that has had no physical activity in past 30 days (#3)) correlated significantly with all three of our measures. The two measures relating to rates of physical and mental health (average number of poor mental health days in past 30 days (#24), and average number of poor physical health days in past 30 days (#27)) correlated significantly with both C_{out} and C_{bal}, but did not correlate significantly with C_{in}.

The four variables relating to fruit and vegetable consumption rates all correlated significantly with all three of our measures, except for median daily intake of vegetables (#18) with caloric output C_{out}. The variables relating to presence of produce in the state (percent of cropland in each state harvested for fruits and vegetables (#33), percent of census tracts with a healthy food retailer within one-half mile (#35), and percent of schools offering fruits and vegetables at celebrations (#31)) were significantly correlated with C_{in} but were not correlated with C_{out} or C_{bal}. Variables relating to local food (number of farmers markets per 100,000 people (#28) and Strolling of the Heifers locavore score(#29)) were not significantly correlated with C_{in}, but were significantly correlated with C_{out}.

Our health and well-being ranking variables included the CNBC quality of life ranking (#5), Gallup Wellbeing ranking (#9), America's Health Ranking overall state rank (#10), life expectancy ranking (#11), Brain Health ranking (#20), Gini index score (#23), and George Mason's overall freedom ranking (#36). Caloric balance correlated with all of these variables except for George Mason's freedom ranking (which did not correlate with any of our three measures). C_{out} correlated significantly with all of these measures except for the Brain Health ranking and the freedom ranking. caloric input C_{in} did not correlate significantly with the CNBC quality of life ranking, Gini index score, or freedom ranking.

Regarding correlations with the *Big Five* personality traits, Pesta *et al.* noted that "Neuroticism...emerged as the only consistent Big Five predictor of epidemiologic outcomes (e.g., rates of heart disease or high blood pressure) and health-related behaviors (e.g., rates of smoking or exercise)" [27]. Additionally, "neuroticism correlates with many health-related variables, including depression and anxiety disorders, mortality, coping skill, death from cardiovascular disease, and whether one smokes tobacco" [27]. Here, in keeping with these observations, we found that neuroticism (#25) was indeed the only *Big Five* personality trait that correlated significantly and negatively with caloric balance.

We also tested our three measures against two measures of socioeconomic status—median income (#17) and percent of state with a bachelor's degree or higher level of education (#21)—and found these correlations were significant for all three of our measures.

Concluding Remarks

Our Lexicocalorimeter has thus, when applied to Twitter, proved to find and demonstrate a range of strong, commonsensical patterns and correlations for the contiguous US. We invite the reader to explore our online instrument, a screenshot of which is shown in Fig. 6.

Given the complex relationships between health, well-being, happiness, and various measures of socioeconomic status, it is rather difficult to say that we are only measuring health or only measuring well-being. We are also measuring socioeconomic status to some extent. However, the correlations between caloric balance and measures of socioeconomic status are not as strong as the correlation of caloric balance with many of the other measures. Given the above, we believe that the caloric content of tweets can be used successfully, along with other well-being and quality of life measures, to help gauge overall well-being in a population.

There are many potential forward directions. A promising avenue is to incorporate tunability to the Lexicocalorimeter by manipulating the dynamic range of C_{in} and C_{bal}. Though a universal approach is

unclear (one constraint is that such an approach must be independent of the particular data set being studied), we may profit from this versatility when focusing on a single demographic. For example, if we are interested in diabetes rates, we could tune the instrument to obtain the best correlation with known levels, and thereby create a real-time estimator. To do so, we would introduce a tunable parameter α into Eq. (1):

$$C_{\text{bal}}^{(\alpha)} = \frac{\alpha C_{\text{out}}}{C_{\text{in}}}, \tag{6}$$

and find the value of α that gives the highest correlation between $C_{\text{bal}}^{(\alpha)}$ and diabetes rates for a given set of populations. Of course, we could use a "black box" method to generate a more optimal fit, but in basing our instrument on food and activity words, we have a far more principled approach that grants us the opportunity not just to mimic but to understand and explain patterns that we find. In particular, our word shifts will be of great use in showing why our hypothetical estimate of diabetes is varying across populations.

We fully recognize that the Twitter population is not the same as the general population; Twitter users differ from the general population in terms of race, age, and urbanity [7]. However, we currently have no reliable way to know, for example, the true age, race, gender, and education level of individual users and as such, are not able to adjust for these factors. While we were able to vet our food and physical activity lists to some extent (as described in Methods and Materials), we could not realistically go through every tweet to be certain that the phrase was being used in the way that we thought. We realize that even if the phrases are being used as we imagine, it does not necessarily mean that the person who tweeted actually performed the physical activity or ate the tweeted-about food (West *et al.* address a similar issue in inferring food consumption from accessing recipes online [18]).

We also currently do not know at what point our metric breaks down at smaller time scales (e.g., months or weeks) or for smaller spatial regions (e.g., city or county) level. Our preliminary research shows that the physical activity metric on its own may be quite effective at the city level, but the food measure may not be accurate on a smaller scale. We have also found the physical activity list to be robust to random partitioning [28], whereas the food list was not. We believe that these preliminary findings may be due to several factors: (a) the size of the food list (just over 1400 phrases) is much smaller than the physical activity phrase list (just over 13,400 phrases); (b) there are generally more tweets about physical activities in our list than the foods in our food list; and (c) the amount of data within a city may not be a large enough sample for any food-based Twitter metric. We note that we have not tried using the metric on counties or Census block or tract groups, and it may be that these are more conducive to the metric.

We propose to use crowdsourcing as a way to build a more comprehensive food phrase list that includes commonly eaten foods with brand names as well as food slang that we did not capture here. Ideally, we would arrive at a food phrase database similar in scale to that of our existing physical activity phrase list. However we move forward, we believe it is clear that the Lexicocalorimeter we have designed and implemented is already of some potency and may be improved substantively in the future.

Methods and Materials

In order to attempt to estimate the "caloric content" of text-extracted phrases [28] relating to food (caloric input) and physical activity (caloric output), we needed comprehensive lists of foods and physical activities and their respective caloric content and expenditure information. Here, we explain in detail how we constructed these phrase lists and assigned calories to each phrase.

We provide all data in the Supporting Information and with the paper's Online Appendices: `http://compstorylab.org/share/papers/alajajian2015a/`.

Calorie estimates for phrases

We used the USDA National Nutrient Database [29] to approximate the caloric content of foods, and the Compendium of Physical Activities from Arizona State University and the National Cancer Institute [30] to approximate average Metabolic Equivalent of Tasks (METs) for physical activities, which we converted to calories expended per hour of activity [30]. Because the foods listed in the USDA National Nutrient Database are not described in a way that people talk about food, we created a list of food phrases used on Twitter by starting with a kernel of basic food terms from the USDA's MyPlate website's food group pages [31]. If the food phrase was not specific, such as "cereal", we chose the most popular version of that food in the United States via an informal Google search at the time of the study (in this instance, Cheerios). If a brand name food was not in the USDA National Nutrient Database, we chose the closest match we could find. (Please note that this means that data in appendix may be inaccurate when searching brand name items.)

This approach yielded examples of foods in the food groups of fruits, vegetables, grains, proteins, dairy, oils, solid fats, and "empty calories" (e.g., junk food), and built up a list of nearly 1400 food phrases used on Twitter. For this study, we did not include sodas or juices in our list. Soups, ice creams, oils and other items that may act as liquids may be excluded in future versions of these analyses.

For physical activity, we used the physical activities listed in the Compendium to build up a list of nearly 14,000 physical activity phrases used on Twitter. The order of magnitude of difference between the length of the two lists exists because of the difference in the number of terms that went into creating each list and the rates at which people tweet about foods vs. physical activities.

Phrase extraction

A major obstacle to the development of the food and physical activity lists is the determination of those phrases used by individuals that most accurately represent a food or physical activity. Various methods exist which may help one ascertain information about the frequency of usage of higher-order lexical units [28]. However, we require one that not only determines reasonable estimates of frequency of usage, but further, does so with nuance regarding context. For example, one should not count the phrase "apple" as having occurred if it appeared within a larger phrase that was recognized as meaningful, such as "you're the apple of my eye." Hence, we impose meaningful partitions of text by utilizing what we call an "inner context model" [32]. In particular, we consider two phrases to be members of a *common context* when they are maximally similar according to a normalized edit distance [33] (allowing for just insertions and deletions). Our basic measure for a phrase's strength of context is then the *local likelihood* (denoted ℓ_s for a phrase s), or rate of occurrence of the phrase amongst all others present in the context. A phrase of length (number of words) N is a member of N contexts (setting $q = 1$ in the context model). Therefore, we assume the phrase's context that occurs most frequently as the one by which likelihood is taken.

To describe our method, *serial partitioning*, we use a mock-up example. Supposing we wish to partition the clause "you're the apple of my eye", assume further that among all the phrases contained within the clause, one knows

$$\ell_{\text{you're}} < \ell_{\text{you're the}} \quad \text{and} \quad \ell_{\text{you're the}} > \ell_{\text{you're the apple}}$$

and that

$$\ell_{\text{apple}} < \ell_{\text{apple of}} < \ell_{\text{apple of my}} < \ell_{\text{apple of my eye}},$$

as this information is sufficient to exemplify the algorithm. We would not have included "you're the apple of my eye" in our food phrase list, as it does not use apple in a way that pertains to eating food.

A complete step-by-step illustration of the serial partitioning algorithm for the clause "you're the apple of my eye" would be:

1. you're the apple of my eye $\qquad$ $\ell_\phi = 0$

2. **you're** the apple of my eye $\qquad$ $\ell_{\text{you're}} > \ell_\phi$

3. **you're the** apple of my eye $\qquad$ $\ell_{\text{the}} > \ell_{\text{you're}}$

4. **you're the apple** of my eye $\qquad$ $\ell_{\text{the apple}} < \ell_{\text{you're the}}$

5. <u>you're the</u> apple of my eye $\qquad$ $\ell_\phi = 0$

6. <u>you're the</u> **apple** of my eye $\qquad$ $\ell_{\text{apple}} > \ell_\phi$

$\vdots$

9. <u>you're the</u> **apple of my eye**

$$\ell_{\text{apple of my eye}} > \ell_{\text{apple of my}}$$

10. <u>you're the</u> <u>apple of my eye</u>.

with each step explained in detail:

1. The first component begins with the null phrase (likelihood 0).

2. The first word is concatenated onto our component, producing the phrase "you're", whose positive likelihood keeps the process going.

3. The word "the" is then concatenated to form "you're the", whose increased likelihood keeps the process going.

4. Once the word "apple" is concatenated onto the growing phrase the likelihood drops.

5. Since the likelihood has dropped the previous phrase "you're the" is partitioned.

6. The process resumes with the null word, and subsequent concatenation forming the new working phrase, "apple".

7–9. Concatenation of the words "of", "my", and "eye" result in increased local likelihood and so perpetuate the process.

10. Since the phrase "apple of my eye" is bounded by punctuation, the phrase is partitioned, whereupon the algorithm moves on to the next clause.

We manually applied the following criteria for constructing both food and exercise phrase lists. For a phrase to be included, it had to be a phrase that used the food or physical activity word(s) in a way that pertained to eating or physical activity; we excluded phrases that were part of hashtags, Twitter user names, song lyrics, or names of organizations or businesses, and phrases that appeared four or fewer times were not included. Misspellings and alternate spellings were included if we happened upon them (for example, "mash potatoes" instead of "mashed potatoes"), but we did not go out of our way to search

for them. We queried questionable phrases to be sure that the majority of their uses were referring to the item of interest. Because we were building up from a small list, some specific versions of foods were included while more general forms were not. For example, because we built phrases up from "strawberry," "strawberry jam" was included while we did not conduct a larger search for "jam". In another example, in building phrases up from "bacon," "bacon wrapped dates" turned up so we included those dates but did not conduct a larger search for all possible "dates". (Note: We removed the physical activities category 'sexual activity' from the study because the task of determining meaning and context was too difficult.)

We searched for phrases containing the physical activities in multiple tenses in order to capture as much information as possible. For example, for the activity type *shoveling snow*, we searched for the forms of shovel, shovel*ing*, and shovel*ed*. Tweets were initially converted to all lowercase text, so we were assured that we were not missing data due to capitalization. To match each food phrase with its closest caloric data, we found the most closely corresponding food from the USDA National Nutrient Database, counting all vegetables and fruits in their raw form unless the phrase indicated otherwise. Similarly, we entered meats as roasted or cooked with dry heat, not fried, unless the phrase indicated otherwise or there was no homemade option. We used the nutrition content of homemade versions of foods (for example, baked goods) rather than store-bought foods unless the phrase indicated otherwise. Our approach, while systematic, was not exhaustive, nor is it the only way of taking on this challenge; there are certainly other methods that we expect to yield similar results.

Finally, we lemmatized the food phrases by their code in the USDA National Nutrient Database. If there were food phrases that were more general in each set of phrases that held the same code, we used the more general phrase as the lemma.

We lemmatized the activity phrases by their METs and activity category. Activity categories were largely the same as listed in the Compendium with slight changes due to items in Compendium being listed in a Miscellaneous category, etc. This yielded instances of physical activity phrases that were in the same activity category but were very different with the same METs being included in the same lemma. From this level of lemmatization, we then used our best judgement to break these lemmas down further until proper phrases were included in each lemma.

Acknowledgments

References

1. (2013). Health-related quality of life: Well-being concepts. Health-related quality of life: Well-being concepts. http://www.cdc.gov/hrqol/wellbeing.htm; Accessed March 29, 2014.

2. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. PLoS ONE 6: e26752.

3. Diener E, Diener M, Diener C (1995) Factors predicting the subjective well-being of nations. Journal of Personality and Social Psychology 69(5): 851-864.

4. State of the states. State of the States. http://www.gallup.com/poll/125066/State-States.aspx; Accessed March 29, 2014.

5. (2013). Stimmungsgasometer. Stimmungsgasometer. http://xn--fhlometer-q9a.de/; Accessed March 29, 2014.

6. Siebens J (2013). Extended measures of well-being: Living conditions in the United Srates: 2011. URL http://www.census.gov/prod/2013pubs/p70-136.pdf.

7. Duggan M, Brenner J (2013). The demographics of social media users—2012. URL http://www.pewinternet.org/files/old-media//Files/Reports/2013/PIP_SocialMediaUsers.pdf.

8. Signorini A, Segre AM, Polgreen PM (2011) The use of twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. PLoS ONE 6(5): e19467.

9. Prieto VM, Matos S, Alvarez M, Cacheda F, Oliveira JL (2014) Twitter: A good place to detect health conditions. PLoS ONE 9(1): e86191.

10. Walpole SC, Prieto-Merino D, Edwards P, Cleland J, Stevens G (2010) The weight of nations: an estimation of adult human biomass. PLoS ONE 5(11): e14118.

11. Paul MJ, Dredze M (2011) You are what you tweet: Aanalyzing Twittier for public health. ICWSM .

12. Watts DJ, Muhamad R, Medina D, Dodds PS (2005) Multiscale, resurgent epidemics in a hierarchcial metapopulation model. Proc Natl Acad Sci 102: 11157–11162.

13. Google Flu Trends, https://www.google.org/flutrends/; accessed March 1, 2015.

14. Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google Flu: Traps in Big Data analysis. Science Magazine 343: 1203–1205.

15. Mitchell L, Frank MR, Dodds PS, Danforth CM (2013) The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. PLoS ONE 8: e64417.

16. Dodds PS, Clark EM, Desu S, Frank MR, Reagan AJ, et al. (2015) Human language reveals a universal positivity bias. Proc Natl Acad Sci 112: 2389–2394.

17. Chunara R, Bouton L, Ayers JW, Brownstein JS (1995) Assessing the online social environment for surveillance of obesity prevalence. PLoS ONE 8(4): e61373.

18. West R, White RW, Horvitz E (2013) From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. Proceedings of the 22nd international conference on World Wide Web International World Wide Web Conferences Steering Committee : 1399-1410.

19. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, et al. (2015) Psychological language on Twitter predicts county-level heart disease mortality. Psychological Science .

20. Abbar S, Mejova Y, Weber I (2014). You tweet what you eat: Studying food consumption through Twitter. Available at: http://arxiv.org/abs/1412.4361.

21. Chew C, Eysenbach G (2012) Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. BMC Public Health 12:439.

22. Golder SA, Macy MW (2011) Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science Magazine 333: 1878–1881.

23. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57(1): 289-300.

24. Centers for disease control and prevention. `http://www.cdc.gov`, Accessed March 15, 2014.

25. (2013). State indicator report on fruits and vegetables. national Center for Chronic Disease Prevention and Health Promotion, Division of Nutrition, Physical Activity, and Obesity. centers for Disease Control and Prevention, US Department of Health and Human Services. `http://www.cdc.gov/nutrition/downloads/State-Indicator-Report-Fruits-Vegetables-2013.pdf`, Accessed March 15, 2014.

26. French MT, Popovici I, Maclean JC (2009) Do alcohol consumers exercise more? Findings from a national survey. Am J Health Promot 24: 2–10.

27. Pesta BJ, Bertsch S, McDaniel MA, Mahoney CB, Poznanski PJ (2012) Differential epidemiology: IQ, neuroticism, and chronic disease by the 50 U.S. states. Intelligence 40: 107-114.

28. Williams JR, Lessard PR, Desu S, Clark E, Bagrow JP, et al. (2014) Zipf's law holds for phrases, not words .

29. (2013). U.S. Department of Agriculture, Agricultural Research Service, USDA National Nutrient Database for Standard Reference, release 25. URL `http://www.ars.usda.gov/ba/bhnrc/ndl`.

30. Ainsworth BE, Haskell WL, Herrmann SD, Meckes N, Bassett Jr DR, et al. (2013). The compendium of physical activities tracking guide. healthy lifestyles research center, college of nursing & health innovation, arizona state university. URL `https://sites.google.com/site/compendiumofphysicalactivities/`.

31. USDA MyPlate food groups. URL `http://www.choosemyplate.gov/food-groups/`. Accessed May 15, 2015.

32. Williams JR, Clark EM, Bagrow JP, Danforth CM, Dodds PS (2015). Identifying missing dictionary entries with frequency-conserving context models. Available online at http://arxiv.org/abs/1503.02120.

33. Levenshtein V (1966) Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 10: 707.

34. blog CSL (2014). Hedonometer 2.0: Measuring happiness and using word shifts.

35. Americas Health Rankings report—State Health Statistics. `http://AmericasHealthRankings.org`, Accessed March 15, 2014.

36. CNBC overall rankings 2012. `http://www.cnbc.com/id/100016697`, Accessed March 15, 2014.

37. State Health Facts—The Henry J. Kaiser Family Foundation. `http://kff.org/statedata`, Accessed March 15, 2014.

38. America's Brain Health Index. `http://www.beautiful-minds.com/AmericasBrainHealthIndex`, Accessed March 15, 2014.

39. US Census American FactFinder. `http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml`, Accessed March 15, 2014.

40. Rentfrow PJ, Gosling SD, Jokela M, Stillwell D, Kosinski M, et al. (2013) Divided we stand: Three psychological regions of the United States and their political, economic, social, and health correlates. Journal of Personality and Social Psychology 105(6): 996-1012.

41. Strolling of the Heiders Locavore Index. `http://www.strollingoftheheifers.com/locavoreindex/`, Accessed March 15, 2014.

42. (2013). Freedom in the 50 states, mercatus center, george mason university. `http://freedominthe50states.org/`, Accessed March 15, 2014.

Figure Legends

Figure 1. Choropleth maps indicating (A) caloric input C_{in}, (B) caloric output C_{out}, and (C) caloric balance C_{bal} in the contiguous United States (including the District of Columbia) based on 50 million geotagged tweets taken from 2011-2012. Darker means higher values as per the color bars. The histograms in Figs. 3, S2, and S3 show the specific rankings according to the three variables. The phrases in **(A)** and **(B)** are those whose increased usage contribute the most to a population's C_{in} and C_{out} differing from the average (see Sec.).

Figure 2. Plots for the contiguous US showing the lack of correlation between caloric input C_{in} and caloric output C_{out}, demonstrating their separate value as they bear different kinds of information. The Pearson correlation coefficient $\hat{\rho}_{\text{p}}$ is -0.13 and the best line of fit slope is $m = -1.69$. Fig. S1 adds plots of C_{bal} as a function of C_{in} and C_{out}.

Figure 3. Histograms of caloric intake C_{in} (food), caloric output C_{out} (activity), and caloric balance C_{bal} for the states of the contiguous US, all ranked by decreasing C_{bal}. Bars indicate the difference in the three quantities from the overall average with colors corresponding to those used in Fig. 1. We provide the same set of histograms re-sorted by C_{in} and C_{out} in Figs. S2 and S3.

Figure 4. Phrase shifts showing which food phrases and physical activity phrases have the most influence on Colorado and Mississippi's top and bottom ranking for caloric balance, when compared with the average for the contiguous United States. Note that phrases are lemmas representing phrase categories. Overall, Colorado scores lower on Twitter food calories (257.4 versus 271.7) and higher on physical activity calories (203.5 versus 161.3) than Mississippi. We provide interactive phrase shifts for as part of the paper's online appendix at `http://compstorylab.org/share/papers/alajajian2015a/` and at `http://panometer.org/instruments/lexicocalorimeter`. We explain phrase (word) shifts in the main text (see Eqs. 4 and 5), and in full depth in [2] and [16] and online at `http://hedonometer.org` [34].

Figure 5. Six demographic quantities compared with caloric balance C_{bal} for the contiguous US. The inset values are the Spearman correlation coefficient $\hat{\rho}_{\text{s}}$, and the Benjamini-Hochberg q-value. See Tab. 1 for a full summary of the 37 demographic quantities studied here.

Figure 6. Screenshot of the interactive dashboard for our prototype Lexicocalorimeter site (taken 2015/07/03). An archived development version can be found as part of our paper's Online Appendices at `http://compstorylab.org/share/papers/alajajian2015a/maps.html`, and a full dynamic implementation will be part of our Panometer project at `http://panometer.org/instruments/lexicocalorimeter`.

Tables:

Demographic quantity	$\hat{\rho}_s$ for C_{bal}	q-val	$\hat{\rho}_s$ for C_{in}	q-val	$\hat{\rho}_s$ for C_{out}	q-val
1. % high blood pressure [35]	-0.78	2.64×10^{-09}	0.29	5.76×10^{-02}	-0.78	2.64×10^{-09}
2. % no physical activity in past 30 days [35]	-0.78	2.64×10^{-09}	0.55	1.63×10^{-04}	-0.66	1.51×10^{-06}
3. % have been physically active in past 30 days [35]	0.78	2.64×10^{-09}	-0.55	1.85×10^{-04}	0.67	1.29×10^{-06}
4. Adult diabetes rate [24]	-0.76	4.80×10^{-09}	0.29	6.17×10^{-02}	-0.77	2.73×10^{-09}
5. CNBC quality of life ranking [36] (lower is better)	-0.76	6.09×10^{-09}	0.27	8.64×10^{-02}	-0.77	3.60×10^{-09}
6. Heart disease death rate [37]	-0.74	1.67×10^{-08}	0.32	3.81×10^{-02}	-0.73	2.07×10^{-08}
7. % adult overweight/obesity [37]	-0.72	3.99×10^{-08}	0.53	2.34×10^{-04}	-0.59	3.07×10^{-05}
8. % adult obesity [24]	-0.72	4.78×10^{-08}	0.53	2.34×10^{-04}	-0.59	2.94×10^{-05}
9. Gallup Wellbeing score [4] (higher is better)	0.72	5.23×10^{-08}	-0.3	5.41×10^{-02}	0.73	3.99×10^{-08}
10. America's Health Rankings, overall [35] (lower is better)	-0.72	3.40×10^{-07}	0.41	7.07×10^{-03}	-0.67	2.77×10^{-06}
11. Life expectancy at birth [37] (higher is better)	0.69	3.52×10^{-07}	-0.38	1.11×10^{-02}	0.65	2.64×10^{-06}
12. % who eat fruit less than once a day [25]	-0.66	1.29×10^{-06}	0.59	2.94×10^{-05}	-0.51	5.74×10^{-04}
13. % child overweight/obesity [37]	-0.64	2.96×10^{-06}	0.25	1.02×10^{-01}	-0.64	3.06×10^{-06}
14. % who eat vegetables less than once a day [25]	-0.61	1.69×10^{-05}	0.49	8.06×10^{-04}	-0.46	1.57×10^{-03}
15. Median daily intake of fruits [25]	0.6	2.36×10^{-05}	-0.6	1.94×10^{-05}	0.41	5.45×10^{-03}
16. Smoking rate [37]	-0.59	2.73×10^{-05}	0.49	8.06×10^{-04}	-0.48	1.08×10^{-03}
17. Median household income [37]	0.5	6.46×10^{-04}	-0.51	5.74×10^{-04}	0.4	8.51×10^{-03}
18. Median daily intake of vegetables [25]	0.49	8.06×10^{-04}	-0.55	1.33×10^{-04}	0.31	4.41×10^{-02}
19. % high cholesterol [35]	-0.49	9.01×10^{-04}	0.22	1.61×10^{-01}	-0.48	9.21×10^{-04}
20. Brain health ranking [38] (lower is better)	-0.49	9.21×10^{-04}	0.6	2.77×10^{-05}	-0.29	5.70×10^{-02}
21. % with bachelor's degree or higher [6]	0.46	1.84×10^{-03}	-0.54	2.05×10^{-04}	0.33	2.85×10^{-02}
22. Colorectal cancer rate [24]	-0.44	4.24×10^{-03}	0.51	6.46×10^{-04}	-0.27	8.44×10^{-02}
23. US Census Gini index score [39] (lower is better)	-0.42	5.05×10^{-03}	-0.04	8.09×10^{-01}	-0.5	6.08×10^{-04}
24. Avg # poor mental health days, past 30 days [35]	-0.42	5.05×10^{-03}	0.11	4.97×10^{-01}	-0.48	1.06×10^{-03}
25. Neuroticism Big Five personality trait [40]	-0.38	1.24×10^{-02}	0.19	2.44×10^{-01}	-0.37	1.42×10^{-02}
26. Binge drinking rate [35]	0.38	1.36×10^{-02}	-0.15	3.41×10^{-01}	0.41	5.94×10^{-03}
27. Avg # poor physical health days, past 30 days [35]	-0.35	2.25×10^{-02}	0.18	2.47×10^{-01}	-0.38	1.15×10^{-02}
28. Farmers markets per 100,000 in pop. [25]	0.34	2.53×10^{-02}	0.05	8.06×10^{-01}	0.42	5.05×10^{-03}
29. Strolling of the Heifers locavore score (lower is better) [41]	-0.3	5.69×10^{-02}	-0.3	5.69×10^{-02}	-0.45	2.94×10^{-03}
30. Extraversion Big Five personality trait [40]	-0.28	6.76×10^{-02}	0.04	8.33×10^{-01}	-0.29	5.69×10^{-02}
31. % schools offering fruit/veg at celebrations [25]	0.24	1.25×10^{-01}	-0.47	1.47×10^{-03}	0.05	7.98×10^{-01}
32. Openness Big Five personality trait [40]	0.23	1.44×10^{-01}	-0.5	6.46×10^{-04}	0.04	8.09×10^{-01}
33. % cropland harvested for fruits/veg [25]	0.19	2.44×10^{-01}	-0.6	2.48×10^{-05}	-0.04	8.09×10^{-01}
34. Conscientiousness Big Five personality trait [40]	-0.12	4.83×10^{-01}	0.18	2.54×10^{-01}	-0.05	8.01×10^{-01}
35. % census tracts, healthy food retailer within 1/2 mile [25]	-0.03	8.41×10^{-01}	-0.51	5.51×10^{-04}	-0.24	1.30×10^{-01}
36. George Mason overall freedom ranking [42] (lower is freer)	-0.03	8.63×10^{-01}	-0.12	4.70×10^{-01}	-0.1	5.64×10^{-01}
37. Agreeableness Big Five personality trait [40]	-0.01	9.66×10^{-01}	0.23	1.40×10^{-01}	0.08	6.47×10^{-01}

Table 1. **Spearman correlation coefficients, $\hat{\rho}_s$, and Benjamini-Hochberg q-values for caloric input C_{in}, caloric output C_{out}, and caloric balance $C_{\text{bal}} = C_{\text{out}}/C_{\text{in}}$ and demographic data related to food and physical activity, Big Five personality traits [40], health and well-being rankings by state, and socioeconomic status, correlated, ordered from strongest to weakest Spearman correlations with caloric balance.** The two breaks in the table indicate significance levels of 0.01 and 0.05 for the Benjamini-Hochberg q of C_{bal}, corresponding to the first 24 demographic quantities and then the next four, numbers 25 to 28. The bottom 9 quantities were not significantly correlated with C_{bal} according to our tests.

Supplementary Figure Legends

Figure S1. Plots for the contiguous US showing the relationships $C_{\mathbf{bal}}$ versus $C_{\mathbf{in}}$ (left), and $C_{\mathbf{bal}}$ versus $C_{\mathbf{out}}$ (right). The Pearson correlation coefficient $\hat{\rho}_{\mathrm{p}}$ is included in each plot, showing a low correlation between C_{in} and C_{out}. With its larger range, caloric output C_{out} is more tightly coupled with the ratio C_{bal}.

Figure S2. Histograms as per Fig. 3 with states sorted by food rank. The bar colors correspond those used in for the choropleth maps in Fig. 1.

Figure S3. Histograms as per Fig. 3 with states sorted by activity rank. The bar colors correspond those used in for the choropleth maps in Fig. 1.

Figure S4. Food phrase shifts for Colorado, broken down into the four ways phrases may contribute to a shift, as described in Sec. . See Fig. 4A for the combined shift.

Figure S5. Food phrase shifts for Mississippi, broken down into the four ways phrases may contribute to a shift, as described in Sec. . See Fig. 4B for the combined shift.

Figure S6. Activity phrase shifts for Colorado, broken down into the four ways phrases may contribute to a shift, as described in Sec. . See Fig. 4C for the combined shift.

Figure S7. Activity phrase shifts for Mississippi, broken down into the four ways phrases may contribute to a shift, as described in Sec. . See Fig. 4D for the combined shift.