A New Source of Health Data: Facebook *Likes*

Steven H. Gittelman – *Mktg, Inc.*
Elaine R. Trimarchi – *Mktg, Inc.*
Victor W. Lange – *Mktg, Inc.*
Eugene B. Lieb – *Custom Decision Support, Inc.*
Satvinder Dhingra – *Centers for Disease Control*
Catherine Okoro – *Centers for Disease Control*
Carol G. Crawford – *Centers for Disease Control*

Online digital data offers opportunities to create models that predict otherwise unavailable information. Models have been used to help predict product usage based on demographic characteristics (Murray and Durrell K., 1999.) However the accuracy of these predictions can be poor in some cases. Online "click-through" data has been used to improve the accuracy of some demographic models (Hu, Jan et. al. 2007). The frequency with which individuals employ Web based news and research is a predictor of their gender, ethnicity and education, providing useful targeting information for ethnicity and income (Goel, Hofman, Sirer, 2012). While there are broad similarities in what various demographic groups do online, such as e-mail and social media, there are some differences that are particularly illuminating, such as the predilection to pursue news and research health. Understanding how online information could be used to predict behavior and personality traits has been a topic of growing interest in recent years (ibid).

Facebook *Likes* are a means by which online users of Facebook can identify their own preferred sites. Facebook *Likes* have been shown to be predictors of a variety of attributes. *Likes* predict intelligence, happiness, ethnicity, religious and political views, sexual orientation, and a spectrum of personality traits (Kosinski, Stillwell, and Graepel, 2013). For example, they correctly predict homosexuality and heterosexuality, African American vs. Caucasian, and Democrat vs. Republican at levels above 85%. They predict the personality trait "Openness" as well as a standard personality test.

It has been proposed that Facebook *Likes* be used as a new behavioral measure in a fashion similar to traditional questionnaires (Kosinski et. al. 2012). A major shortcoming of questionnaires is the problem of 'measurement error' – the unavoidable possibility that even in attempting to answer honestly, respondents' provide responses that may not correlate well with their actual behavior. The power of *Likes* is that they *represent* behavior. An understanding of *Like* posting behavior has already proven to be a powerful tool for marketers who seek to target purchasing populations. The potential of social media information is only beginning to be mined and understood. We may be on the threshold of understanding of how *Likes* can be applied in marketing and commerce.

Different websites attract persons of different personalities (Kosinski, et. al. 2013). Personality traits can be the drivers behind human preference and behavior (Allport, 1962). Personality measurement has traditionally been performed by the administration of a questionnaire. Personal websites are giving us a view into the personalities of their creators (Marcus et. al. 2006). Facebook profiles, (Quercia et. al., 2012) and Facebook *Likes* (Kosinski, et. al. 2012) as well as the strokes on a keyboard–mouse (click-through) (Khan et. al. 2008) all bring us closer to understanding the personalities of our subjects. We are learning who is online (Goel 2012) and the large sample sizes by which we can measure them put new dimension on the term quantitative. In one recently reported study (Goel, ibid) 250,000 respondents formed the sample frame and in another there were 180,000 (Bachrach et.al, 2012). Here we use summary data on billions of *Likes*, each of which could be considered to be a behavior. We seek to apply them to an understanding of the personalities that drive the determinants of health.

Diabetes, cancer, obesity, asthma, and heart attacks are at epidemic proportions. To deal with these health outcomes, we need the data to understand the behaviors that generate them. While morbidity and mortality data are generally available at the local level, measures of behaviors and personality that may predict these outcomes are generally not.

A sufficient quantity of data from the Behavioral Risk Factor Surveillance Study (BRFSS) and other similar survey efforts is not available at the county level in over 90 percent of American counties. The median sample size by county is 66. Aggregating data over time does not supply sufficient granularity and buries trends. Other national data (i.e. www.Countylevelrankings.org) are model based and/or aggregated from the sparsely available federal data.

The rallying call that "all health is local" has become increasingly loud (Luck et. al. 2006). Data collection efforts have lagged behind. Through funding by the Robert Woods Foundation, the Institute for Health Metrics and Evaluation (IHME) has provided county level data modeled from the BRFSS and other sources, but it is sorely lacking in behavioral data for the vast majority of American counties. The differences between counties can be stark, even within a state: in Fayette County, Georgia (a suburb of Atlanta), the average life expectancy is 80.6 years, while a few hours east in Jefferson County, the average is 8 years less. While these outcome data are available at the community level, the behavioral factors that cause these disparate outcomes are not.

At the same time, measuring health behavior is more important than ever before. Obesity, diabetes, and chronic heart conditions are not the communicable diseases that haunted us in the past. The driving determinants of these ailments may be behavior, genetics and community structure. Diet and exercise are behavioral determinants of diabetes, and the availability of fresh foods and places to exercise represent components of community structure that can make appropriate behavior more difficult to actuate.

We need new sources of local health data, and the Big Data revolution may provide a partial answer. Social Networks, such as Facebook, have expanded to include over half of the US population, allowing for interesting data on respondent lifestyle in virtually every area of the country. These data are not explicitly health-related, but statistical analysis shows that when taken together, the 'network' of an individual's *Likes* are predictive of many types of health behaviors and outcomes, regardless of the source.

We view Facebook *Likes* as a new class of data that can help us understand health conditions at a community level. To do this, the data we derive from Facebook *Likes* must be relevant to the health metrics we seek to address. Firstly, *Likes* must predict life expectancy, the ultimate outcome of one's quality of health. Predicting intermediary causes of a shortened lifespan, such as obesity and diabetes, is also a worthwhile stepping stone to that goal. But in order to specifically target the causes of these conditions, *Likes* must also be able to predict the behavioral determinants of those outcomes. If the Facebook characteristics of a region can predict exercise, smoking, and health maintenance, then a strong argument can be made in favor of the use of these data to target and correct behaviors of concern.

The demand for data is high. We hypothesize that the behaviors that drive the determinants of modern disease are behavior, lifestyle, and personality and that Facebook *Likes* are potentially a way to quantify regional patterns of these characteristics. We hypothesize that:

1. *Likes* provide a means of categorizing communities (counties)
2. *Likes* can be used as an indicator of mortality.
3. *Likes* can be used as an indicator of disease outcomes (obesity, diabetes, chronic heart disease).
4. *Likes* can be used as an indicator of the behaviors that impact disease.
5. The categorization of communities according to their *Likes* will suggest different strategies for behavioral modification.

Data & Methods:

Data for the analysis was collected from a number of sources. Health survey data was aggregated by county from the Behavioral Risk Factor Surveillance System (BRFSS). Though BRFSS data is not sampled in such a way as to ensure a representative distribution at any geographic level finer than state, at 500,000 observations collected by telephone, it represents the largest health study conducted on an annual basis. As such, it is the most appropriate source available for county-level health behavior, despite an insufficiency of sample in many areas.

The most rigorously modeled data come from the Center for Disease Control's Diabetes Data & Trends system, which uses Bayesian multilevel modeling on BRFSS data to estimate prevalence in counties where data is sparse. Estimates for obesity, diabetes, and lack of exercise were obtained from this source. For the remaining BRFSS variables, data from CountyHealthRanks.org were used, which are an aggregate of weighted BRFSS data from the past six years. However, in two cases, raw BRFSS data were the only available source for variables we deemed conceptually important and thus we made use of these estimates where necessary.[1]

Health outcomes data (life expectancy, mortality & % low birth weight) were collected from the National Vital Statistics System (NVSS) which provides population data on deaths and births in the United States. Unlike health data collected by the BRFSS, these data represent as complete a body of information on these statistics as can be achieved. As such, they can be considered to be the most reliable estimates employed by this study.

Facebook data were collected using the Facebook Advertising API, which aggregates the number of users who express interest in certain categories of items by zip code. This zip code data were then aggregated to the county level to allow for direct comparisons to the health data. Out of 127 categories, forty were selected for the model from the 'super-categories' of activities, interests, and retail & shopping. Due to rounding performed automatically by the API that routinely led to overestimates, counties with fewer than 1,000 profiles overall were excluded from the analysis. Facebook *Likes* were scored as a percentage of completed profiles in an area. Finally, in order to reduce multicollinearity caused by variation in levels of Facebook usage by county, values were divided by the average percentage of *Likes* across all categories. The resulting variables can be characterized as a measure of popularity relative to other categories.[2]

[1] Measures of "last routine checkup" and "did not receive treatment due to cost" were both based directly off of 2010 BRFSS data

[2] Though the individual variables resulting from this transformation were sometimes entirely uncorrelated with the originals, estimates using the raw and transformed variables correlated at R=0.9. Thus, we conclude that the results of the proceeding analyses are not an artifact of this transformation.

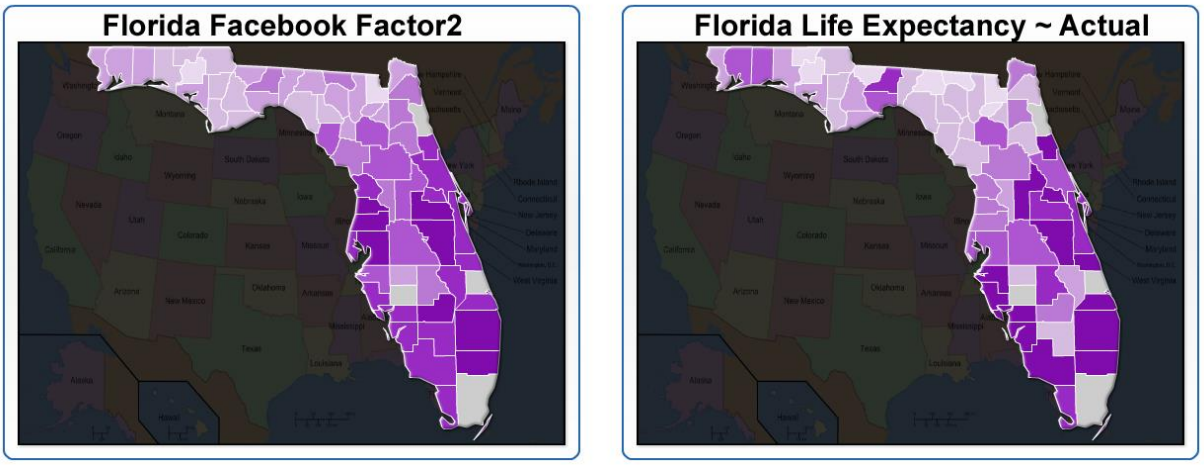
Population data, such as average income, median age, and sex ratio, was collected using the 2010 census, broken into county aggregates. Supporting county-level statistics unrelated to health were collected using "USA Counties Information" provided by the Census Bureau. (http://www.census.gov/support/USACdata.html) Overall, 1,928 counties in the continental United States contained sufficient data for all variables in the analysis.

Analysis & Discussion

The first stage in the analysis was to establish that health outcomes could indeed be determined by Facebook *Likes*. Through principal components analysis, the forty categories were reduced to nine factors[3] (varimax rotation). Due to the complex structure of *Likes* contributing to these factors, we have resisted the urge to attempt to describe the meaning of these factors. Instead, each is merely numbered in accordance with the amount of variance it explains. The full matrix of loadings for the analysis can be found in Appendix 2.

Figure 1 displays the geographic patterns associated with the prevalence of each factor in Florida. Just as one would expect given the very different populations on the panhandle and the peninsula, scores on Factor2 (our most strongly health-related factor) vary a great deal between the state's two regions. Beside the map of the factor is shown a map of life expectancy. There are substantial differences, but the general contrast between the Northern and Southern portions of the state is evident in each. This is a confirmation of our first hypothesis.

Figure 1: Distribution of a dimension of Facebook *Likes* is shown next to Life Expectancy in Florida (darker counties are higher in their respective measure).



In order to test our second hypothesis, that Facebook *Likes* can be used to predict mortality on their own, we used OLS regression. The results, as shown in the Facebook Only column of Table 1, were quite strong (model $R^2 = .69$). This is a confirmation of our second hypothesis. Though this result is encouraging, it is not sufficient to prove the

[3] These were identified through examination of the scree plot that explained 85% of the variance.

value of Facebook-based methods. Ultimately, Facebook only has value insofar as it provides predictive value beyond reliable data that is already available through the census or other means. Regression results for demographic information on average age and socioeconomic status (as embodied by average household income, unemployment rate and % with bachelor's degree) are shown in the SES only column of Table 1. There is, unsurprisingly, a very strong relationship to be found there as well (though notably less strong than for Facebook factors alone). Finally, the two groups of variables are combined in the last column of Table 1, displaying that while a great deal of the variance in life expectancy is shared by both the Facebook and SES variables, the addition of Facebook improves the model fit above and beyond readily available socioeconomic measures.

Table 1: OLS regression coefficients for life expectancy (All independent variables are standardized)

	Facebook Only	SES Only	Facebook & SES
FB Factor 1	-0.4586**	-	-0.0302
FB Factor 2	1.2112**	-	0.8461**
FB Factor 3	-0.9336**	-	-0.3356**
FB Factor 4	0.4112**	-	0.5662**
FB Factor 5	0.4947**	-	0.3774**
FB Factor 6	0.1934**	-	-0.0411
FB Factor 7	-0.0511**	-	-0.0713**
FB Factor 8	0.2269**	-	0.1337**
FB Factor 9	-0.1147**	-	-0.0085
Age	-	0.3268**	0.0330
Income	-	0.8257**	0.7105**
Education	-	0.7158**	0.4419**
Unemployment	-	-0.3074**	-0.1084**
Constant	77.1254**	77.1254**	77.1254**
R^2	0.69	0.56	0.80

*-p<.05 **-p<.01

Table 2 indicates the percent improvement in variance explained by the inclusion of Facebook *Likes* when added to SES compared to the SES alone. There are three conclusions we can draw from this model. First, Facebook *Likes* do prove to be an effective identifier of some disease outcomes. Second, there is a persistent added value of Facebook *Likes*, though its magnitude varies wildly.

Table 2: Facebook *Likes*' impact on model fit

Dependent Variable	Source	SES (R^2)	SES + Facebook (R^2)	% Improvement with Facebook
Life Expectancy	*NVSS*	0.57	0.8	40%
Mortality	*NVSS*	0.43	0.63	47%
Low Birthweight	*NVSS*	0.17	0.57	235%
Obesity	*Diabetes Data & Trends*	0.46	0.61	33%
Diabetes	*Diabetes Data & Trends*	0.41	0.65	59%

Heart Attack	*BRFSS*	0.1	0.11	10%
Lack of Exercise	*Diabetes Data & Trends*	0.49	0.69	41%
Uninsured	*CountyHealthRankings.org*	0.27	0.66	144%
Poor or Fair Health	*CountyHealthRankings.org*	0.21	0.33	57%
Smoker	*CountyHealthRankings.org*	0.039	0.11	182%
Last Checkup	*BRFSS*	0.033	0.21	536%
Declined Treatment	*BRFSS*	0.18	0.27	50%

The third result we find is that our model fit seems highly contingent on the source of our dependent variable. We achieve an impressive fit on all non-survey variables, those being life expectancy, mortality, and percent underweight births. Modeled survey variables that have been estimated by the CDC (obesity, diabetes, and lack of exercise) also provide good fits. However, unmodeled survey variables, whether using CountyHealthRankings.org or raw BRFSS 2010 data, do not generally provide very good model fit. While it is possible that the shortcoming lies with our predictors, it is at least equally as likely that these measures are themselves flawed. If the latter possibility is true, it only supports the need for innovative, non-survey methods of measuring health-related behaviors.

Our fourth hypothesis stated that Facebook *Likes*, as a measure of personality or behavior, should be able to determine the behaviors that drive health outcomes. Though the Facebook *Likes* factors had a sizeable impact in the predictive models of all tested health-related behaviors, in some cases such as health insurance and exercise, the total model fit was surprisingly good.

We can take this relationship one step further and show that much of the non-SES related impact that Facebook *Likes* have as an indicator of health outcomes such as life expectancy may be through their ability to predict behaviors. A correlated error term often indicates an underlying structure (Gerbing & Anderson, 1984). By extension, eliminating or diminishing the significance of this error term indicates that a portion of this underlying structure has been identified.

Table 3 shows the results of a structural equation model (the model fits are reported in Appendix 2) where Facebook Factors and Life Expectancy were simultaneously predicted using a sequentially increasing set of variables. First, an intercept-only model was run as a baseline. Then sequentially, age, SES (income, unemployment, and education), and finally behavior (exercise and health insurance) were added, with the covariance of error for the dependent variables of life expectancy and Facebook *Likes* factors measured at each step.

In eight out of nine cases, the correlation between the errors of each Factor and life expectancy is significant in an intercept only model, as would be expected with no predictors. In seven of the remaining eight cases, the addition of SES to the model significantly diminished this covariance, and in six cases, the addition of behavioral variables helped explain this shared error beyond the effect of socioeconomic status[4].

This finding supports the argument that Facebook *Likes* predicts differences in health outcomes not only due to serving as indicators of socioeconomic status, but as indicators of behavior as well.

Table 3: Structural Equation Modeling shows the marginal impact of behavior on models for each of the nine dimensions of Facebook *Likes*.

[4] In the case of factor 4, this was not an improvement compared to the baseline.

Covariance with Life Expectancy Error

	Intercepts Only	+Age	+SES	+Behavior
Factor 1 Error	-0.4584238**	-0.407**	-0.00699	0.008561
Factor 2 Error	1.210571**	1.199**	0.732399**	0.515878**
Factor 3 Error	-0.9331782**	-0.9231**	-0.20287**	-0.04233**
Factor 4 Error	0.4109835**	0.40541**	0.571163**	0.42469**
Factor 5 Error	0.4944573**	0.486915**	0.338354**	0.132159**
Factor 6 Error	0.1933138**	0.191003**	-0.05753**	-0.10763**
Factor 7 Error	-0.0511053*	-0.05277*	-0.06161**	-0.03399
Factor 8 Error	0.2267512**	0.226721**	0.128323**	0.047423*
Factor 9 Error	-0.1146117**	-0.11634**	-0.02849	0.001073

*-p<.05 **-p<.01

Ideally, we would be able to further advance this claim by displaying the existence of other behavioral relationships. However, as we found with predicting diseases, variables derived either from the BRFSS directly or through CountyHealthRankings.org's six year aggregates are not predicted as well by the model, and thus play only a very small part of the shared variance between Facebook and Life Expectancy.

With the exception of exercise and insurance, the model fits for most measures of health-related behaviors were only moderate. Without accurate measurement across a range of behaviors, we cannot make the more nuanced argument of which health-related behaviors Facebook *Likes* can predict, and whether they can successfully distinguish between counties with similar health outcomes but differing behavioral antecedents. As such, we are unable to reject H_0 for our fifth hypothesis, with the caveat that such relationships may exist with additional modeling.

The Predictive Model

In the preceding analysis, the predictive viability of Facebook *Likes* was demonstrated across a range of health-related variables. At this point it is difficult to deny that a correlative relationship exists, whatever the mechanism through which it operates. However, though we feel Facebook's utility has been demonstrated, there are some shortcomings in its data (for example, the rounding error mentioned earlier) that leave a great deal of its potential untapped.

We have already established that there is a need for better estimates of health needed in small communities that is not derived from surveys. We believe a statistical model can be used for this purpose that incorporates the Facebook *Likes*, but it is not necessary that Facebook *Likes* be the dominant force in the model. Until our data on Facebook *Likes* are refined further, we must bolster our predictive ability by making use of data for which our estimates are more reliable. A number of the variables used as dependent variables previously are extremely reliable, and when used as predictors of measures with less widespread data, we can increase model fit above and beyond what Facebook *Likes* and SES can do.

The results of a predictive model are shown in Table 4. In order to cross-validate the analysis (2 fold cross-validation), these results were only conducted on a random subsample of half (N=1,035) of available counties.

Cross-validation is an important component of verifying the relationships determined in a predictive model to avoid data fitting.[5] Once again, the model R^2 is quite high for our variables obtained through the NVSS and CDC's Diabetes Data & Trends system, while many estimates derived from CountyHealthRankings.org or the BRFSS are less accurate and may not be useful as a predictive model. As previously noted this may be because the BRFSS data is not of a high enough quality and as such, it makes little sense to pursue a model that can successfully predict incorrect estimates. The inclusion of vitality statistics reduces but does not eliminate the contribution of Facebook *Likes* to the model. Although we would expect demographics and vitality statistics to be very effective at predicting "healthy" versus "unhealthy" communities, we believe that the additional data provided by Facebook *Likes* should help to clarify the finer distinctions between communities with similar general levels of health.[6]

Table 4: OLS regression results for prediction of diabetes (All independent variables standardized)

Variables	β	SE
FB Factor 1	-0.0070697	0.03145
FB Factor 2	-0.5715777**	0.036641
FB Factor 3	0.4476107**	0.037421
FB Factor 4	-0.4139983**	0.034799
FB Factor 5	-0.2839315**	0.028475
FB Factor 6	0.0600914*	0.028775
FB Factor 7	-0.0001286	0.024898
FB Factor 8	-0.1300811**	0.025713
FB Factor 9	0.1320938**	0.026149
Life Expectancy	-0.5801971**	0.095772
Mortality	0.0932389	0.068009
Low Birth weight	0.1968717**	0.041725
Average Income	-0.1710613**	0.047091
Education	-0.1781927**	0.057903
Unemployment	0.0148643	0.03033
Age	0.6218488**	0.029532
Constant	10.2976**	0.024601
$R^2 = 0.74$		

Figure 3 shows a graphical comparison of estimates versus source data in a state where data is generally available (South Carolina), dynamically shaded from light to dark in accordance with the % incidence of diabetes[7]. As should be apparent visually, the fit is generally good – 90% of errors in the model fall inside of ±1.5%, or 0.7 standard deviations from CDC estimated values. The same process is repeated for lack of exercise in Figure 4. Finally, for general health[8], where there is a substantial amount of missing data, we show how this predictive model can

[5] The average error of the resulting model was 8% and predictions correlated at R=.994

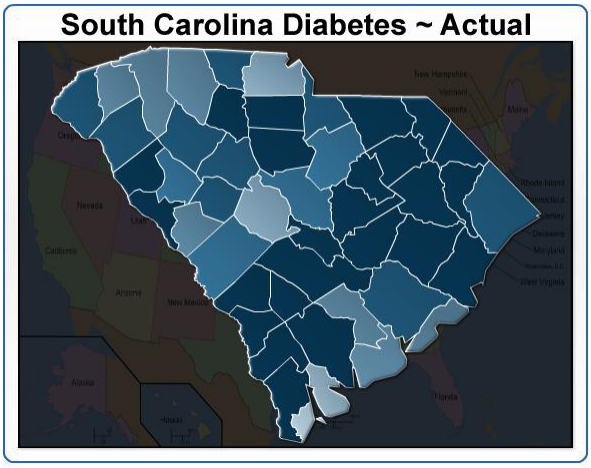
[6] The correlation between predictions of two diseases only dropped from .89 to .86 with the addition of Facebook likes (t-test = 1.36, p<.1) providing only marginal support for this theory.

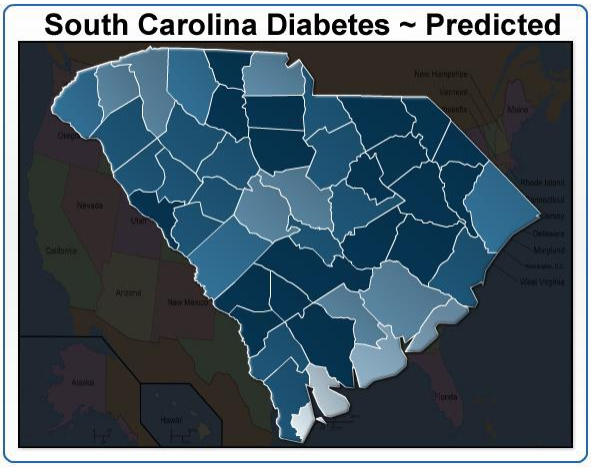
[7] As this model was calculated using data available in all counties, some counties not present in earlier models were included.

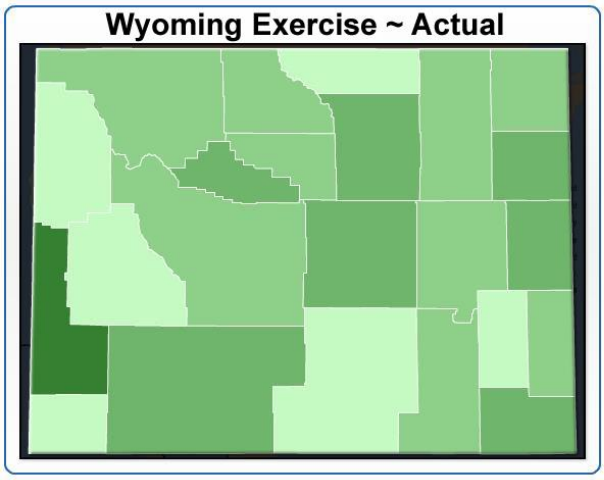
[8] While CountyHealthRankings.org provides an approximation of general health, we chose to make use of raw BRFSS data instead due to a better predictive model.

create imputed values in a state with a large number of low-population counties, thus effectively increasing our level of knowledge.

Figures 3 & 4: Actual statistics (left) and predictions (right) of diabetes (top) and exercise(bottom) generated by our predictive model.







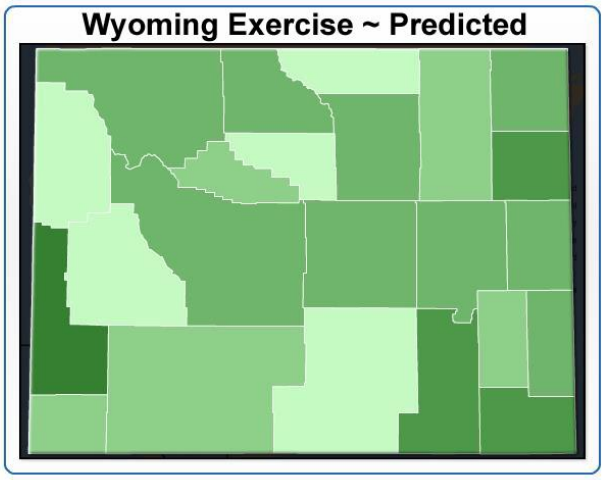
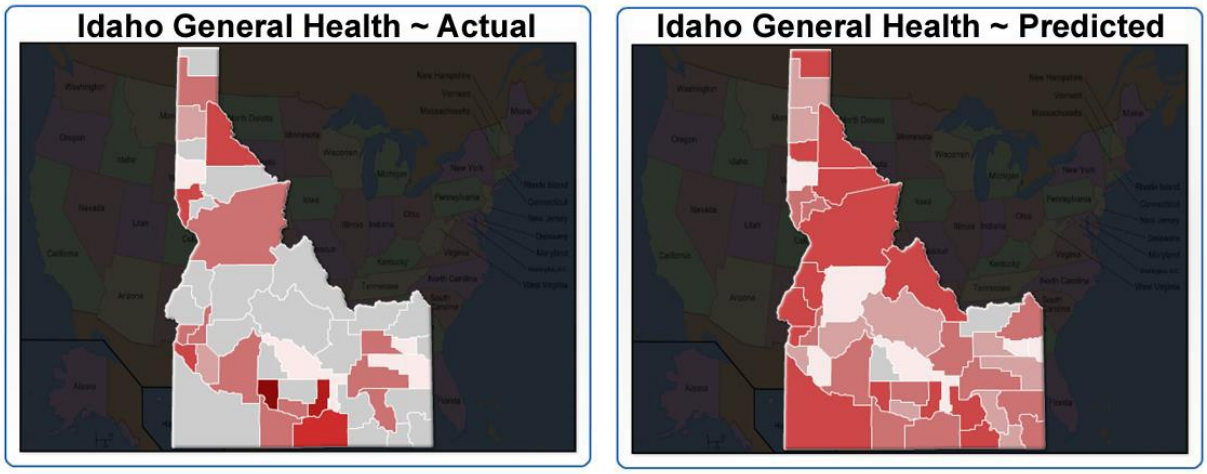


Figure 5: Imputed Data from the Predictive model (right) is used to fill where the BRFSS collected an insufficient number of respondents.



Conclusion:

We have demonstrated the effectiveness of Facebook as a predictor across a range of health-related variables. Furthermore, we have shown that its effectiveness as a predictor is owed in part to behavior and it is this latter finding that we deem more significant, even while further research is required.

Given that *Likes*, by their nature, revolve around commercial activities, whether they take the form of entertainment, food, or packaged goods, it is unavoidable that a great deal of what we see in the variation of *Likes* is driven by socioeconomic status. In parallel, given the nature of our healthcare system, wealth and health outcomes are inextricably linked. Nevertheless, while this socioeconomic component of health will be difficult to alter, there is a further element that is linked to voluntary behavior that can improve one's health in the areas of nutrition, exercise, and health maintenance. Unfortunately, at this point, we feel that our measures of all but exercise are too weak to form a complete picture of how well Facebook could potentially predict these other behaviors.

When we first undertook this research, it was our expectation that the larger part of the measurement error that would impact our results would come through the imprecise categorization and geographic aggregation of the Facebook Data. But while there are some exceptions, the pattern of fit across data sources is inescapable. When our models do so well in predicting verifiable health statistics and almost as well with modeled health data, their failure casts more doubt on the raw survey data collected from other sources than on Facebook.

Thus, while we argue that Facebook may serve an intermediary role in augmenting sparse data at a community level, we see this as an intermediate goal. In the shorter term, the survey measures we use for health-related behaviors must first be improved, either through the Bayesian modeling method employed for diabetes by the Center of Disease Control or through alternative forms of data collection. Once reliable measures exist for enough of the US population, a "Big Data" augmented model can be estimated that will serve to fill in the gaps and diminish the need for costly, periodic updates to this data.

There are two primary implications of the finding that a behavioral element serves as the means by which Facebook *Likes* can predict health behavior. The first is that when properly elaborated upon, the method promises

to eventually allow for distinctions between communities that go beyond simple 'good' or 'bad' health. Unfortunately the current data set does not allow for this extension of the analysis.

The ultimate goal of our analysis of Facebook *Likes* is to establish the potential contribution of Big Data to research that directly impacts government spending and public policy. At a fraction of the cost of traditional research, data that might seem on its face to have little to do with health can predict life expectancy and epidemic-level health problems such as diabetes and obesity. With the effectiveness of RDD interviewing methods falling into doubt, and with the importance of local-level samples becoming apparent, the emergence of this potential source comes just at the right time.

Whether this data ultimately comes from Facebook or not is of little importance; the online landscape may change, and it may be a different source of data that proves more viable in five years. Even if Facebook does prove to endure as a social institution, there is still room for a great deal of improvement on the models presented here. With cooperation from the social media outlets themselves, we may be able to attain better estimates in categories that align better with our needs.

Ultimately, we do not see the potential of such research beginning and ending with health. With data that can stand in for measures of social conditions, behavior, and personality, Big Data can become a great deal more than the target of privacy advocates that it is today. In its place, we could have a trove of valuable local data for the purposes of not only of market researchers, but for social scientists and policymakers, as well.

Appendix 1: Variable Descriptions

Control Variables	Source	Question Wording or Description
Average Household Income	2010 Census	Mark the "Yes" box for each income source received during 2009 to a maximum of $999,999.
Median Age	2010 Census	What is this person's age and what is this person's date of birth?
Percent with bachelors degree	2010 Census	What is the highest degree or level or school this person has COMPLETED?
% Unemployed	Bureau of Labor Statistics (2010)	% in Labor Force without a job
Life Expectancy	National Vital Statistics System (2009)	Average age of death in a county
Adjusted Mortality	National Vital Statistics System (2009)	Age-Adjusted Death Rates
% of Underweight Births	National Vital Statistics System (2009)	% of babies born underweight
Obesity	CDC Diabetes Data & Trends - based on BRFSS (2009)	BMI > 30 based on self-reported height and weight
Diabetes	CDC Diabetes Data & Trends - based on BRFSS (2009)	Has a doctor, nurse, or other health professional ever told you you had Diabetes? (Diabetes caused by pregnancy excluded)
Lack of Exercise	CDC Diabetes Data & Trends - based on BRFSS (2009)	No to: During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calistehenics, golf, gardening, or walking for exercise?
Uninsured	CountyHealthRankings - Based on BRFSS (2010)	No to: Do you currently have health insurance?
General Health/Poor or Fair Health	CountyHealthRankings - Based on BRFSS (2010)	In general, would you say your health is Excellent, Very Good, Good, Fair or Poor?
Smokes Every Day	CountyHealthRankings - Based on BRFSS (2010)	(To those who have smoked 100 cigarettes) Do you now smoke cigarettes every day, some days, or not at all?
Last Checkup	BRFSS (2010)	About how long has it been since you last visited a doctor for a routine checkup?
Declined Treatment	BRFSS (2010)	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?

Appendix 2: Ancillary Tables

Table Ia & Ib: Rotated (Orthogonal Varimax) Factors & Loadings

	Eigenvalue	Difference	Proportion	Cumulative %

Factor1	8.15809	3.39113	0.2205	0.2205
Factor2	4.76697	0.80242	0.1288	0.3493
Factor3	3.96455	0.48799	0.1072	0.4565
Factor4	3.47656	0.38719	0.094	0.5504
Factor5	3.08937	0.12598	0.0835	0.6339
Factor6	2.9339	0.43231	0.0801	0.714
Factor7	2.53109	0.98687	0.0684	0.7824
Factor8	1.54422	0.50612	0.0417	0.8242
Factor9	1.031	.	0.0281	0.8522

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Auto Intenders	0.9617	-0.0206	0.0506	-0.0335	0.0001	-0.0725	-0.003
Automobiles	0.3698	0.177	0.4983	0.4453	0.2792	0.0582	0.0489
Beauty	0.537	-0.1706	0.5095	-0.2078	-0.2586	0.0561	0.1883
Beer/Wine/Spirits	0.2143	0.7312	0.0987	0.0406	0.0798	-0.1622	0.3439
Charity	-0.1008	-0.8091	0.244	-0.038	-0.1489	0.1103	0.0665
Electronics	-0.4048	-0.0684	0.7577	-0.176	-0.1939	-0.1786	0.1409
Cooking	-0.0201	0.3403	-0.2026	0.1151	0.2414	0.7686	0.1273
Dancing	0.2056	-0.0511	0.0176	-0.0877	-0.0092	-0.1315	-0.9444
Do-It-Yourselfing	0.2259	0.2603	-0.0591	0.4813	0.3427	0.5401	0.1425
Teaching	0.8527	-0.0852	-0.046	-0.1278	-0.0368	-0.0866	-0.4094
Television	-0.6677	0.3328	0.1216	0.0645	0.2408	0.2605	0.2731
Environment	0.2074	0.3268	-0.3263	0.6258	0.4052	0.1468	0.0214
Planning	0.9613	-0.0424	0.0369	-0.0378	0.0068	-0.0732	-0.0121
Fashion	-0.6427	-0.1994	0.4196	-0.2189	-0.1012	-0.2475	0.2634
Fast Food	0.1112	-0.136	0.0866	-0.082	-0.9261	-0.1337	0.0497
Food & Dining	0.0419	0.0082	0.0021	-0.0072	0.0928	0.9079	0.0815
Frequent Casual Diners	-0.1064	-0.1554	0.0172	-0.0608	-0.9344	-0.1388	0.0831
Game Consoles	0.0494	0.2899	0.4897	-0.0439	0.0436	-0.203	0.0553
Social Gaming	-0.0266	0.0822	0.8227	0.2308	0.0454	-0.1446	0.1124
Gardening	0.8088	0.0324	-0.0031	0.3319	0.192	0.171	0.0439
Health & Wellness	0.014	0.4426	-0.4525	0.4023	0.4108	0.2875	0.0867
Home & Garden	0.1721	0.1869	0.0922	0.2671	0.2818	0.2568	-0.0433
Literature	-0.0926	0.1469	-0.1568	0.4665	0.3935	0.4739	0.0782
Luxury Items	0.3351	0.099	-0.1855	-0.7809	-0.2401	-0.0426	0.1214
News	-0.3251	0.3166	-0.3513	-0.1588	0.1599	0.2863	0.097
Outdoor Activities & Fitness	-0.0683	0.0995	-0.1691	0.1884	0.1308	-0.0445	-0.9235
Pets	0.1006	0.5345	0.0872	0.3108	0.1958	0.3842	0.1424
Cats	0.9165	0.1689	0.103	-0.0442	0.0054	0.1107	0.0205
Dogs	0.7732	0.4081	0.0819	0.1486	0.1385	0.0239	0.0569
Photo Uploads	0.8038	0.0517	0.0238	-0.1711	-0.2023	0.0203	0.0273
Photography	0.7768	0.1544	-0.3233	0.0586	0.1081	0.2214	0.065

Politics	-0.2038	-0.7978	-0.088	-0.359	-0.0424	-0.2274	0.1694
Conservative Politics	-0.06	-0.7883	-0.0923	0.1785	-0.1902	-0.3164	0.143
Liberal Politics	-0.0113	-0.2381	-0.0724	-0.8902	0.1221	-0.0034	0.0546
Nonpartisan Politics	-0.2404	0.8188	-0.0362	0.2682	0.1204	0.1603	0.0523
Pop Culture	-0.3853	0.2409	0.5176	0.0767	-0.0714	0.0621	0.3114
Travel	-0.2209	0.0504	-0.8341	-0.0663	0.0519	-0.1225	0.056

Variable	Factor8	Factor9	Uniqueness
Auto Intenders	0.0731	-0.0683	0.0557
Automobiles	-0.2184	0.1008	0.2437
Beauty	-0.2353	0.1445	0.198
Beer/Wine/Spirits	-0.0481	-0.2966	0.1668
Charity	-0.2328	0.1228	0.1662
Electronics	0.0683	0.1537	0.1088
Cooking	0.1536	-0.0495	0.1383
Dancing	0.0196	0.0245	0.0369
Do-It-Yourselfing	0.2333	-0.0669	0.1577
Teaching	-0.0632	-0.0638	0.0627
Television	-0.264	0.0027	0.1543
Environment	0.1359	0.0562	0.1443
Planning	0.0517	-0.0868	0.0555
Fashion	0.1254	-0.0359	0.1654
Fast Food	-0.1002	0.0468	0.0647
Food & Dining	0.0603	-0.0443	0.153
Frequent Casual Diners	-0.0871	-0.0549	0.0506
Game Consoles	0.2291	0.6379	0.1662
Social Gaming	0.1367	0.1547	0.1843
Gardening	0.1765	-0.1569	0.1109
Health & Wellness	0.0981	0.0649	0.1645
Home & Garden	0.7878	0.1349	0.0696
Literature	-0.0708	0.1615	0.311
Luxury Items	-0.1079	0.1557	0.1236
News	0.4123	0.0895	0.3506
Outdoor Activities & Fitness	0.0254	-0.045	0.0467
Pets	-0.1453	0.2468	0.3119
Cats	-0.0046	0.1111	0.0939
Dogs	-0.059	0.1952	0.1423
Photo Uploads	-0.0699	0.0941	0.2656
Photography	0.0952	0.1413	0.1708
Politics	-0.0715	-0.2392	0.0409
Conservative Politics	-0.0303	-0.2351	0.1217
Liberal Politics	-0.0451	-0.0374	0.1242

Nonpartisan Politics	-0.0224	0.0484	0.1527
Pop Culture	-0.4353	0.1707	0.1952
Travel	0.0899	0.1453	0.1985

Table II: Model Fit of Structural Components (full model)

	R^2
avglifeexp	0.689406
factor1	0.232181
factor2	0.221529
factor3	0.59206
factor4	0.144479
factor5	0.352521
factor6	0.162388
factor7	0.0085
factor8	0.046014
factor9	0.05321

Works Cited

Allport, G. W. (1962). The general and the unique in psychological science. *Journal of personality* 30, 405-422.

Bachrach, Y., Pushmeet, K., Kosinski, M., Stillwell, D., and Graepel, T. (2012). Personality and Patterns of Facebook Usage. *Web Science'12*. June 22-24, Evanston, IL.

Gerbing, D. and Anderson, J. (1984) On the meaning of within-Factor Correlated Measurement Errors. *Journal of Consumer Research* 11, 1, 572-580.

Goel, S., Hofman, J., Sirer, I. (2012). Who Does What on the Web: A Large-Scale Study of Browsing Behavior. *Association for the Advancement of Artificial Intelligence*. July 27-28, Toronto, Ontario.

Hu, J., Zeng, H.-J., Li, H., Niu, C. and Chen, Z. (2007). Demographic Prediction Based on User's Browsing Behavior. In *WWW*, 151-160.

Khan, I.A., Brinkman, W.P., Fine, N. Hieruns, R.M. (2008) Measuring Personality from Keyboard and Mouse Use. *European Conference on Cognitive Ergonomics*. September 16-19, Madeira, Portugal.

Kosinski, M., Stillwell, D., and Graepel, T. (2012). Private traits and attributes are predicable from digital records of human behavior. www.pnas.org/cgi/doi/10.1073/pnas.1218772110

Luck, J., C. Chang, Brown, E.R., and Lumpkin, J. (2006). Using local health information to promote public health. *Health Affairs.* 25, 4, 979-991.

Marcus, B., Machilek, F., and Schutz, A. (2006). Personality in cyberspace: Personal web sites as media for personality expressions and impressions. *Journal of Personality and Psychology* 90, 6, 1014-1031.

Murray, D., and Durrell, K. (2000). Inferring demographics attributes of anonymous internet users *Lecture Notes in Computer Science,* 1836, 7-20.

Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., and Crowcroft, J. (2012). The personality of popular Facebook users. *Association for Computing, Computer Supported Cooperative Work* February 11-15, Seattle, Washington.