Use #BigData to #UnderstandSociety

Mehrdad Yazdani

Culture Analytics Long Workshop
Institute for Pure and Applied Mathematics at the University of California Los Angeles

March 16, 2016

OpenMedicine INSTITUTE

QUALCOMM INSTITUTE

UC San Diego

The potential for online social networks

- Everyday hundreds of millions of users voluntarily share thoughts, feelings, and opinions at scales never seen

- Can we use this Big Data scale of thoughts, feelings, and opinions as a "lens" to gain insight into society?

Overview

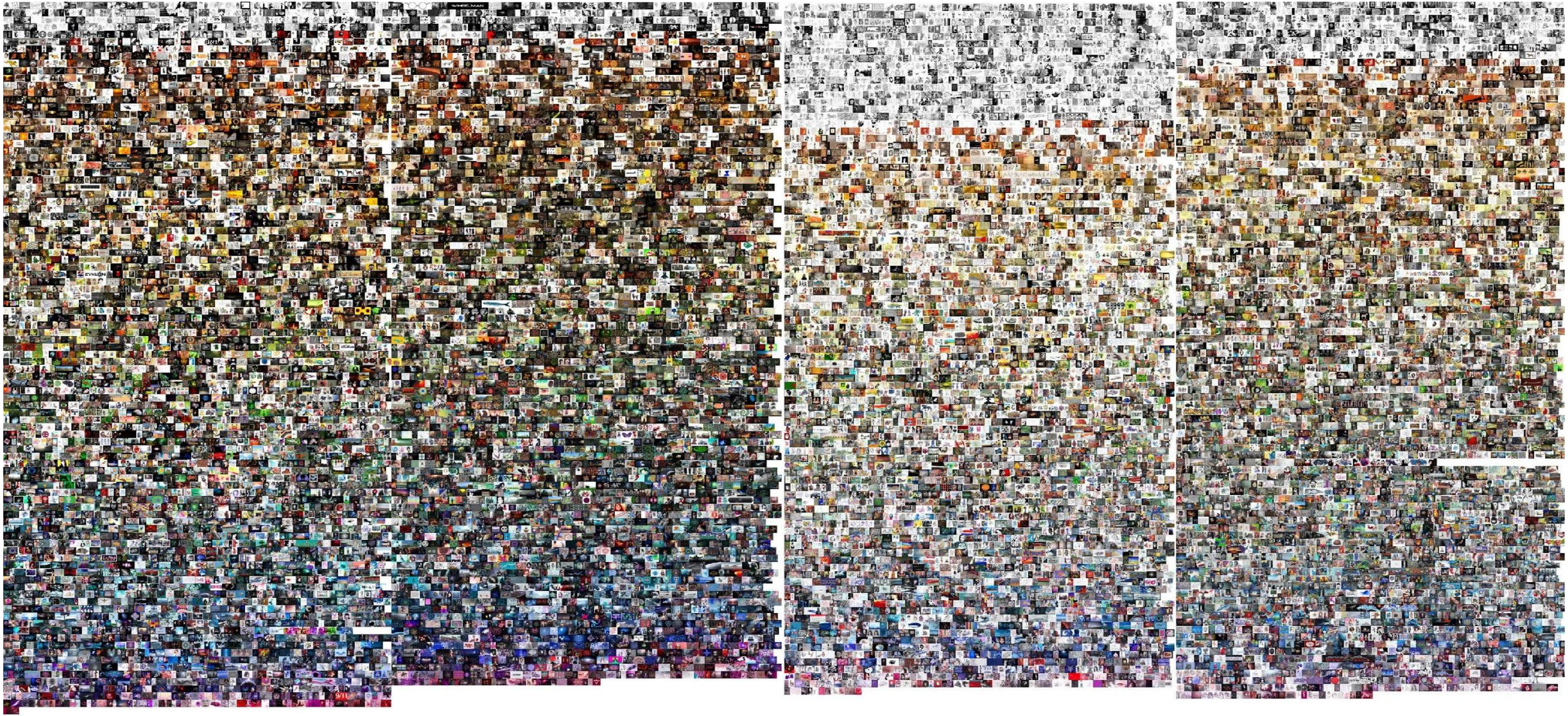
- Understanding Culture: DeviantArt

- Understanding Society: Twitter

DeviantArt

- Online community for sharing artistic works (amateurs and professionals)

- Study the temporal changes of 270,000 digital and traditional artworks from 2001 to 2010



Digital Art 2004

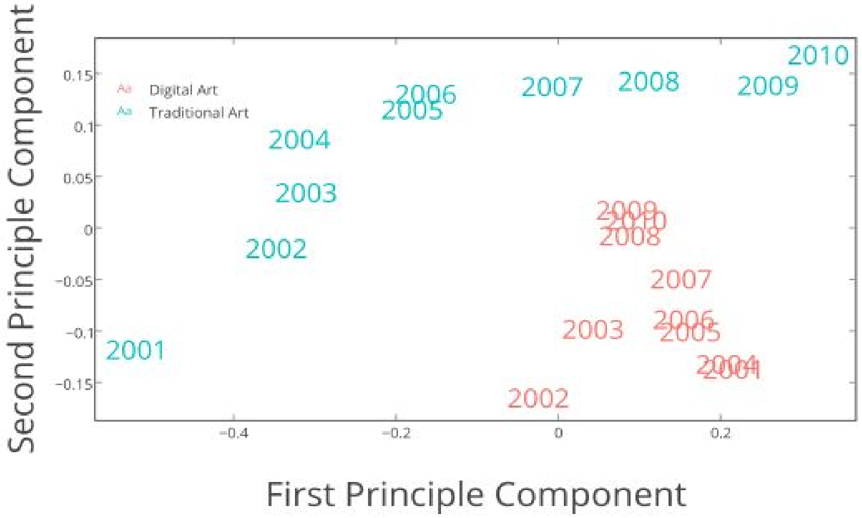
Digital Art 2010

Traditional Art 2004

Traditional Art 2010

Apply Quantitative Methods

- Extract *aggregated* color histograms per year for both categories (Digital and Traditional Art)



Bin	Weights: Both Categories	Weights: Digital Art	Weights: Traditional Art
1	3.11E-05	1.20E-04	**9.20E+01***
2	**1.66E+03***	**2.32E+03***	**4.78E+02***
3	**2.75E+03***	**1.46E+09***	2.22E-08
4	6.96E-04	4.86E-04	2.75E-08
5	7.28E-04	**8.80E+02***	4.46E-08
6	8.05E-04	1.89E-03	1.17E-08
7	8.19E-04	**5.78E+03***	1.05E-08
8	2.14E-04	5.91E-04	**2.78E+02***

Table 3. Changes in Hue histograms vs time differences calculated using a metric learned from Equation 4.

Overview

- Understanding Culture: DeviantArt
- Understanding Society: Twitter

Do social networks provide a clear enough lens?

Important questions to keep in mind:

- Do users only share the banal?
- Is social media only for the narcissist?
- Is there a sample bias to youth?

Example investigations

- “Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures” by Golder and Macy (2008)

- “The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place” by Mitchell et. al (2013)

- “Psychological language on Twitter predicts county-level heart disease mortality” by Eichstaedt et. al (2015)

What about images?

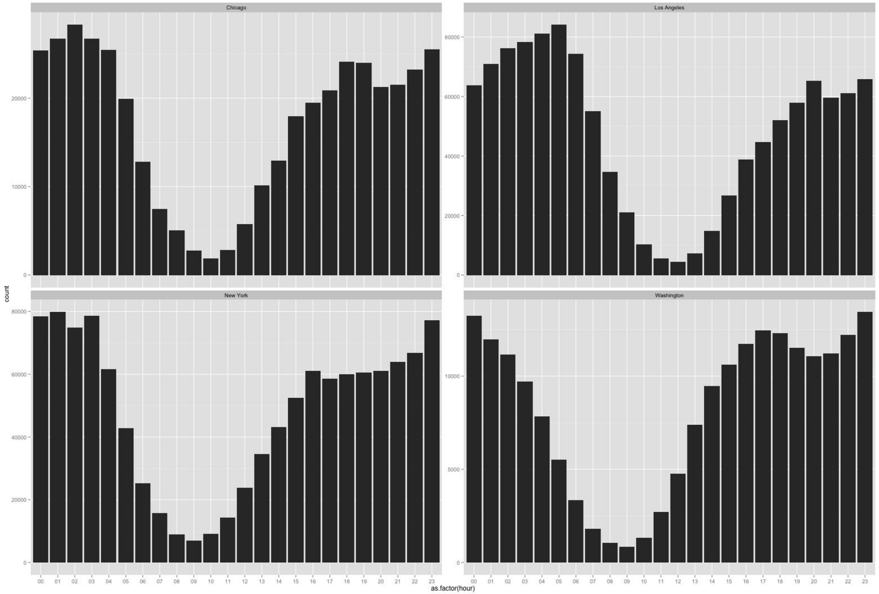
- Text limits us to specific language
- Increasingly, social media users share content beyond just text
- We propose that images compliment text and together can be used to form stronger signals in measuring the well-being of society

Challenges with social media images

- How do we actually go about measuring features that are relevant for determining social well-being?

- First step: look at metadata

Volume of tweeted images per hour for 4 different cities for a single day

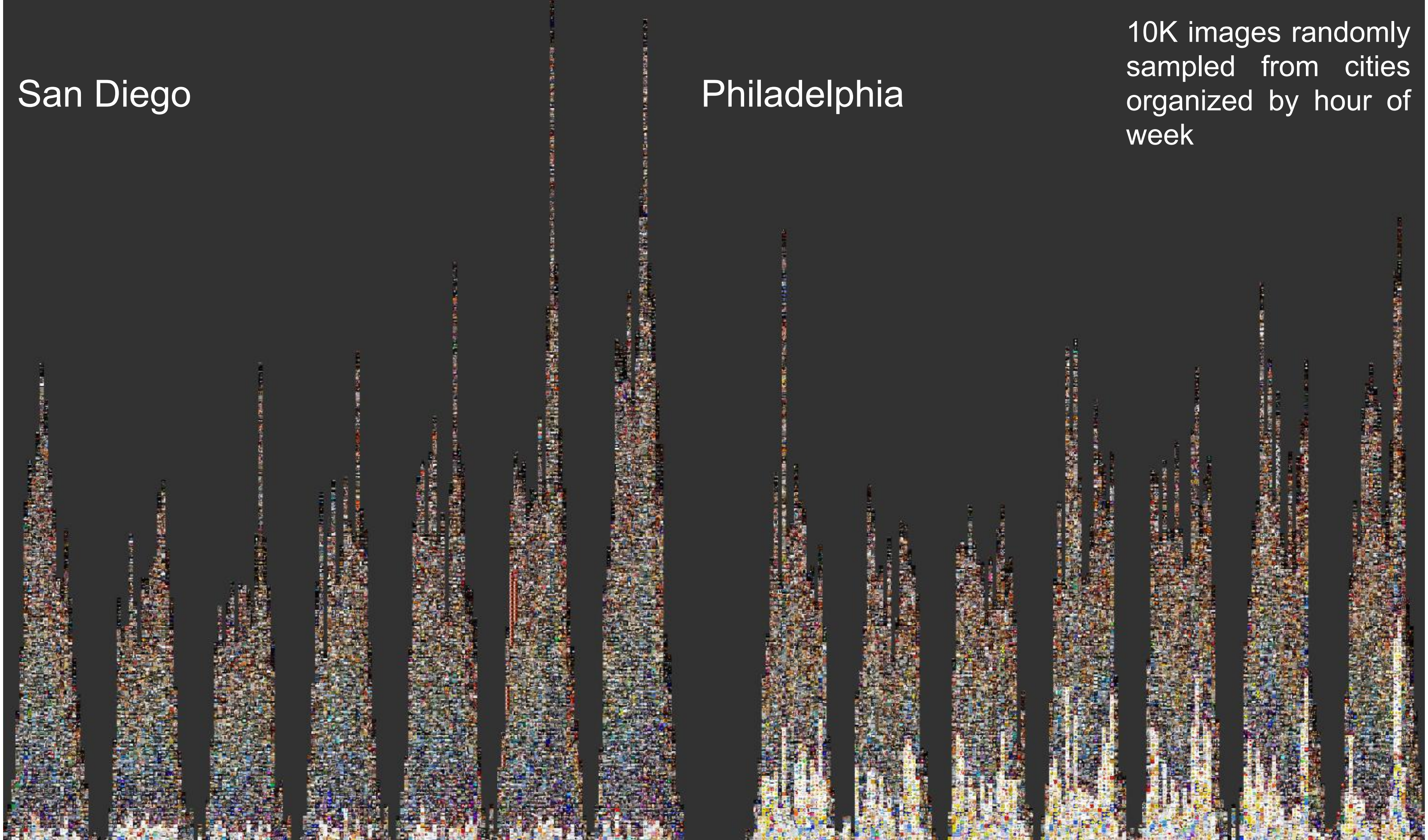


What about content of images?

New York

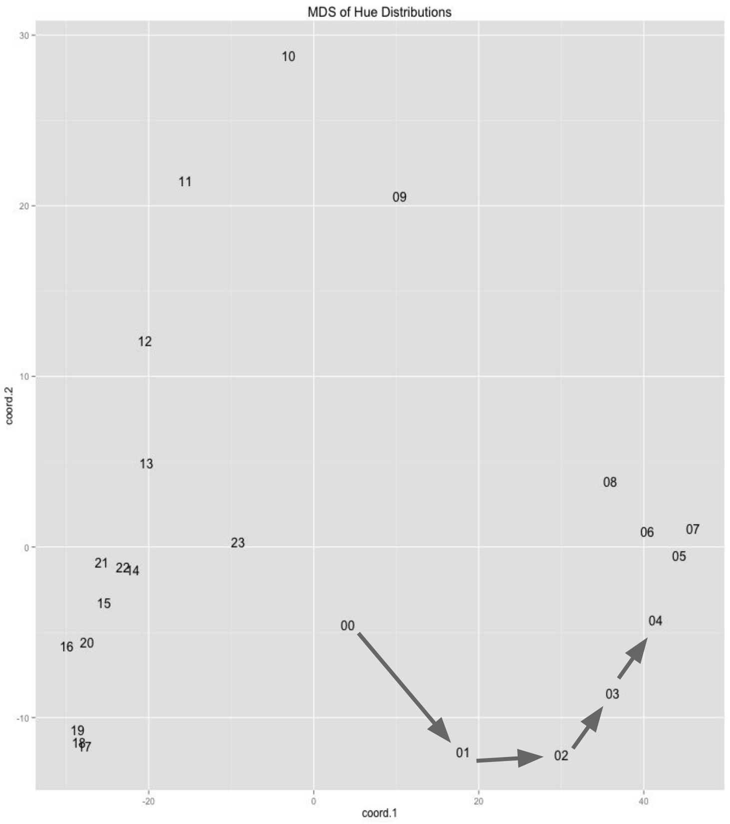
Tokyo

Hochman, Chow, Manovich
Phototrails.net


San Diego
Philadelphia
10K images randomly sampled from cities organized by hour of week

MDS of color distributions

- A trajectory of aggregated color distributions
- Does each city have a specific trajectory?
- Does the unique trajectory for each city suggest something about cultural and societal differences?



Can we systematically study the content of images?

- Recent advances in deep learning allow us to classify content of images at high accuracies

- GoogLeNet convolutional neural network won the ImageNet challenge in 2014 reported to have an error rate ~6% (human error rate ~5%)

- Available for free as open source through the Caffe framework provided by UC Berkeley

Non-photographic images

- Take random 50K sample of images from the top 20 populous cities from the lower 48 of the United States

- In our data, the most popular category (out of 1,000 categories) is the category for “web site, website, internet site, site”

- We refer to these images as non-photographic images “image-texts”

City	Volume	City	Volume
New York	1034643	Jacksonville	79850
Los Angeles	810046	Seattle	78139
Houston	405051	Milwaukee	75941
Chicago	334422	Mesa	73567
Dallas	290407	Detroit	71079
Fort Worth	271916	Cleveland	71055
Washington	238254	New Orleans	69473
Philadelphia	229252	Tucson	58937
San Antonio	228038	Baltimore	56520
San Diego	227794	Sacramento	53649
San Francisco	192470	Raleigh	53624
Boston	186484	Wichita	52635
Phoenix	177377	Minneapolis	51944
Austin	167255	Tulsa	50996
Arlington	132146	Omaha	50814
Long Beach	122521	Oakland	50283
Las Vegas	119437	Louisville	50236
Columbus	111506	Memphis	49207
San Jose	109444	Fresno	44687
Tampa	109387	Riverside	44557
Nashville	102341	Virginia Beach	43278
Atlanta	98322	St. Louis	41098
Anaheim	96452	Albuquerque	40291
Denver	96151	Bakersfield	39582
Oklahoma City	94246	Lexington	39100
Charlotte	94024	Corpus Christi	34199
Kansas City	93991	El Paso	32547
Portland	93729	Colorado Springs	30502
Indianapolis	84863	Santa Ana	25750
Miami	83999	Aurora	22048

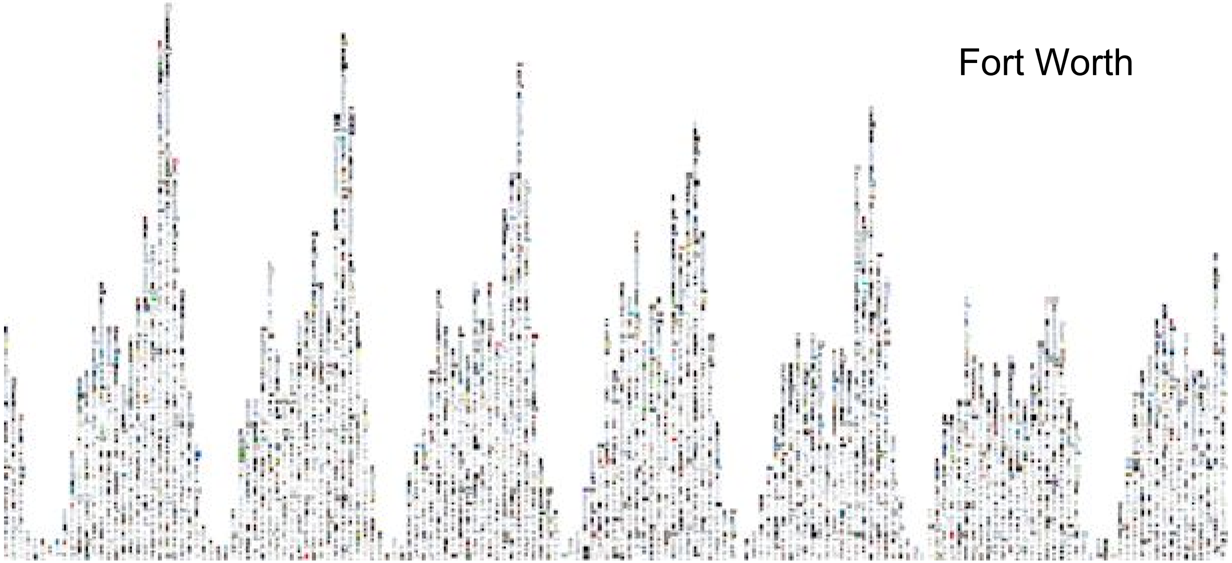
TABLE I. 60 U.S. CITIES SORTED BY NUMBER OF GEOLOCATED IMAGES PUBLICALY SHARED ON TWITTER IN 2013. THE TOP 20 CITIES USED IN OUR CITY ARE HIGHLIGHTED IN **BOLD**.

F.O.O.D.
Fans Of One Direction
CRIME ALERT
When someone says they don't like Zac Efron.
FRESH SUMMER
HAIRSPRAY

Compute features from non-photographic images

Different cities have different **proportions** of non-photographic images.

Are these differences indicative of socio-cultural difference between these cities?



We quantify the temporal distributions of non-photographic images with the **entropy** of their hourly distributions.

$$\mathbf{X24}_g(h) = \frac{1}{K_h^g} \sum_{k=1}^{K_h^g} I(l_k^{g,t_h} = l^*)$$

Different cities have different *temporal* distributions.

Are these differences indicative of socio-cultural difference between these cities?

Indicator	Correlation	P-value
Median Housing Price	-0.5638	0.007735
Rate of Bachelor's Degree	-0.6413	0.001623
Average Income	-0.4772	0.01805
Social well-being	0.56100	0.001623

TABLE II. PEARSON CORRELATIONS BETWEEN THE PROPORTION OF IMAGES CLASSIFIED AS IMAGE-TEXTS AND FOUR SOCIO-ECONOMIC VARIABLES (FIGURE 2).

Sources of socio-economic variables:

1. Median housing price (Zillow)
2. Bachelor's degree rate (Census)
3. Average income (Census)
4. Social well-being (Gallup survey)

Indicator	Correlation	P-value
Median Housing Price	-0.5332	0.007735
Rate of Bachelor's Degree	-0.62451	0.001623
Average Income	-0.4709	0.01805
Social well-being	0.5381	0.001623

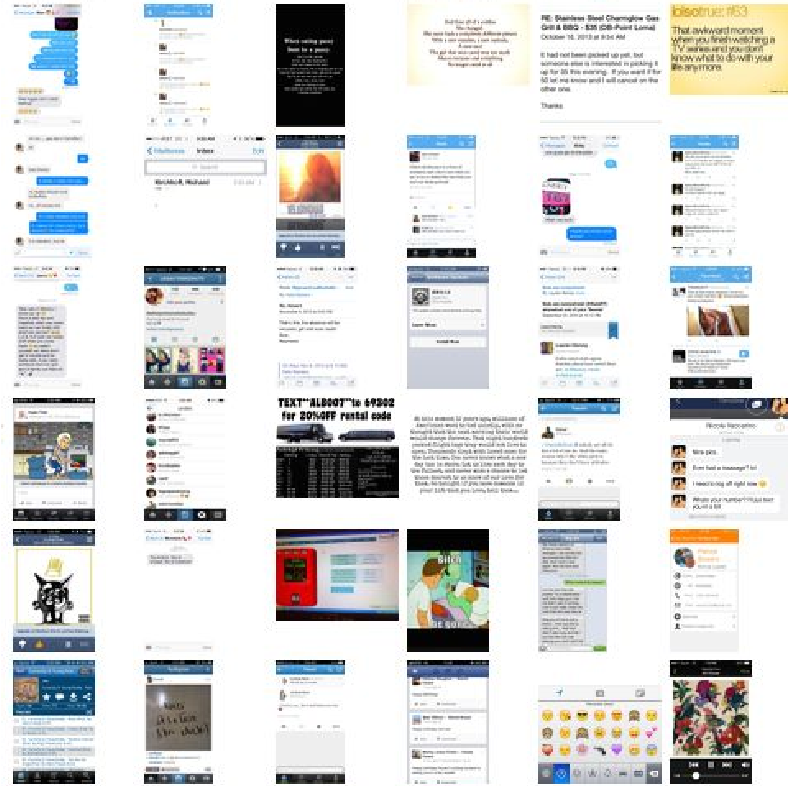
TABLE III. PEARSON CORRELATIONS BETWEEN THE ENTROPY MEASURES COMPUTED FROM THE SERIES IN EQUATIONS 1 AND 2 AND FOUR SOCIO-ECONOMIC INDICATORS.

Similar results using other measures for correlation (eg, Spearman Rank)

Image-texts positively correlate with social well-being

- Cities that report being more *socially* satisfied, tend to also share *more* image-texts

- This may be linked to the fact that one of the most consistent sub-categories in image-texts are screenshots of text message conversations



Summary

- Our work suggests that images in social media have features that relate to scoio-economic variables
- Other content types should also be investigated (including texts on images)
- Future work should combine features from both images and texts to form a more complete picture

Use #BigData to #UnderstandSociety

Mehrdad Yazdani

@crude2refined

myazdani@ucsd.edu