Methods and Initial Results

MACS 30200
Leoson Hoay

The methodology deployed in this study attempts to define predictive models of *reader linger time* and *average site visits* of online science news outlets, based on input variables that measure the aesthetics of articles that originate from the outlet in question. Another question of interest that may be explored with the same set of data is whether these measures can also *classify* the media outlet that the articles belong to. The exogenous variables for the models are similar.

Regression

Table 1: Model variables

Variables	Symbol	Exogenous/Endogenous
Average Site Visits	V	Endogenous
Average Linger Time	L	Endogenous
Semantic Consistency	c	Exogenous
Kolmogorov Complexity	ω	Exogenous
Shannon Entropy	α	Exogenous
Edge Density	x_1	Exogenous
Colorfulness	x_2	Exogenous
Screen Equilibrium	x_3	Exogenous
Screen Symmetry	x_4	Exogenous

This study defines the two multivariate polynomial regression models as follows (beta values are estimated separately for each model):

$$V = \beta_1 c + \beta_2 \omega + \beta_3 \alpha + \beta_4 x_1 + \beta_5 x_2 + \beta_5 x_2^2 + \beta_6 x_3 + \beta_7 x_4 \tag{1}$$

$$L = \beta_1 c + \beta_2 \omega + \beta_3 \alpha + \beta_4 x_1 + \beta_5 x_2 + \beta_5 x_2^2 + \beta_6 x_3 + \beta_7 x_4 \tag{2}$$

Take note that a squared term for Colorfulness is included in the model, as previous studies have shown that some level of colorfulness in a website does credit to its aesthetic appeal, but very high levels of colorfulness generally result in lower aesthetic ratings (Reinecke et. al 2013 and others).

The average monthly site visits and average linger time data were obtained from *SEM-Rush.com*, a proprietary service that provides traffic estimates based on search engine crawls. Bear in mind that these are not exact estimates, but was the most available data obtainable within a limited timeframe. Below are bar charts of the distributions of each dependent variable of 5 of these popular science media websites.

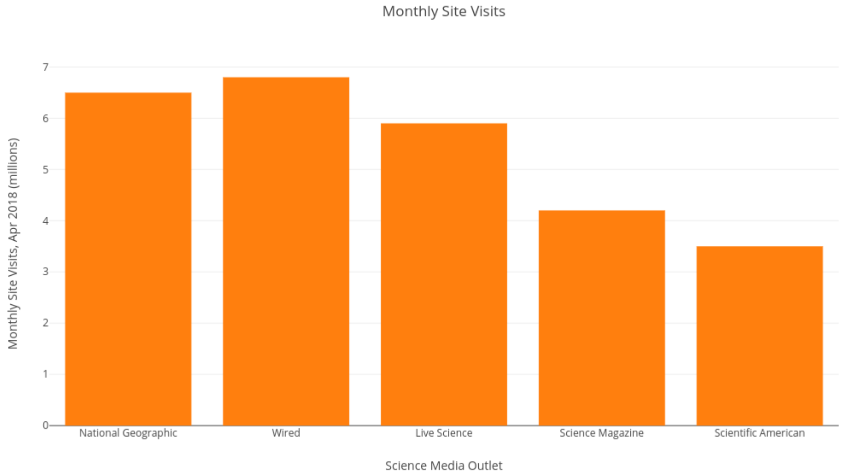

Figure 1

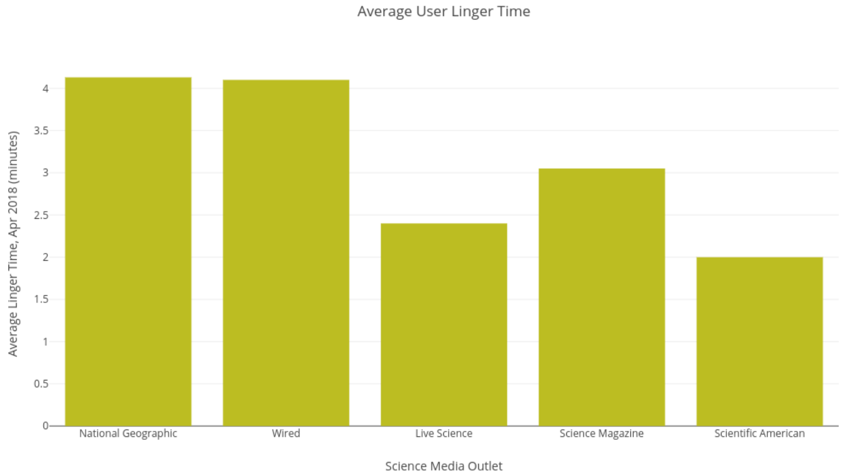

Figure 2

As of this writeup, 18 news sites have been considered in the dataset (the eventual numerical aim is 25), and they are categorized into three sets based on disparities in dates of establishment and monthly visit estimations. The means of monthly visits for the sets are *18.55 million, 4.43 million, and 457.7 thousand* respectively. The standard deviations of visits are *15.75 million, 1.89 million, and 272.36 thousand* respectively. The means of linger times for the sets are *3.88, 2.74, and 2.60 minutes* respectively. The standard deviations of times are *0.71, 0.97, and 0.56 minutes* respectively. It is readily observable that Set 1 has a large standard deviation for visits, likely due to the inclusion of *The Guardian*, which pulls about 4.5 times more visitors per month (45.8 million) than the next runner-up. At this point, the outlet is considered an outlier and will not be included in the analysis.

Classification

An RF model is used to initialize the classification of unseen articles (and their parameters) into the various media outlets. Random subsets of the variables are taken to define the best split at each node of the classification trees, created by bootstrapping samples on the training data. Performance is internally assessed with prediction errors using the Leave-One-Out cross-validation method. For this particular study, the forest is grown in Python using the scikit-learn package. The dependent variables are the various science media outlets, and the RF is trained to classify articles to these outlets based on the abovementioned independent variables. The number of trees is set to 100.

Extracting Data and Computing Aesthetic Measures

The extraction of the text data was performed by adapting the *boilerpipe* Java library (Kohlschtter, 2009), a library developed for Boilerplate removal and text extraction from HTML pages. URLs of science media articles are passed into the code, which then pulls the main article text from a page. An example output of running the code on a National

Geographic article is shown below, and example code is also available on Github [1].

> New Zealand's newest sinkhole may be one of its largest. It extends down more than six stories. From end-to-end, it measures just about the length of two football fields. The sinkhole is so large it even exposed 60,000-year-old volcanic soil.
>
> New Zealand volcanologist Brad Scott told a local news outlet it was the largest sinkhole he had ever seen and that it had potential to get even bigger. The feature appeared on the country's North Island after a long period of record heavy rains. A local farmworker rounding up cows discovered the opening, narrowly avoiding falling into it while riding his bike. Speaking with New Zealand's Newshub outlet shortly after the sinkhole appeared, farm manager Colin Tremain said he plans to erect a fence to prevent cows from falling in.
>
> It's not the first sinkhole to recently open up in the region. Nine additional sinkholes have formed there in the past few years. New Zealand has several major fault lines running the length of the country, and sinkholes are thought to be more likely to occur near fault lines where soluble rocks wash through. The new sinkhole opened up over pumice terrain, but it's only one of many sites around the world where the ground is prone to collapsing underfoot.
>
> Where Sinkholes Happen
>
> According to the U.S. Geological Survey , sinkholes open up when groundwater doesn't drain from the surface and dissolves the rock lying underneath. Limestone and salt beds are often the sources of sudden sinkholes. While New Zealand's new sinkhole has tall, vertical walls that make it look more like the Grand Canyon, some sinkholes form a rounder bowl shape. Their formations are often dramatic. As water erodes the ground underneath sinkholes over time, the weight of the land that remains on top eventually becomes too much, and it collapses. In regions where people have built homes or businesses, the sudden collapse can be deadly. Entire homes have been swallowed up by the Earth.
>
> In the U.S., Florida, Texas, Alabama, Missouri, Kentucky, Tennessee, and Pennsylvania have the terrain most conducive to forming a sinkhole, says the USGS's website.
>
> (This is how one sinkhole opened up in this Florida community.)
>
> While less common, sinkholes can also be caused by human land development. Activities like construction and pumping groundwater can make the ground less stable and a sinkhole more likely.
>
> Around the world, any region that has easily dissolvable terrain could potentially develop sinkholes. Mexico and Belize are built on an abundance of soluble rock. Parts of Italy, Slovenia, Croatia, and Russia are also prone to sinkholes.
>
> China is home to the world's largest cluster of sinkholes and the region's terrain has, over thousands of years, yielded enormous rock towers and one of the world's largest systems of caves. [2].

Extracting image data of the websites was performed using the *PhantomJS* Javascript API. PhantomJS enables the automation of screenshot capturing by operating as a headless browser. There were several difficulties that had to be resolved during this process, one of which was some websites that had overlays or advertisements that would obstruct and grey out the content of interest, resulting in 'shadowed' screenshots. When building the javascript code, it was straightforward to develop a workaround for some of the websites, but not for the others. Hence, to standardize, screenshots with inverted backgrounds were 're-inverted' using color manipulation in PhantomJS, resulting in a white background and readable black text. This occured in 2 out of the 18 domains. Example code and output for

[1] https://github.com/LeosonH/MACS30200proj

[2] https://news.nationalgeographic.com/2018/05/new-zealand-sinkholes-volcanic-fault-lines-science-spd/

this process is also on Github, but running it requires pre-installing PhantomJS dependencies on one's machine.

Text aesthetics is operationalized as document congruence for each article , which is implemented using the *Gensim* Python module. For each of the articles from a specific news media outlet, a corpus is built out of all the other sampled articles from the same source, excluding the one in question. The corpus and the article are then vectorized, and a similarity score computed using TF-IDF for each article against the corpus. The similarity scores are then averaged to obtain a document consistency measure for each news media outlet.

Image aesthetics is conceptualized as the various constructs mentioned at the beginning of the writeup, and computed using a combination of *OpenCV* in Python and *EBimage* in R, both of which are open source image analysis and computer vision tools. Due to the drawn-out process of implementing the algorithms, only Screen Equilibrium (x_3), Shannon Entropy (α), and Colorfulness (x_2) have been implemented. Screen Equilibrium is operationalized as:

$$(x_c, y_c) = (\frac{\sum_i a_i x_i}{\sum_i a_i}, \frac{\sum_i a_i y_i}{\sum_i a_i}) \tag{3}$$

, where (x_c, y_c) is the center of the screen, a_i is the area of an element i on the screen, and (x_i, y_i) is the position of the element i on the screen. The Euclidean distance between the computed value and the center of the screen is then standardized between 0 to 1 to form a measure of equilibrium.

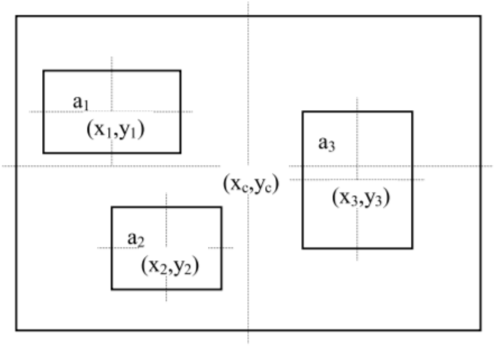

Figure 3: A screen in equilibrium (Ngo et. al, 2000).

Shannon Entropy is operationalized by adapting the classical entropy model as follows:

$$Entropy = \sum_k p_k \log p_k \tag{4}$$

, where p_k is the probability that the difference between the grayscale DN values of 2 adjacent pixels is equal to i. Colorfulness is operationalized by first defining the opposing color spaces:

$$rg = R - G \tag{5}$$

$$yb = \frac{R - G}{2} - B \tag{6}$$

Then define the standard deviation(μ) and mean(σ), before computing the colorfulness metric C (Hassler and Susstrunk, 2003):

$$\mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{7}$$

$$\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} \tag{8}$$

$$C = \sigma_{rgyb} + 0.3 \times \mu_{rgyb} \tag{9}$$

The computed values are then standardized to a value between 0 and 1 based on the total range of colorfulness scores to obtain the input values for the final model.

After all the remaining measures have been implemented, PCA will be performed to assess any overlap and covariations between the existing measures before feeding them into the final models.

Descriptives and Initial Results

Since the algorithms for the image analysis have not been completely implemented, running the Random Forest Classifier at this time is infeasible. However, given that the bulk of the complexity in this study lie with data preparation and measure implementation, it is beneficial to allocate more time to this part of the process before running the final models. Presented below are figures that represent descriptive data based on the currently collected measures of interest. Due to space constraints, only data for the 5 outlets presented above will be displayed in the figures to follow, except Figure 4, which shows the top 6 rated by Colorfulness.

Table 2: Document Consistency

Media Outlet	Average Document Consistency
National Geographic	0.76
Wired	0.63
Live Science	0.73
Science Magazine	0.54
Scientific American	0.71

Document consistency scores might have some variation depending on factors such as the number of writers in each media outlet, how diverse their working team and areas of expertise are, and whether or not there has been recent changes in the editorial make-up of the company. However, it is not unreasonable to assume that ”consistency should stay relatively consistent” over time, given that the science writers of a specific organization are embedded within the same company environment and brand, and are likely to have to go through similar training and screening processes.

In the case of screen equilibrium, the scores should not vary much from article to article, since most articles from one website should be formatted in a similar manner. This was indeed the case. The variation between media outlets was also much smaller than that of document consistency, which also makes sense, since web article formatting is more likely to be more similar between outlets than writing style.

Table 3: Screen Equilibrium

Media Outlet	Average Screen Equilibrium
National Geographic	0.79
Wired	0.86
Live Science	0.76
Science Magazine	0.76
Scientific American	0.82

Differences in colorfulness *within* a particular media outlet is definitely more significant (due to inclusion of different pictures, page elements in each article), and hence it is of paramount concern to average out these differences.

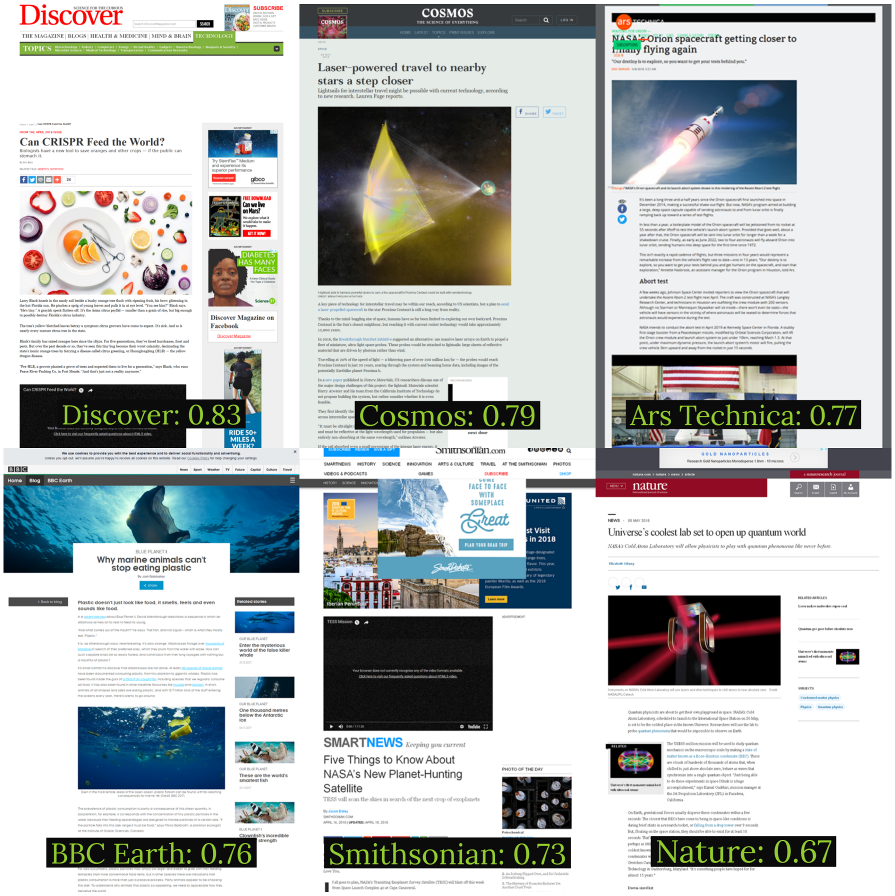
Discover
Can CRISPR Feed the World?
Discover: 0.83
COSMOS
Laser-powered travel to nearby stars a step closer
Cosmos: 0.79
ars technica
NASA's Orion spacecraft getting closer to finally flying again
Ars Technica: 0.77
BBC Earth
Why marine animals can't stop eating plastic
BBC Earth: 0.76
Smithsonian.com
SMARTNEWS
Five Things to Know About NASA's New Planet-Hunting Satellite
Smithsonian: 0.73
nature
Universe's coolest lab set to open up quantum world
Nature: 0.67

Figure 4: Standardized Average Colorfulness Scores, Top 6