THE AESTHETICS OF KNOWLEDGE CONSUMPTION:

[A Study of Textual and Graphical Forms in Online Science Communication]

Project Proposal

Leoson Hoay

RESEARCH QUESTION

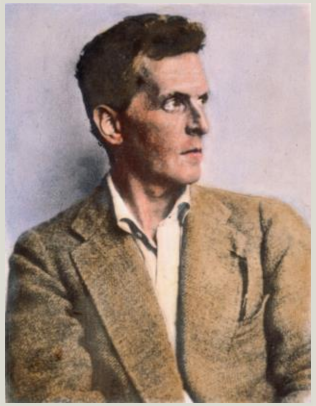
- Can **aesthetic measures** of science web articles predict the **readership** and **reader linger time** of the publication/website that the articles belong to? (Observational)

…and/or:

- Can aesthetic measures on web articles predict readers' **ratings of scientific content/websites**, and their **interest in the aforementioned content?** (Survey/Experimental)

FOUNDATIONS

- “Ethics and Aesthetics are one.”/”Knowledge is in the end based on acknowledgement.” – Ludwig Wittgenstein (1914 - 1916, 1953)
  - ***Value* and Aesthetics are inextricable** (Gombrich, 1960)
  - *Build on previous studies in HCI (Human-Computer Interaction) and knowledge aesthetics*
- **Defining and Quantifying “Aesthetics”**
  - “Formal notions” relating a reader to the content
    - **Form** and Function
  - **Text Aesthetics:** Semantic Consistency (Tang, Qin and Liu, 2015)
  - **Layout Aesthetics:** HCI/UX Literature – Pixel Fields, Screen Balance, Entropy, Complexity, Gestalt Unity, Edge Density, etc. (Machado et. al 2015, Rigau et. al 2007, Ngo et. al 2000 and others)



FONTS MATTER
You'll always be mine...
YOU'LL ALWAYS BE MINE...

BITS AND PIECES

- **Text Aesthetics:** Semantic Consistency
- **Layout Aesthetics (6 measures):** (+Color Distribution, +Edge Density)**:**

$$M_D(r) = 1 - avg_{1i<jr}\{NID(i,j)\}$$

Kolmogorov Complexity

$$(x_c, y_c) = \left(\frac{\sum_i a_i x_i}{\sum_i a_i}, \frac{\sum_i a_i y_i}{\sum_i a_i}\right)$$

Screen Equilibrium

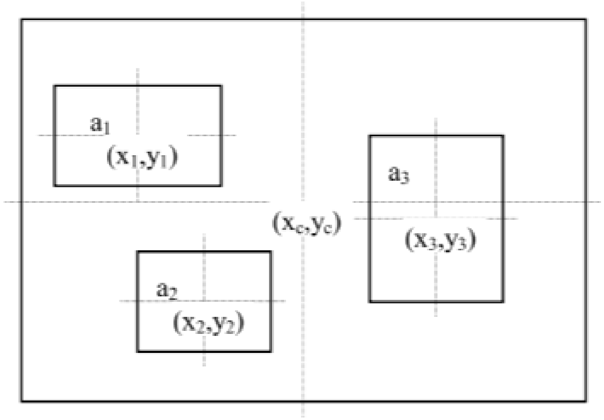


Fig. 2. A screen layout in equilibrium.

Gravitational model of screen equilibrium (Ngo and others 2000, 2002)

$$M_I(r) = \frac{I(X_l, \hat{Y}_r)}{H(X_l)}$$

Shannon Entropy

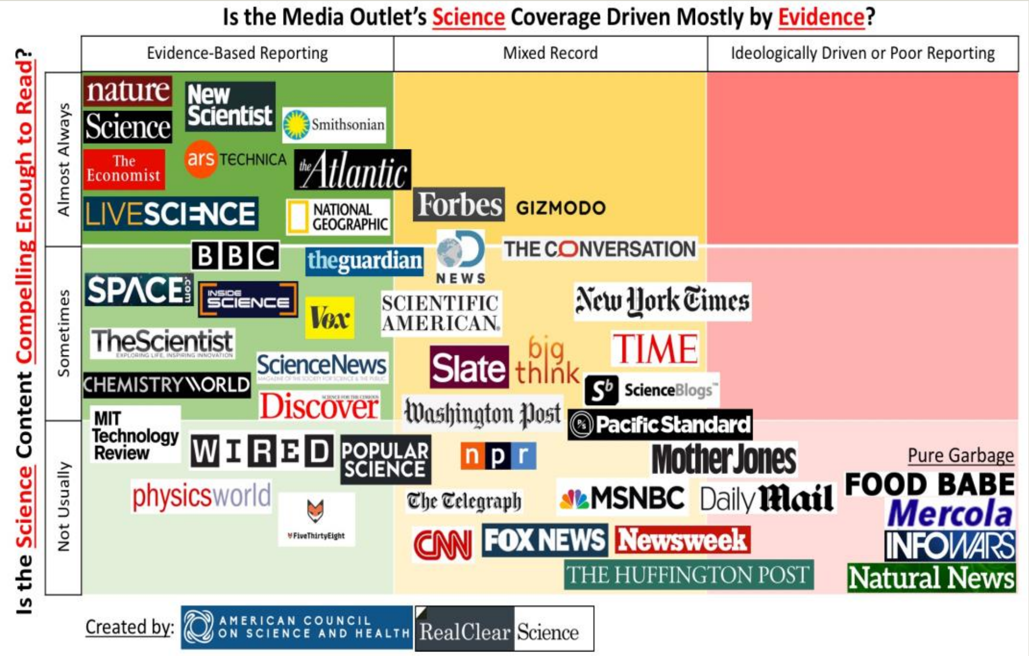
$$sqm = (p, d)$$

Screen Sequence

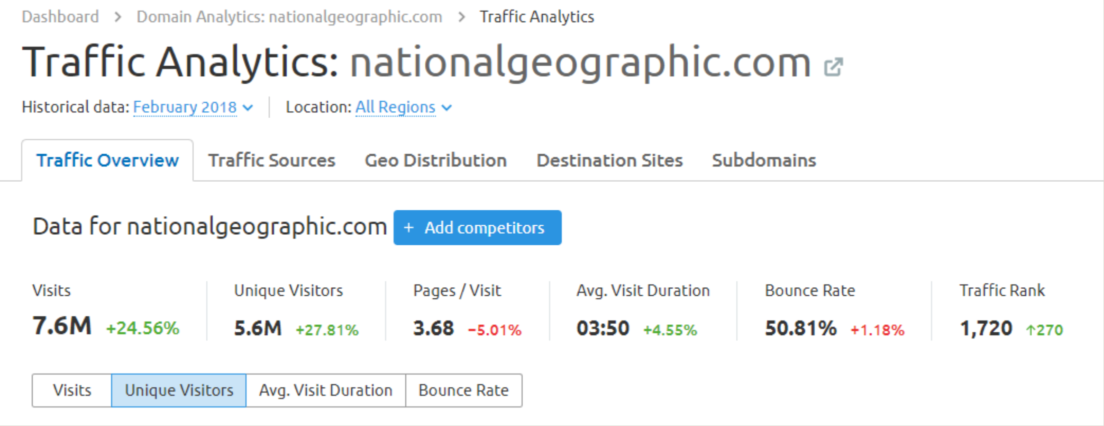
DATA SOURCE(S)

- Aesthetics: Science Website Article Layouts and Text Content
  - *Popular American web publications – National Geographic, BBC Earth, Nature, WIRED, New Scientist, etc. (Include global publications?)*
  - *Layouts: **PhantomJS** to scrape screenshots of article pages*
- Readership, Linger Time: Website Metrics
  - *Domain Data*
  - ***Estimated Data (SimilarWeb, SEMRush)***

DATA SOURCES



American Council on Science and Health, RealClearScience (2017)



Traffic Analytics: nationalgeographic.com

Historical data: February 2018 | Location: All Regions

Data for nationalgeographic.com

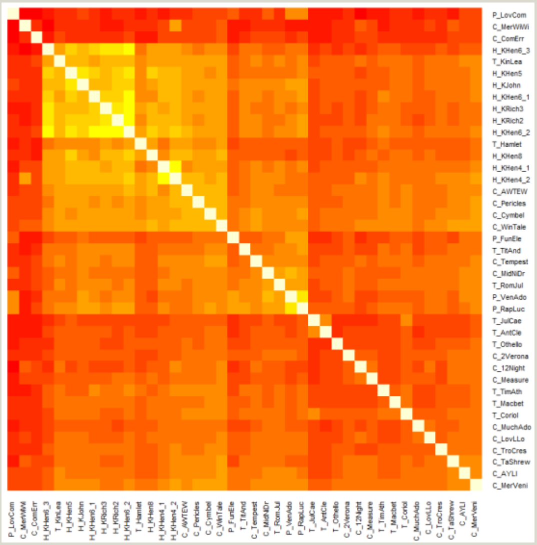
Visits	Unique Visitors	Pages / Visit	Avg. Visit Duration	Bounce Rate	Traffic Rank
7.6M +24.56%	5.6M +27.81%	3.68 -5.01%	03:50 +4.55%	50.81% +1.18%	1,720 ↑270

SEMRush.com

33 articles x 30 publications = 990 data points

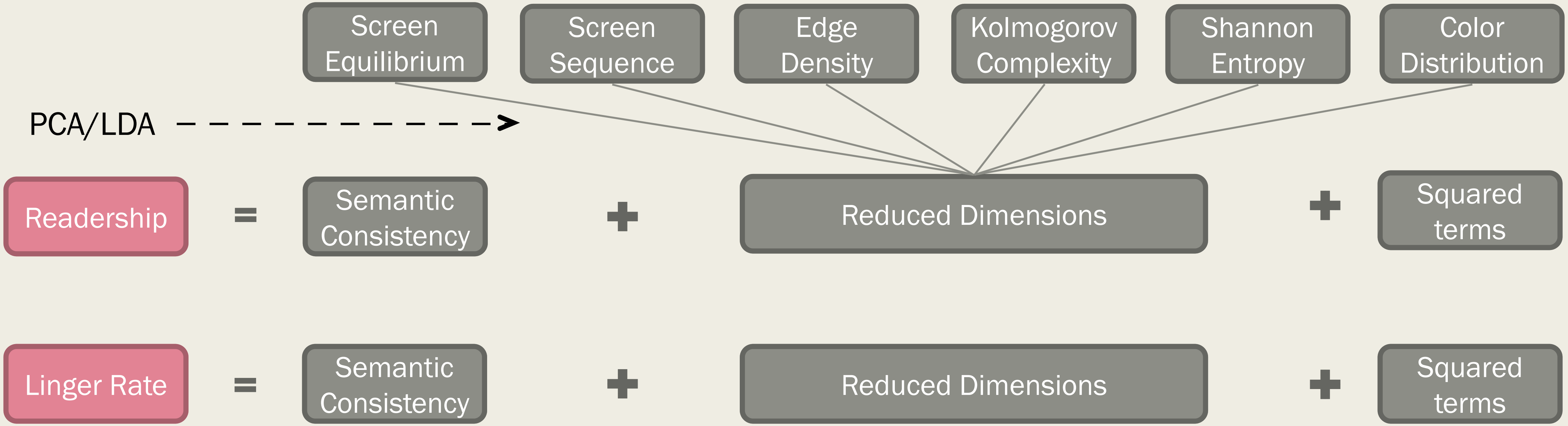
METHODS

- **Article Text Consistency:** Gensim + Doc2Vec/Word2Vec
  - ***Trained vector space of documents*** *from each individual science media outlet used to calculate individual article similarity*
  - ***'Document Congruence'*** *for each science article formulated as the inverse of the document distance from the vector space*
  - *Other models: Cosine Similarity, WMD (Word Mover's Distance)*
- **Article Webpage Aesthetics:**
  - ***EBImage*** *in R for pixel analysis,* ***scikit-image*** *for clustering,* ***OCRopus*** *for layout analysis,* ***OpenCV*** *for almost everything else*



METHODS

- Supervised Learning



Screen Equilibrium

Screen Sequence

Edge Density

Kolmogorov Complexity

Shannon Entropy

Color Distribution

PCA/LDA

Readership = Semantic Consistency + Reduced Dimensions + Squared terms

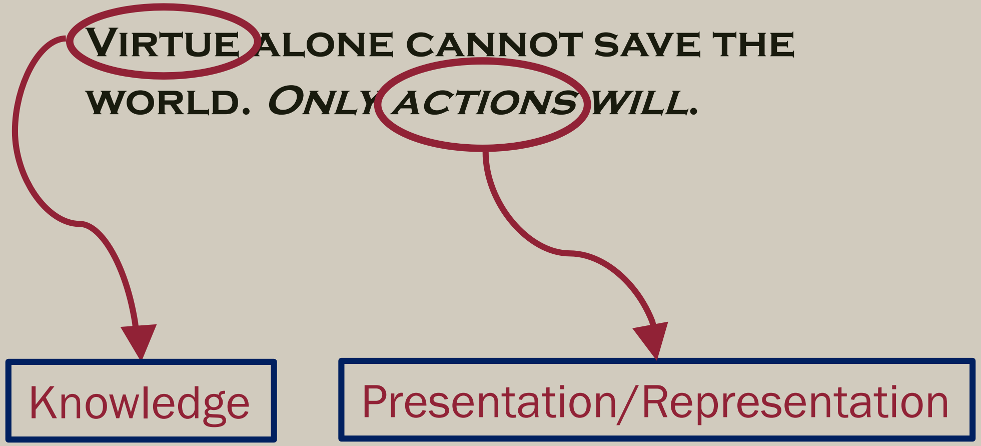
Linger Rate = Semantic Consistency + Reduced Dimensions + Squared terms

EXPECTED FINDINGS

- Model should be able to predict readership relatively well, maybe not as well for linger rate
  - *While observing that higher aesthetic scores are usually correlated with higher readership and longer visit duration*
  - ***Expected Challenges:*** *Unknown requirements for n-power (larger dataset may be needed), extreme non-linearity, inaccuracies in traffic estimation*

- Survey/Experimental: Aesthetic scores based on HCI and UI principles should predict readers' ratings of content/websites and initial interest level well

Virtue alone cannot save the world. Only actions will.



VIRTUE ALONE CANNOT SAVE THE WORLD. *ONLY ACTIONS WILL.*

Knowledge

Presentation/Representation

This presentation was presented at:
Saieh 247, The University of Chicago
Chicago, Illinois
United States of America
Earth
The Solar System
The Milky Way
The Local Group
Virgo Supercluster
The Known Universe