

Some Emerging Data Sources and Their Statistical Implications

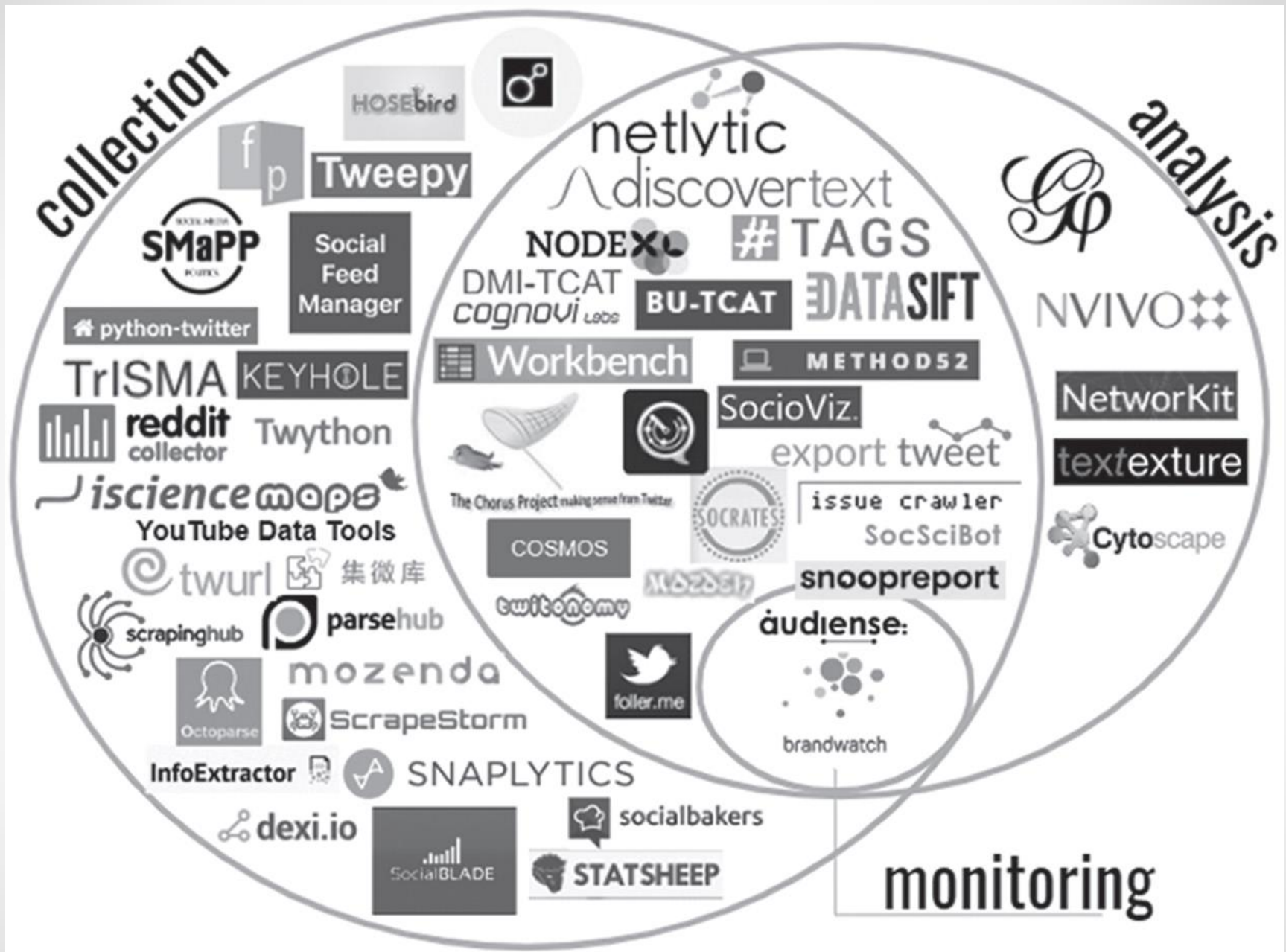
Hugh McCague
Institute for Social Research
York University

Social Media

- increasing interest in using information/data gathered from social media (e.g.. Twitter, Facebook) for research, monitoring and decision making
- can be less expensive and have a larger sample size than a traditional probability sample

Social Media

- can be combined and compared with survey results, and administrative and geographic data
- new methods with various methodological issues and concerns to be justified and resolved (e.g., biases and representativeness)

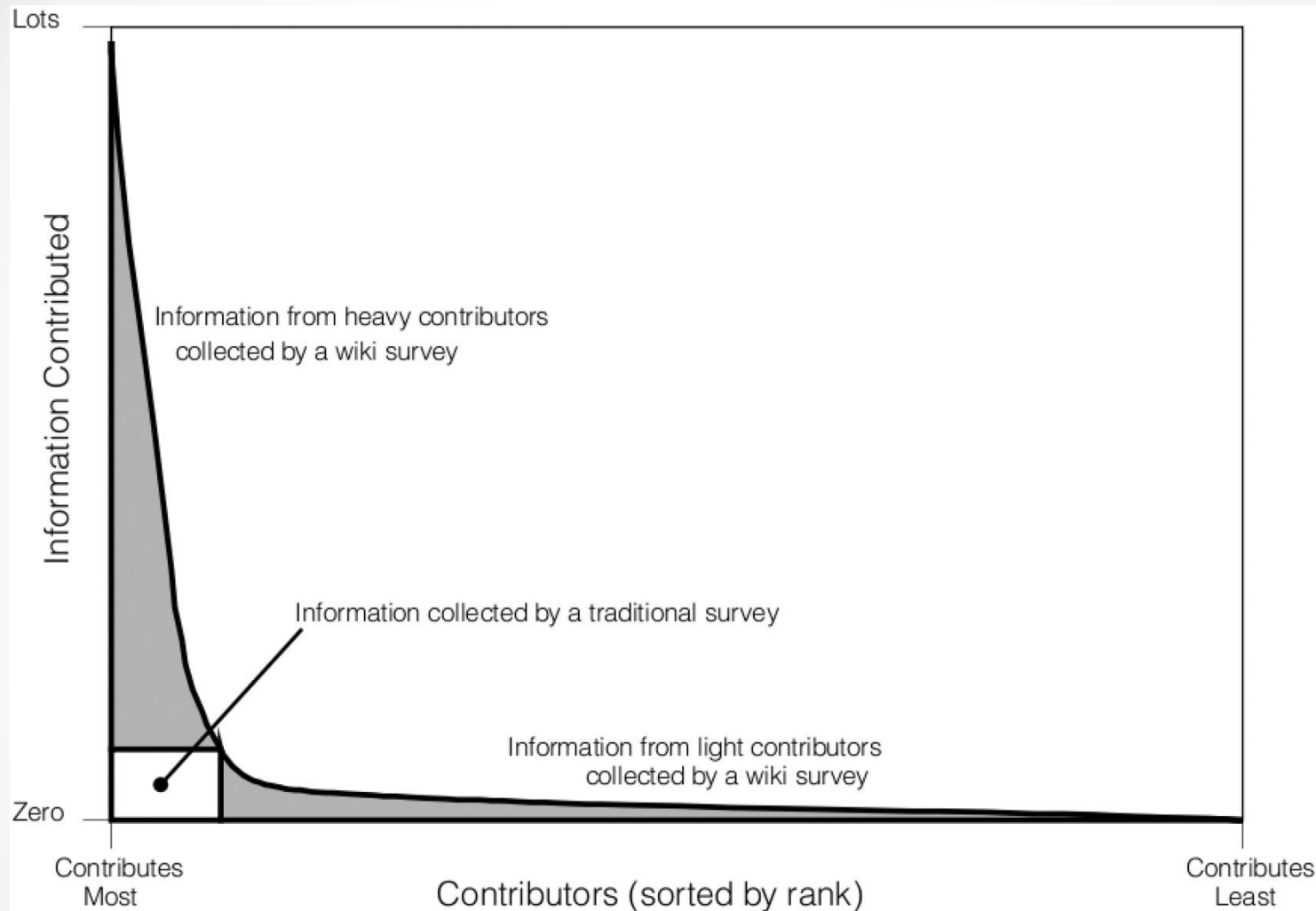


Source: Fig.10, p.12, <http://uk.sagepub.com/sites/default/files/sage-whitepaper-the-ecosystem-of-technologies-for-social-science-research.pdf>

Wiki Surveys

- Open source
- Available at www.allourideas.org
- “Adaptive”
- “Collaborative”
- “Greedy”

Schematic of rank order plot of contributions to successful online information aggregation projects.




Salganik MJ, Levy KEC (2015) Wiki Surveys: Open and Quantifiable Social Data Collection. PLOS ONE 10(5): e0123483.
<https://doi.org/10.1371/journal.pone.0123483>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123483>

Response and results interfaces at www.allourideas.org.

A



The screenshot shows the Plan NYC website interface for submitting ideas. At the top left is the Plan NYC logo. To its right are links for "Cast Votes", "View Results", and "About this page". The main heading asks, "Which do you think is a better idea for creating a greener, greater New York City?". Below this, there are two blue buttons representing different ideas: "Revise green roof tax abatement to promote gardening in green roofs" and "Increase safety on all bike lanes: protected lanes with barriers between bikes and traffic, bike bridges over roadways, bike parking." Below these buttons is a grey button labeled "I can't decide" and a green button labeled "Add your own idea". At the bottom right, it says "23038 votes on 269 ideas".

B



The screenshot shows the Plan NYC website interface displaying the results of a survey. At the top left is the Plan NYC logo. To its right are links for "Cast Votes", "View Results", and "About this page". The main heading asks, "Which do you think is a better idea for creating a greener, greater New York City?". Below this is a table with two columns: "Idea" and "Score". Each row in the table includes a description of an idea, a score with a question mark, and a horizontal bar chart showing the relative number of votes.

Idea	Score
Keep NYC's drinking water clean by banning fracking in NYC's watershed.	84 [?]
Invest in multiple modes of transportation and provide both improved infrastructure and improved safety	81 [?]
Plug ships into electricity grid so they don't idle in port - reducing emissions equivalent to 12000 cars per ship.	78 [?]
Implement congestion pricing in lower Manhattan	74 [?]
Continue enhancing bike lane network, to finally connect separated bike lane systems to each other across all five boroughs.	73 [?]
Composting! Provide municipal support for composting!!	73 [?]
Support and protect community gardens and create mechanisms to create new gardens and open space	72 [?]
Provide long-term leases for organic farms in unused public spaces, a garden at every public school and public housing development	72 [?]
Provide better transit service outside of Manhattan	72 [?]
Create a network of protected bike paths throughout the entire city	71 [?]

Salganik MJ, Levy KEC (2015) Wiki Surveys: Open and Quantifiable Social Data Collection. PLOS ONE 10(5): e0123483.

<https://doi.org/10.1371/journal.pone.0123483>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123483>

Summary of data analysis plan.

Responses

Respondent	Response	Pair	
1	1	[item 1]	item 9
1	2	item 3	[item 2]
1	3	[item 4]	item 3
2	4	[item 8]	item 5
2	5	item 4	[item 2]
⋮	⋮	⋮	⋮

Opinion matrix

$$\begin{bmatrix} \hat{\theta}_{1,1} & \hat{\theta}_{1,2} & \dots & \hat{\theta}_{1,K} \\ \hat{\theta}_{2,1} & \hat{\theta}_{2,2} & \dots & \hat{\theta}_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\theta}_{J,1} & \hat{\theta}_{J,2} & \dots & \hat{\theta}_{J,K} \end{bmatrix}$$

Scores

$$[\hat{s}_1 \quad \hat{s}_2 \quad \dots \quad \hat{s}_K]$$

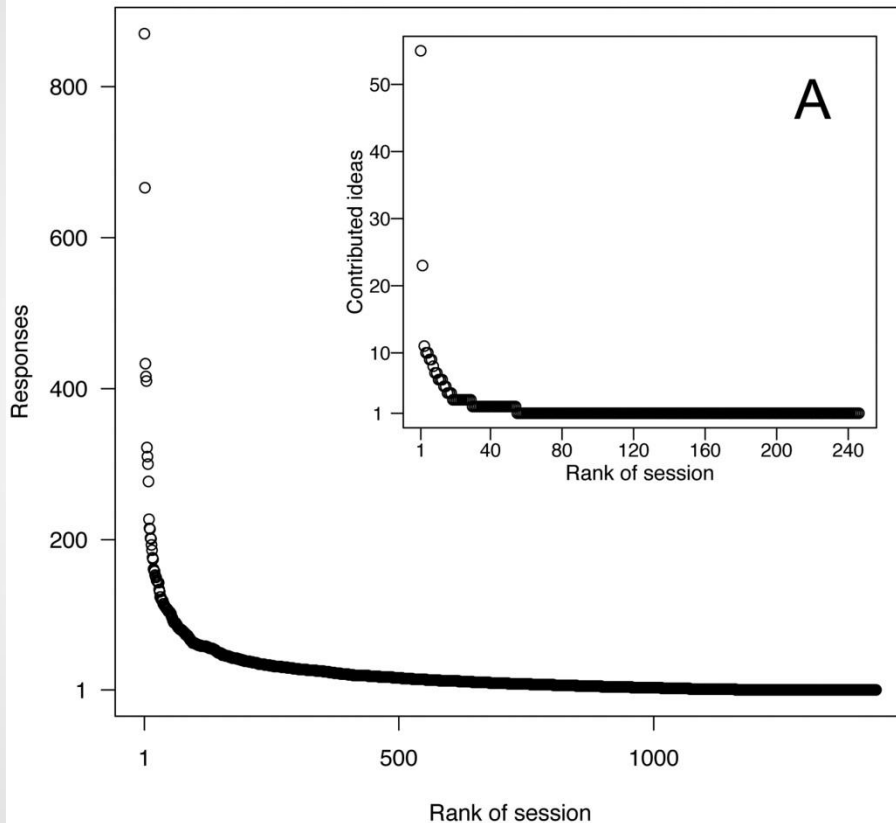
Salganik MJ, Levy KEC (2015) Wiki Surveys: Open and Quantifiable Social Data Collection. PLOS ONE 10(5): e0123483.

<https://doi.org/10.1371/journal.pone.0123483>

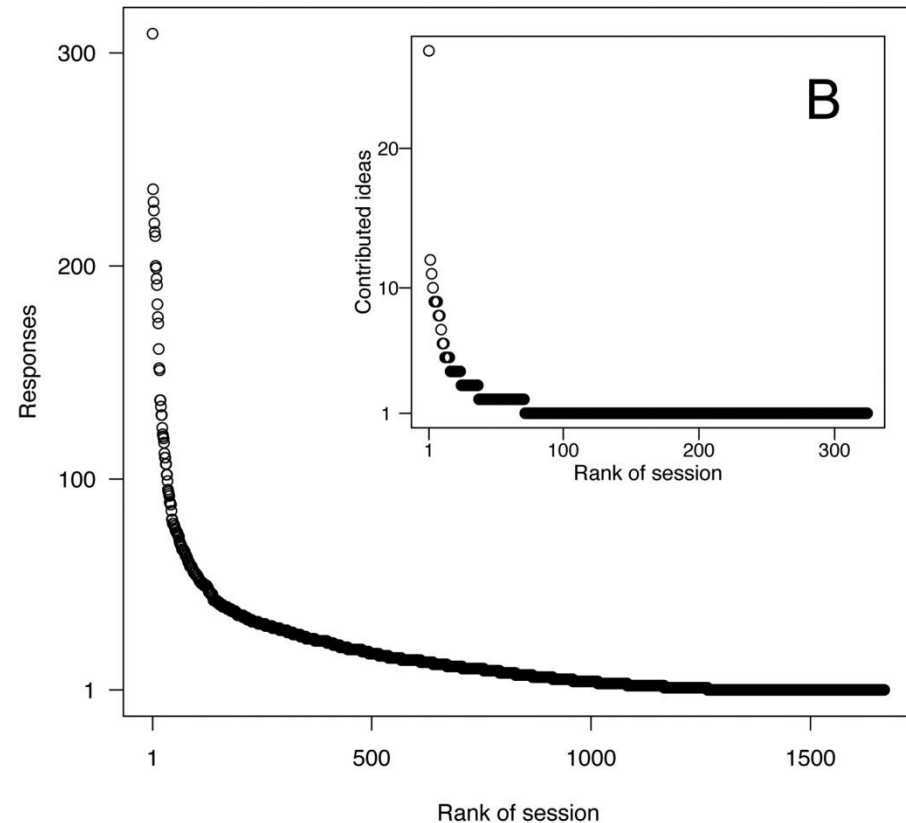
<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123483>

Distribution of contribution per respondent for PlaNYC [A] and OECD [B].

Which do you think is a better idea for creating a greener, greater New York City?



Which is the more important action we need to take in education today?



Salganik MJ, Levy KEC (2015) Wiki Surveys: Open and Quantifiable Social Data Collection. PLOS ONE 10(5): e0123483.

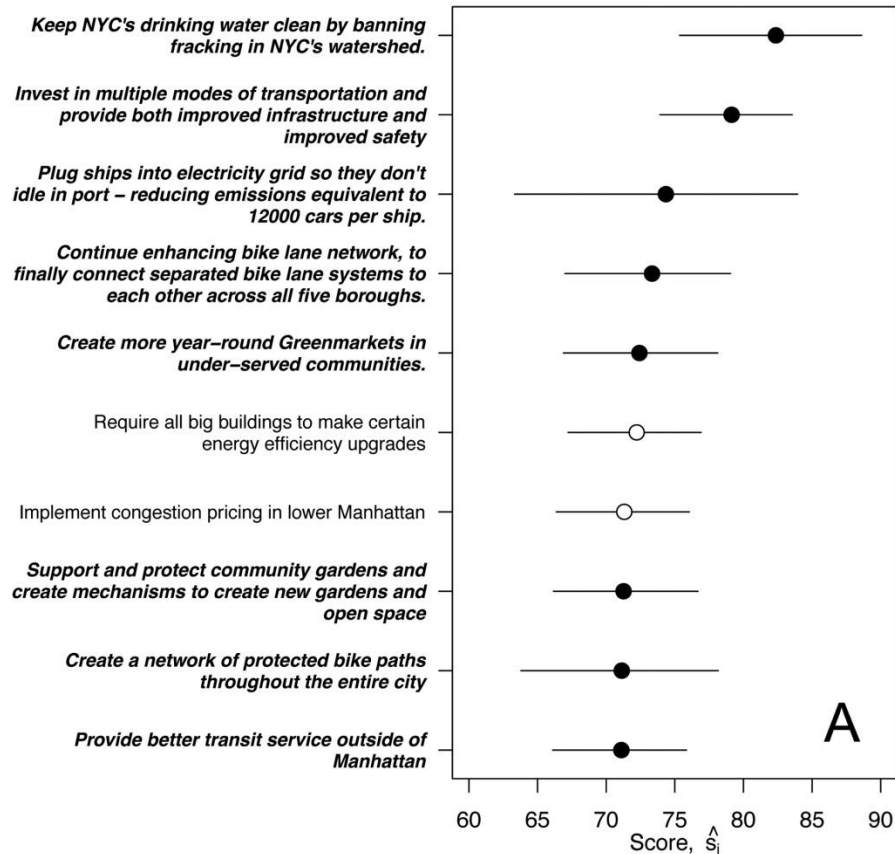
<https://doi.org/10.1371/journal.pone.0123483>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123483>

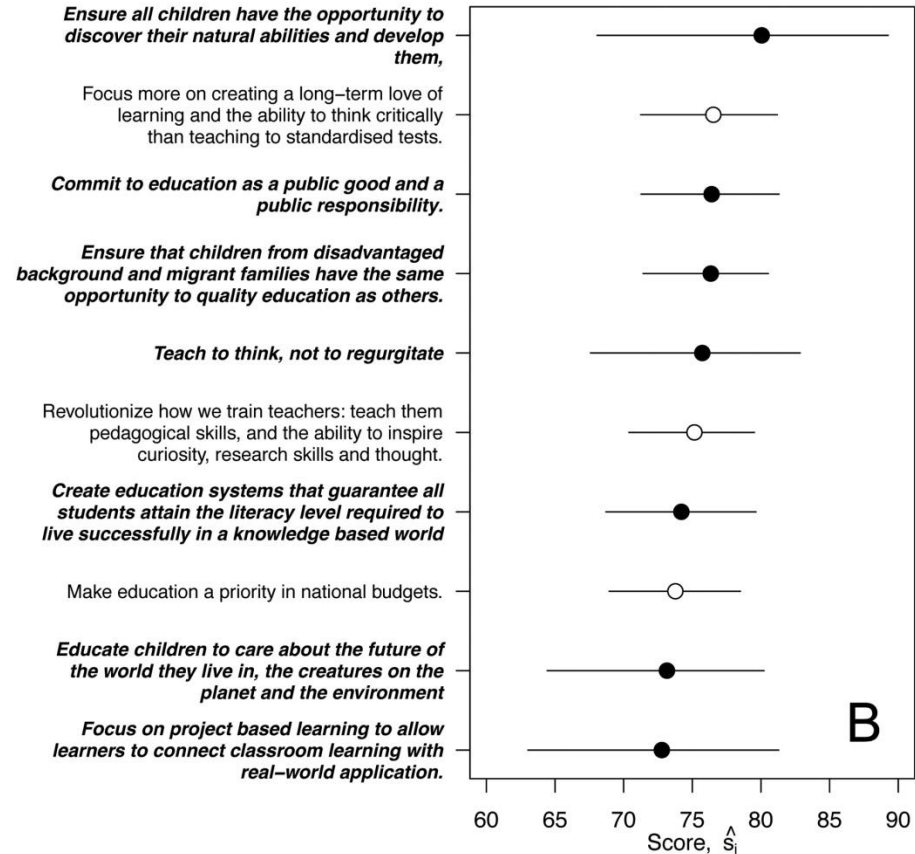
Ten highest-scoring ideas for PlanNYC [A] and OECD [B].

Ideas in **bold** were contributed by the survey participants.

Which do you think is a better idea for creating a greener, greater New York City?



Which is the more important action we need to take in education today?

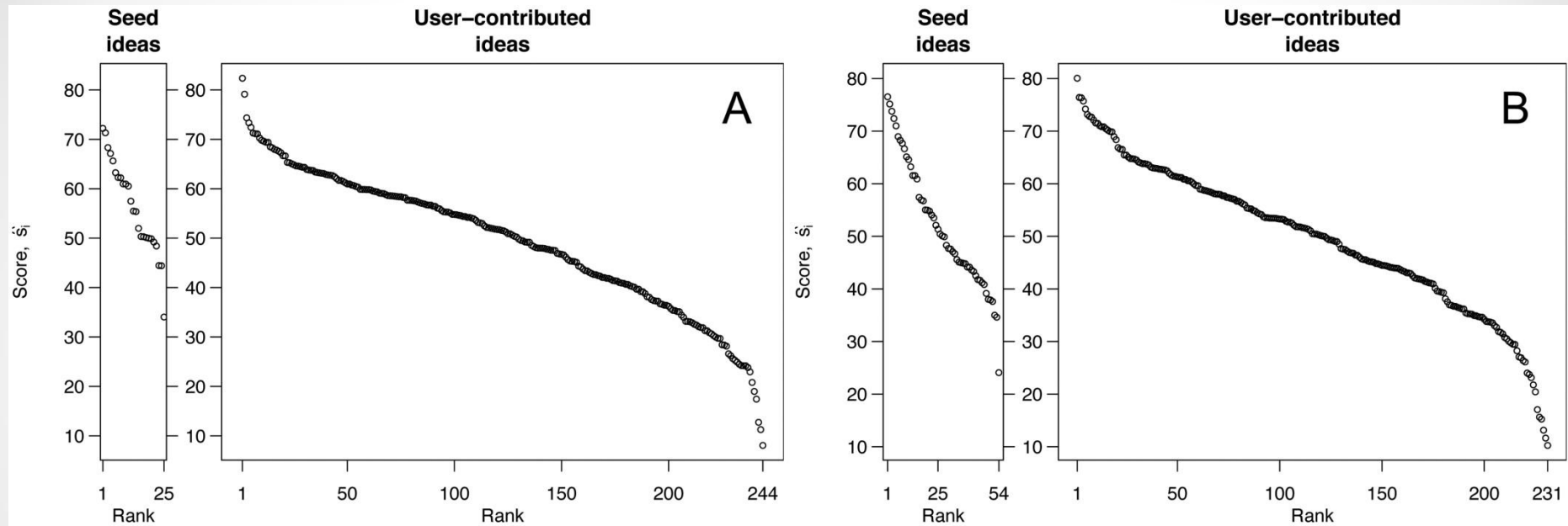


Salganik MJ, Levy KEC (2015) Wiki Surveys: Open and Quantifiable Social Data Collection. PLOS ONE 10(5): e0123483.

<https://doi.org/10.1371/journal.pone.0123483>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123483>

Fig 7. Distribution of scores of seed ideas and user-contributed ideas for PlaNYC [A] and OECD [B].



Salganik MJ, Levy KEC (2015) Wiki Surveys: Open and Quantifiable Social Data Collection. PLOS ONE 10(5): e0123483.

<https://doi.org/10.1371/journal.pone.0123483>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123483>

Google Trends

- launched in 2006
- relative frequencies (0 to 100) of a specified Google search
- can confine to a specific country and top provinces/states and cities
- can specify time frame
- no fee

Google Correlate

- launched in 2011
- the reverse of Google Trends
- provides a list of Google searches that have higher correlated time series with a specified time series or Google search
- had no fee
- discontinued Dec.15, 2019
- Google Trends does provide related queries

Google AdWords

- launched in 2000
- has some features similar to Google Trends, but provides exact frequencies (counts) and more flexible specifications
- fee

Google Trends and AdWords

- provide related queries similar to Google Correlate
- may reveal more frankly what people are thinking than a traditional survey (e.g., less stigma bias)
- a very flexible and modifiable method
- this data can be combined with other social and geographic data to gain insights

Google Trends and Ads

- new method that has objections and issues (e.g., biases and representativeness) to address in order to become more widely viable
- promising scholarly work on assessing, comparing and testing this method is underway
- Seth Stephen-Davidowitz did his PhD on this topic and has written a popular book *Everybody Lies: Big Data, New Data ...* (sethsd.com)

Google Trends Example

<https://trends.google.com/trends/>

Compare

● panic attack
Search term

● anxiety attack
Search term

+ Add comparison

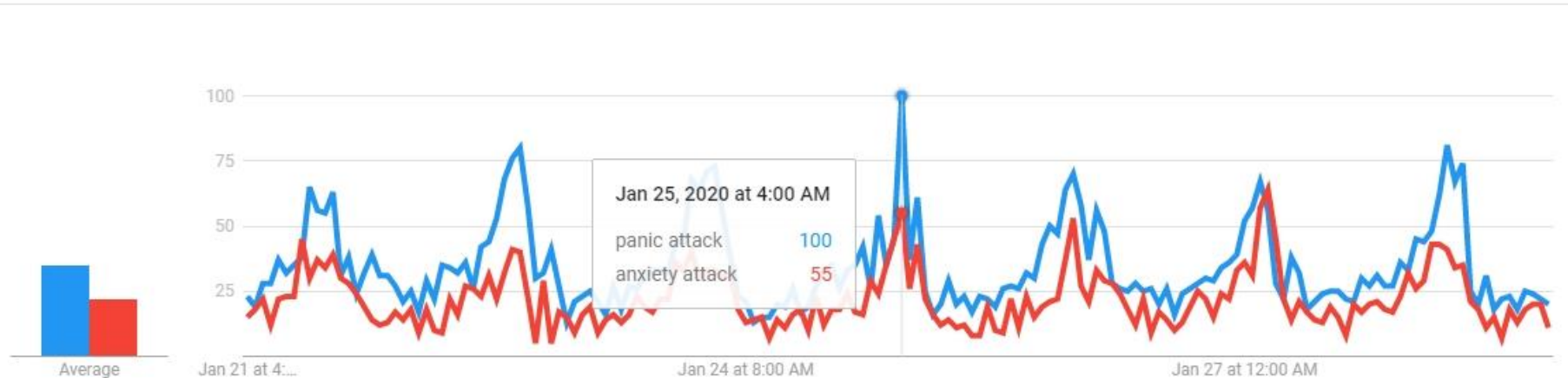
Canada ▾

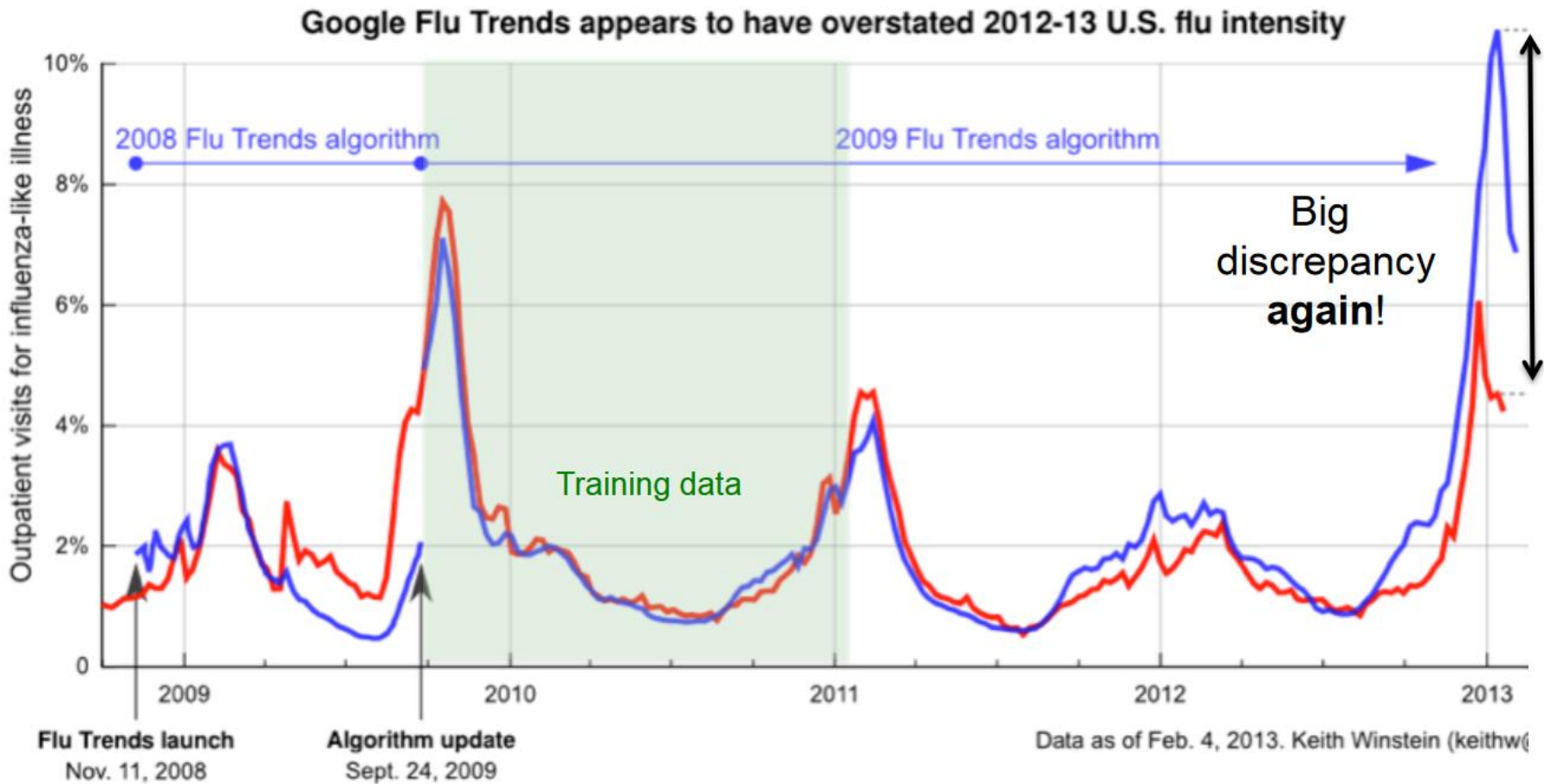
Past 7 days ▾

All categories ▾

Web Search ▾

Interest over time ⓘ





Fixes reported in: Cook et al. (2011). Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: <http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html>

Source: <http://stat.psu.edu/rao-prize-presentations-2017/big-data-google-and-disease-detection-the-statistical-story>

ARGO (AutoRegression with GOogle search data)

— our model

$y_t = \text{logit}(\text{CDC reported flu activity at time } t)$

— transform from $[0, 1]$ to real line

$X_{i,t} = \log(\text{Google search frequency of term } i \text{ at time } t + 0.5)$

— transform from $[0, 100]$ to real line

- y_t evolves according to AR(N)

$$y_t = \mu_y + \sum_{j=1}^N \alpha_j y_{t-j} + \epsilon_t, \quad \epsilon_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- $\mathbf{X}_t = (X_{1,t}, X_{2,t}, \dots)$ depends only on the flu activity at t
— intuition: online searches are in response of real flu occurrences

$$\mathbf{X}_t \mid y_t \sim \mathcal{N}_K(\mu_x + y_t \beta, \mathbf{Q})$$

- (vector) hidden Markov structure

$$\begin{array}{ccccccc} y_{1:N} & \rightarrow & y_{2:(N+1)} & \rightarrow & \dots & \rightarrow & y_{(T-N+1):T} \\ \downarrow & & \downarrow & & & & \downarrow \\ \mathbf{X}_N & & \mathbf{X}_{N+1} & & & & \mathbf{X}_T \end{array}$$

ARGO (AutoRegression with GOogle search data)

— our model

$y_t = \text{logit}(\text{CDC reported flu activity at time } t)$
— transform from $[0,1]$ to real line

$X_{i,t} = \log(\text{Google search frequency of term } i \text{ at time } t + 0.5)$
— transform from $[0,100]$ to real line

Predictive distribution $P(y_t | y_{1:(t-1)}, X_{1:t})$ normal, mean **linear**:

$$y_t = \mu_y + \sum_{j=1}^N \alpha_j y_{t-j} + \sum_{i=1}^K \beta_i X_{i,t} + \epsilon_t$$

Model **training**:

- $N = 52$ (weeks), $K = 100$ terms (from Google Correlate/Trends)
- Use L_1 and L_2 penalization to estimate $\alpha = (\alpha_1, \dots, \alpha_{52}), \beta = (\beta_1, \dots, \beta_{100})$

$$\text{minimize} \quad \sum_t \left(y_t - \mu_y - \sum_{j=1}^{52} \alpha_j y_{t-j} - \sum_{i=1}^{100} \beta_i X_{i,t} \right)^2 \\ + \lambda_\alpha \|\alpha\|_1 + \eta_\alpha \|\alpha\|_2^2 + \lambda_\beta \|\beta\|_1 + \eta_\beta \|\beta\|_2^2$$

- Use a two-year (104 weeks) rolling window for dynamic training

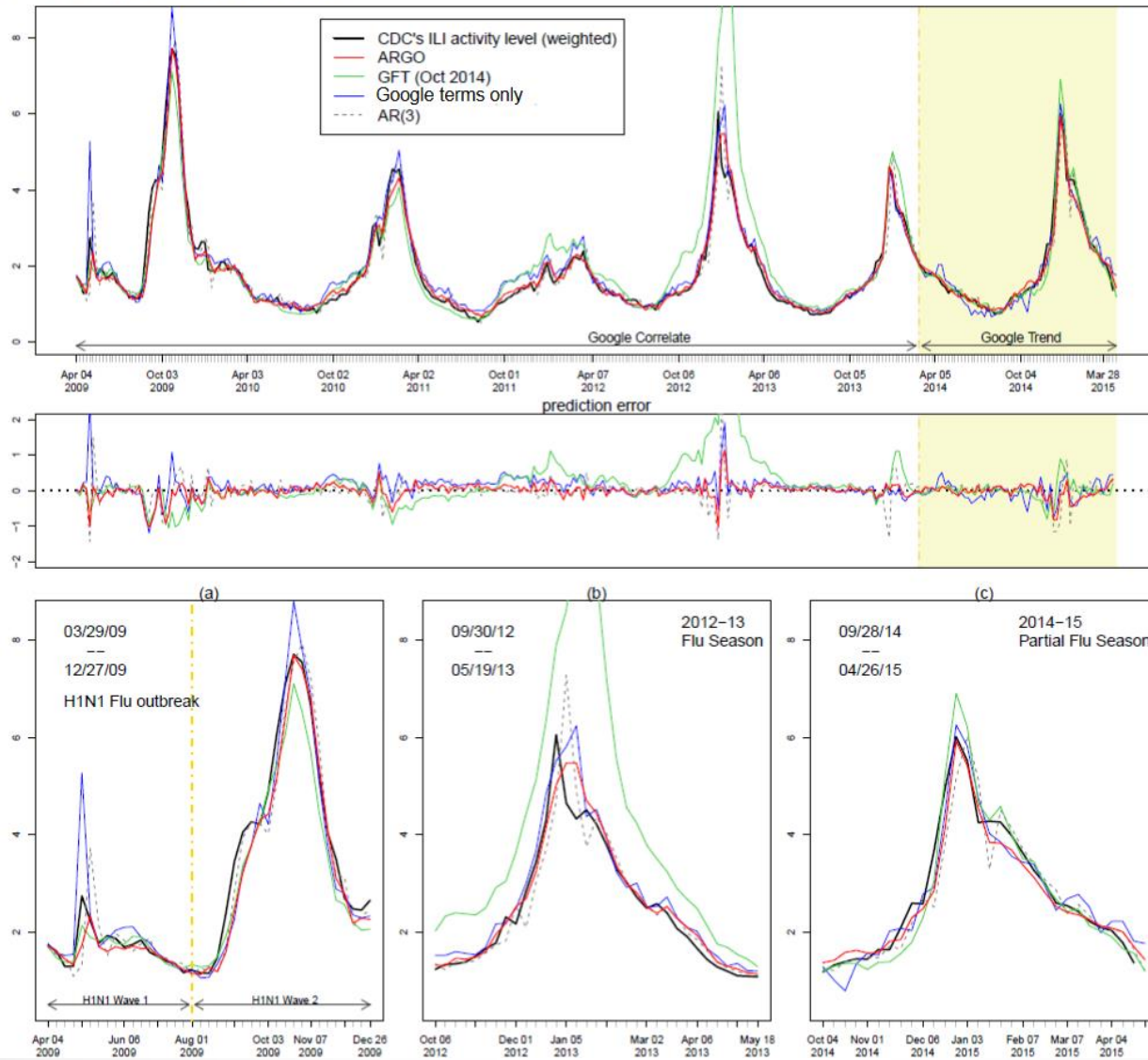
ARGO (AutoRegression with GOogle search data)

— our model

Features of ARGO:

- dynamically incorporate new CDC numbers when they become available
- incorporate seasonality of flu activity
- automatically select the most useful Google search queries for estimation
- two-year rolling window gives **dynamic** fitting; capture the changes in search behavior and pattern

ARGO in action



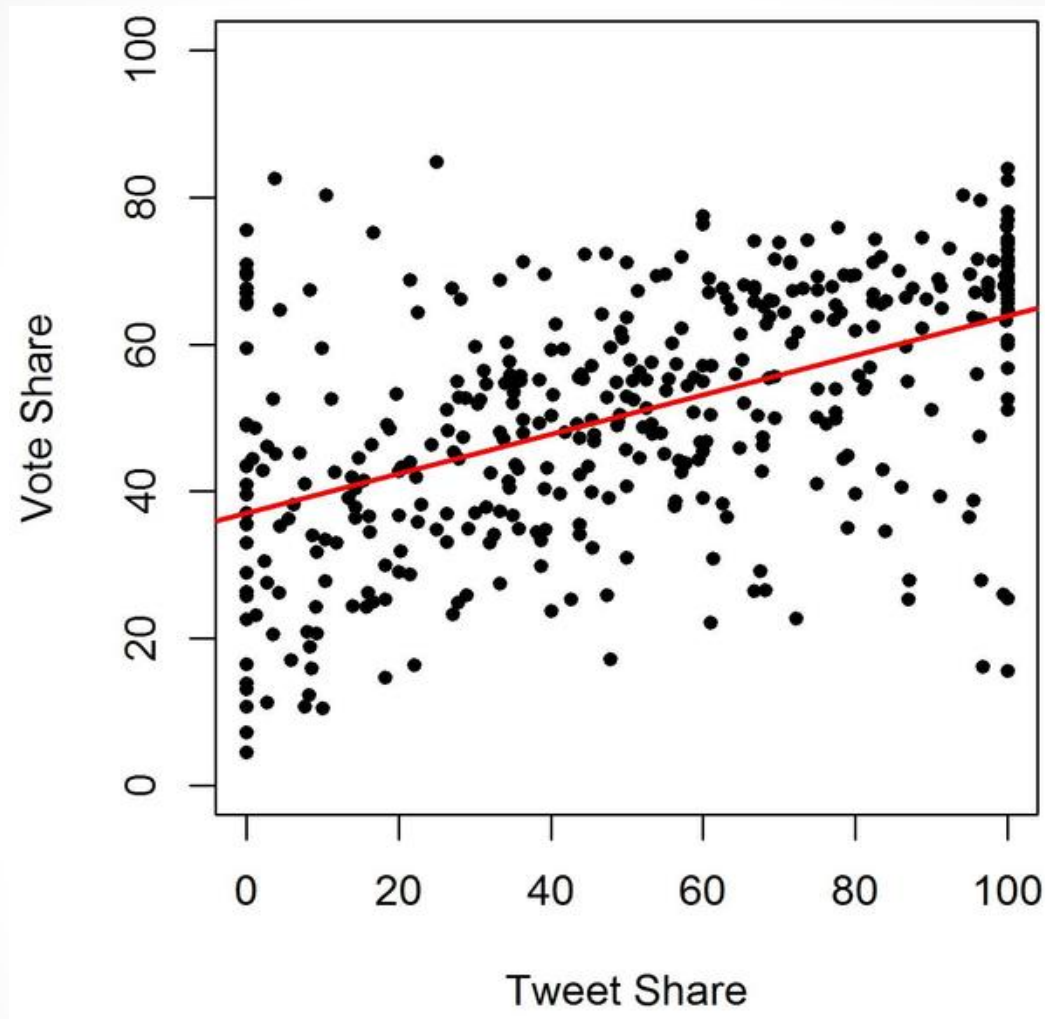
Tweet Share

$$tw_S(i) = \frac{tw_R(i)}{tw_D(i) + tw_R(i)} \times 100$$

Vote Share

$$v_S(i) = \frac{v_R(i)}{v_D(i) + v_R(i)} \times 100$$

2010 Republican Tweet Share vs. Vote Share.



DiGrazia J, McKelvey K, Bollen J, Rojas F (2013) More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. PLOS ONE 8(11): e79449. <https://doi.org/10.1371/journal.pone.0079449>
<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079449>

Results for Regression of Republican Vote Share on Tweet Share with Controls.

Variable	Bivariate (SE)	Full Model (SE)
Republican Tweet Share	0.268 (0.022)***	0.022 (0.01)*
Republican Incumbent		11.06 (0.66)***
% McCain		0.776 (0.03)***
Median Age		0.012 (0.09)
% White		0.129 (0.02)***
% College Educated		-0.004 (0.05)
Median HH Income		0.016 (0.03)
% Female		0.089 (0.30)
CNN share		0.002 (0.01)
<i>Const</i>	37.042 (1.35)	-4.07 (15.04)
<i>N</i>	406	406
R_{adj}^2	.26	.92

Explaining Republican vote share with the proportion of tweets that included a Republican candidate during the three months before the 2010 election. The share of Republican tweets remains significant after adding controls. Standard error (SE) is in parentheses.

*($p < .05$).

** ($p < .01$).

***($p < .001$).

doi:10.1371/journal.pone.0079449.t001

DiGrazia J, McKelvey K, Bollen J, Rojas F (2013) More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. PLOS ONE 8(11): e79449. <https://doi.org/10.1371/journal.pone.0079449>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079449>

Conclusion

- early stages of develop where successes, failures and much learning are taking place
- expanding interest in scholarly work, teaching, business and government
- promising scholarly work assessing, comparing and testing

Conclusion

- combining different data sources becoming more common which can greatly augment any one dataset
- helpful to have some understanding of machine learning and statistical modelling to determine which or both are advisable