Some Emerging Data Sources and Their Statistical Implications

Hugh McCague
Institute for Social Research
York University

Social Media

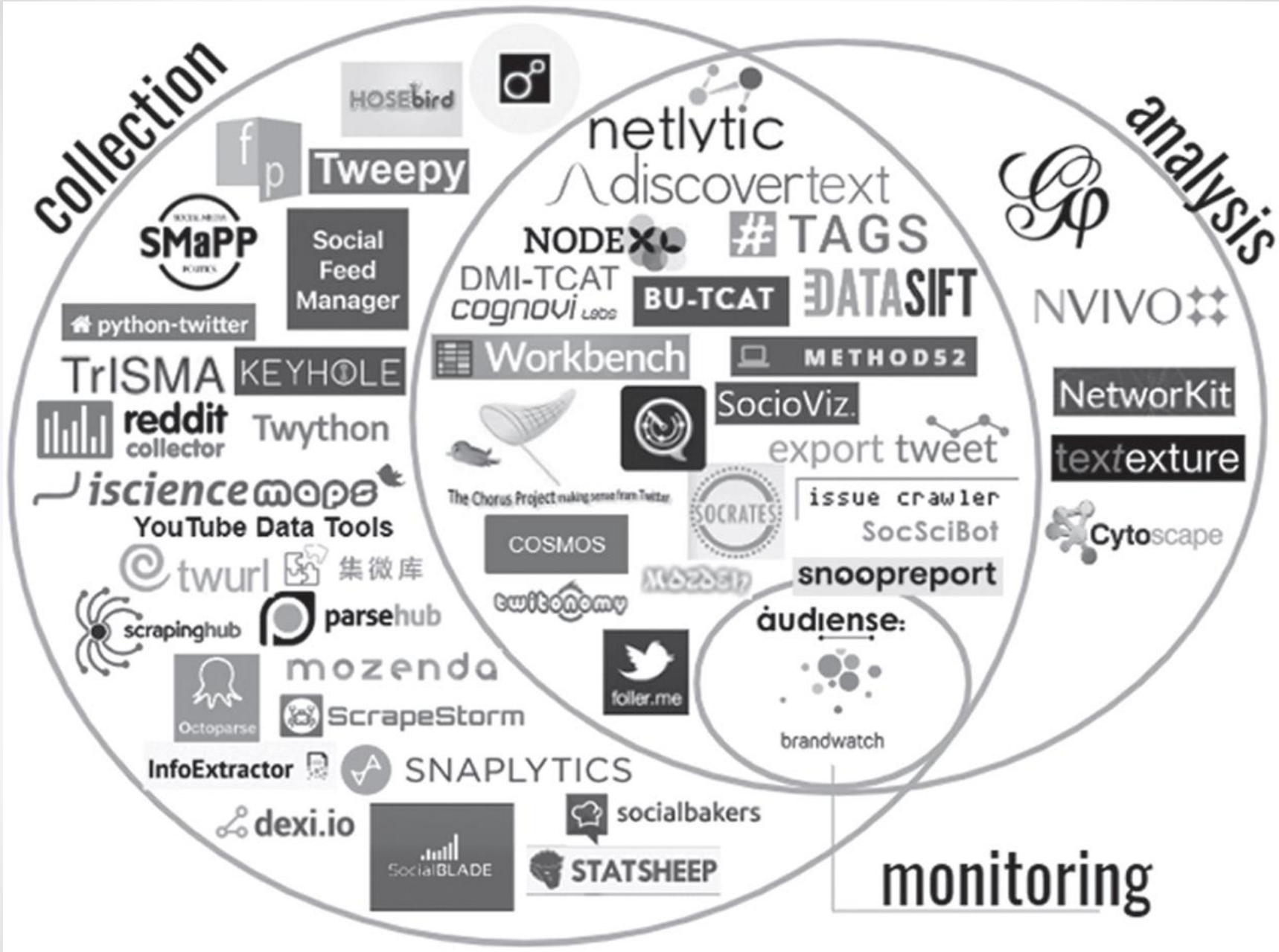
- increasing interest in using information/data gathered from social media (e.g.. Twitter, Facebook) for research, monitoring and decision making

- 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)

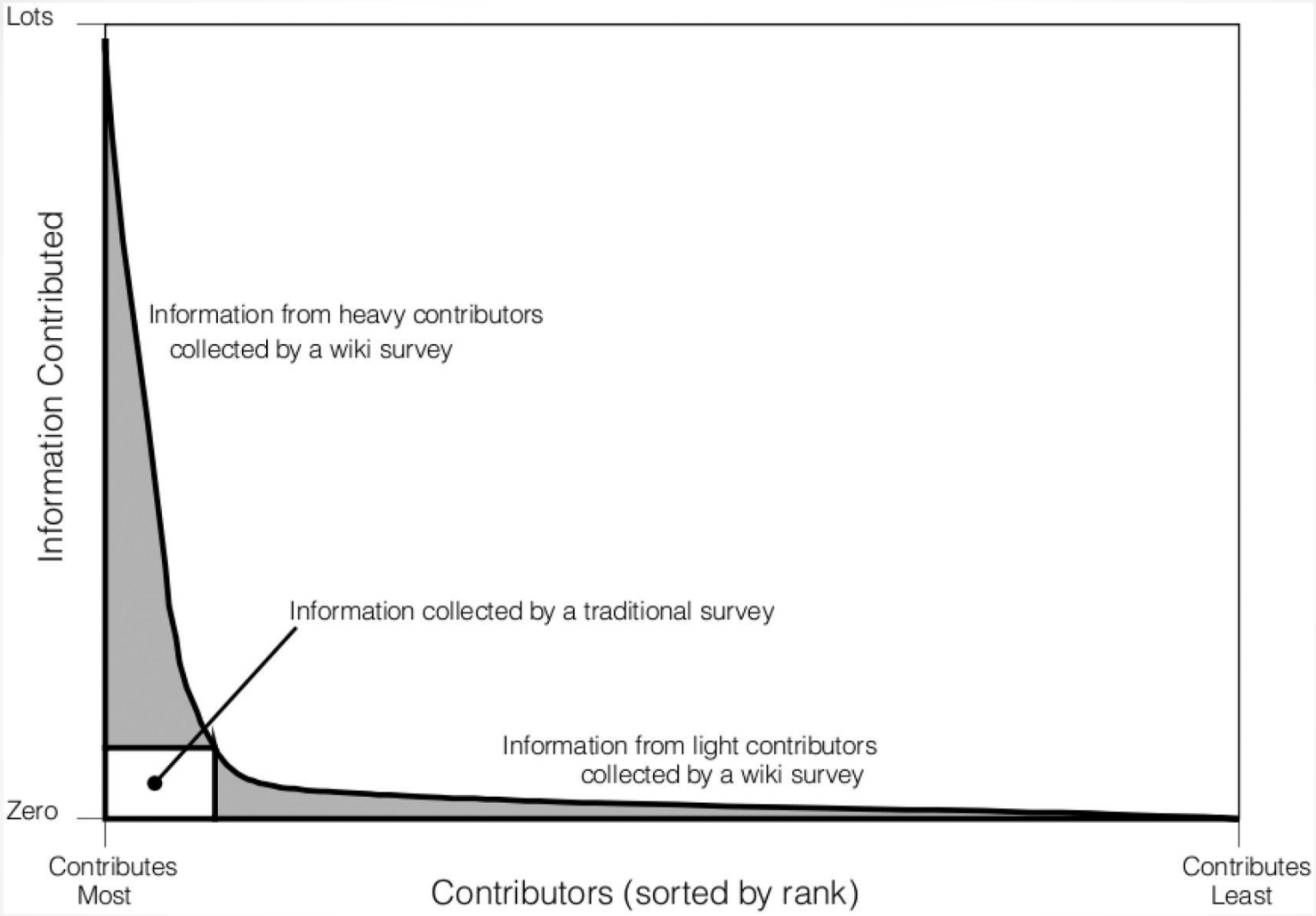


Source: Fig.10, p.12, http://uk.sagepub.com/sites/default/files/sage-whitepaper-the-ecosystem-of-technologies-for-social-science-research.pdf

Wiki Surveys

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

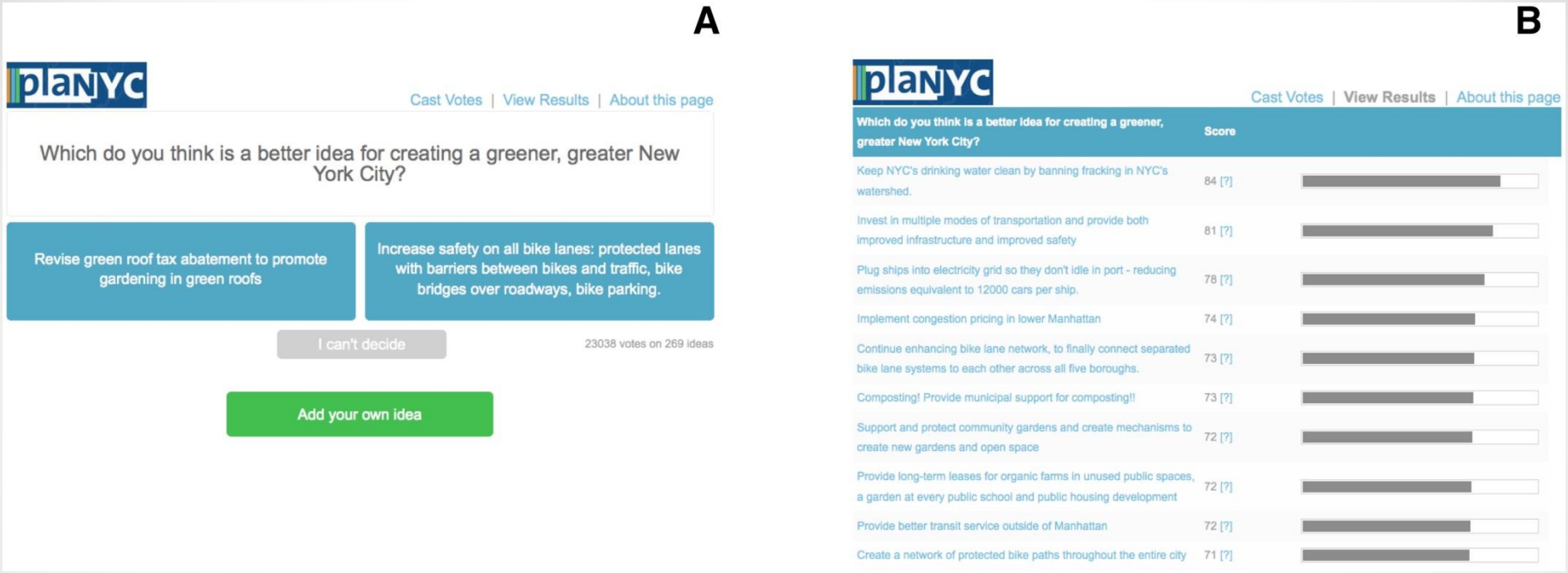
Schematic of rank order plot of contributions to successful online information aggregation projects.



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

PLOS ONE

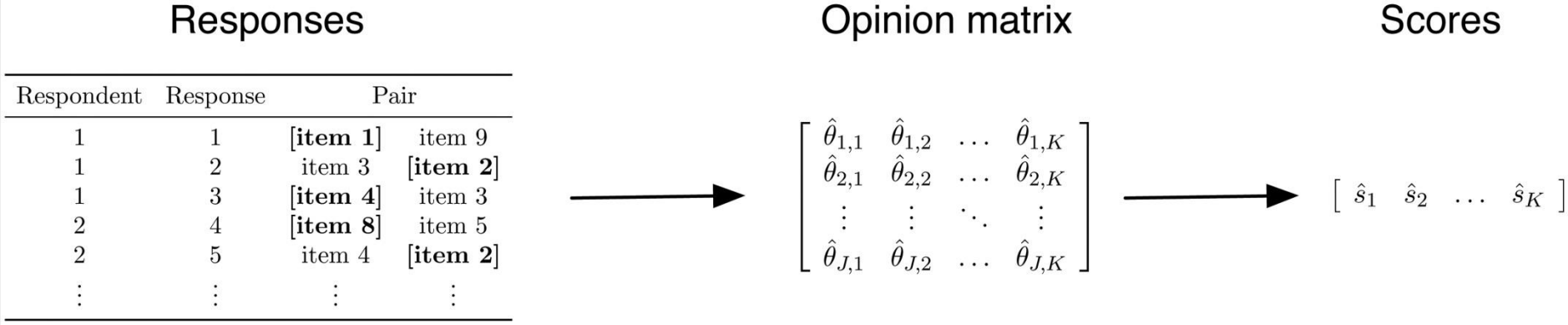
Response and results interfaces at www.allourideas.org.



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

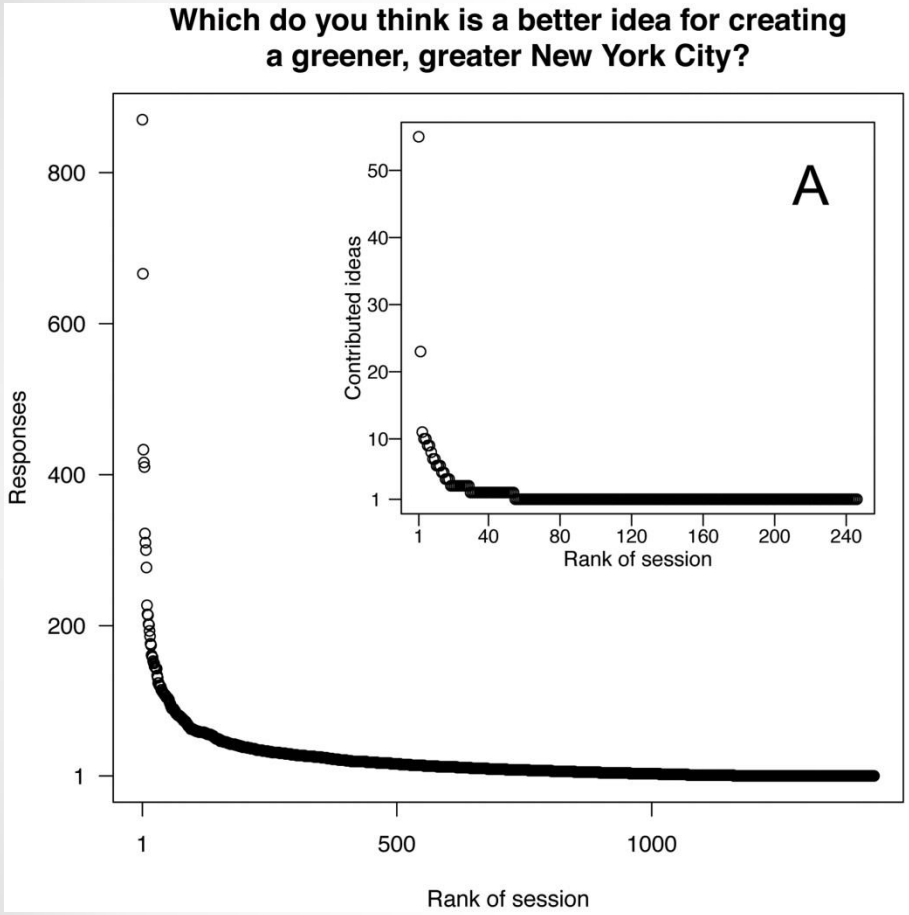
PLOS ONE

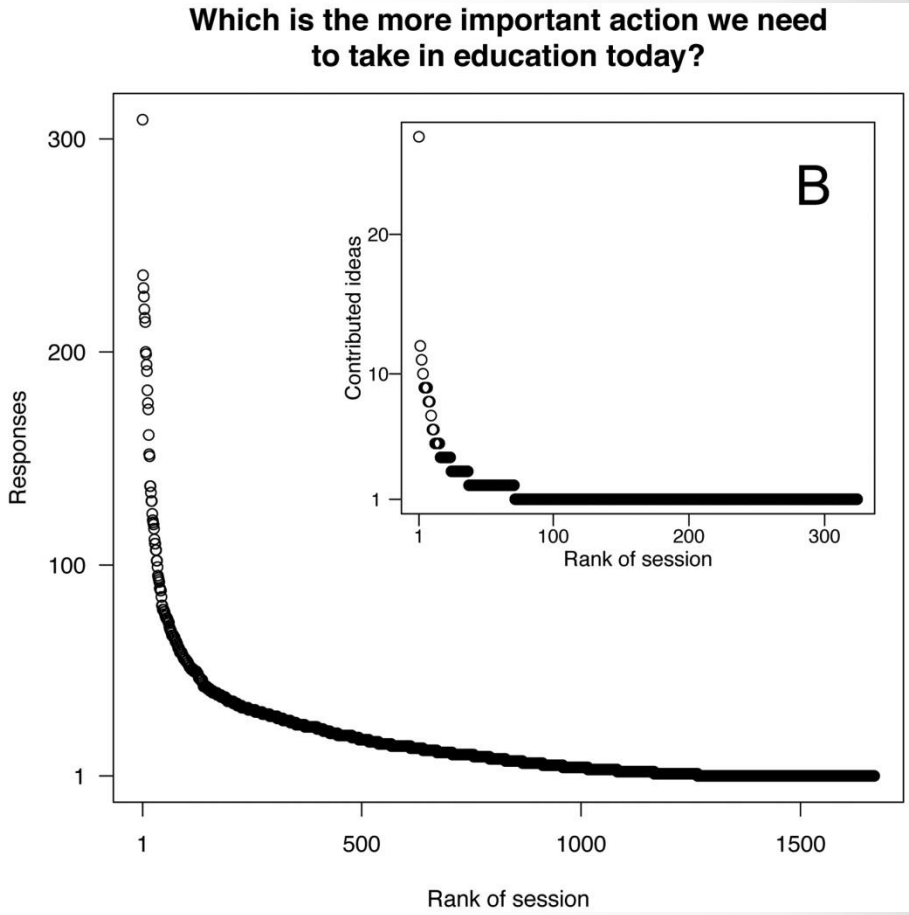
Summary of data analysis plan.



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].



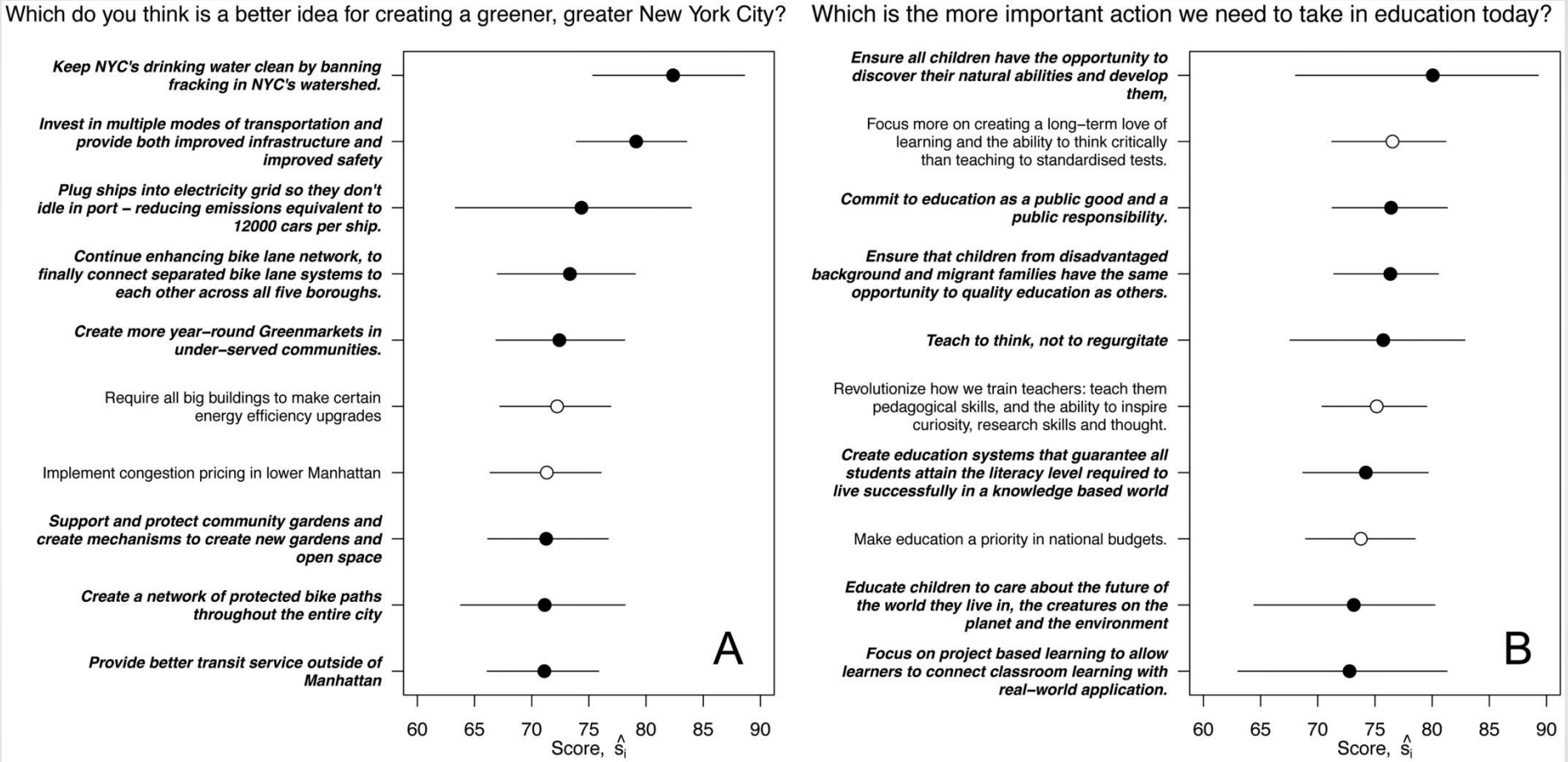


Salganik MJ, Levy KEC (2015) Wiki Surveys: Open and Quantifiable Social Data Collection. PLOS ONE 10(5): e0123483. https://doi.org/10.1371/journal.pone.0123483
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123483

PLOS ONE

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

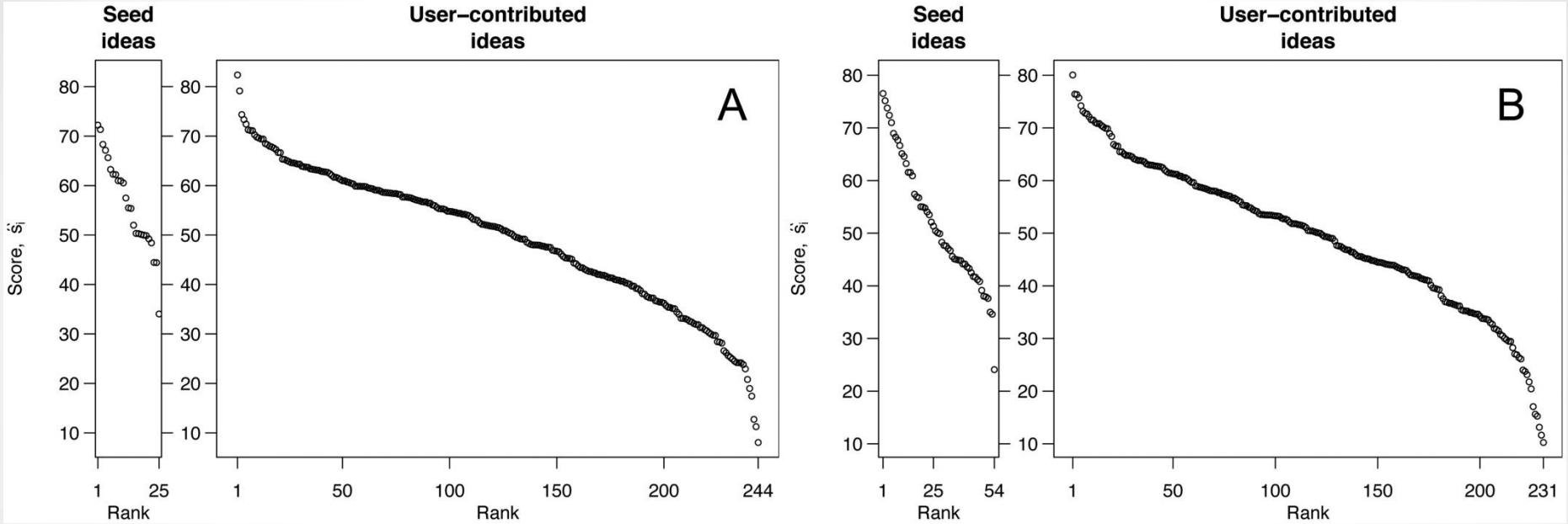
Ideas in **bold** were contributed by the survey participants.



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

PLOS ONE

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

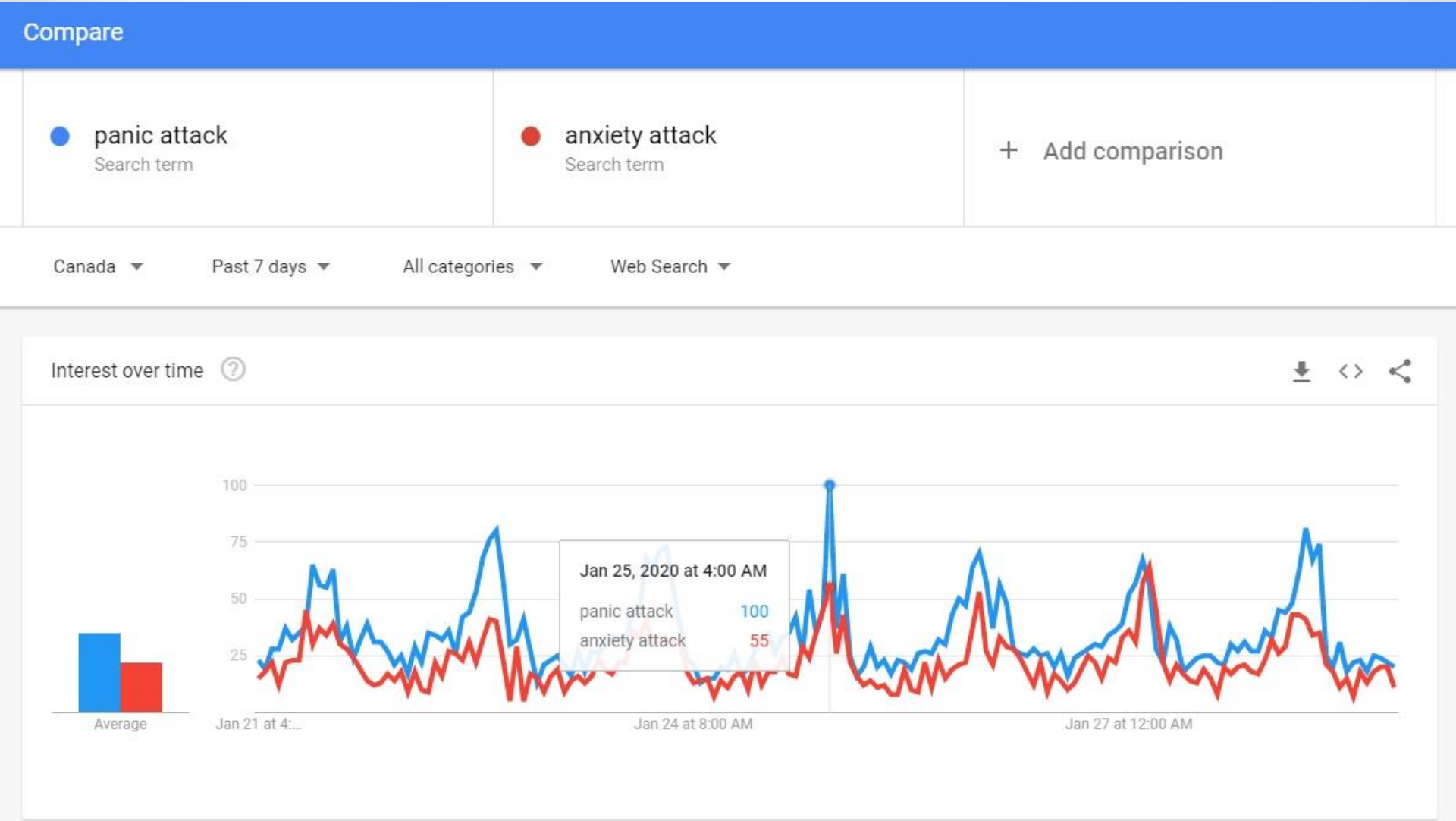
- new method that has objections and issues (e.g., biases and representativeness) to address in order to become more widely viable

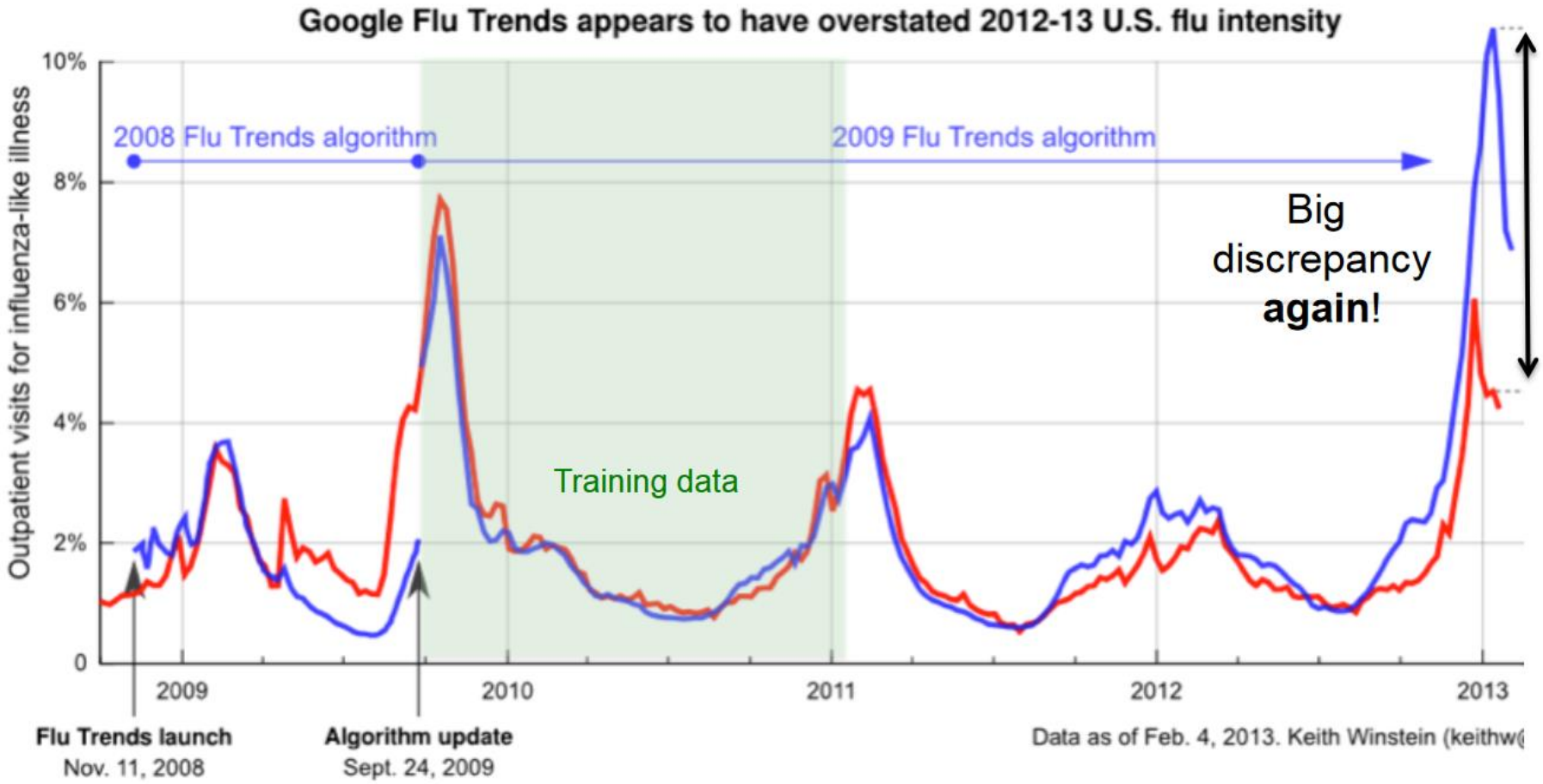
- promising scholarly work on assessing, comparing and testing this method is underway

- Seth Stephen-Davidowitz did his PhD on this topic and has written a popular book *Everbody Lies: Big Data, New Data …* (sethsd.com)

Google Trends Example

https://trends.google.com/trends/





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 = logit(CDC reported flu activity at time t)
— transform from [0,1] to real line

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

- y_t evolves according to AR(N)

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

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

$$\boldsymbol{X}_t \mid y_t \sim \mathcal{N}_K(\boldsymbol{\mu}_x + y_t \boldsymbol{\beta}, \boldsymbol{Q})$$

- (vector) hidden Markov structure

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

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 = logit(CDC reported flu activity at time t)
— transform from [0,1] to real line

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

Predictive distribution $\mathrm{P}(y_t \mid y_{1:(t-1)}, \boldsymbol{X}_{1:t})$ 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 $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{52}), \boldsymbol{\beta} = (\beta_1, \ldots, \beta_{100})$

minimize

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

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

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

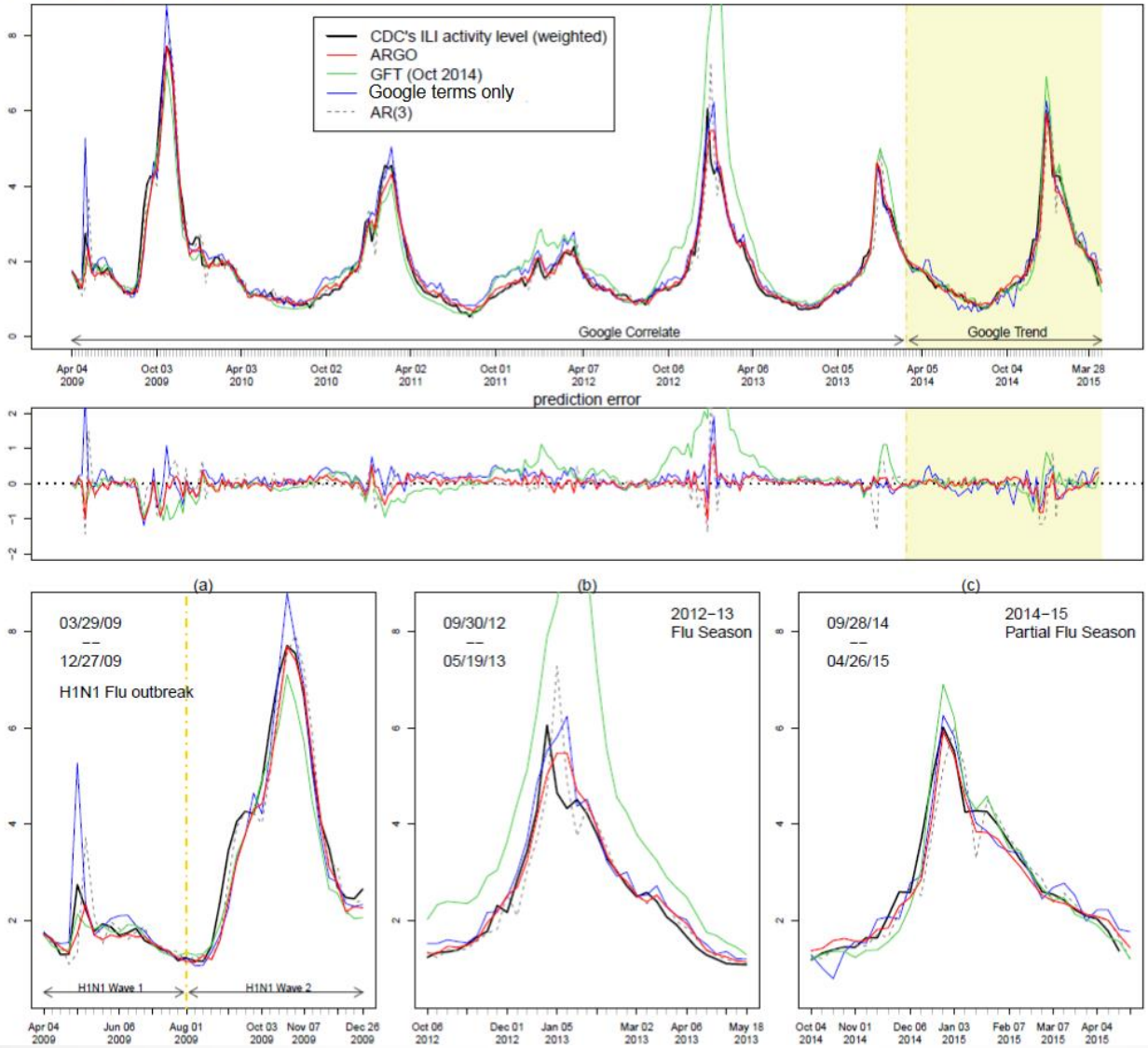
ARGO (AutoRegression with GOogle search data) — our model

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

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

ARGO in action



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

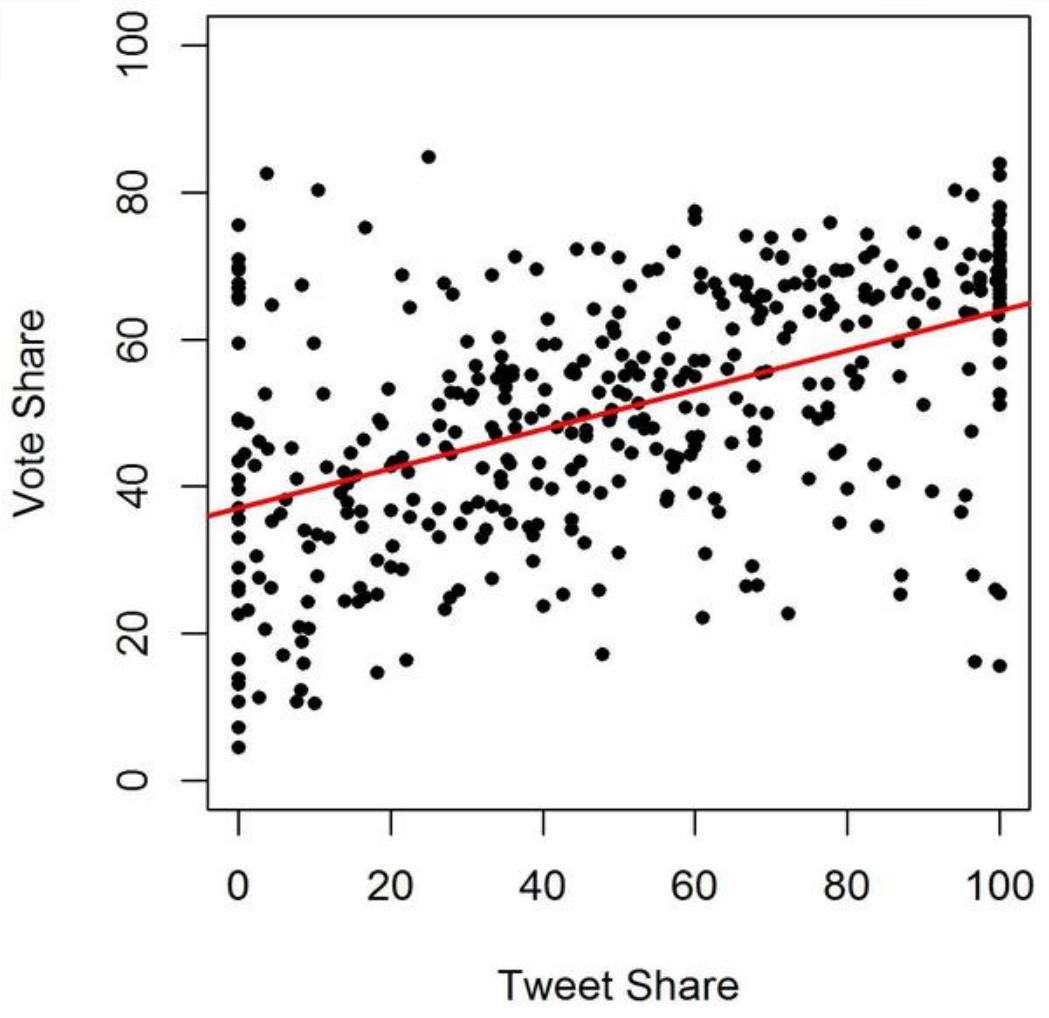
Tweet Share

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

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

PLOS ONE

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)
Const	37.042 (1.35)	−4.07 (15.04)
N	406	406
R^2_{adj}	.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