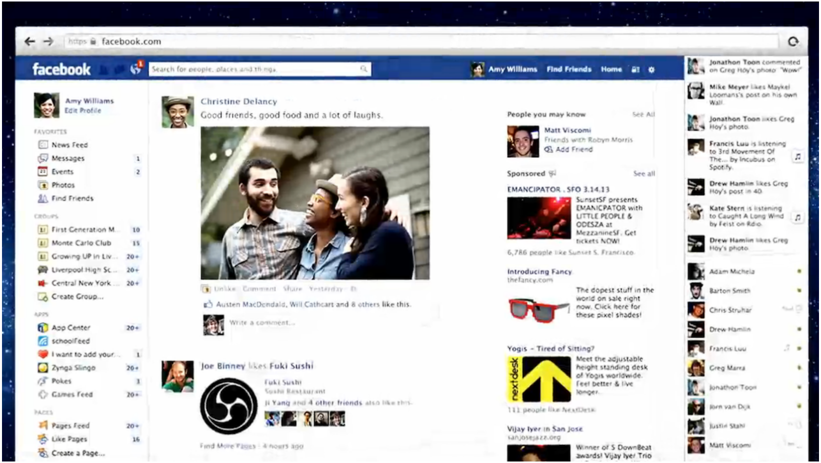
Social Science in the Age of Big Data

Matthew J. Salganik
Department of Sociology, Princeton University
& Cornell Tech

SIAM
July 15, 2016

https://commons.wikimedia.org/wiki/File:EscombrosBelAir5.jpg

http://www.businessdailyafrica.com/Corporate-News/Safaricom-cuts-calling-rates-to-Rwanda-by-60pc-/-/539550/2470628/-/thw2o5/-/index.html



Christine Delancy

Good friends, good food and a lot of laughs.

Austen MacDondald, Will Cathcart and 8 others like this.

Joe Binney likes Fuki Sushi

Fuki Sushi
Sushi Restaurant

Ji Yang and 4 other friends also like this.

Matt Viscomi
Friends with Robyn Morris

EMANCIPATOR . SFO 3.14.13

SunsetSF presents EMANICPATOR with LITTLE PEOPLE & ODESZA at MezzanineSF. Get tickets NOW!

6,786 people like Sunset S. Francisco.

Introducing Fancy
thefancy.com

The dopest stuff in the world on sale right now. Click here for these pixel shades!

Yogis - Tired of Sitting?

Meet the adjustable height standing desk of Yogis worldwide. Feel better & live longer.

111 people like NextDesk.

Vijay Iyer in San Jose
sanjosejazz.org

Winner of 5 DownBeat awards! Vijay Iyer Trio

Jonathon Toon commented on Greg Hoy's photo: "Wow!"

Mike Meyer likes Maykel Loomans's post on his own Wall.

Jonathon Toon likes Greg Hoy's photo.

Francis Luu is listening to 3rd Movement Of The... by Incubus on Spotify.

Drew Hamlin likes Greg Hoy's post in 40.

Kate Stern is listening to Caught A Long Wind by Feist on Rdio.

Drew Hamlin likes Greg Hoy's photo.

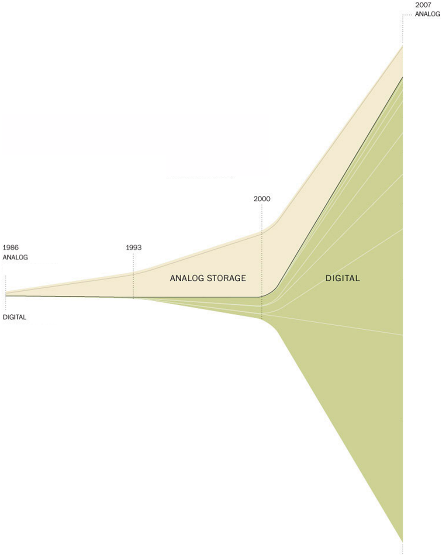
http://www.technobuffalo.com/2013/03/07/facebook-announces-newly-designed-news-feeds/

- Bengtsson et al. (2011) "Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti" *PLOS Medicine*.
- Blumenstock et al. (2015) "Predicting Poverty and Wealth from Mobile Phone Metadata" *Science*.
- Kramer et al. (2014) "Experimental evidence of massive-scale emotional contagion through social networks" *PNAS*.

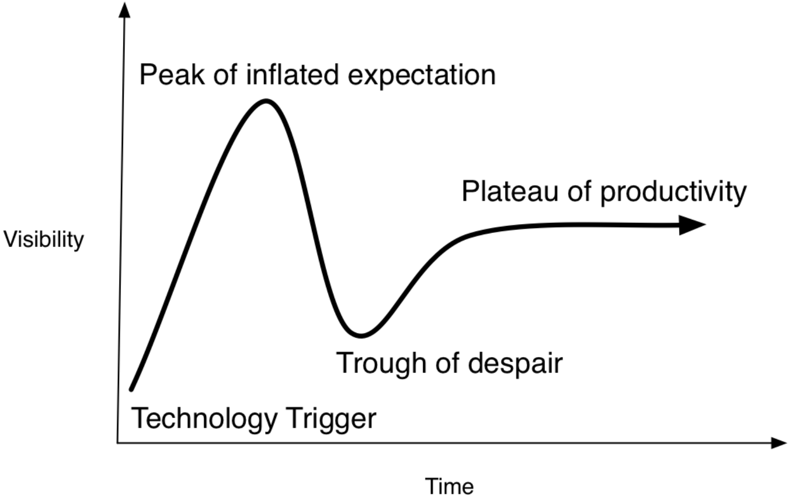
Bit by Bit:
Social Research in the Digital Age

Social Scientists ⟷ Data Scientists

Isn't computational social science a fad?



`http://www.washingtonpost.com/wp-dyn/content/graphic/2011/02/11/GR2011021100614.html`, based on Hilbert and Lòpez (2011)



Fenn and Raskino (2008)

Two important developments

Two important developments

- Online

Two important developments

- Online → Everywhere

Two important developments

- Online → Everywhere
- Found data

Two important developments

- Online → Everywhere
- Found data → Designed data

Four main research designs:

Four main research designs:

- Observing behavior

Four main research designs:

- Observing behavior
- Asking questions

Four main research designs:

- Observing behavior
- Asking questions
- Running experiments

Four main research designs:

- Observing behavior
- Asking questions
- Running experiments
- Creating mass collaboration

Four main research designs:

- Observing behavior
- Asking questions
- Running experiments
- Creating mass collaboration

Why should I care about surveys?

Why should I care about surveys in the age of big data?

We will always need to ask

- limitations of digital exhaust (fubu vs. nufu-nubu)

We will always need to ask

- limitations of digital exhaust (fubu vs. nufu-nubu)
- internal states vs. external states

We will always need to ask

- limitations of digital exhaust (fubu vs. nufu-nubu)
- internal states vs. external states
- inaccessibility of digital exhaust

We will always need to ask

- limitations of digital exhaust (fubu vs. nufu-nubu)
- internal states vs. external states
- inaccessibility of digital exhaust

But how we are going to ask is going to change

	Sampling	Interviews
1st era	Area probability	Face-to-face

	Sampling	Interviews
1st era	Area probability	Face-to-face
2nd era	Random digital dial probability	Telephone

	Sampling	Interviews
1st era	Area probability	Face-to-face
2nd era	Random digital dial probability	Telephone
3rd era		

	Sampling	Interviews
1st era	Area probability	Face-to-face
2nd era	Random digital dial probability	Telephone
3rd era	Non-probability	

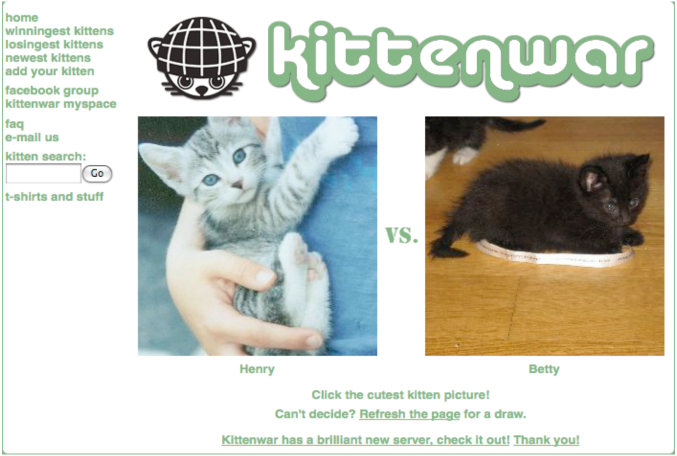
	Sampling	Interviews
1st era	Area probability	Face-to-face
2nd era	Random digital dial probability	Telephone
3rd era	Non-probability	Computer-administered

	Sampling	Interviews	Data environment
1st era	Area probability	Face-to-face	Stand-alone
2nd era	Random digital dial probability	Telephone	Stand-alone
3rd era	Non-probability	Computer-administered	Linked

	Sampling	Interviews	Data environment
1st era	Area probability	Face-to-face	Stand-alone
2nd era	Random digital dial probability	Telephone	Stand-alone
3rd era	Non-probability	Computer-administered	Linked

Human-administered → Computer-administered

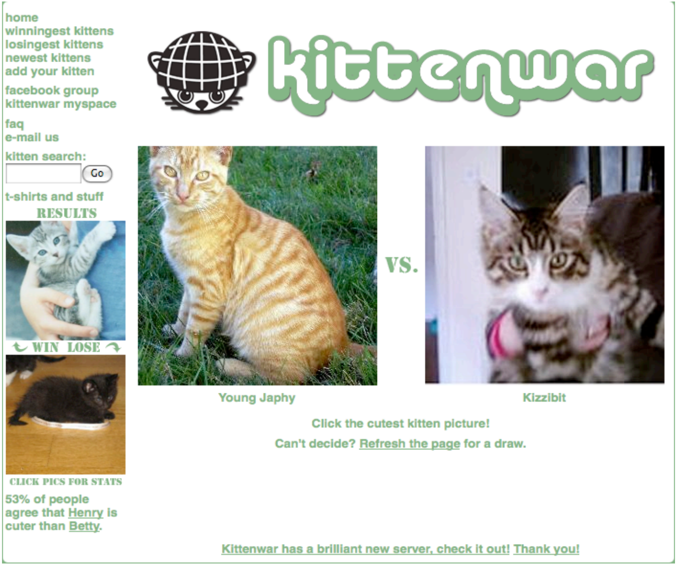
- enables change
- requires change



Click the cutest kitten picture!

Can't decide? Refresh the page for a draw.

Kittenwar has a brilliant new server, check it out! Thank you!

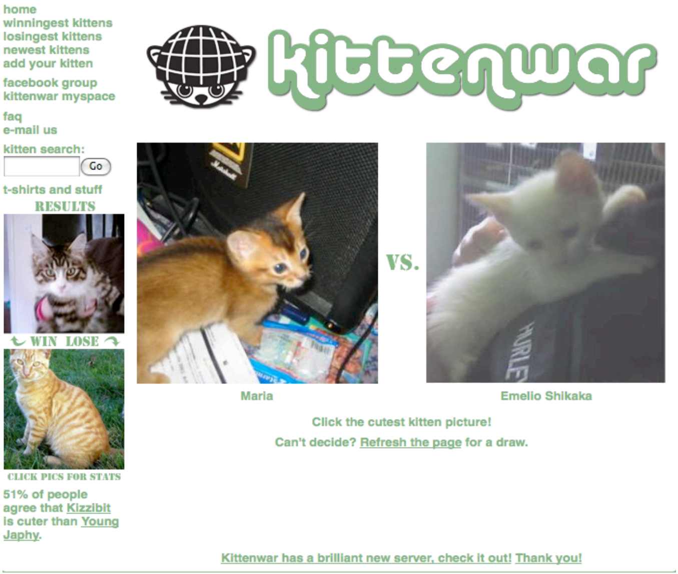


Click the cutest kitten picture!

Can't decide? Refresh the page for a draw.

53% of people agree that Henry is cuter than Betty.

Kittenwar has a brilliant new server, check it out! Thank you!



51% of people agree that Kizzibit is cuter than Young Japhy.

Click the cutest kitten picture!

Can't decide? Refresh the page for a draw.

Kittenwar has a brilliant new server, check it out! Thank you!

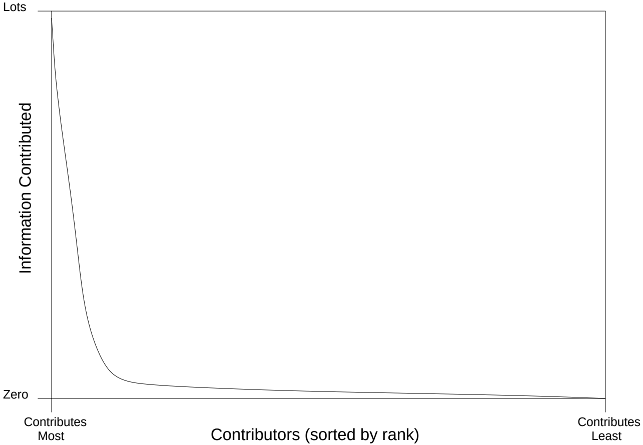
quantification or openness

quantification + openness =
wiki surveys

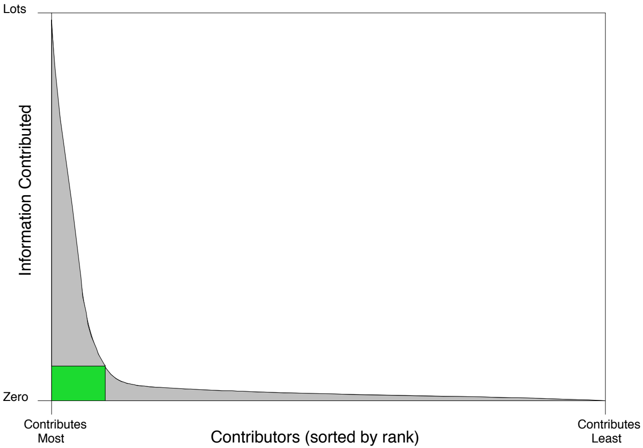
General principles of wiki surveys:

- greedy

Good web-based systems use the fat-head and the long-tail



Surveys don't use the fat-head or the long-tail



General principles of wiki surveys:

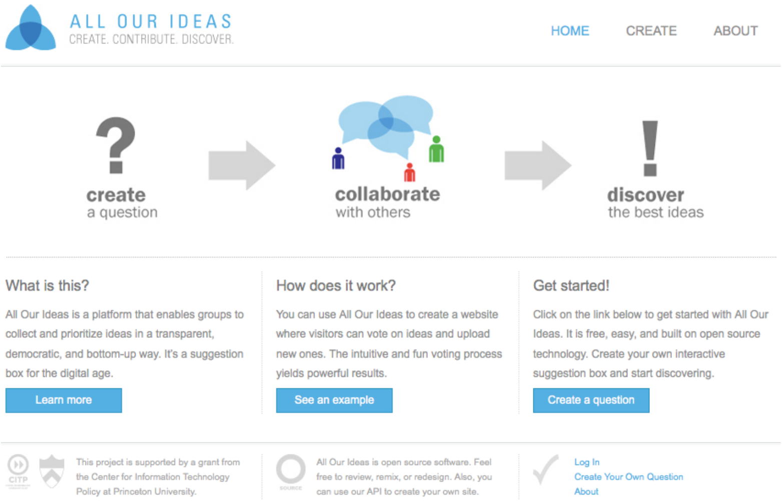
- greedy

General principles of wiki surveys:

- greedy
- collaborative

General principles of wiki surveys:

- greedy
- collaborative
- adaptive



What is this?

All Our Ideas is a platform that enables groups to collect and prioritize ideas in a transparent, democratic, and bottom-up way. It's a suggestion box for the digital age.

How does it work?

You can use All Our Ideas to create a website where visitors can vote on ideas and upload new ones. The intuitive and fun voting process yields powerful results.

Get started!

Click on the link below to get started with All Our Ideas. It is free, easy, and built on open source technology. Create your own interactive suggestion box and start discovering.

This project is supported by a grant from the Center for Information Technology Policy at Princeton University.

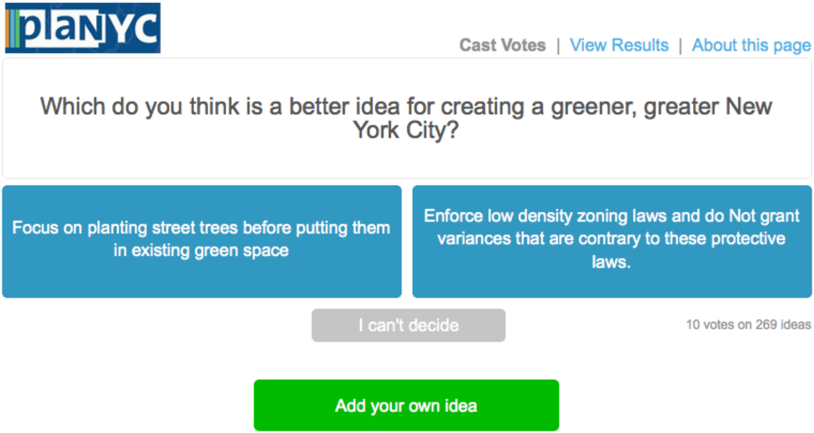
All Our Ideas is open source software. Feel free to review, remix, or redesign. Also, you can use our API to create your own site.

better.yahoo.com/planet
YAHOO! GREEN TAXI
better.yahoo.com/pl
DETROIT
YAHO
AXI FARE
No Charge
Learn how we are creating a greener,
greater New York City.

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

Seeded the wiki survey with 25 ideas:

- Require all big buildings to make certain energy efficiency upgrades
- Increase targeted tree plantings in neighborhoods with high asthma rates
- Establish a New York City Energy Planning Board

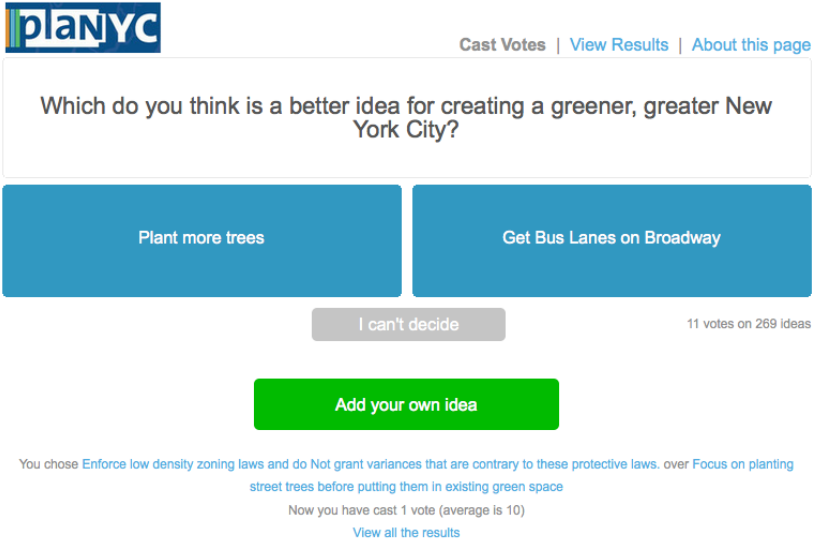


Which do you think is a better idea for creating a greener, greater New York City?

Focus on planting street trees before putting them in existing green space

Enforce low density zoning laws and do Not grant variances that are contrary to these protective laws.

10 votes on 269 ideas


planyc
Cast Votes | View Results | About this page
Plant more trees
Get Bus Lanes on Broadway
I can't decide
Add your own idea

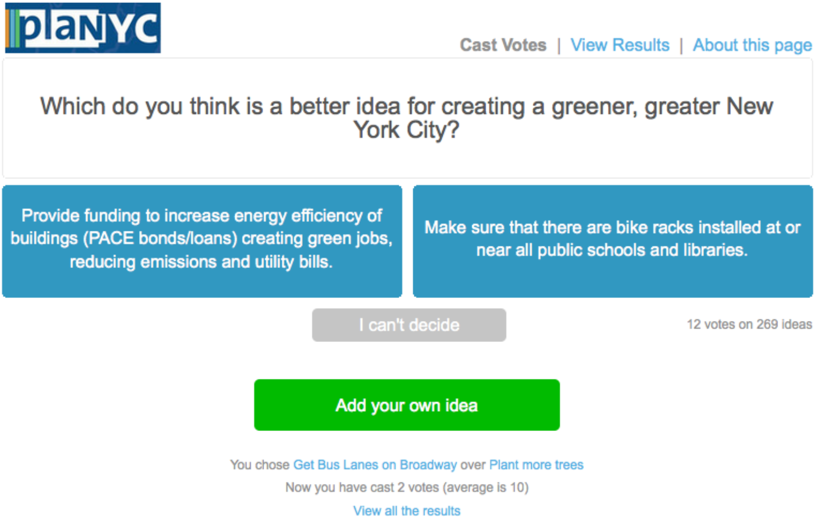
Which do you think is a better idea for creating a greener, greater New York City?

11 votes on 269 ideas

You chose Enforce low density zoning laws and do Not grant variances that are contrary to these protective laws. over Focus on planting street trees before putting them in existing green space

Now you have cast 1 vote (average is 10)

View all the results



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

Provide funding to increase energy efficiency of buildings (PACE bonds/loans) creating green jobs, reducing emissions and utility bills.

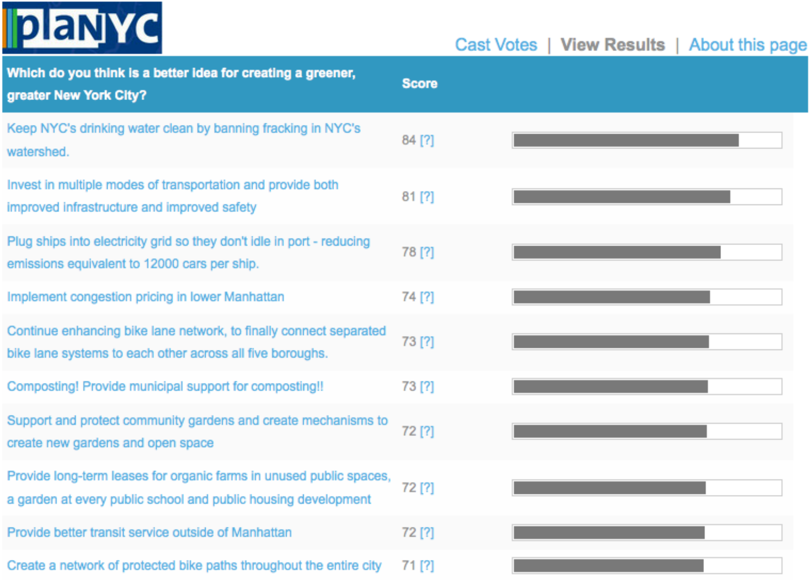
Make sure that there are bike racks installed at or near all public schools and libraries.

12 votes on 269 ideas

You chose Get Bus Lanes on Broadway over Plant more trees

Now you have cast 2 votes (average is 10)

View all the results



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

What are we trying to estimate?

Data

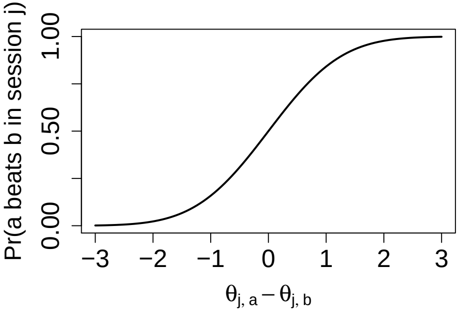
Vote	Session	Prompt	
1	1	**item 4**	item 1
2	1	item 3	**item 1**
3	1	**item 4**	item 3
4	2	**item 3**	item 4
5	2	item 4	**item 2**
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Opinion matrix

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,K} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{J,1} & \theta_{J,2} & \cdots & \theta_{J,K} \end{bmatrix}$$

$\theta_{j,k}$: how much respondent j likes item k

$$Pr[a \text{ beats } b \text{ in session } j] = \Phi(\theta_{j,a} - \theta_{j,b})$$



$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,K} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{J,K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{J,1} & \theta_{J,2} & \cdots & \theta_{J,K} \end{bmatrix}$$

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,K} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{J,K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{J,1} & \theta_{J,2} & \cdots & \theta_{J,K} \end{bmatrix}$$

$$\theta_{j,1} \sim N(\mu_1, \sigma)$$

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,K} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{J,K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{J,1} & \theta_{J,2} & \cdots & \theta_{J,K} \end{bmatrix}$$

$$\theta_{j,1} \sim N(\mu_1, \sigma)$$

$$\theta_{j,2} \sim N(\mu_2, \sigma)$$

$$\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,K} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{J,K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{J,1} & \theta_{J,2} & \cdots & \theta_{J,K} \end{bmatrix}$$

$$\theta_{j,1} \sim N(\mu_1, \sigma)$$

$$\theta_{j,2} \sim N(\mu_2, \sigma)$$

$$\vdots$$

$$\theta_{j,K} \sim N(\mu_K, \sigma)$$

$$p(\boldsymbol{\theta}, \boldsymbol{\mu} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\mu_0}, \boldsymbol{\tau_0^2}, \sigma) \propto$$

$$\begin{aligned}
&\prod_{i=1}^{V} \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta})^{y_i} (1 - \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta}))^{1-y_i} \\
&\times \prod_{j=1}^{J} \prod_{k=1}^{K} N(\theta_{j,k} \mid \mu_k, \sigma) \\
&\times \prod_{k=1}^{K} N\left(\mu_k \mid \mu_{0[k]}, \tau_{0[k]}^2\right)
\end{aligned}$$

What are we trying to estimate?

Opinion matrix

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

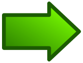
$\theta_{j,k}$: how much
respondent j likes item k

What are we trying to estimate?

Opinion matrix

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

$\theta_{j,k}$: how much respondent j likes item k



Score

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

s_k: probability that item k beats a randomly chosen item for a randomly chosen session

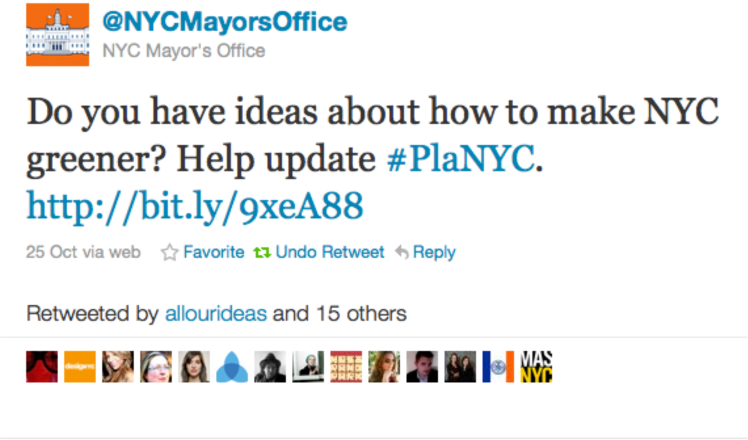
Learn how we are creating a greener, greater New York City.

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

Seeded the wiki survey with 25 ideas:

- Require all big buildings to make certain energy efficiency upgrades
- Increase targeted tree plantings in neighborhoods with high asthma rates
- Establish a New York City Energy Planning Board

Recruited participants through Twitter, Facebook, blogs, etc.



@NYCMayorsOffice
NYC Mayor's Office

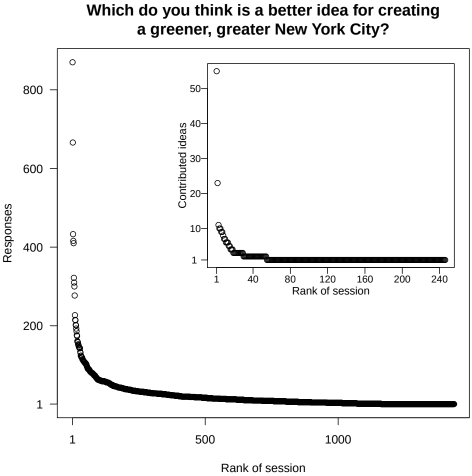
Do you have ideas about how to make NYC greener? Help update #PlaNYC. http://bit.ly/9xeA88

25 Oct via web

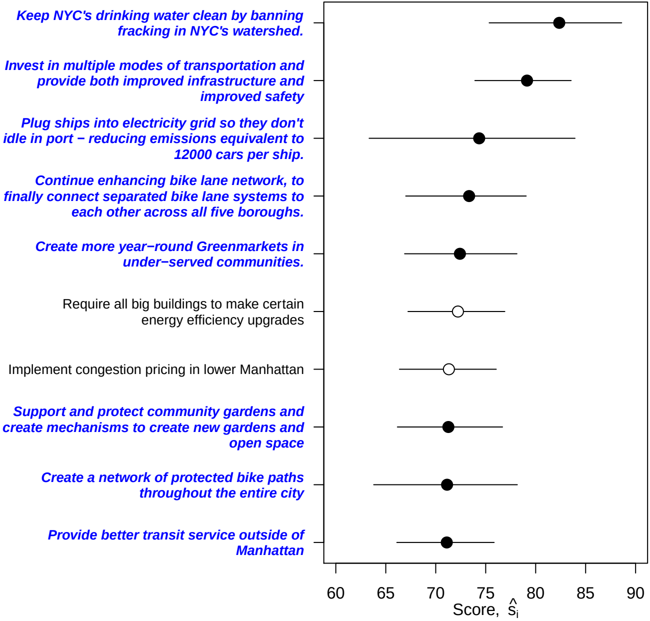
Retweeted by allourideas and 15 others

This is not a random sample, but random samples are possible

- 31,893 responses
- 464 ideas uploaded

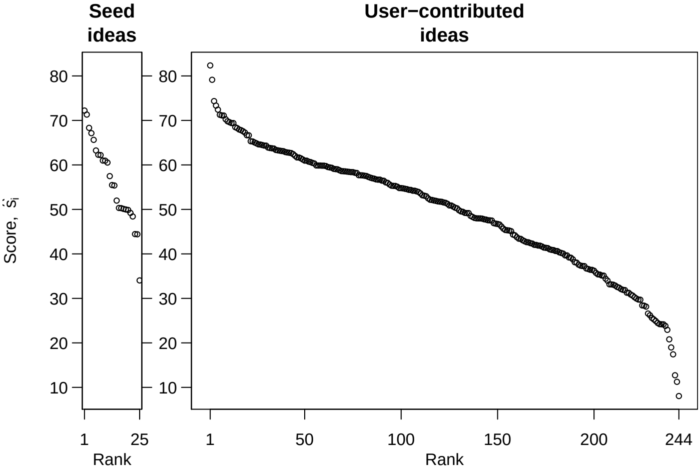


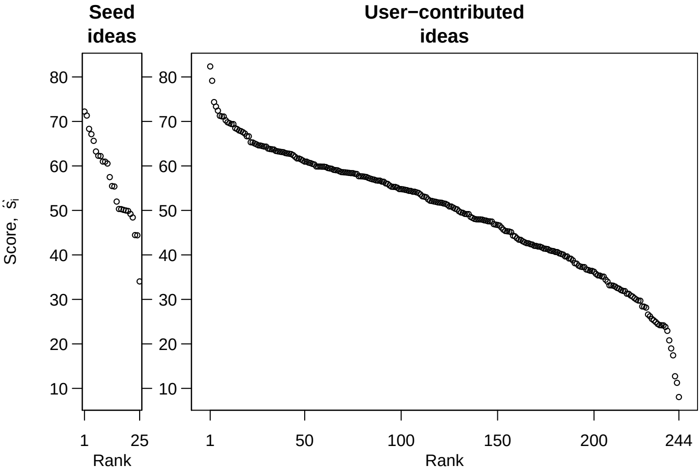
Which do you think is a better idea for creating a greener, greater New York City?



- Alternative framings: “Keep NYC’s drinking water clean by banning fracking in NYC’s watershed”

- Alternative framings: “Keep NYC’s drinking water clean by banning fracking in NYC’s watershed”
- Novel information: “Plug ships into electricity grid so they don’t idle in port - reducing emissions equivalent to 12000 cars per ship.”

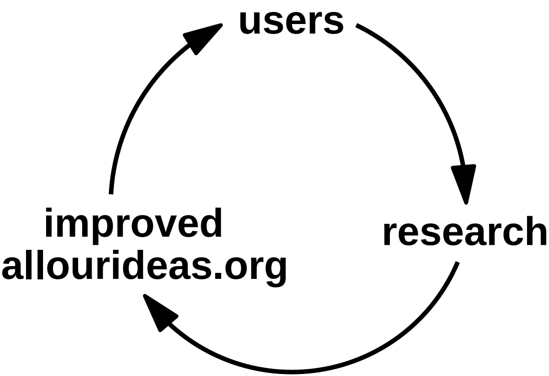

Seed
ideas
User−contributed
ideas
Score, $\hat{s}_i$
80
70
60
50
40
30
20
10
1
25
Rank
1
50
100
150
200
244
Rank



variance + volume → extreme cases

Currently hosting:
8,500 wiki surveys with 450,000 ideas and 12 million votes

e_democracia
participação virtual, cidadania real
OccupyWallStreet
The revolution continues worldwide!
LEGG MASON
GLOBAL ASSET MANAGEMENT
HARVARD
BUSINESS
PUBLISHING



users

research

improved allourideas.org

Virtuous cycle accelerated by openness

Salganik and Levy (2015) “Wiki surveys: Open and quantifiable social data collection” *PLoS ONE*.

Salganik and Levy (2015) “Wiki surveys: Open and quantifiable social data collection” *PLoS ONE*.

The All Our Ideas team, current and past:

- Peter Lubell-Doughtie, Adam Sanders, Pius Uzamere, Dhruv Kapadia, Chap Ambrose, Calvin Lee, Dmitri Garbuzov, Brian Tubergen, Peter Green, Luke Baker, Paul Yuan
- Josh Weinstein, Nadia Heninger, Bill Zeller, Bambi Tsui, Dhwani Shah, Karen Levy

More information:

- `http://opr.princeton.edu/archive/ws/`
- `http://www.allourideas.org`
- `http://blog.allourideas.org`
- `http://github.com/allourideas` (open source)
- Follow us 🐦: @allourideas

	Sampling	Interviews	Data environment
1st era	Area probability	Face-to-face	Stand-alone
2nd era	Random digital dial probability	Telephone	Stand-alone
3rd era	Non-probability	Computer-administered	Linked

Four main research designs:

- Observing behavior
- Asking questions
- Running experiments
- Creating mass collaboration

Concluding thoughts:

Concluding thoughts:

- We don’t need to look through people’s trash

Concluding thoughts:

- We don't need to look through people's trash
- With great power comes great responsibility

Concluding thoughts:

- We don't need to look through people's trash
- With great power comes great responsibility
- Technology is going to continue to change faster than researchers

Bit by Bit:
Social Research in the Digital Age

Social Scientists ⟷ Data Scientists

Open Review begins August 18

http://bitbybitbook.com