


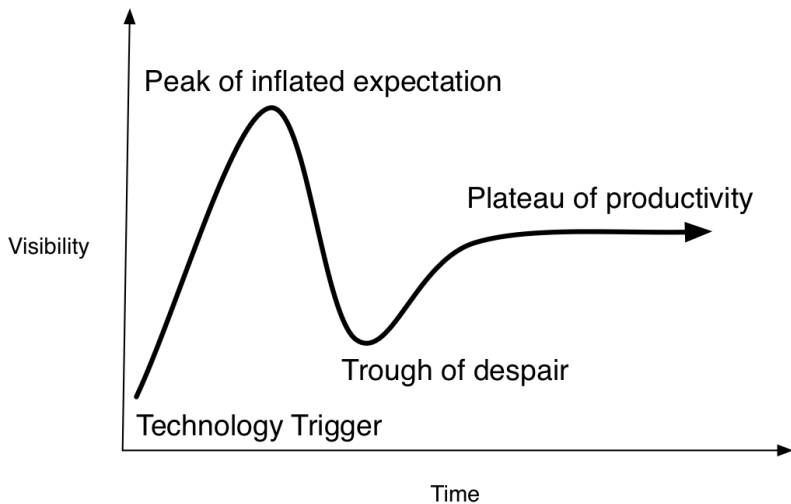
# Survey research in the digital age

Matthew J. Salganik  
Department of Sociology  
Princeton University  
 msalganik

AAPOR Webinar  
February 21, 2019



Isn't “big data” a fad?





Key abstraction is research design

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

Social Scientists  $\longleftrightarrow$  Data Scientists







## Readymades



Readymades





Readymades



Custommades



Readymades

+



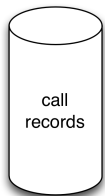
Custommades

[https://commons.wikimedia.org/wiki/File:Duchamp\\_Fontaine.jpg](https://commons.wikimedia.org/wiki/File:Duchamp_Fontaine.jpg)

[https://commons.wikimedia.org/wiki/File:%27David%27\\_by\\_Michelangelo\\_JBU0001.JPG](https://commons.wikimedia.org/wiki/File:%27David%27_by_Michelangelo_JBU0001.JPG)

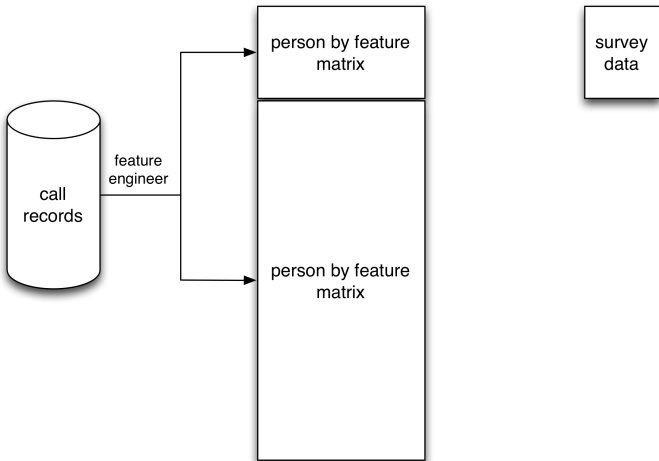
# Predicting poverty and wealth from mobile phone metadata

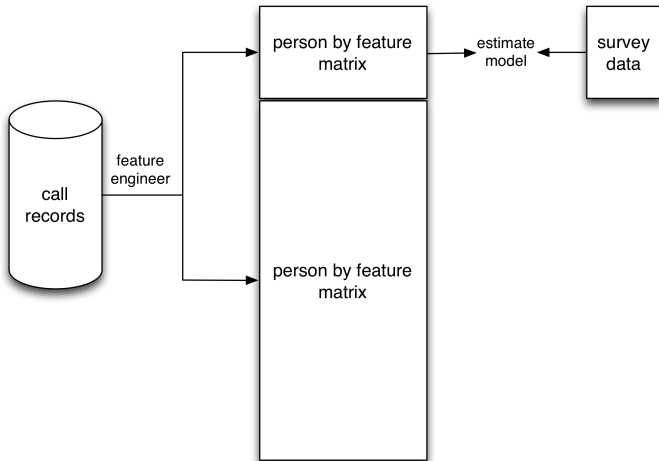
Joshua Blumenstock,<sup>1\*</sup> Gabriel Cadamuro,<sup>2</sup> Robert On<sup>3</sup>

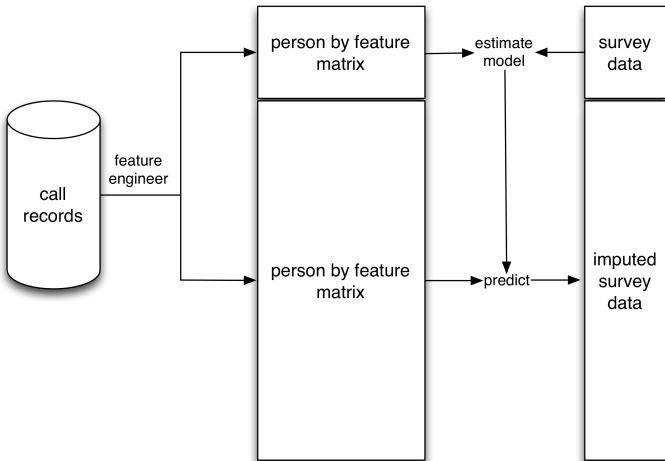


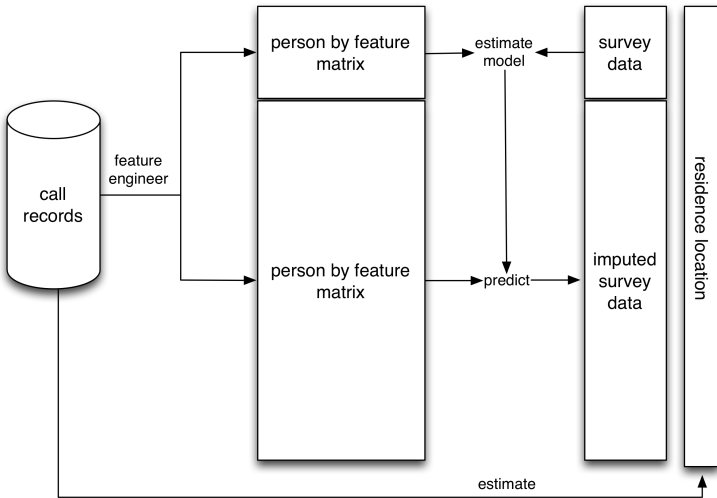


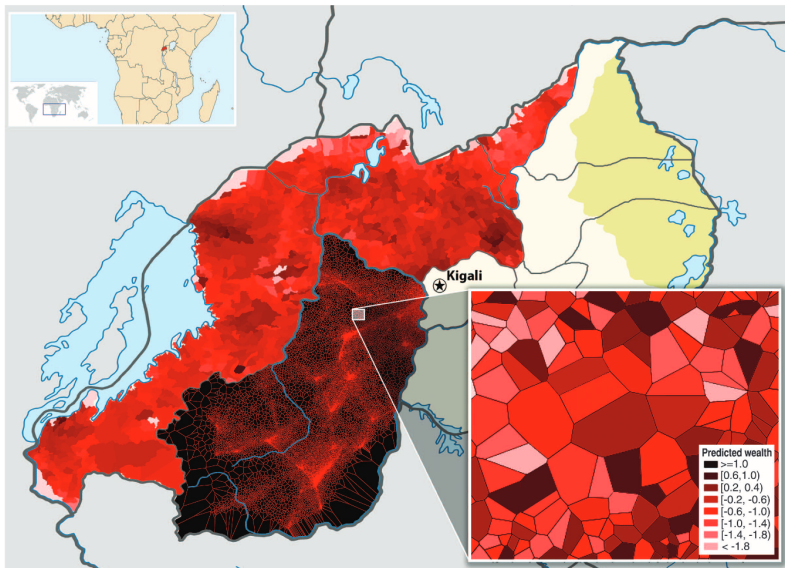




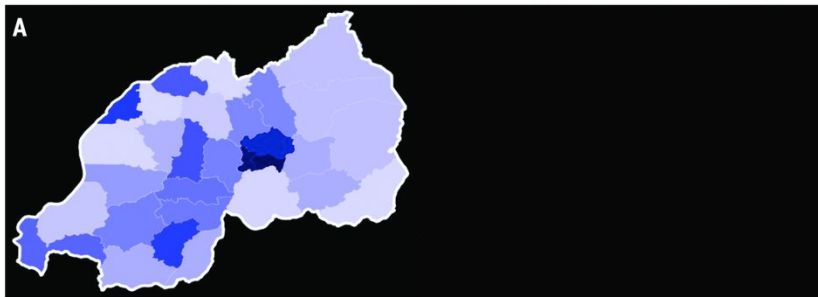




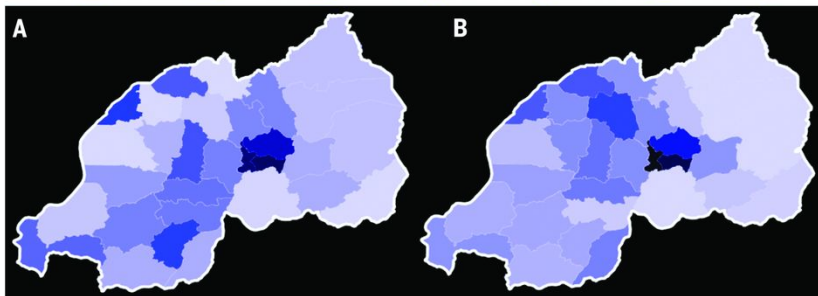




■  $\geq 0$  ■  $[-0.2, -0.0)$  ■  $[-0.4, -0.2)$  ■  $[-0.6, -0.4)$  ■  $[-0.8, -0.6)$  ■  $[-1.0, -0.8)$  ■  $[-1.2, -1.0)$  ■  $< -1.2$

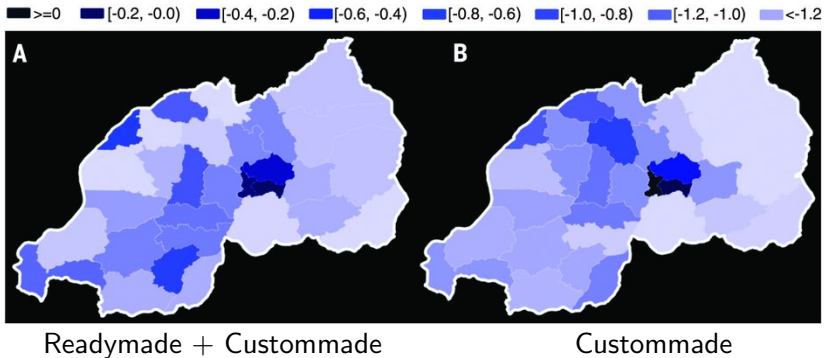


Readymade + Custommade



Readymade + Custommade

Custommade



- ▶ 10 times faster
- ▶ 50 times cheaper





Readymades

+



Custommades

[https://commons.wikimedia.org/wiki/File:Duchamp\\_Fontaine.jpg](https://commons.wikimedia.org/wiki/File:Duchamp_Fontaine.jpg)

[https://commons.wikimedia.org/wiki/File:%27David%27\\_by\\_Michelangelo\\_JBU0001.JPG](https://commons.wikimedia.org/wiki/File:%27David%27_by_Michelangelo_JBU0001.JPG)



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

We will always need to ask

- ▶ limitations of big data (fubu vs. nufu-nubu)

We will always need to ask

- ▶ limitations of big data (fubu vs. nufu-nubu)
- ▶ internal states vs. external states

We will always need to ask

- ▶ limitations of big data (fubu vs. nufu-nubu)
- ▶ internal states vs. external states
- ▶ inaccessibility of big data

We will always need to ask

- ▶ limitations of big data (fubu vs. nufu-nubu)
- ▶ internal states vs. external states
- ▶ inaccessibility of big data

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	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

	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

## Probability Samples

$$P(u_i) = \frac{p_i}{(N-1) \cdots (N-n+1)} \binom{N-1}{n-1} (n-1)! \\ + \sum_{j \neq i} \frac{p_j}{(N-1) \cdots (N-n+1)} \binom{N-1}{n-1} (n-1)! \frac{n-1}{N-1},$$

which upon simplification becomes

$$(19) \quad P(u_i) = \frac{N-n}{N-1} p_i + \frac{n-1}{N-1}, \quad (i = 1, 2, \dots, N).$$

Similarly, it may be shown that for this case

$$(20) \quad P(u_i u_j) = \frac{n-1}{N-1} \left[ \frac{N-n}{N-2} (p_i + p_j) + \frac{n-2}{N-2} \right], \\ (i \neq j; i, j = 1, 2, \dots, N).$$

## Non-Probability Samples



<https://www.chicagotribune.com/news/opinion/commentary/ct-truman-defeats-dewey-1948-flashback-perspec-1113-md-20161111-story.html>

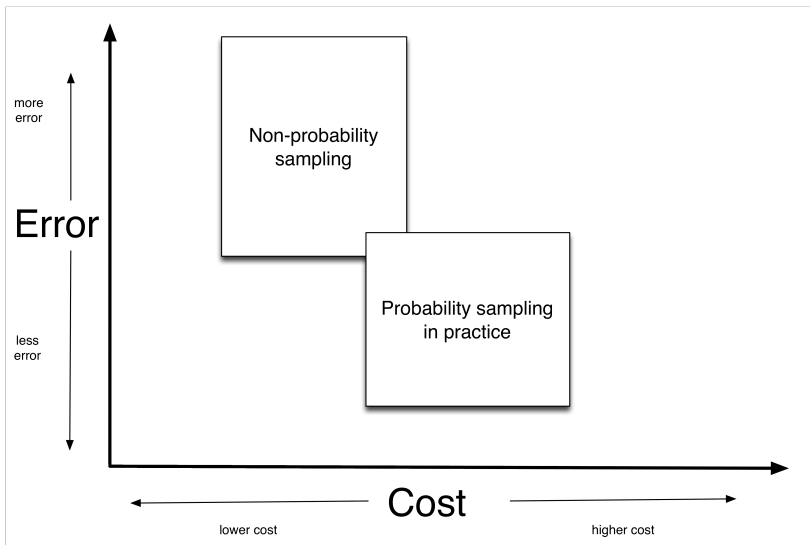
## Probability Samples

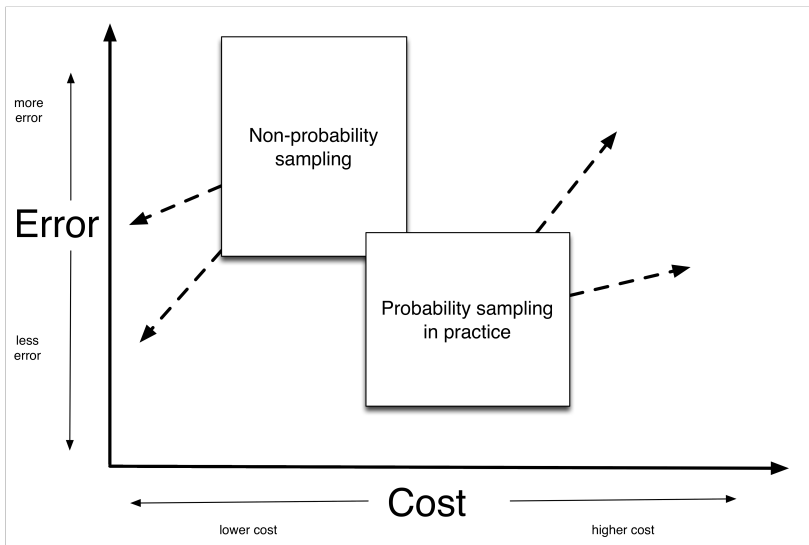
unknown sampling process  
weighting based on unverifiable assumptions

## Non-Probability Samples

unknown sampling process  
weighting based on unverifiable assumptions







	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

RESEARCH ARTICLE

# Wiki Surveys: Open and Quantifiable Social Data Collection

**Matthew J. Salganik<sup>1</sup>, Karen E. C. Levy<sup>2</sup>**

**1** Department of Sociology, Center for Information Technology Policy, and Office of Population Research, Princeton University, Princeton, NJ, USA, **2** Information Law Institute and Department of Media, Culture, and Communication, New York University, New York, NY, USA and Data & Society Research Institute, New York, NY, USA

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

[home](#)  
[winningest kittens](#)  
[losingest kittens](#)  
[newest kittens](#)  
[add your kitten](#)  
[facebook group](#)  
[kittenwar myspace](#)

[faq](#)  
[e-mail us](#)

kitten search:

Go

[t-shirts and stuff](#)



# kittenwar



Henry

VS.



Betty

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!](#)

<http://kittenwar.com>

home  
winningest kittens  
losingest kittens  
newest kittens  
add your kitten  
facebook group  
kittenwar myspace  
faq  
e-mail us

kitten search:

Go

t-shirts and stuff  
**RESULTS**



WIN LOSE

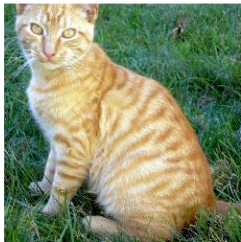


CLICK PICS FOR STATS

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

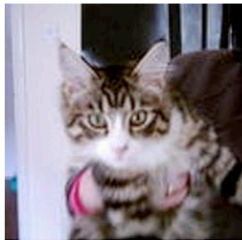


# kittenwar



Young Japhy

VS.



Kizzibit

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!](#)

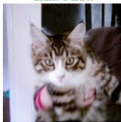
home  
winningest kittens  
losingest kittens  
newest kittens  
add your kitten  
facebook group  
kittenwar myspace

faq  
e-mail us

kitten search:

Go

t-shirts and stuff  
RESULTS



WIN LOSE

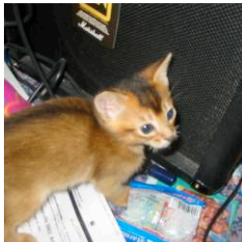


CLICK PICS FOR STATS

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

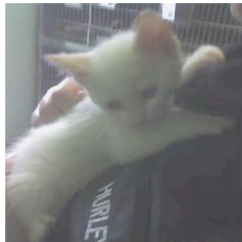


# kittenwar



Marla

VS.



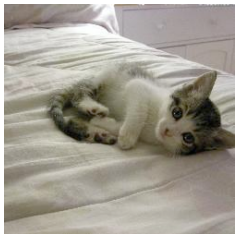
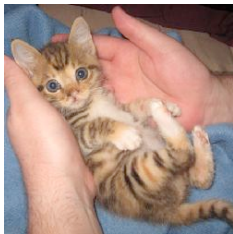
Emelio Shikaka

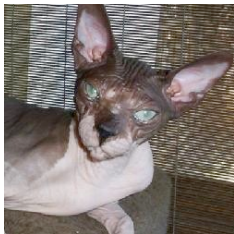
Click the cutest kitten picture!

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



## Bringing survey research into the digital age.

Mix core ideas from survey research with new insights from crowdsourcing. Add a heavy dose of statistics. Stir in a bit of fresh thinking. Enjoy.

[Try a Wiki Survey](#)[Create a Wiki Survey](#)

### HOW A WIKI SURVEY WORKS



#### Create

Start with a question and some seed ideas, and you can create a wiki survey in moments.



#### Participate

The participants you invite will enjoy our simple process of voting and adding new ideas.



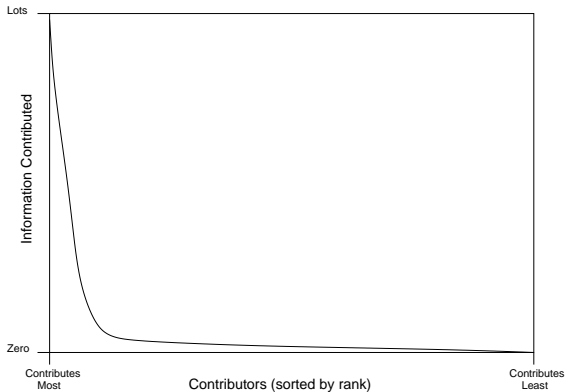
#### Discover

The best ideas will bubble to the top using our system that is open, transparent, and powerful.

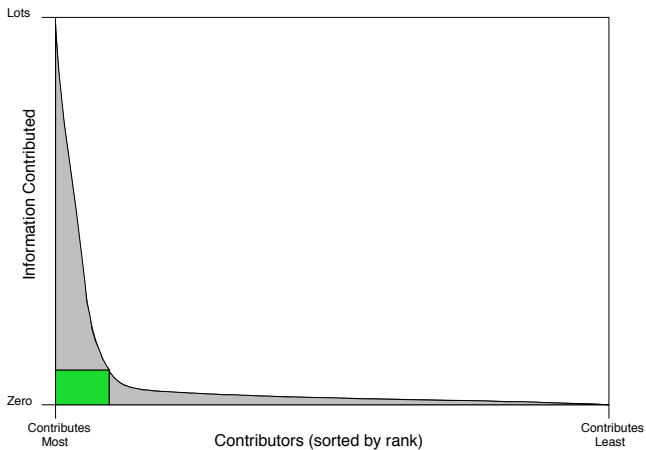
General principles of wiki surveys:

- ▶ 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



## Bringing survey research into the digital age.

Mix core ideas from survey research with new insights from crowdsourcing. Add a heavy dose of statistics. Stir in a bit of fresh thinking. Enjoy.

[Try a Wiki Survey](#)[Create a Wiki Survey](#)

### HOW A WIKI SURVEY WORKS



#### Create

Start with a question and some seed ideas, and you can create a wiki survey in moments.



#### Participate

The participants you invite will enjoy our simple process of voting and adding new ideas.



#### Discover

The best ideas will bubble to the top using our system that is open, transparent, and powerful.



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



[Cast Votes](#) | [View Results](#) | [About this page](#)

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.

I can't decide

10 votes on 269 ideas

Add your own idea



[Cast Votes](#) | [View Results](#) | [About this page](#)

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

Plant more trees

Get Bus Lanes on Broadway

I can't decide

11 votes on 269 Ideas

Add your own idea

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](#)





[Cast Votes](#) | [View Results](#) | [About this page](#)

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.

I can't decide

12 votes on 269 Ideas

Add your own idea

You chose [Get Bus Lanes on Broadway](#) over [Plant more trees](#)

Now you have cast 2 votes (average is 10)

[View all the results](#)



[Cast Votes](#) | [View Results](#) | [About this page](#)

**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 [?]



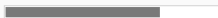
Composting! Provide municipal support for composting!!

73 [?]



Support and protect community gardens and create mechanisms to create new gardens and open space

72 [?]



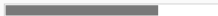
Provide long-term leases for organic farms in unused public spaces, a garden at every public school and public housing development

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	<b>item 4</b>	item 1
2	1	item 3	<b>item 1</b>
3	1	<b>item 4</b>	item 3
4	2	<b>item 3</b>	item 4
5	2	item 4	<b>item 2</b>
$\vdots$	$\vdots$	$\vdots$	$\vdots$



Opinion matrix

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

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

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>

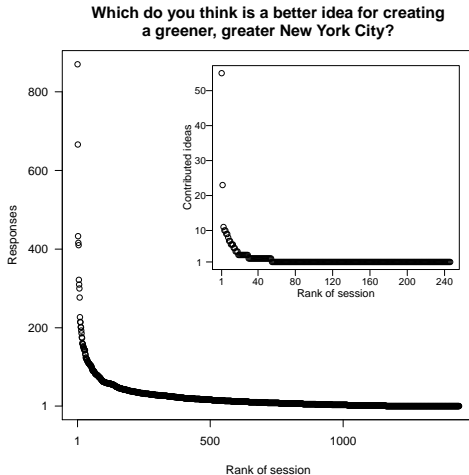
25 Oct via web ☆ Favorite ↻ Undo Retweet ↻ Reply

Retweeted by [allourideas](#) and 15 others

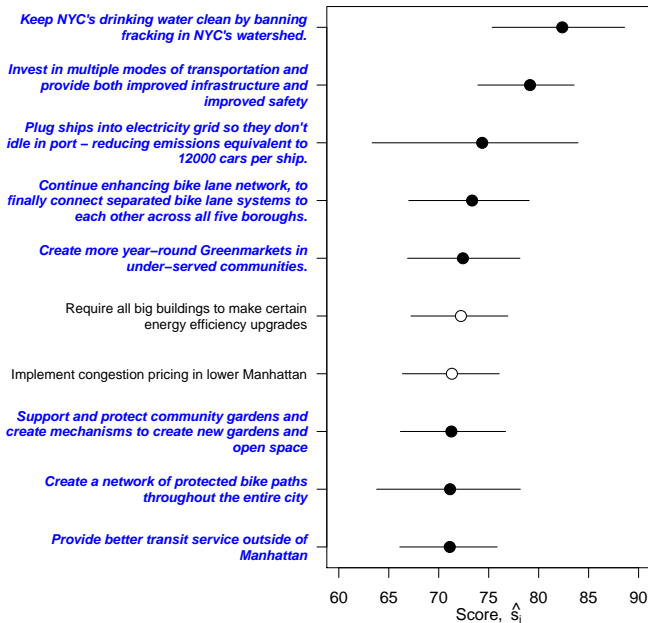


This is **not a random sample**, but random samples are possible

- ▶ 31,893 responses
- ▶ 464 ideas uploaded



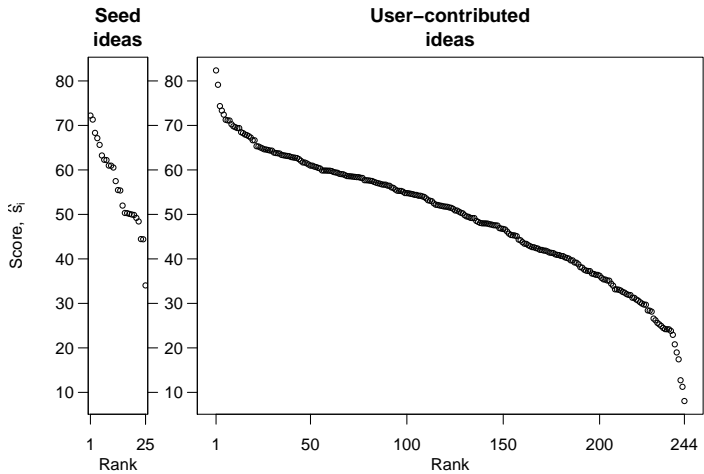
Which do you think is a better idea for creating a greener, greater New York City?

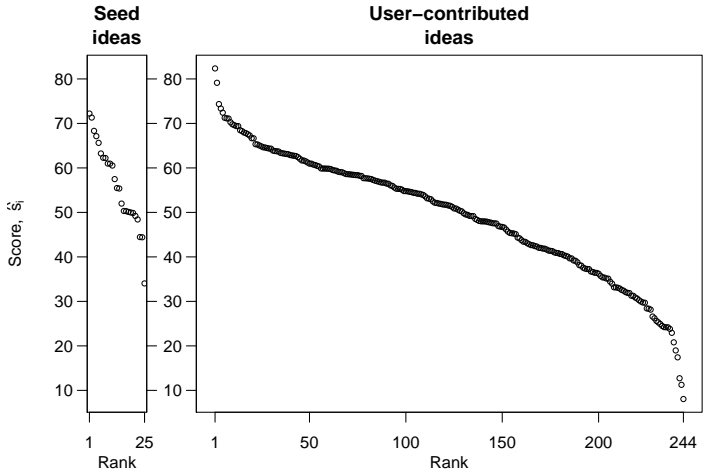


- ▶ Alternative framings: “Keep NYC’s drinking water clean by banning fracking in NYC’s watershed”



- ▶ 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.”



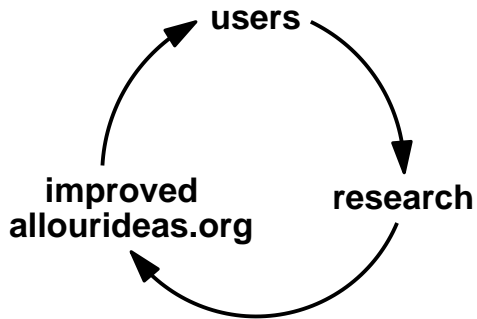


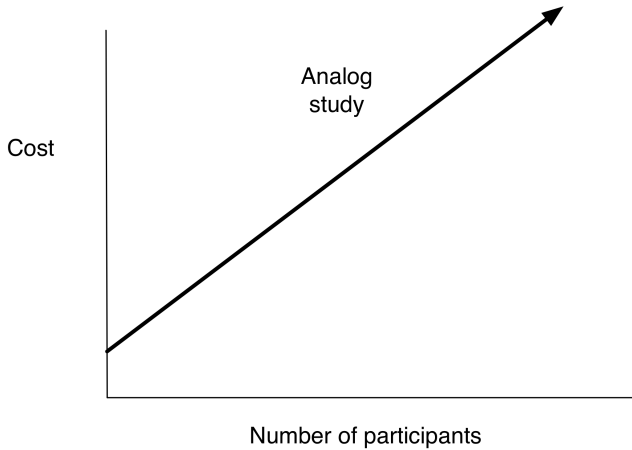
variance + volume  $\rightarrow$  extreme cases

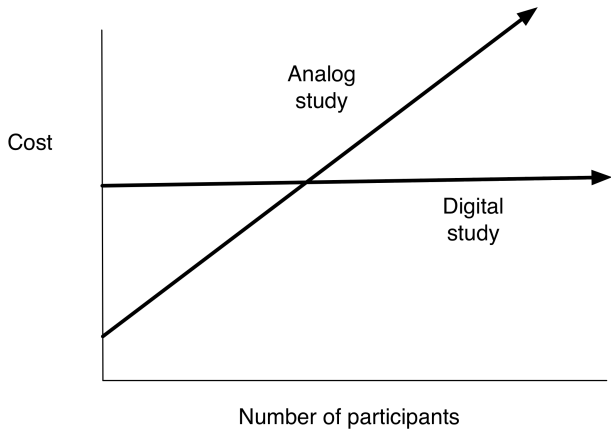
Currently hosting:

14,000 wiki surveys with 700,000 ideas and 21 million votes









	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

Will big data kill surveys?

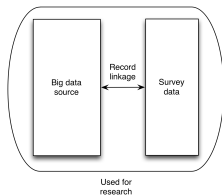


<http://schlitterblog.com/wp-content/uploads/2014/05/peanutbutterlover.jpg>

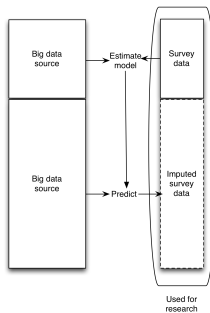


<http://schlitterblog.com/wp-content/uploads/2014/05/peanutbutterlover.jpg>

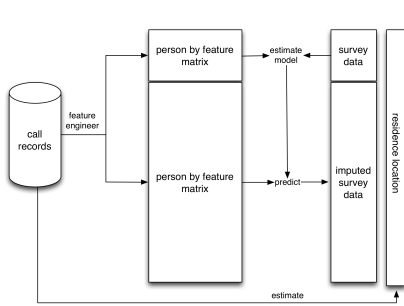
### Enriched asking



### Amplified asking



Note the different role of the big data in each case



## Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,<sup>1\*</sup> Gabriel Cadamuro,<sup>2</sup> Robert On<sup>3</sup>

<http://dx.doi.org/10.1126/science.aac4420>

# Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty

Muhammad Raza Khan<sup>1</sup>, Joshua Mittal<sup>2</sup>, Arshad Singh<sup>3</sup>, Joshua Hammerback<sup>4</sup>  
Information School, University of Washington, Seattle, WA, USA  
Email: {khanmraz@u.washington.edu, jmittal@u.washington.edu, arshad@u.washington.edu, jhammer@u.washington.edu}

**Behavioral Modeling for Churn Prediction:  
Early Indicators and Accurate Predictors of Custom Defection and Loyalty**

Muhammad Raza Khan<sup>1</sup>, Joshua Mittal<sup>2</sup>, Arshate Singh<sup>3</sup>, Joshua Blumenstock<sup>4</sup>  
Information School, University of Washington, Seattle, WA, USA  
Email: {khanmrazan@cs.washington.edu, jmittal@cs.washington.edu, arshate@cs.washington.edu, joshblum@uw.edu}

**Calling for Better Measurement:**

Estimating an Individual's Wealth and Well-Being  
from Mobile Phone Transaction Records

Joshua E. Blumenstock  
University of Washington  
Seattle, WA  
joshblum@uw.edu



**Behavioral Modeling for Churn Prediction:  
Early Indicators and Accurate Predictors of Custom Defection and Loyalty**

Muhammad Raza Khan<sup>1</sup>, Joshua Mittal<sup>2</sup>, Arshad Singh<sup>3</sup>, Joshua Blumenstock<sup>4</sup>  
Information School, University of Washington, Seattle, WA, USA  
Email: {mrazan@uw.edu, jmittal@uw.edu, arshadsingh@uw.edu, joshblum@uw.edu}

**Calling for Better Measurement:**

Estimating an Individual's Wealth and Well-Being  
from Mobile Phone Transaction Records

Joshua E. Blumenstock  
University of Washington  
Seattle, WA  
joshblum@uw.edu

**Predicting poverty and wealth from  
mobile phone metadata**

Joshua Blumenstock,<sup>1\*</sup> Gabriel Cadamuro,<sup>2</sup> Robert Oni<sup>3</sup>

**Behavioral Modeling for Churn Prediction:  
Early Indicators and Accurate Predictors of Custom Defection and Loyalty**

Muhammad Raza Khan<sup>1</sup>, Joshua Mittal<sup>2</sup>, Arshad Shafiq<sup>3</sup>, Joshua Blumenstock<sup>4</sup>  
Information School, University of Washington, Seattle, WA, USA  
Email: {mrazan@cs.washington.edu, jmittal@cs.washington.edu, arshadshafiq@cs.washington.edu, joshblum@uw.edu}

**Calling for Better Measurement:**

Estimating an Individual's Wealth and Well-Being  
from Mobile Phone Transaction Records

Joshua E. Blumenstock  
University of Washington  
Seattle, WA  
joshblum@uw.edu

**Predicting poverty and wealth from  
mobile phone metadata**

Joshua Blumenstock,<sup>1\*</sup> Gabriel Cadamuro,<sup>2</sup> Robert Oni<sup>3</sup>

The beginning is not the end . . . .

# Combining satellite imagery and machine learning to predict poverty

Neal Jean,<sup>1,2\*</sup> Marshall Burke,<sup>3,4,5\*</sup>† Michael Xie,<sup>1</sup> W. Matthew Davis,<sup>4</sup>  
David B. Lobell,<sup>3,4</sup> Stefano Ermon<sup>1</sup>

# Combining satellite imagery and machine learning to predict poverty

Neal Jean,<sup>1,2\*</sup> Marshall Burke,<sup>3,4,5\*\*</sup> Michael Xie,<sup>1</sup> W. Matthew Davis,<sup>4</sup>  
David B. Lobell,<sup>3,4</sup> Stefano Ermon<sup>1</sup>

## Artificial Intelligence Is Predicting Human Poverty From Space

August 18, 2016 // 02:00 PM EST

<http://dx.doi.org/10.1126/science.aaf7894>

[https://motherboard.vice.com/en\\_us/article/artificial-intelligence-is-predicting-human-poverty-from-space](https://motherboard.vice.com/en_us/article/artificial-intelligence-is-predicting-human-poverty-from-space)

Daytime satellite images are available, but most researchers had been using night lights



[https://www.nasa.gov/multimedia/imagegallery/image\\_feature\\_2480.html](https://www.nasa.gov/multimedia/imagegallery/image_feature_2480.html)

Prior research:

Nightlights + survey data  $\rightarrow$  estimates of wealth in places without surveys

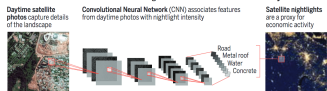
Jean et al. (2016):

Day pictures + Nightlights + survey data  $\rightarrow$  estimates of wealth in places without surveys

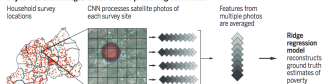
## Predicting poverty

Satellite images can be used to estimate wealth in remote regions.

**Neural network learns features in satellite images that correlate with economic activity**



**Daytime satellite images can be used to predict regional wealth**



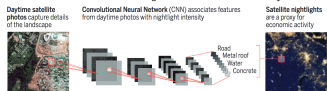
- Start with CNN pretrained on ImageNet



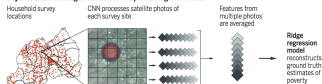
## Predicting poverty

Satellite images can be used to estimate wealth in remote regions.

### Neural network learns features in satellite images that correlate with economic activity



### Daytime satellite images can be used to predict regional wealth

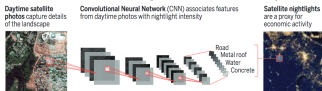


- ▶ Start with CNN pretrained on ImageNet
- ▶ Train CNN to predict nightlights from day pictures (lots of training data)

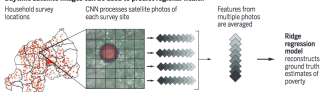
## Predicting poverty

Satellite images can be used to estimate wealth in remote regions.

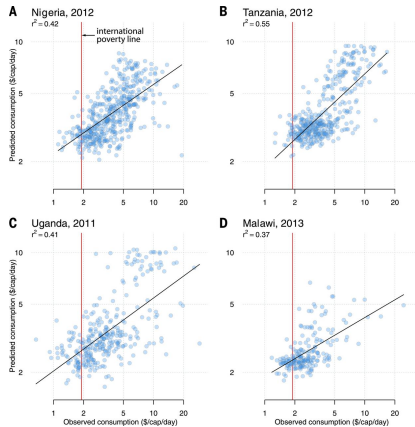
Neural network learns features in satellite images that correlate with economic activity



Daytime satellite images can be used to predict regional wealth

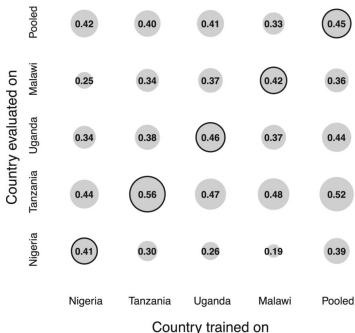


- ▶ Start with CNN pretrained on ImageNet (e.g. hampsters and weasels)
- ▶ Train CNN to predict nightlights from day pictures (lots of training data)
- ▶ Take features from CNN and train ridge regression to predict cluster mean survey response

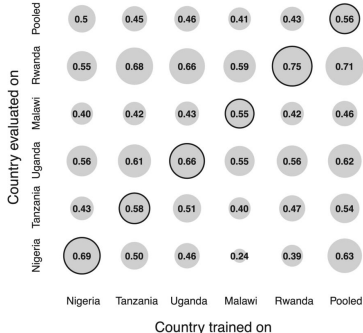


<http://dx.doi.org/10.1126/science.aaf7894>

## A Consumption expenditures



## B Assets



Two patterns:

- ▶ Performance decreases when train on one country and test on another
- ▶ Performance varies by the quantity being estimated (assets seems easier to estimate than consumption expenditures)

nealjean / predicting-poverty

Watch 22 Star 110 Fork 60

Code Issues 1 Pull requests 0 Projects 0 Wiki Insights

Combining satellite imagery and machine learning to predict poverty

18 commits 1 branch 0 releases 4 contributors MIT

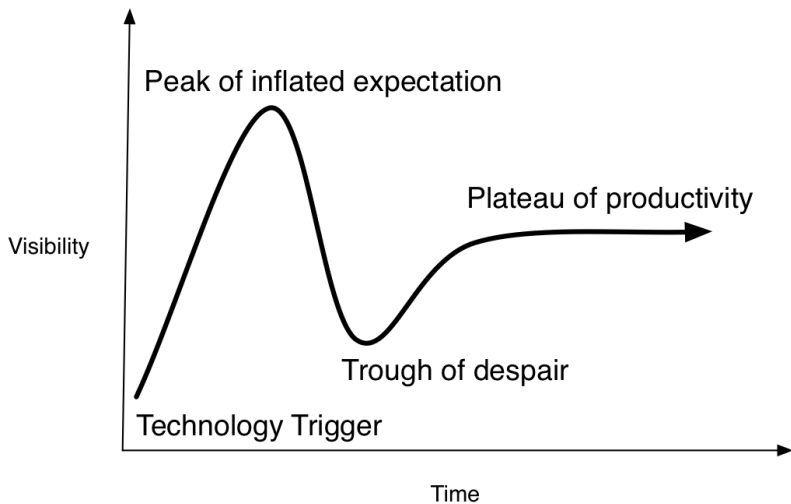
Branch: master New pull request Create new file Upload files Find file Clone or download

lmthexie select middle of pixel	Latest commit 975f6dc on Mar 27
data/input	Clean replication code 10 months ago
figures	Fixing cluster prefix in fig_utils.py 7 months ago
model	Clean replication code 10 months ago
scripts	select middle of pixel 3 months ago
.gitignore	Clean replication code 10 months ago
LICENSE	MIT License 6 months ago
README.md	Update README.md 8 months ago
requirements.txt	Clean replication code 10 months ago

<https://github.com/nealjean/predicting-poverty>

	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







- ▶ Read: <http://www.bitbybitbook.com>
- ▶ Teach: <http://www.bitbybitbook.com/en/teaching/>