

# Linking survey data to other data (03-05)

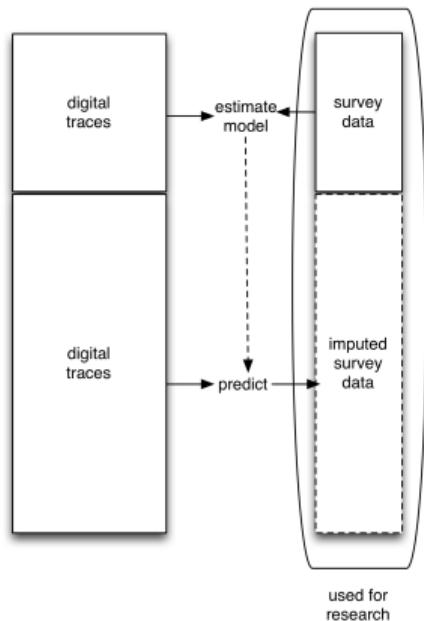
Matthew J. Salganik  
Department of Sociology  
Princeton University

Soc 596: Computational Social Science  
Fall 2016

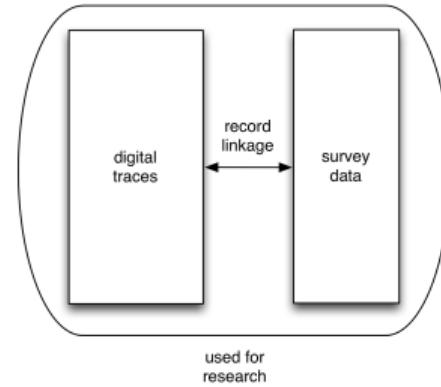


	Sampling	Interviews	Data environment
1st era	Area probability	Face-to-face	Stand-alone
2nd era	Random digital dial probability	Telephone	Stand-alone
3rd era	Non-probability	Computer-administered	<a href="#">Linked</a>

## Amplified asking



## Enriched asking

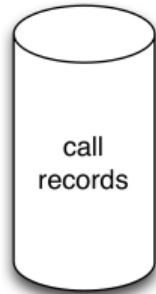


Note the different role of the big data in each case

# Predicting poverty and wealth from mobile phone metadata

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

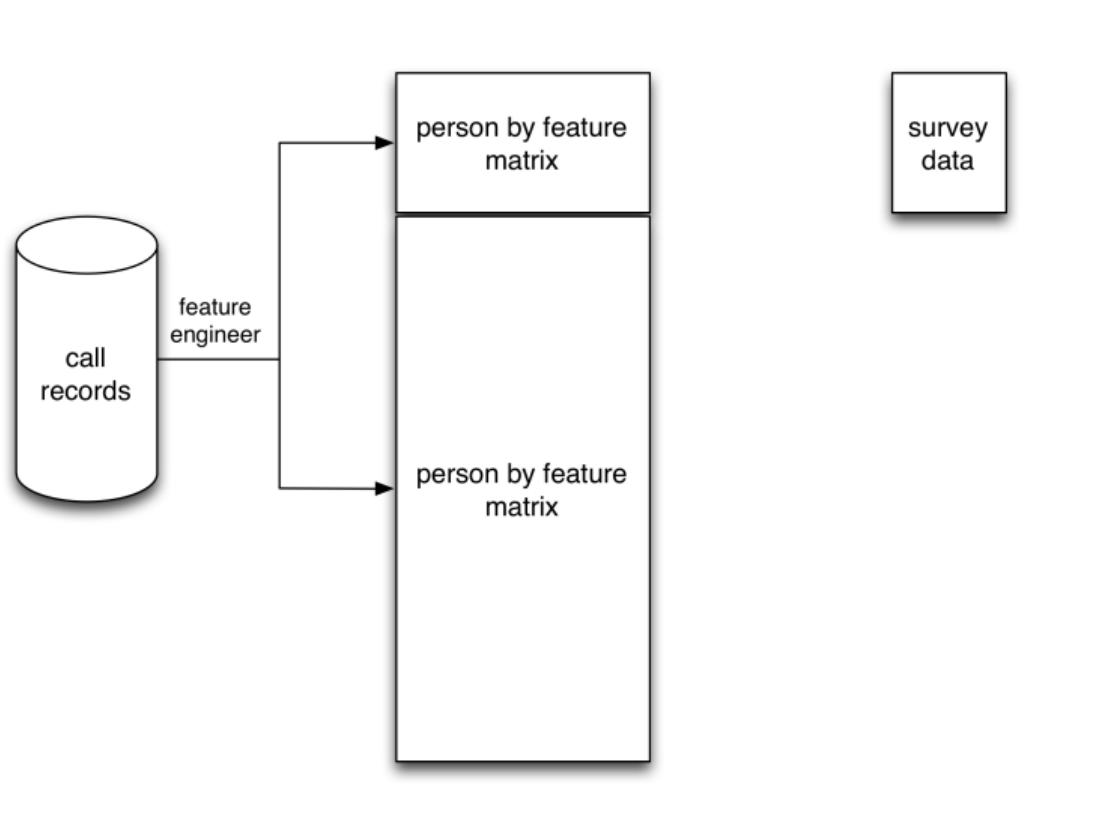


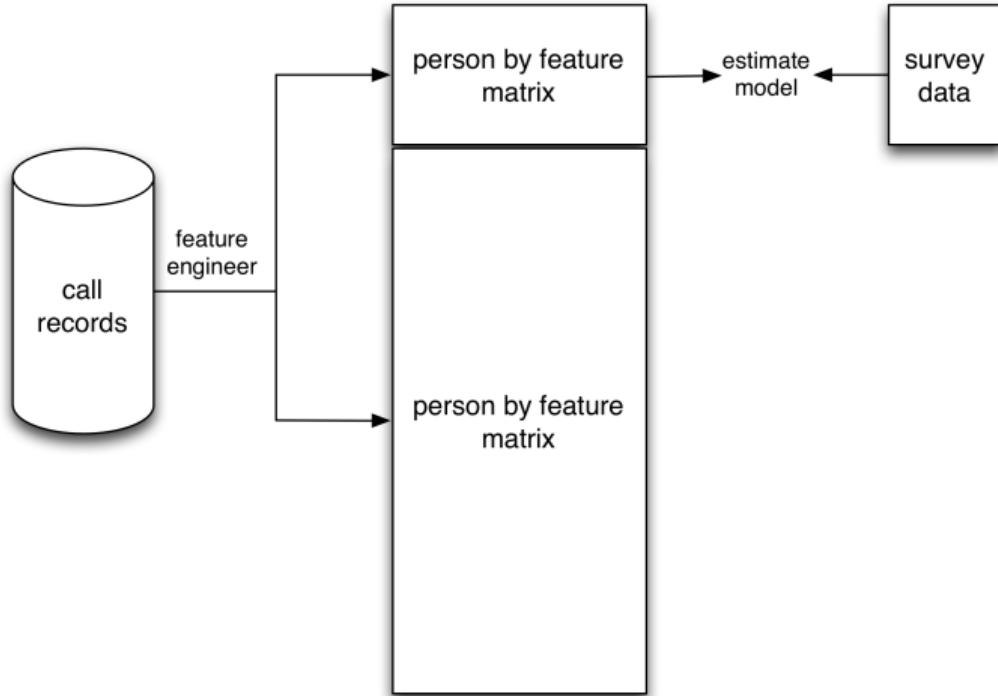


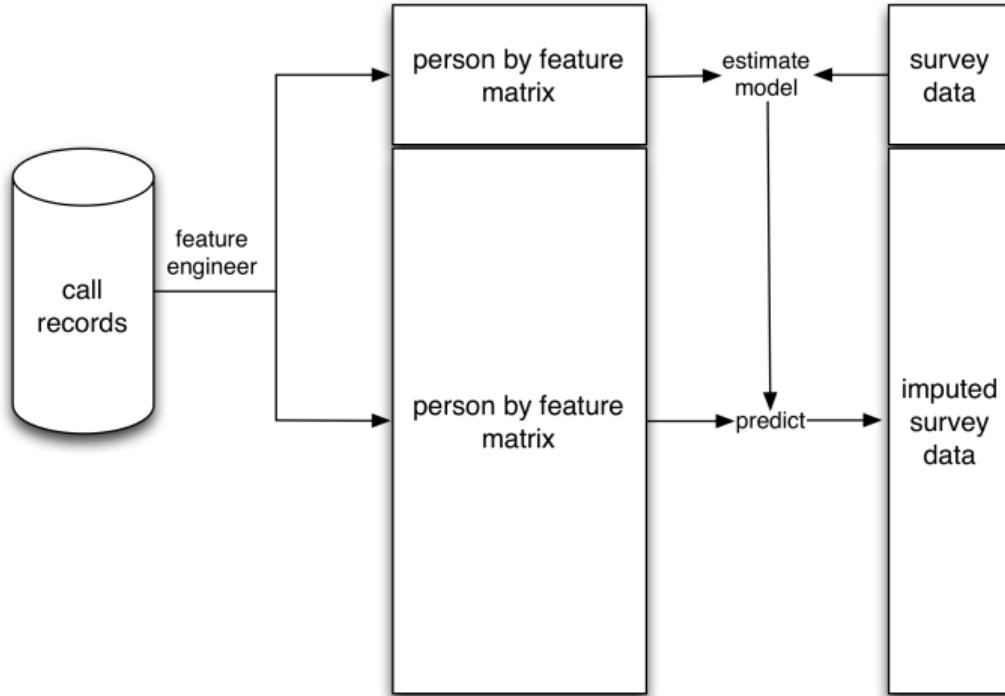
call  
records

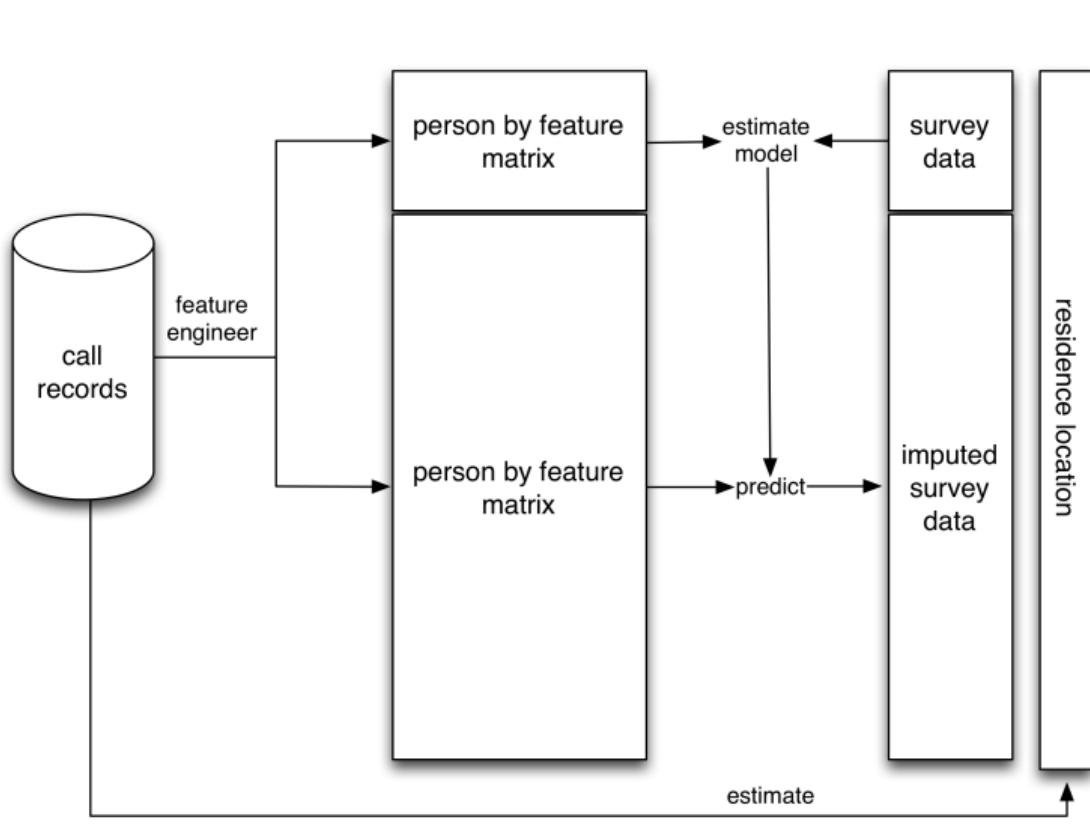


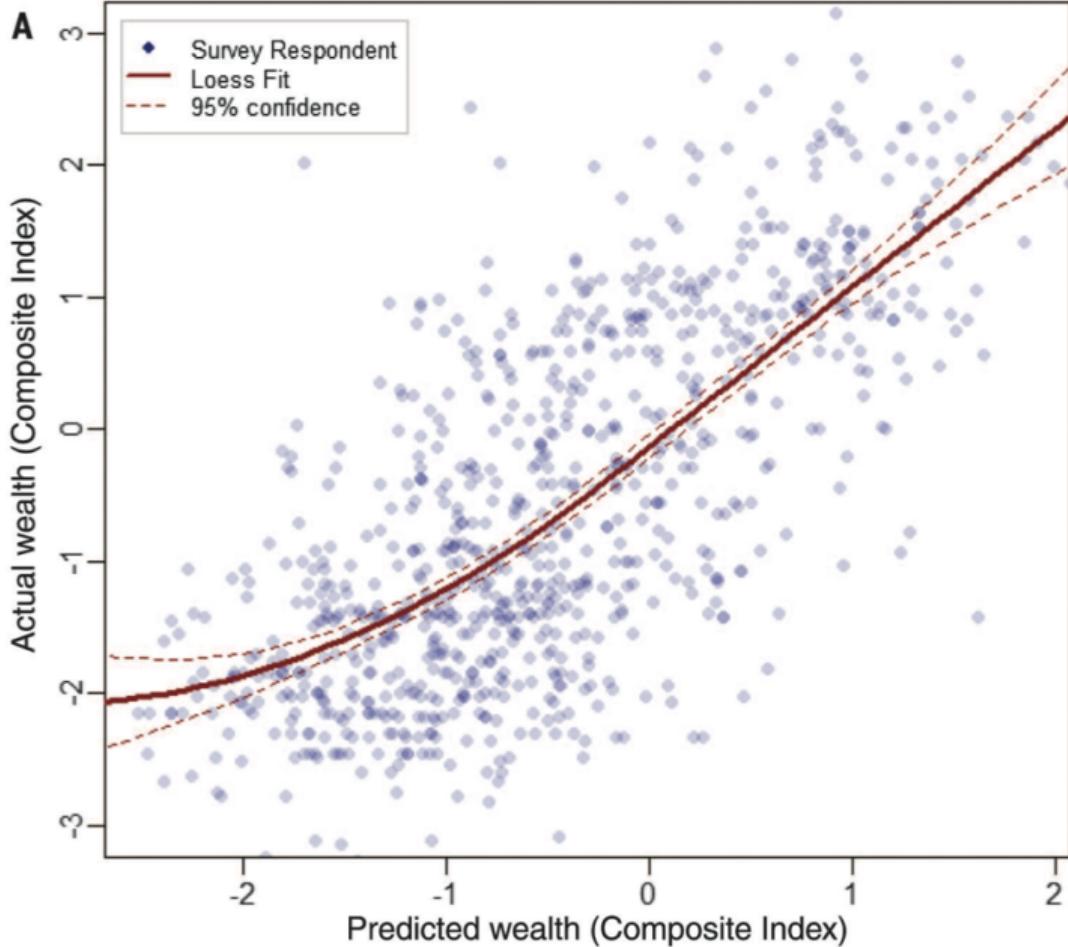
survey  
data

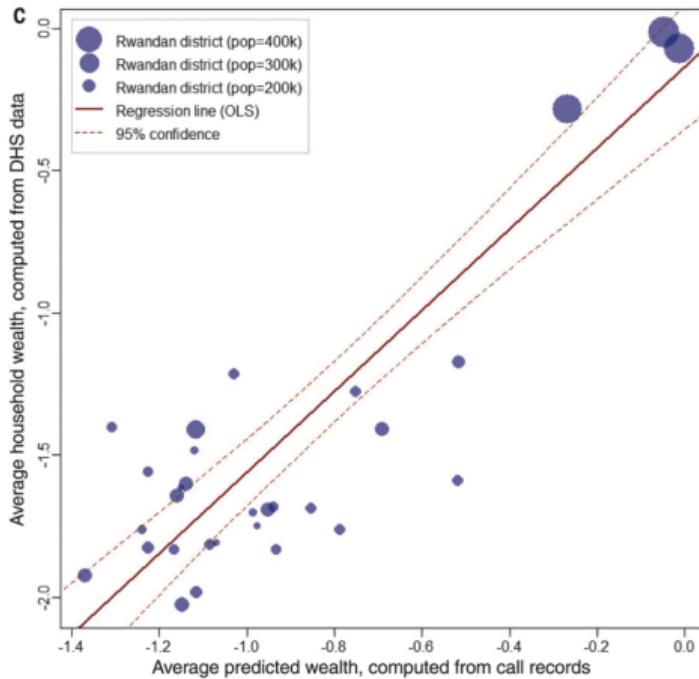


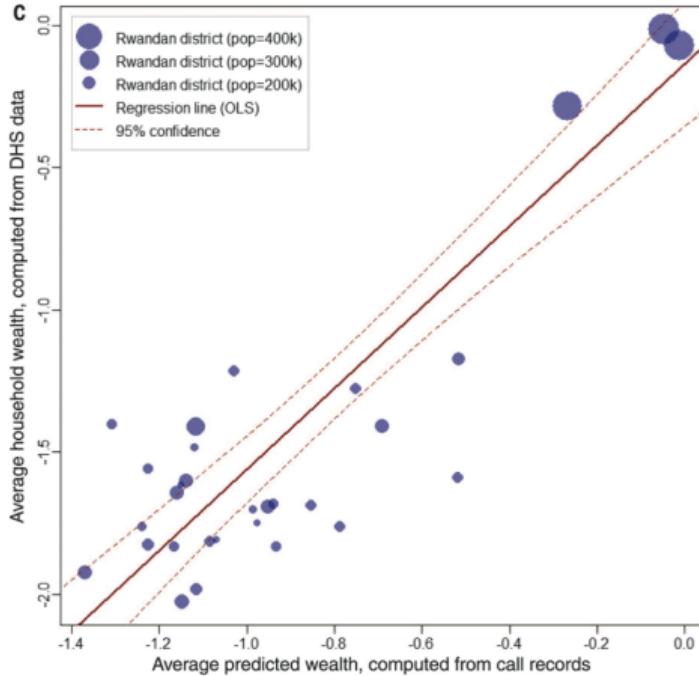












- ▶ 10 times faster
- ▶ 50 times cheaper



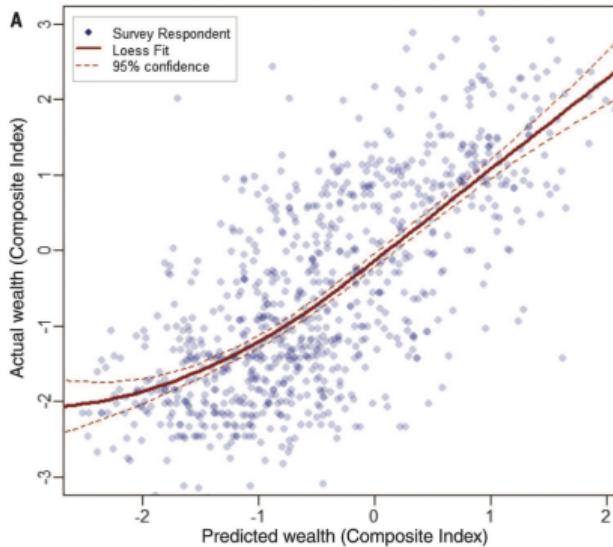
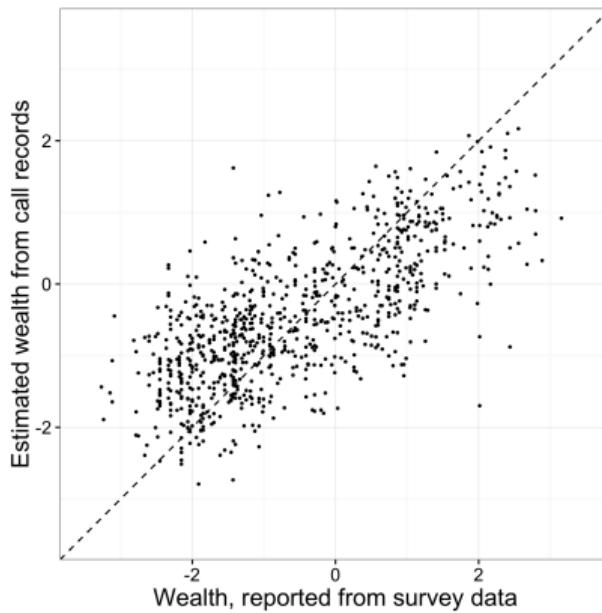
Readymades

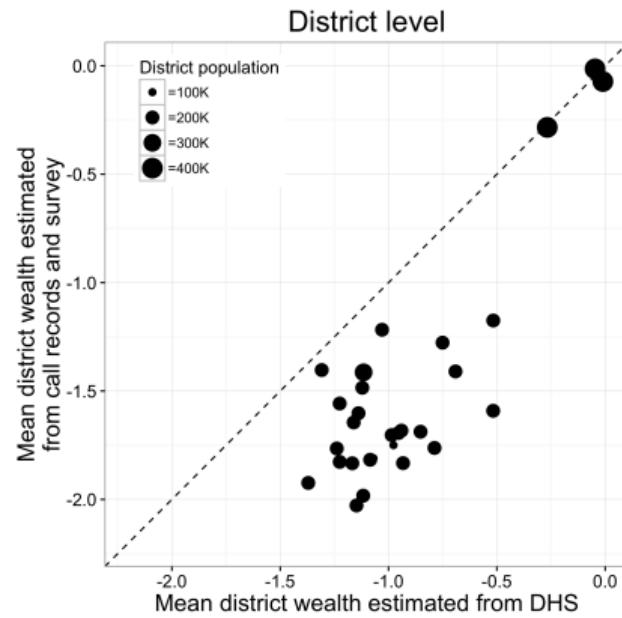
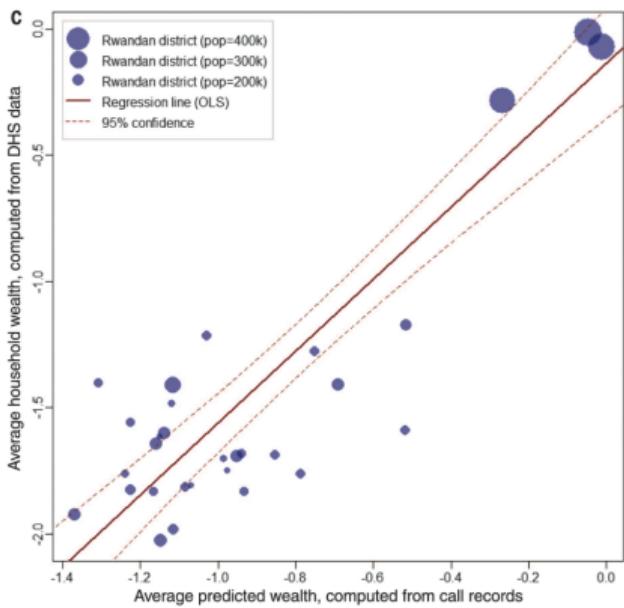
+



Custommades

## Individual level





# Predicting poverty and wealth from mobile phone metadata

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

## 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 Customer Defection and Loyalty

Muhammad Raza Khan<sup>1</sup>, Joshua Manoj<sup>2</sup>, Aniket Singh<sup>3</sup>, Joshua Blumenstock<sup>4</sup>  
<sup>1</sup>Information School, University of Washington, Seattle, WA, USA  
Email: [mraza@uw.edu](mailto:mraza@uw.edu), [joshuam@uw.edu](mailto:joshuam@uw.edu), [aniket@uw.edu](mailto:aniket@uw.edu), [joshblum@uw.edu](mailto:joshblum@uw.edu)

# Predicting poverty and wealth from mobile phone metadata

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

## 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](mailto:joshblum@uw.edu)

# Predicting poverty and wealth from mobile phone metadata

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

## 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](mailto:joshblum@uw.edu)

- ▶ better feature engineering
- ▶ new outcome variable
- ▶ out-of-sample test

## **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](mailto:joshblum@uw.edu)

## **Behavioral Modeling for Churn Prediction:**

Early Indicators and Accurate Predictors of Customer Defection and Loyalty

Muhammad Raza Khan<sup>1</sup>, Joshua Manoj<sup>2</sup>, Anikate Singh<sup>3</sup>, Joshua Blumenstock<sup>4</sup>

<sup>1</sup>*Information School, University of Washington, Seattle, WA, USA*

Email: <sup>1</sup>[mraza@uw.edu](mailto:mraza@uw.edu), <sup>2</sup>[joshuacm@uw.edu](mailto:joshuacm@uw.edu), <sup>3</sup>[aniksing@uw.edu](mailto:aniksing@uw.edu), <sup>4</sup>[joshblum@uw.edu](mailto: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](mailto:joshblum@uw.edu)

## **Behavioral Modeling for Churn Prediction:**

Early Indicators and Accurate Predictors of Customer Defection and Loyalty

Muhammad Raza Khan<sup>1</sup>, Joshua Manoj<sup>2</sup>, Anikate Singh<sup>3</sup>, Joshua Blumenstock<sup>4</sup>

<sup>1</sup>*Information School, University of Washington, Seattle, WA, USA*

Email: <sup>1</sup>[mraza@uw.edu](mailto:mraza@uw.edu), <sup>2</sup>[joshuacm@uw.edu](mailto:joshuacm@uw.edu), <sup>3</sup>[aniksing@uw.edu](mailto:aniksing@uw.edu), <sup>4</sup>[joshblum@uw.edu](mailto:joshblum@uw.edu)

- ▶ outcome in data vs collected in a survey

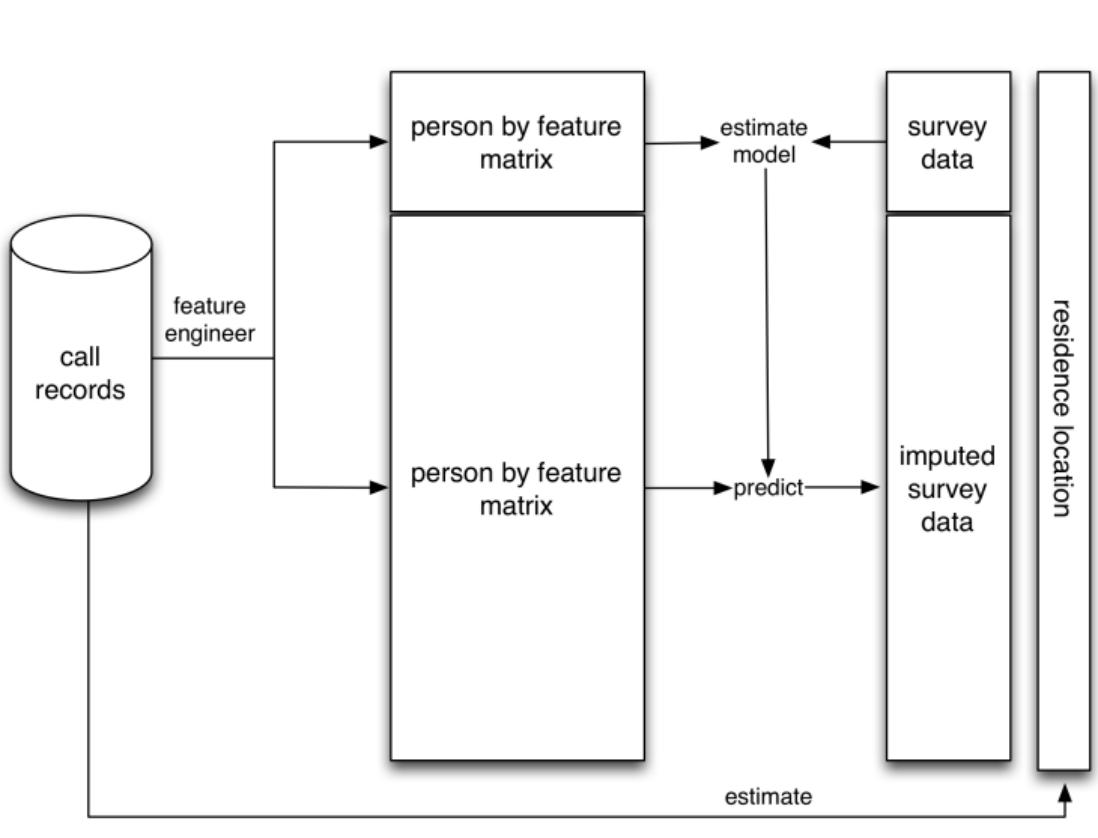
# Combining satellite imagery and machine learning to predict poverty

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>

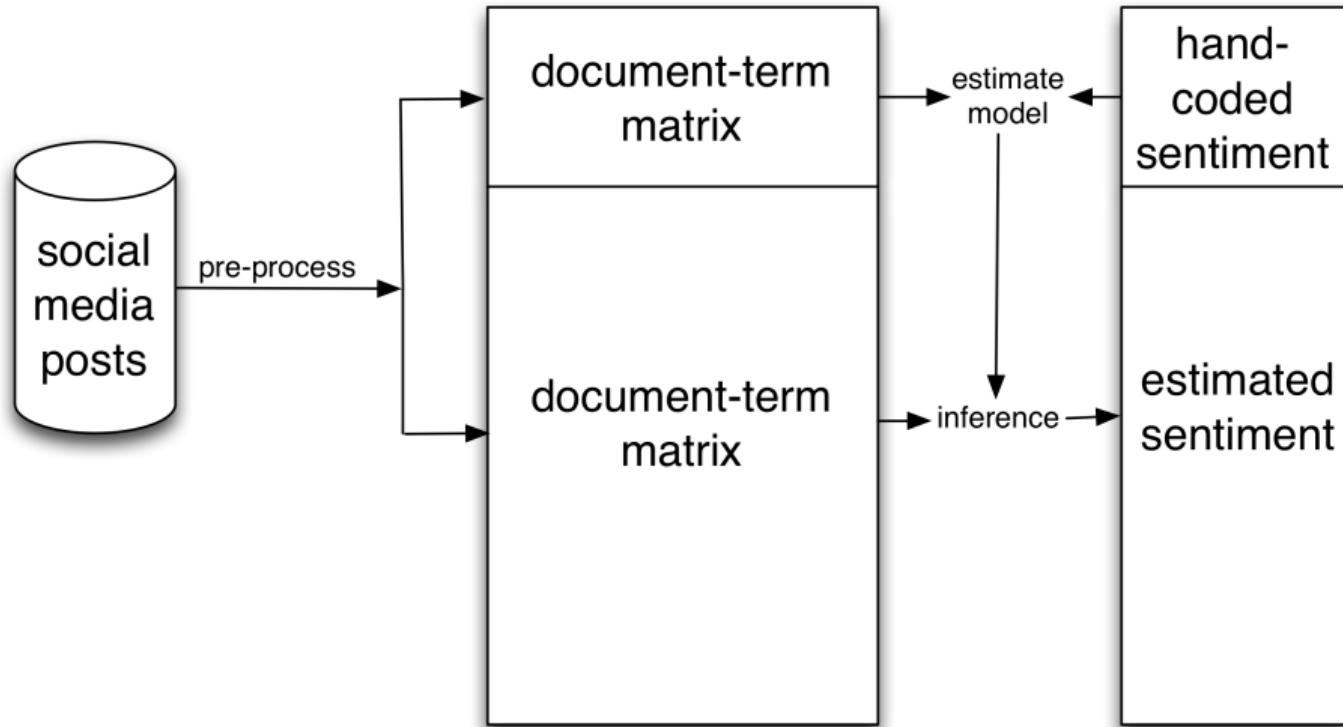
This paper is amazing and surprising. First a digression . . . .

Supervised learning:

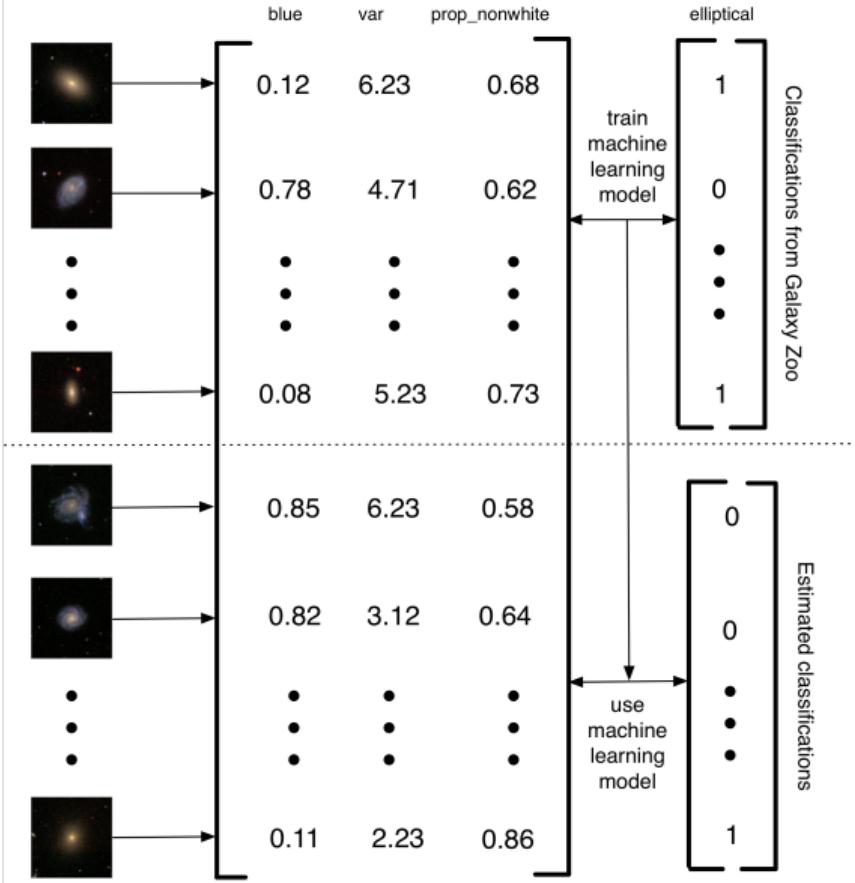
Lots of input-output pairs; goal is to develop a function that will predict the output from the input



See Chapter 3 of Salganik (2016)



See Chapter 2 of Salganik (2016)



See Chapter 5 of Salganik (2016)

Supervised learning:

Lots of input-output pairs; goal is to develop a function that will predict the output from the input

What if rather than engineering the features you could “learn” them automatically?

# Deep learning

Yann LeCun<sup>1,2</sup>, Yoshua Bengio<sup>3</sup> & Geoffrey Hinton<sup>4,5</sup>

<http://dx.doi.org/10.1038/nature14539>

“Have you tried deep learning?”

Deep learning + poverty + space

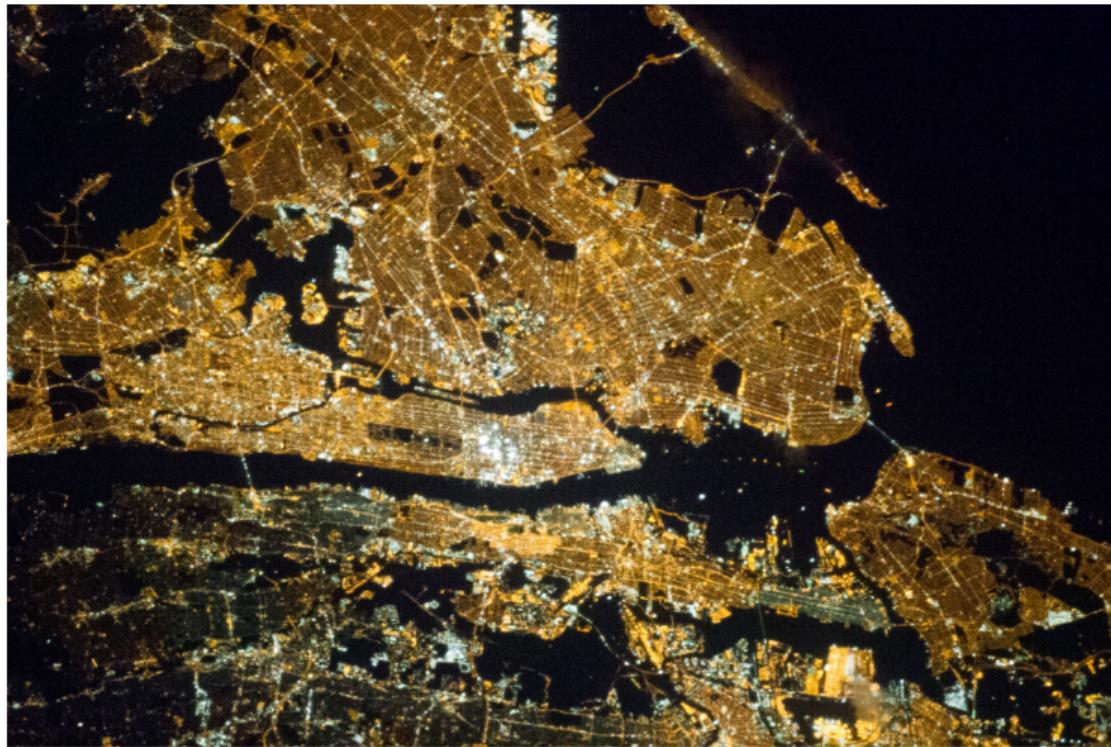
# Combining satellite imagery and machine learning to predict poverty

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>3</sup>

Live demo:

<https://www.google.com/maps/place/Kigali,+Rwanda/@-1.9546259,30.0345059,26517m/data=!3m2!1e3!4b1!4m5!3m4!1s0x19dca4258ed8e797:0xf32b36a5411d0bc8!8m2!3d-1.9705786!4d30.1044288>

But, most people 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)

## Nightlights + survey data to estimate wealth in places without surveys

Jean et al. (2016):  
Day pictures + Nightlights + survey data to estimate wealth in places without surveys

## Predicting poverty

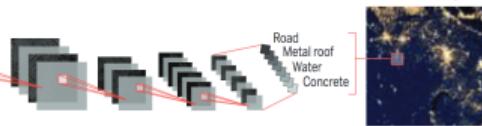
Satellite images can be used to estimate wealth in remote regions.

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

Daytime satellite photos capture details of the landscape



Convolutional Neural Network (CNN) associates features from daytime photos with nightlight intensity

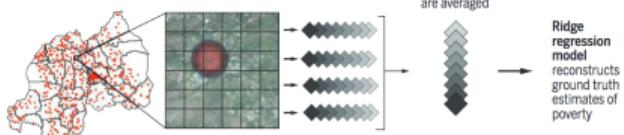


Satellite nightlights are a proxy for economic activity

Daytime satellite images can be used to predict regional wealth

Household survey locations

CNN processes satellite photos of each survey site



- ▶ Start with CNN pretrained on ImageNet (e.g. hamsters and weasels)

## Predicting poverty

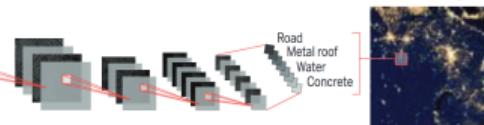
Satellite images can be used to estimate wealth in remote regions.

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

Daytime satellite photos capture details of the landscape



Convolutional Neural Network (CNN) associates features from daytime photos with nightlight intensity



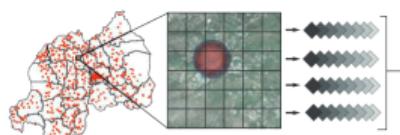
Satellite nightlights are a proxy for economic activity



Daytime satellite images can be used to predict regional wealth

Household survey locations

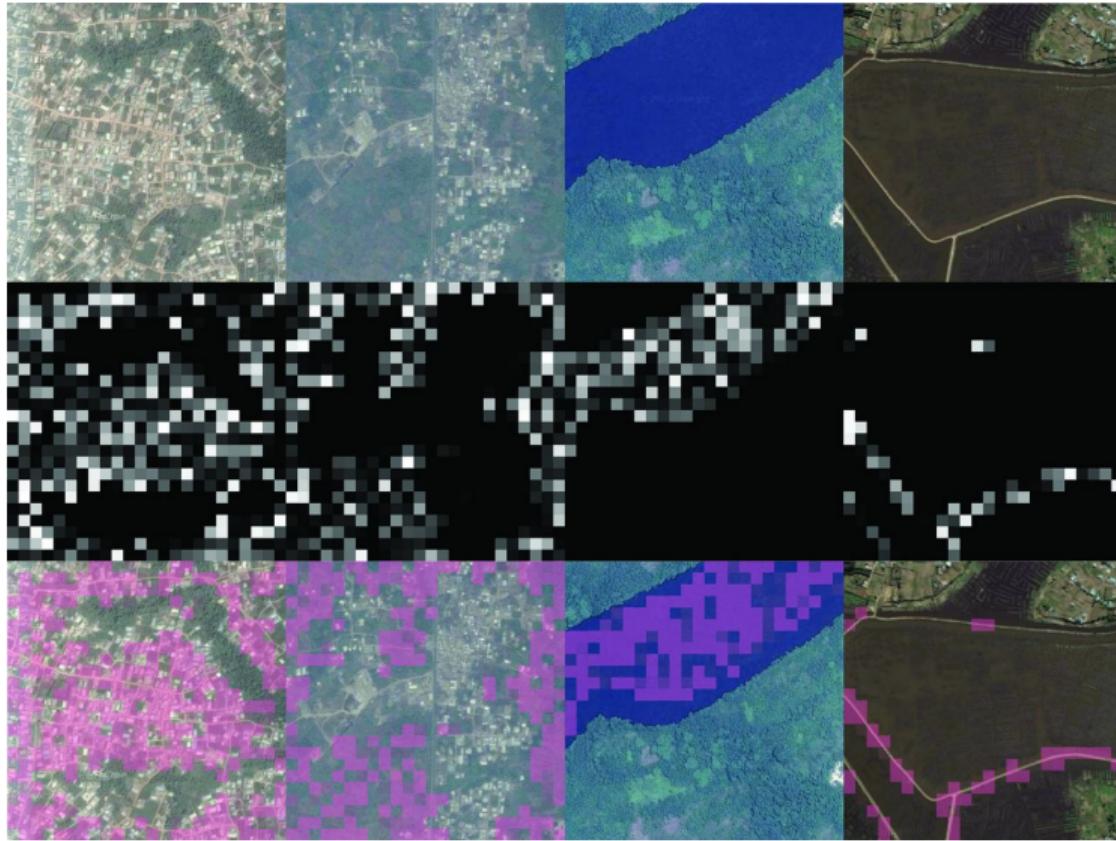
CNN processes satellite photos of each survey site

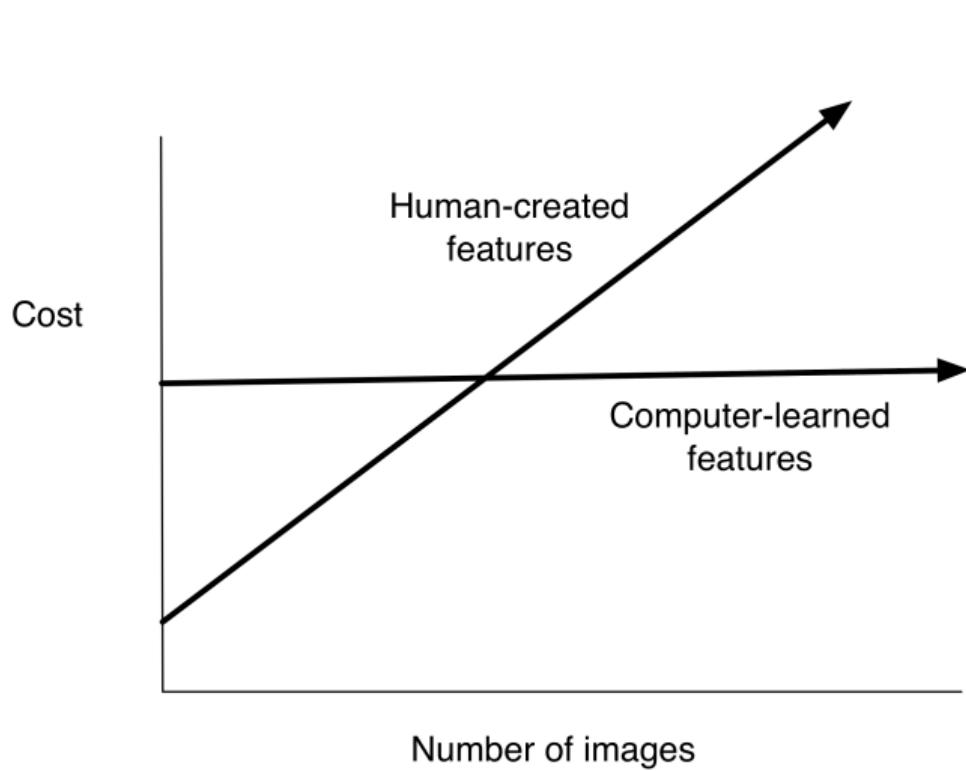


Features from multiple photos are averaged

Ridge regression model reconstructs ground truth estimates of poverty

- ▶ Start with CNN pretrained on ImageNet (e.g. hamsters and weasels)
- ▶ Train CNN to predict nightlights from day pictures (lots of training data)

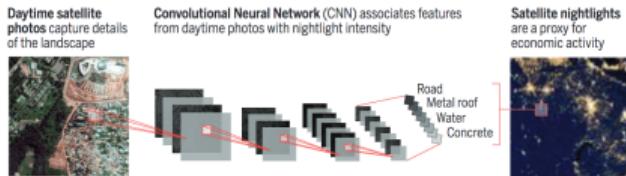




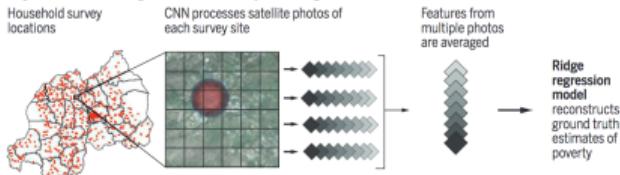
## Predicting poverty

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

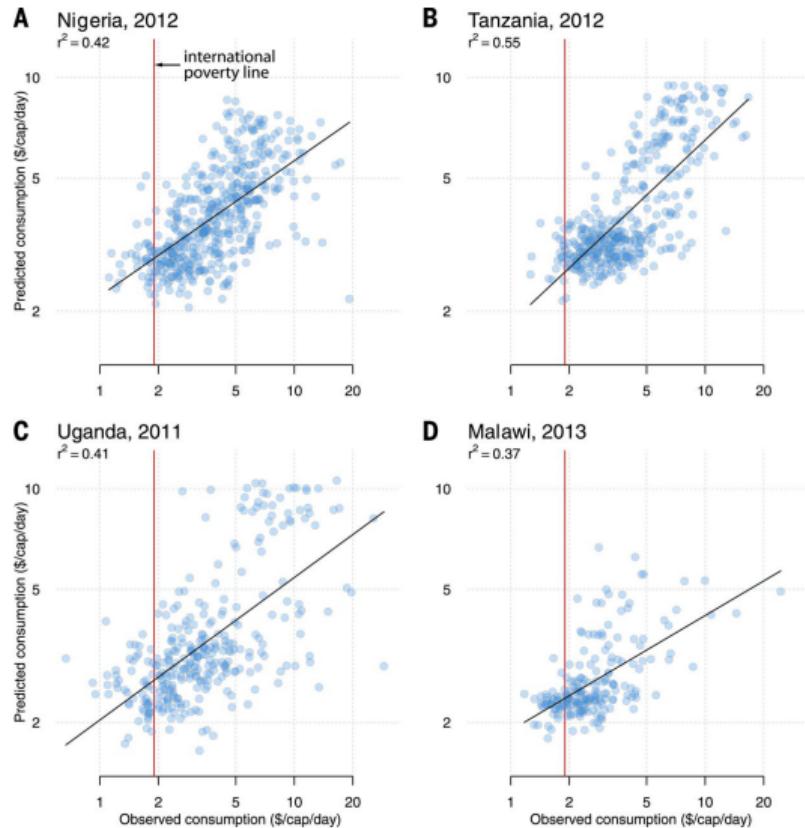
Neural network learns features in satellite images that correlate with economic activity



Daytime satellite images can be used to predict regional wealth

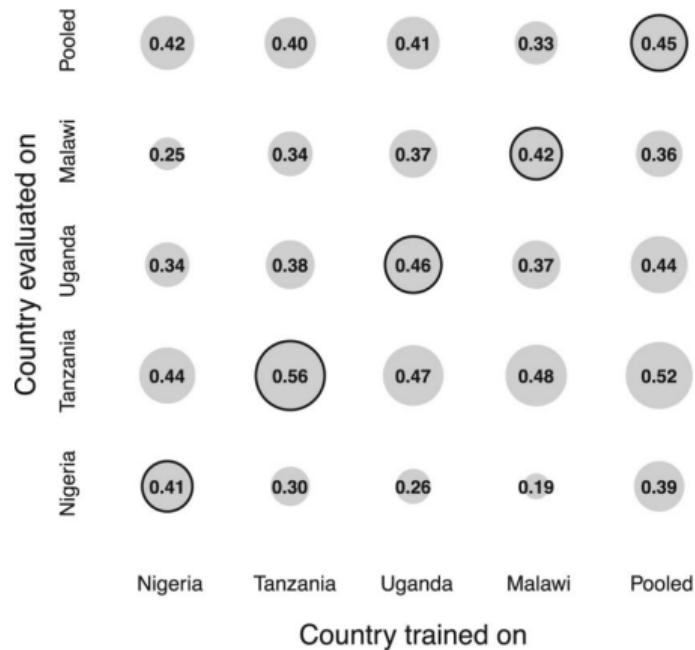


- ▶ Start with CNN pretrained on ImageNet (e.g. hamsters and weasels)
- ▶ Train CNN to predict nightlights from day pictures (lots of training data)
- ▶ Take features from CNN and train ridge regression to predict cluster mean survey response

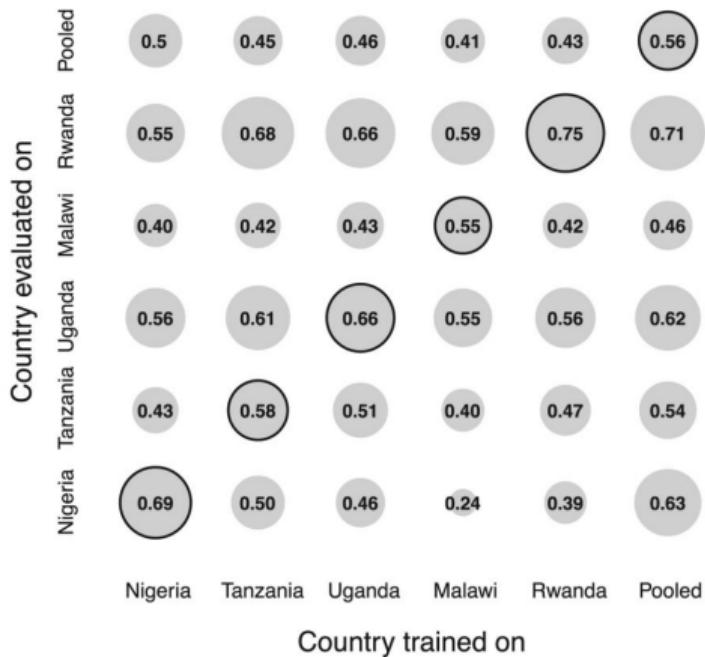


How could this figure be improved?

## A Consumption expenditures



## B Assets



# Predicting poverty and wealth from mobile phone metadata

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

Readymade + Custommade

# 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>

Readymade + Readymade + (Quasi)  
Custommade

Both Blumenstock et al (2015) and Jean et al (2016) estimate poverty in Africa?  
Compare the strengths and weaknesses of the approaches.

## 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>

Uses Readymade linked to  
Researcher-Custommade

## Detecting influenza epidemics using search engine query data

Jeremy Ginsberg<sup>1</sup>, Matthew H. Mohebbi<sup>1</sup>, Rajan S. Patel<sup>1</sup>, Lynnette Brammer<sup>2</sup>, Mark S. Smolinski<sup>1</sup> & Larry Brilliant<sup>1</sup>

Uses Readymade linked to  
Researcher-Custommade

Is there a role for individual researchers collecting data in the age of the Readymades and Quasi Custommades?

How should governments collect large general purpose datasets if they will be combined with Readymades?