

Combining surveys and big data

Matthew J. Salganik
Department of Sociology
Princeton University

Summer Institutes in Computational Social Science
2020

The Summer Institutes in Computational Social Science is supported by grants from the Russell Sage Foundation and the Alfred P. Sloan Foundation.



	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?

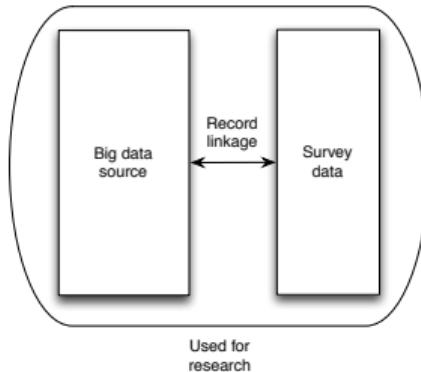


<http://schlitterblog.com/wp-content/uploads/2014/05/peanutbutterlover.jpg>

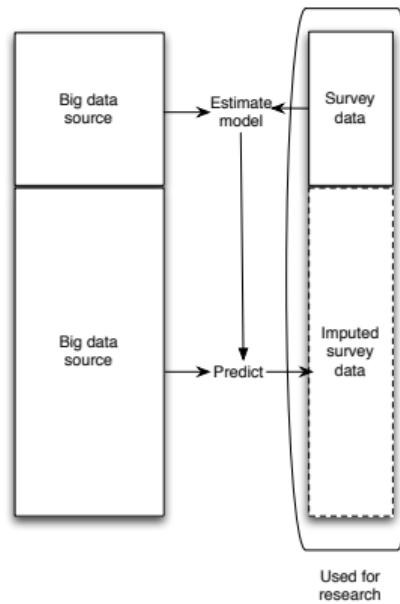


<http://schlitterblog.com/wp-content/uploads/2014/05/peanutbutterlover.jpg>

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,^{1*} Gabriel Cadamuro,² Robert On³

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

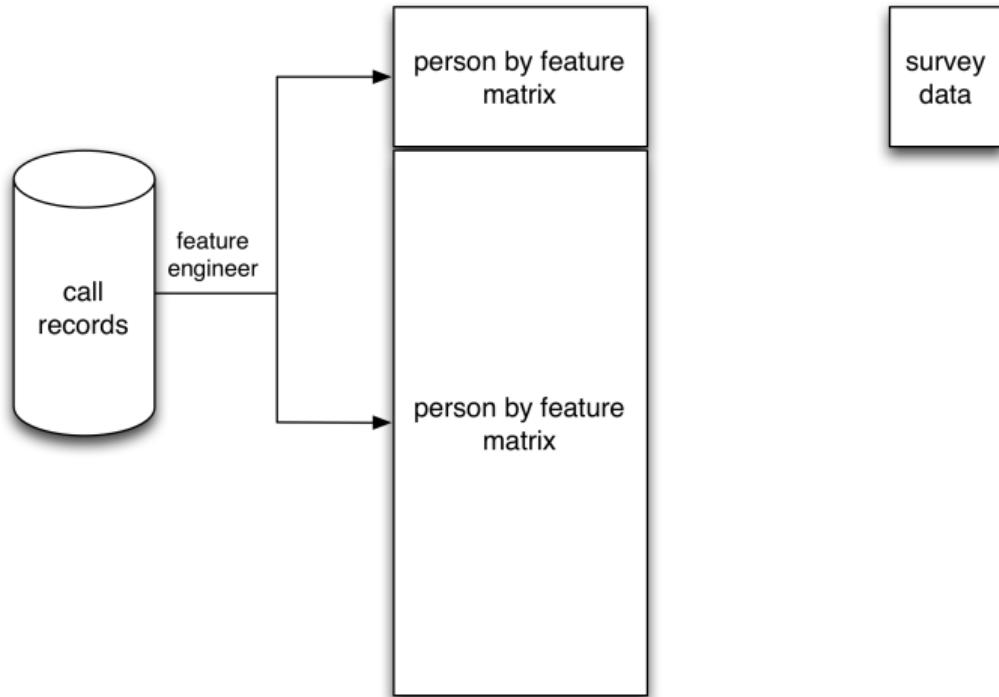


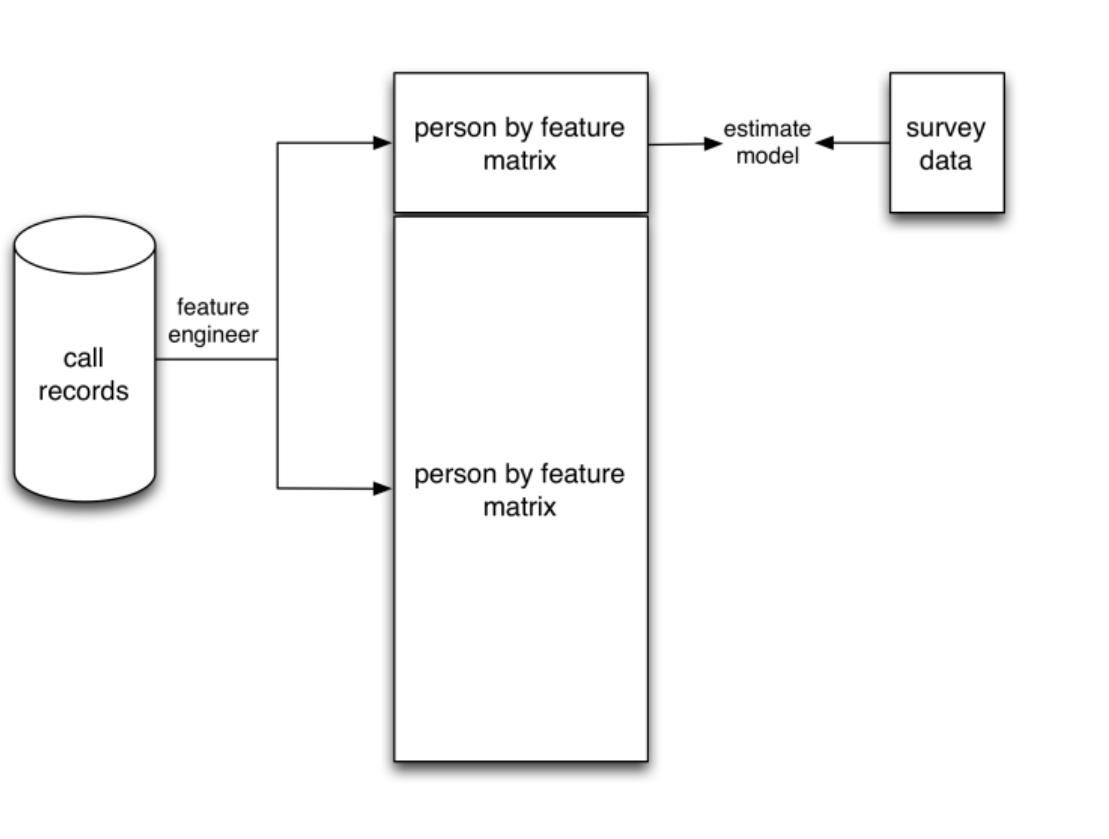


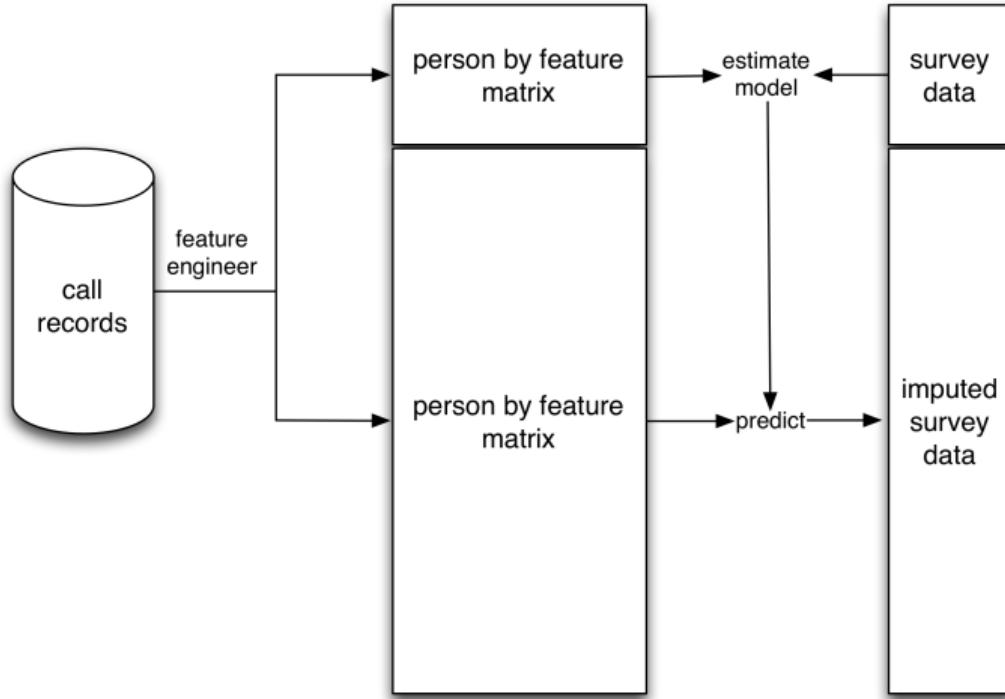
call
records

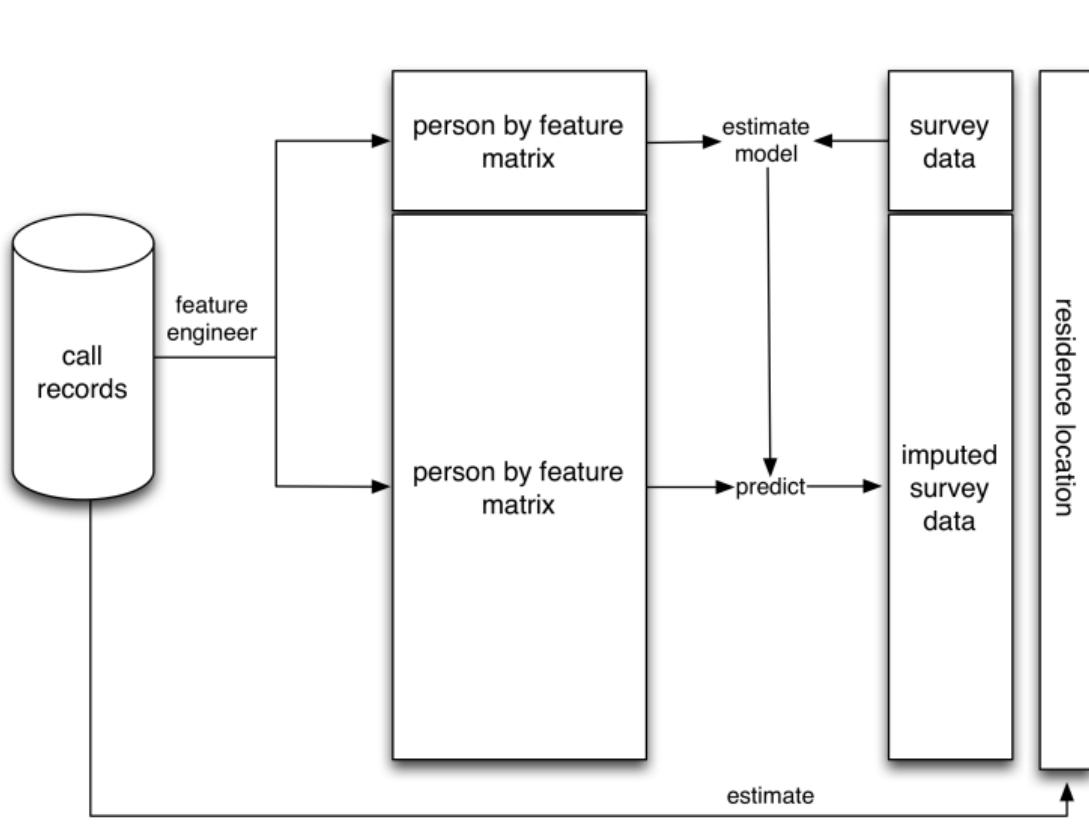


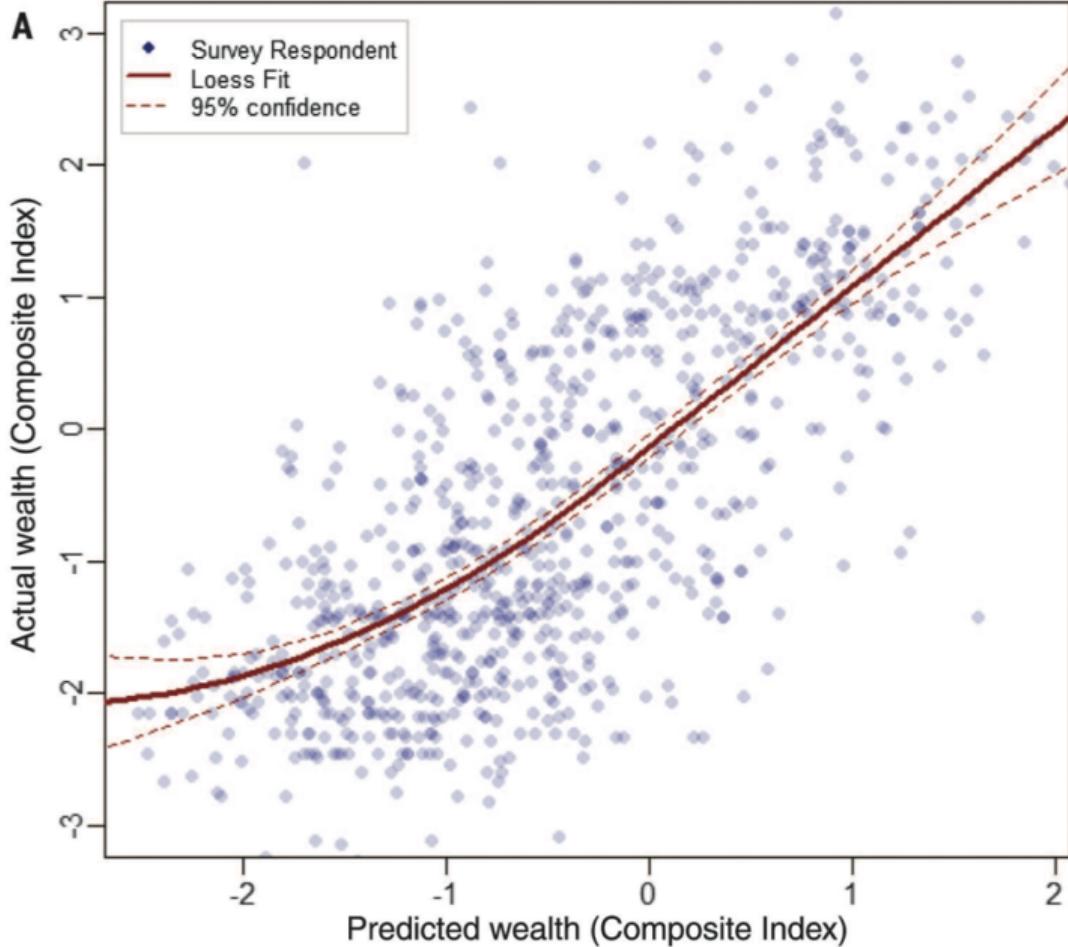
survey
data

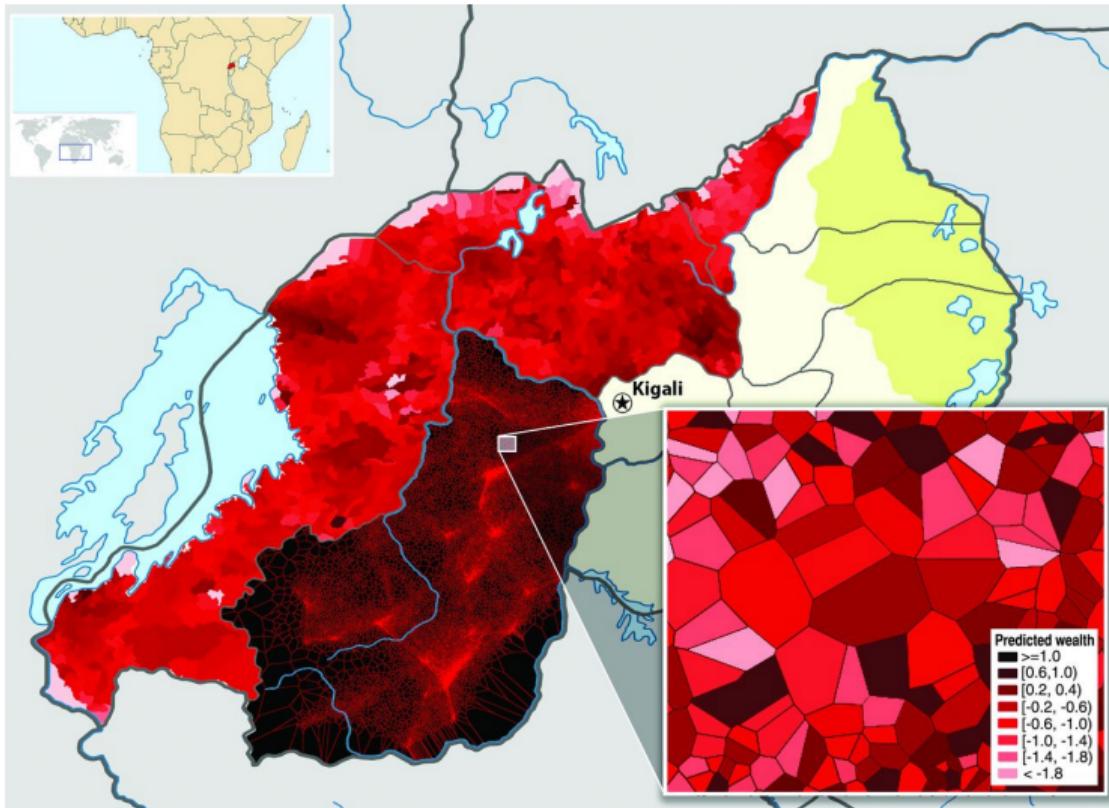


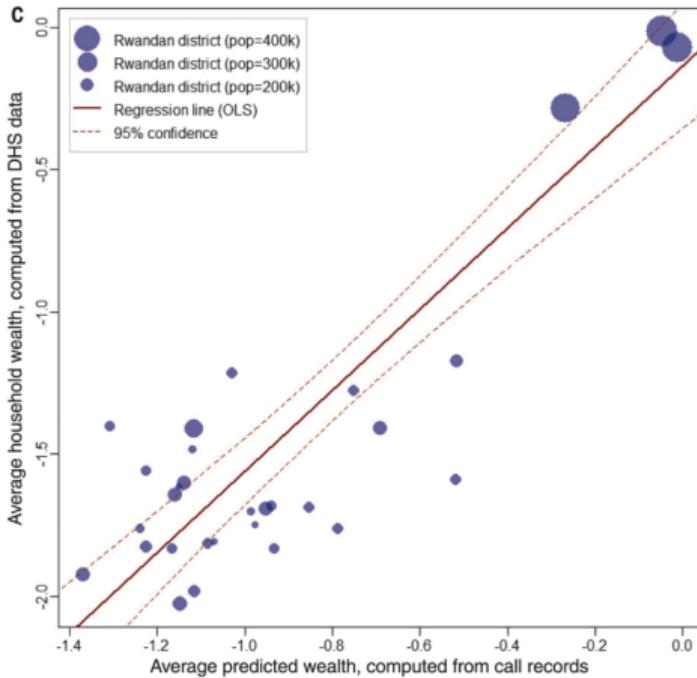


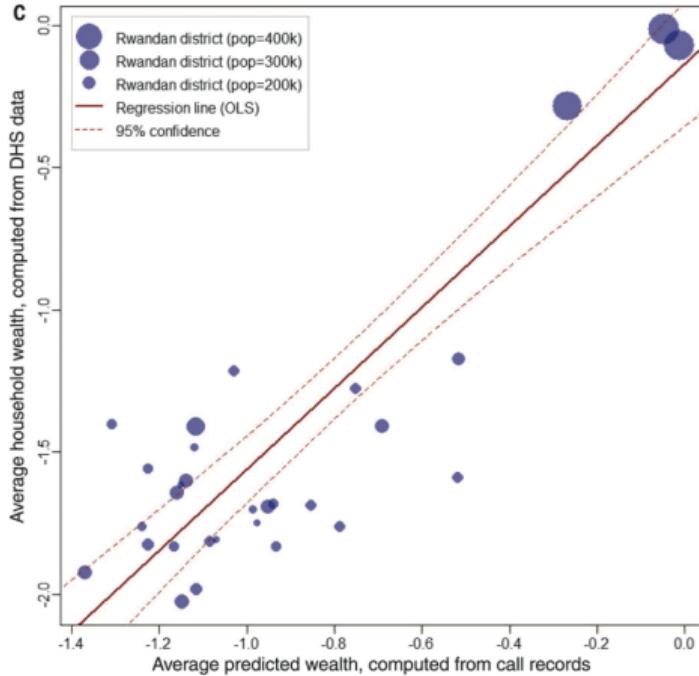












- ▶ 10 times faster
- ▶ 50 times cheaper



Readymades

+



Custommades



Behavioral Modeling for Churn Prediction:

Early Indicators and Accurate Predictors of Customer Defection and Loyalty

Muhammad Raza Khan¹, Joshua Mano², Ankikate Singh³, Joshua Blumentrock⁴

¹Information School, University of Washington, Seattle, WA, USA

Emails: mraza@uw.edu, joshua@uw.edu, ankikating@uw.edu, joshbl@uw.edu

Behavioral Modeling for Churn Prediction:

Early Indicators and Accurate Predictors of Custom Defection and Loyalty

Muhammad Raza Khan¹, Joshua Mano², Ankikate Singh³, Joshua Blumenstock⁴

¹Information School, University of Washington, Seattle, WA, USA

Emails: mraza@uw.edu, joshua@uw.edu, ankikating@uw.edu, joshbl@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

joshbl@uw.edu

Behavioral Modeling for Churn Prediction:

Early Indicators and Accurate Predictors of Custom Defection and Loyalty

Muhammad Raza Khan¹, Joshua Mano², Ankikate Singh³, Joshua Blumenstock⁴

¹Information School, University of Washington, Seattle, WA, USA

Emails: mraza@uw.edu, joshua.mano@uw.edu, ankikating@uw.edu, joshbl@u.washington.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
joshbl@uw.edu

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,^{1,*} Gabriel Cadamuro,² Robert On³

The beginning is not the end

Behavioral Modeling for Churn Prediction:

Early Indicators and Accurate Predictors of Custom Defection and Loyalty

Muhammad Raza Khan¹, Joshua Mano², Ankikate Singh³, Joshua Blumenstock⁴

¹Information School, University of Washington, Seattle, WA, USA

Emails: mraza@uw.edu, joshua@uw.edu, ankikating@uw.edu, joshbl@u.washington.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
joshbl@uw.edu

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,^{1,*} Gabriel Cadamuro,² Robert On³

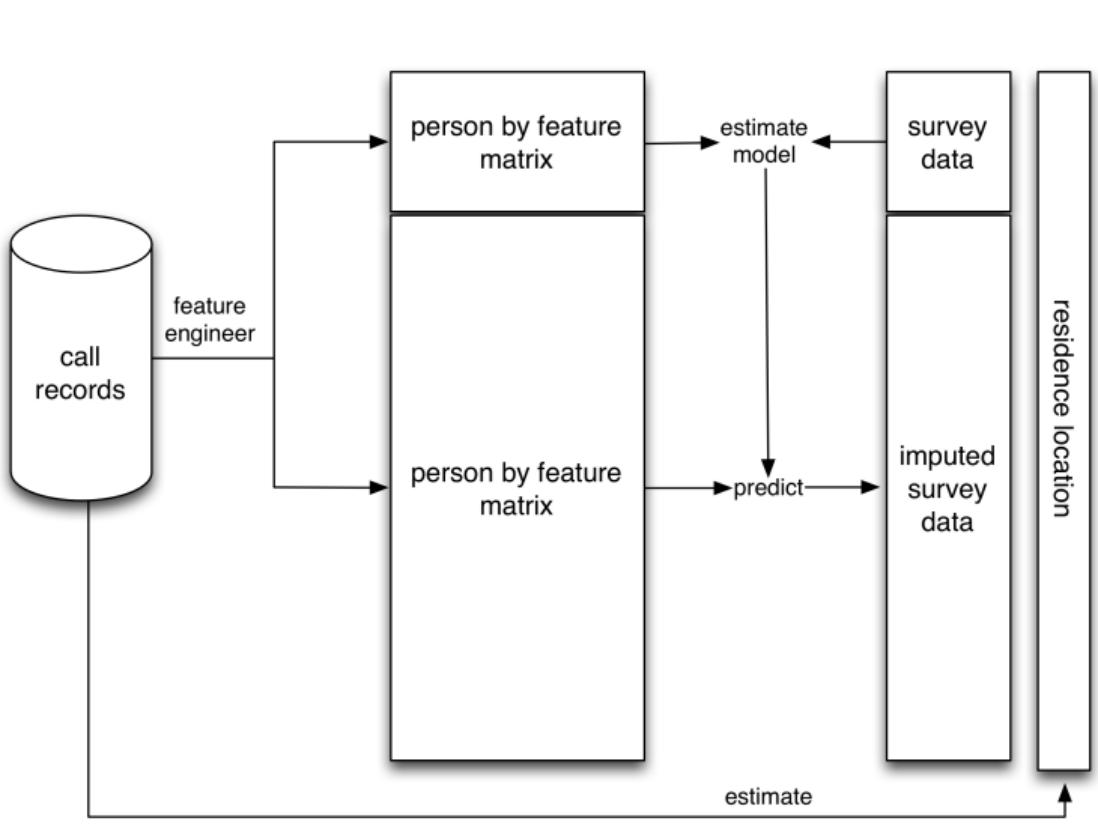
Combining satellite imagery and machine learning to predict poverty

Neal Jean,^{1,2*} Marshall Burke,^{3,4,5*} Michael Xie,¹ W. Matthew Davis,⁴
David B. Lobell,^{3,4} Stefano Ermon¹

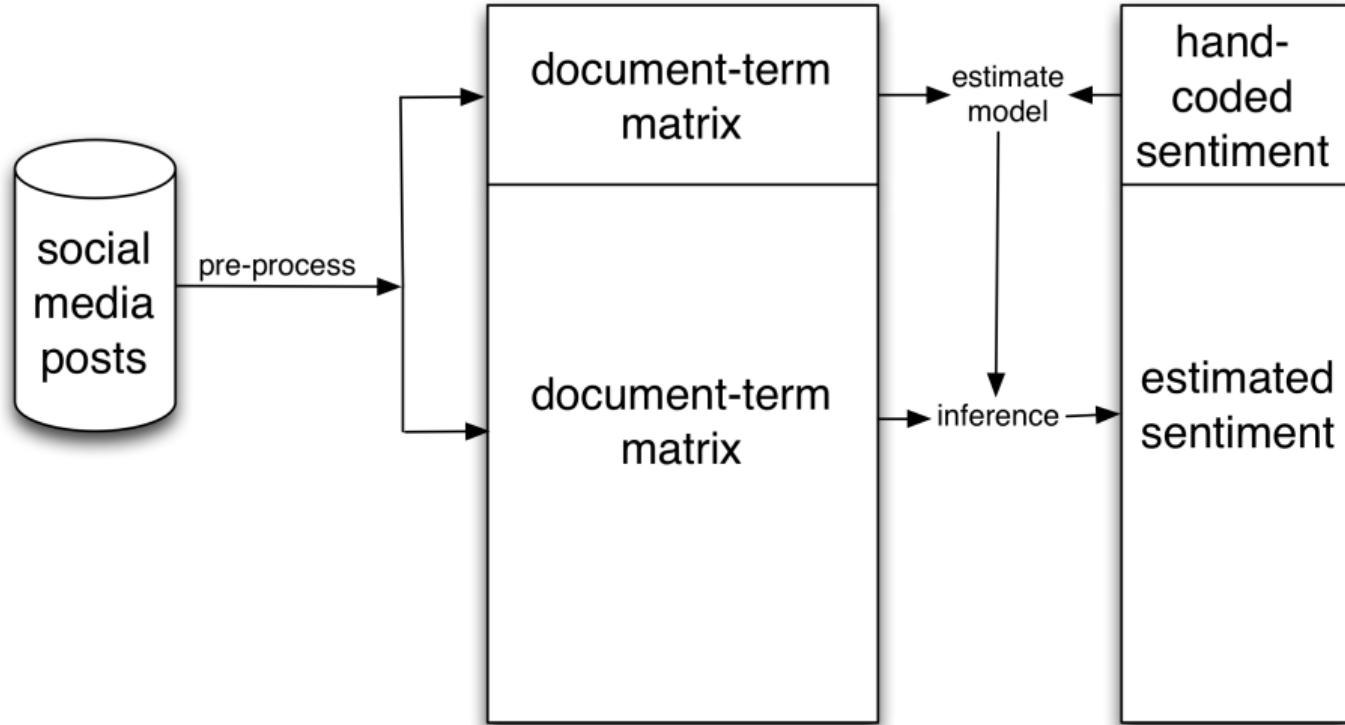
This paper is amazing and surprising (to me). First a digression

Supervised learning:

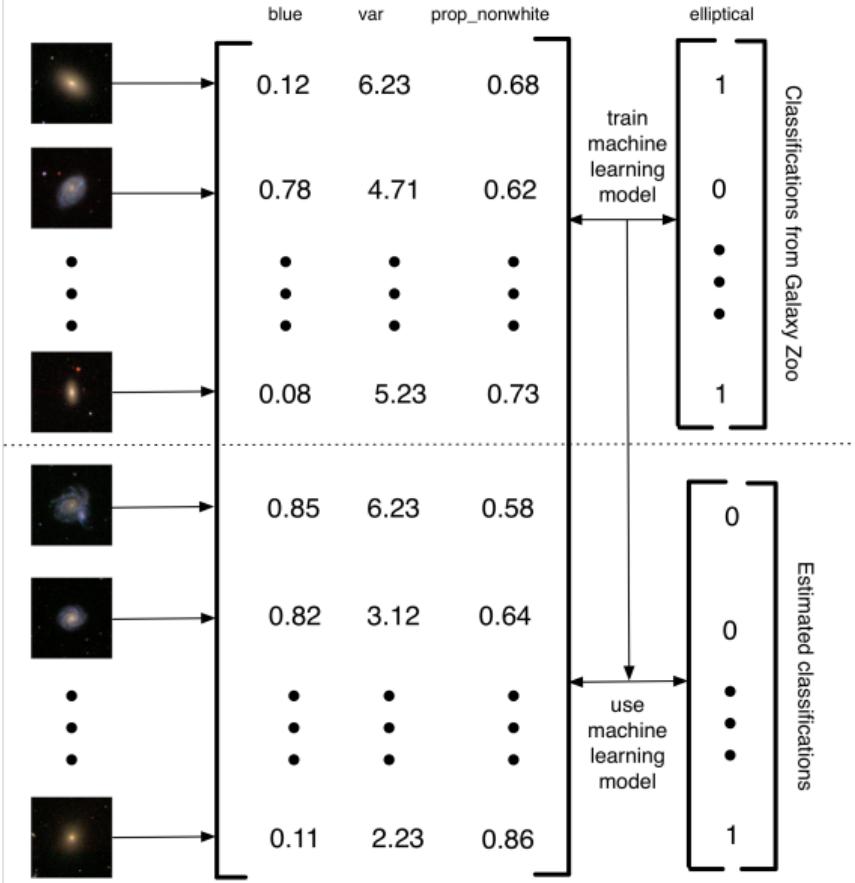
Lots of input-output pairs; goal is to develop a function that will predict the output from the input



See Chapter 3 of Salganik (2018)



See Chapter 2 of Salganik (2018)



See Chapter 5 of Salganik (2018)

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^{1,2}, Yoshua Bengio³ & Geoffrey Hinton^{4,5}

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

Combining satellite imagery and machine learning to predict poverty

Neal Jean,^{1,2*} Marshall Burke,^{3,4,5*}† Michael Xie,¹ W. Matthew Davis,⁴
David B. Lobell,^{3,4} Stefano Ermon¹

Combining satellite imagery and machine learning to predict poverty

Neal Jean,^{1,2*} Marshall Burke,^{3,4,5*}† Michael Xie,¹ W. Matthew Davis,⁴
David B. Lobell,^{3,4} Stefano Ermon¹

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

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

Prior research:

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

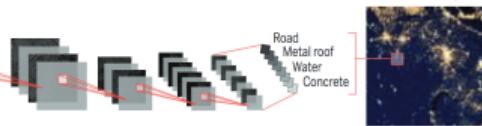
Satellite images can be used to estimate wealth in remote regions.

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

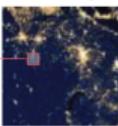
Daytime satellite photos capture details of the landscape



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

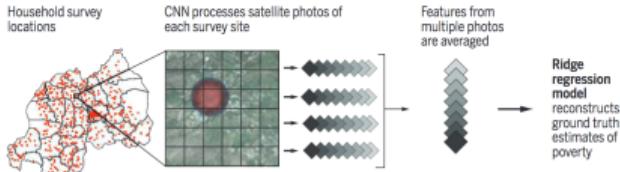


Satellite nightlights are a proxy for economic activity



Daytime satellite images can be used to predict regional wealth

Household survey locations



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)

Predicting poverty

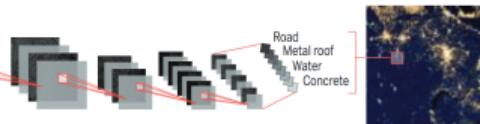
Satellite images can be used to estimate wealth in remote regions.

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

Daytime satellite photos capture details of the landscape



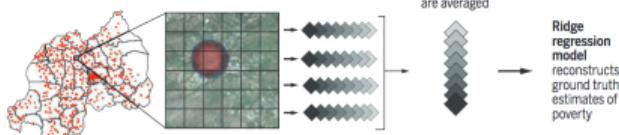
Convolutional Neural Network (CNN) associates features from daytime photos with nightlight intensity



Daytime satellite images can be used to predict regional wealth

Household survey locations

CNN processes satellite photos of each survey site

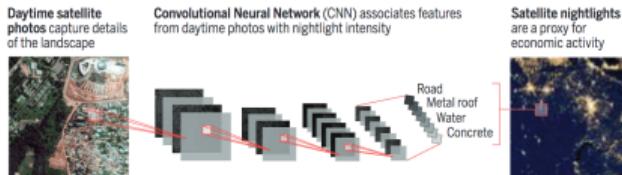


- ▶ Start with CNN pretrained on ImageNet (e.g. hamsters and weasels)
- ▶ Train CNN to predict nightlights from day pictures (lots of training data)

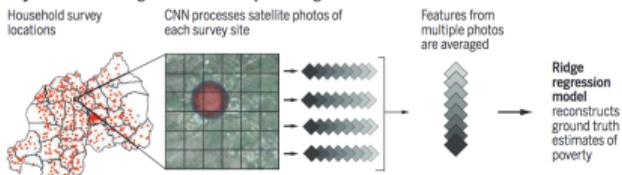
Predicting poverty

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

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

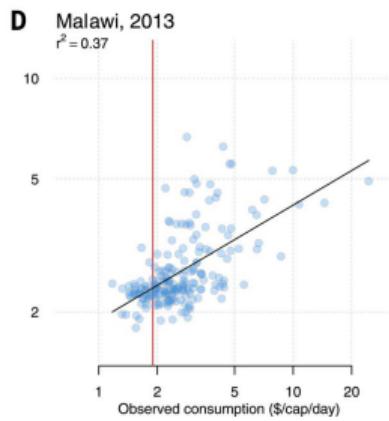
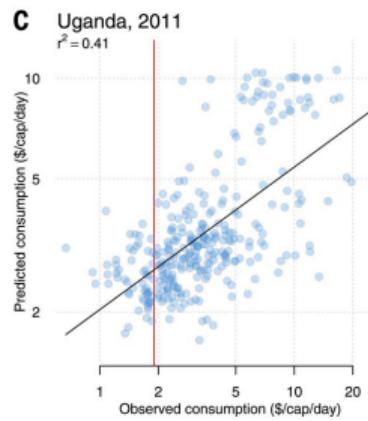
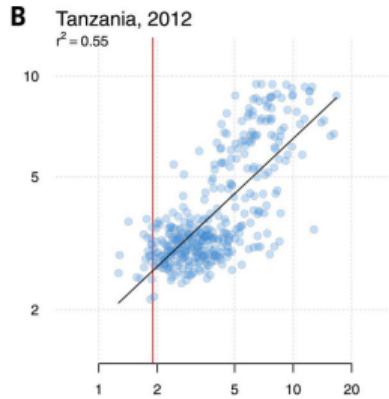
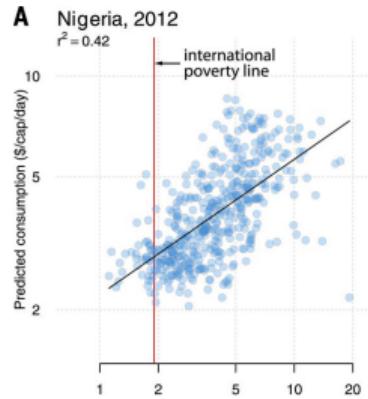


Daytime satellite images can be used to predict regional wealth

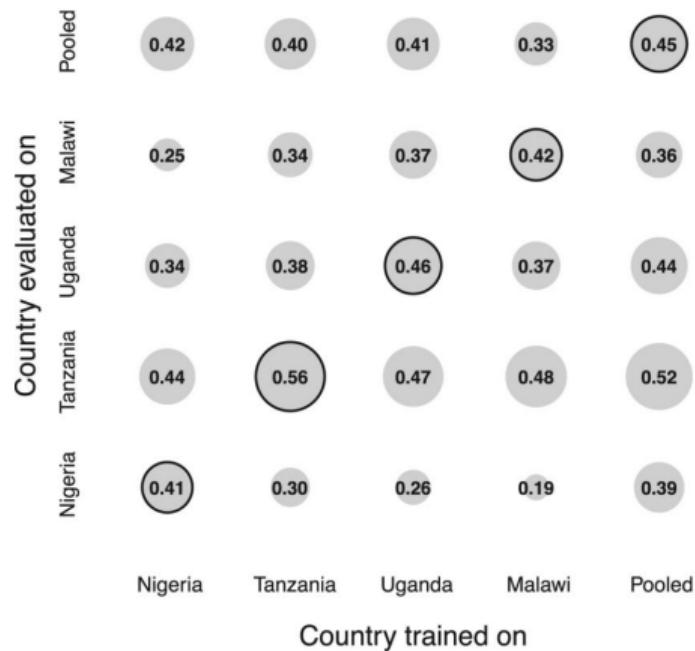


- ▶ Start with CNN pretrained on ImageNet (e.g. hamsters and weasels)
- ▶ Train CNN to predict nightlights from day pictures (lots of training data)
- ▶ Take features from CNN and train ridge regression to predict cluster mean survey response

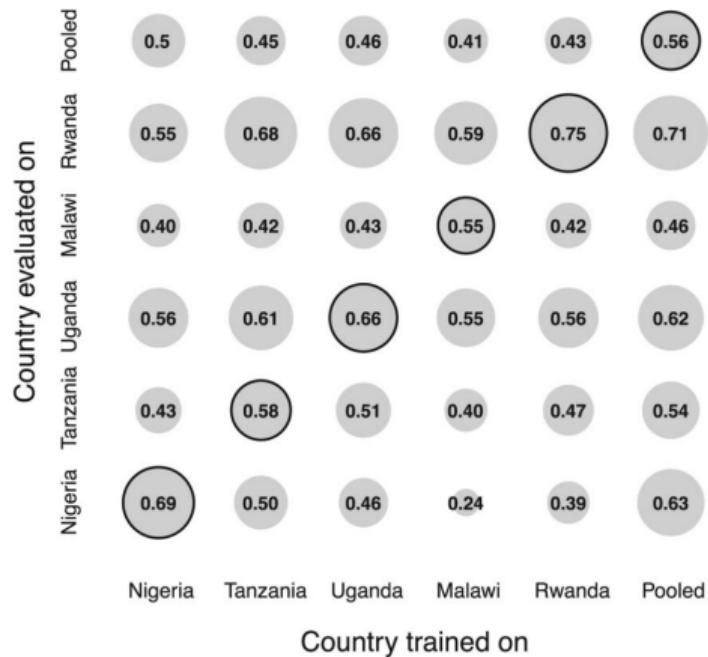
<http://dx.doi.org/10.1126/science.aah5217>

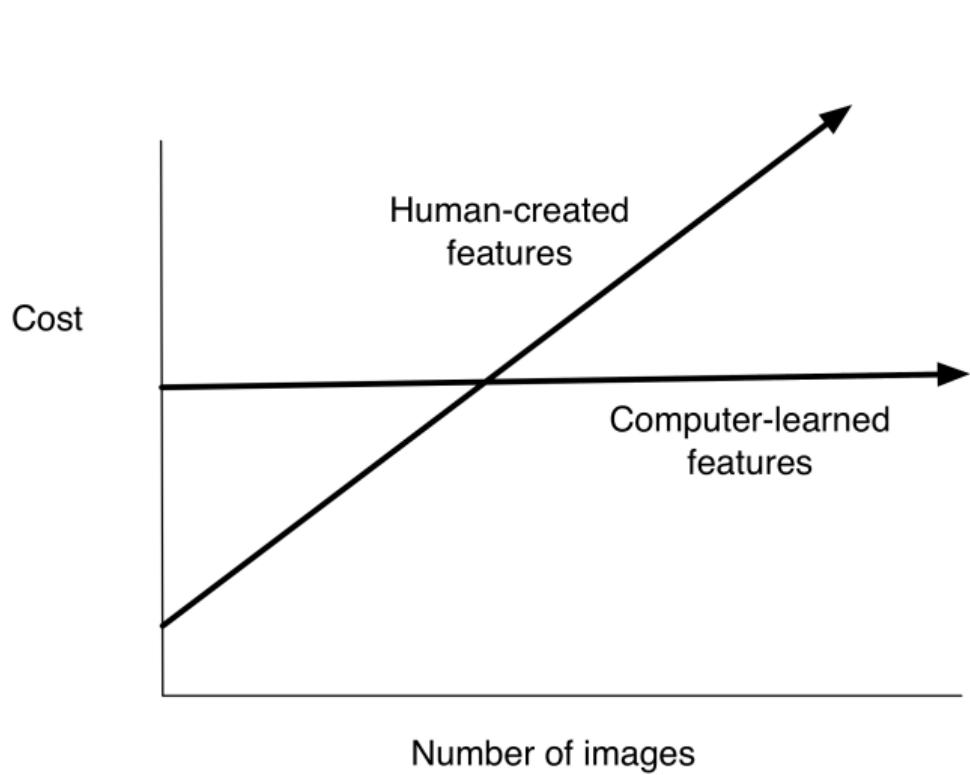


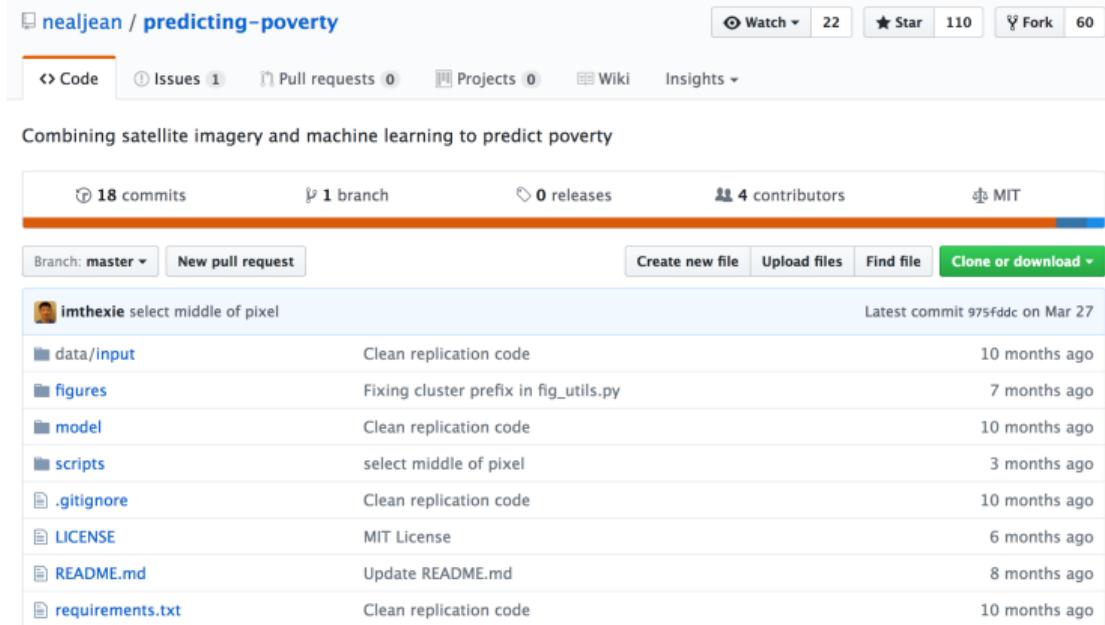
A Consumption expenditures



B Assets





A screenshot of a GitHub repository page. The repository name is "nealjean / predicting-poverty". It has 22 stars, 60 forks, and 18 commits. The branch is "master". The latest commit was on March 27. The commits are listed below.

File	Message	Time Ago
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>

Wrap-up:

- ▶ Surveys and big data are compliments not substitutes

Wrap-up:

- ▶ Surveys and big data are compliments not substitutes
- ▶ Sometime we do “enriched asking” and sometimes “amplified asking” (role of big data source is different in both cases)

Wrap-up:

- ▶ Surveys and big data are compliments not substitutes
- ▶ Sometime we do “enriched asking” and sometimes “amplified asking” (role of big data source is different in both cases)
- ▶ Learn more: see “what to read next” in Ch 3 of Bit by Bit.