Combining surveys and big data

Matthew J. Salganik
Department of Sociology
Princeton University

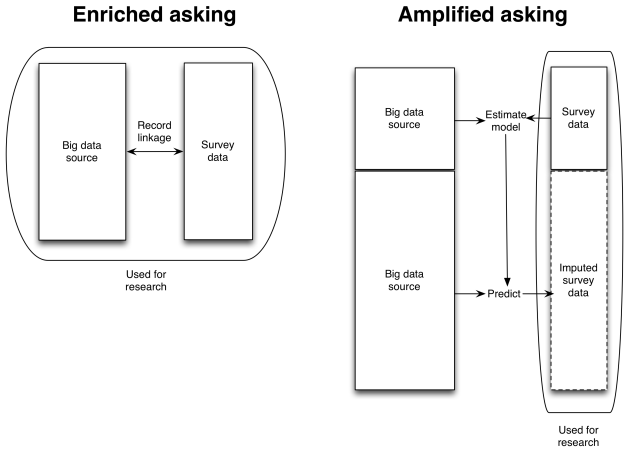
Summer Institute in Computational Social Science
June 22, 2017

	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



Note the different role of the big data in each case

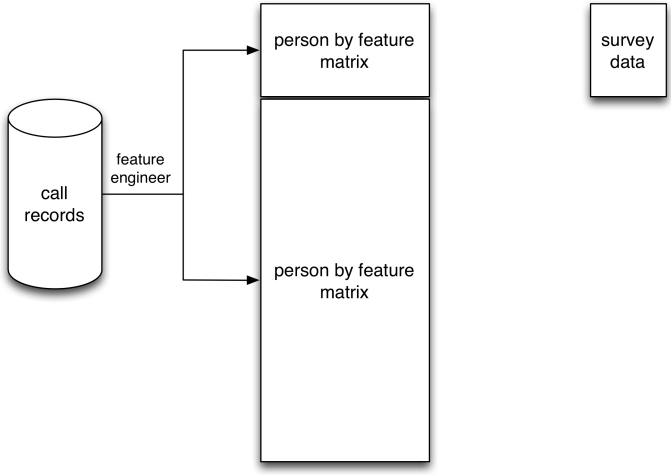
Predicting poverty and wealth from mobile phone metadata

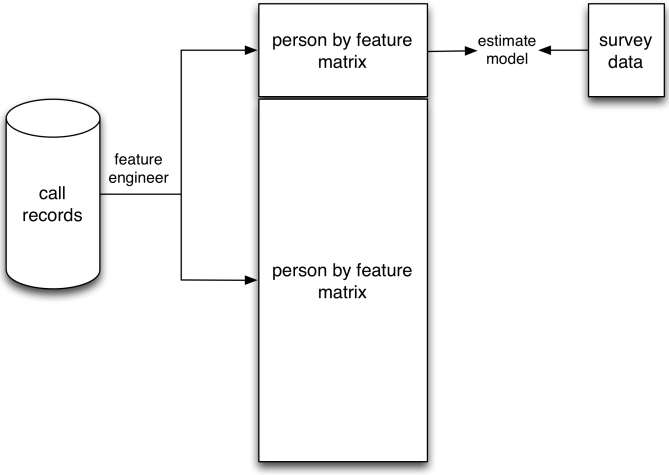
Joshua Blumenstock,[1]* Gabriel Cadamuro,[2] Robert On[3]

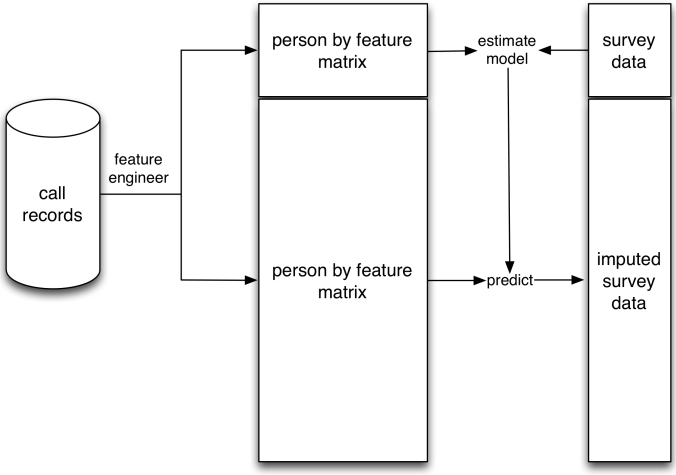

call
records

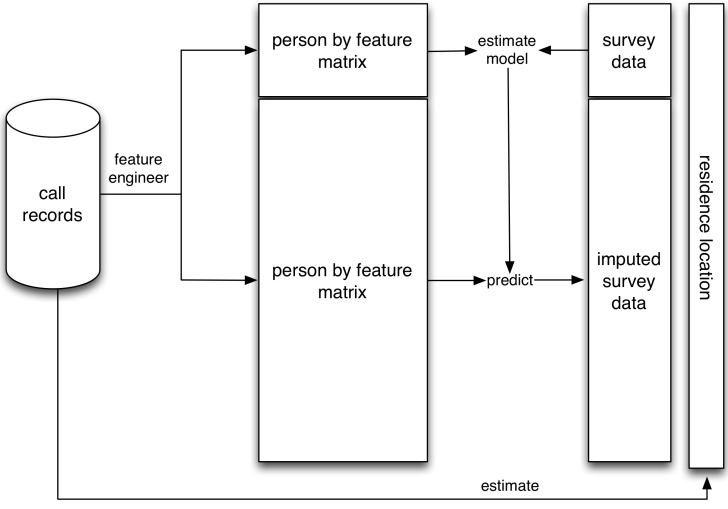

call
records

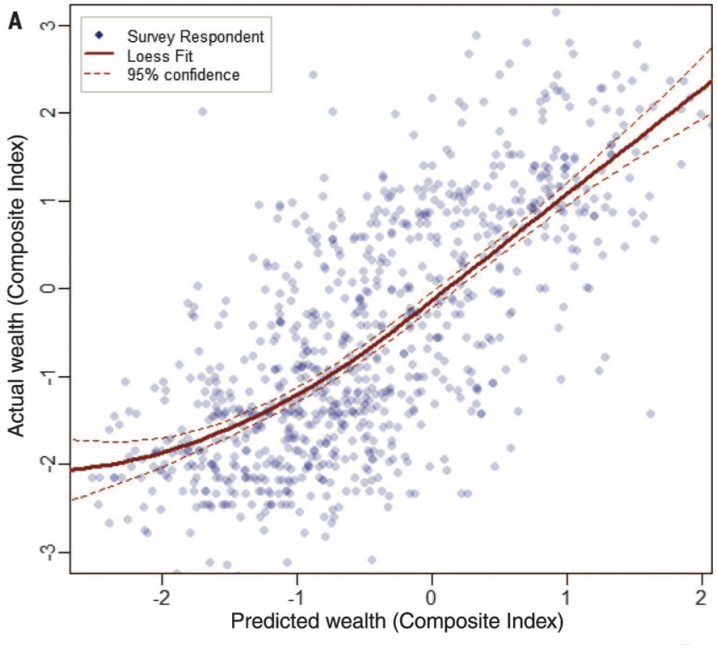

survey
data

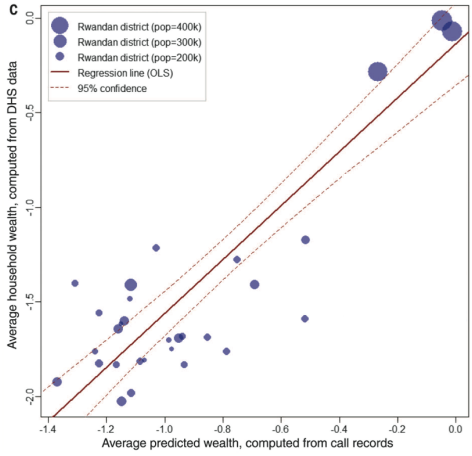

person by feature matrix
survey data
call records
feature engineer
person by feature matrix

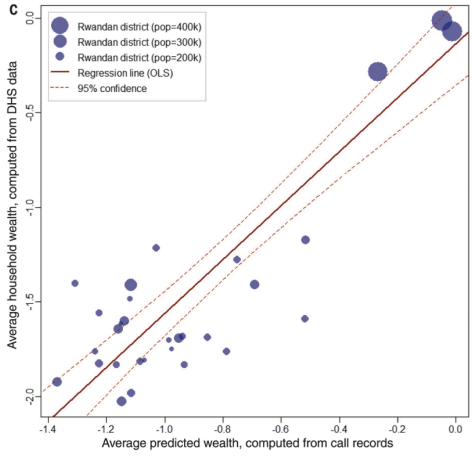

person by feature matrix
estimate model
survey data
call records
feature engineer
person by feature matrix


person by feature matrix
estimate model
survey data
call records
feature engineer
person by feature matrix
predict
imputed survey data


person by feature matrix
estimate model
survey data
call records
feature engineer
person by feature matrix
predict
imputed survey data
residence location
estimate


A
Survey Respondent
Loess Fit
95% confidence
Actual wealth (Composite Index)
Predicted wealth (Composite Index)
3
2
1
0
-1
-2
-3
-2
-1
0
1
2


c
Rwandan district (pop=400k)
Rwandan district (pop=300k)
Rwandan district (pop=200k)
Regression line (OLS)
95% confidence
Average household wealth, computed from DHS data
0.0
-0.5
-1.0
-1.5
-2.0
-1.4
-1.2
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
Average predicted wealth, computed from call records



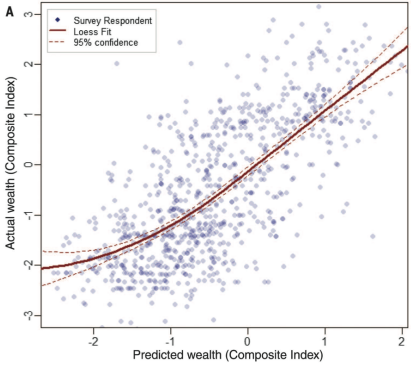
- 10 times faster
- 50 times cheaper

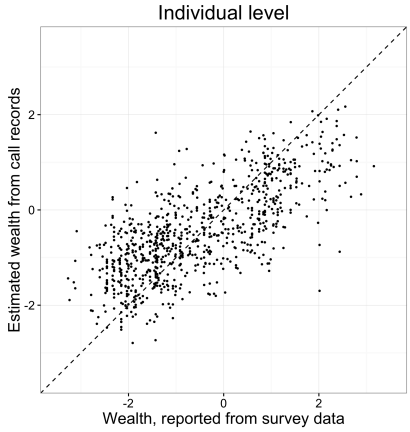
Readymades

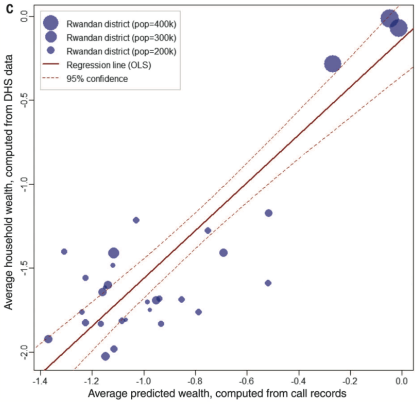
+

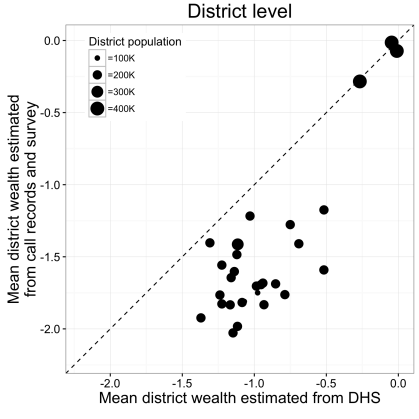
Custommades

https://commons.wikimedia.org/wiki/File:Duchamp_Fountaine.jpg
https://commons.wikimedia.org/wiki/File:%27David%27_by_Michelangelo_JBU0001.JPG


A
Survey Respondent
Loess Fit
95% confidence
Actual wealth (Composite Index)
Predicted wealth (Composite Index)
3
2
1
0
-1
-2
-3
-2
-1
0
1
2


Individual level
Estimated wealth from call records
Wealth, reported from survey data
2
0
-2
-2
0
2


c
Rwandan district (pop=400k)
Rwandan district (pop=300k)
Rwandan district (pop=200k)
Regression line (OLS)
95% confidence
Average household wealth, computed from DHS data
Average predicted wealth, computed from call records
0.0
-0.5
-1.0
-1.5
-2.0
-1.4
-1.2
-1.0
-0.8
-0.6
-0.4
-0.2
0.0


District level
District population
=100K
=200K
=300K
=400K
Mean district wealth estimated from call records and survey
Mean district wealth estimated from DHS
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1*] Gabriel Cadamuro,[2] Robert On[2]

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

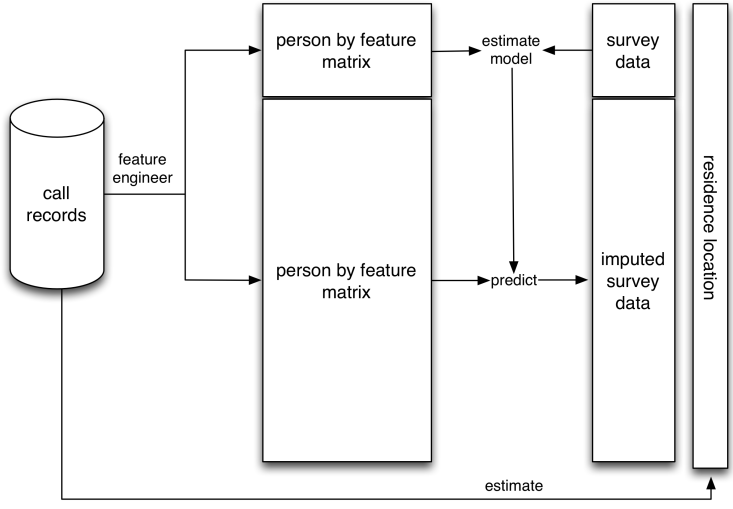
Muhammad Raza Khan[1], Joshua Manoj[1], Anikate Singh[1], Joshua Blumenstock[1]
[1]*Information School, University of Washington, Seattle, WA, USA*
Email: 'mrazaQuw.edu, 'joshuacmQuw.edu, 'anikatsQuw.edu, 'joshblumQuw.edu

Combining satellite imagery and machine learning to predict poverty

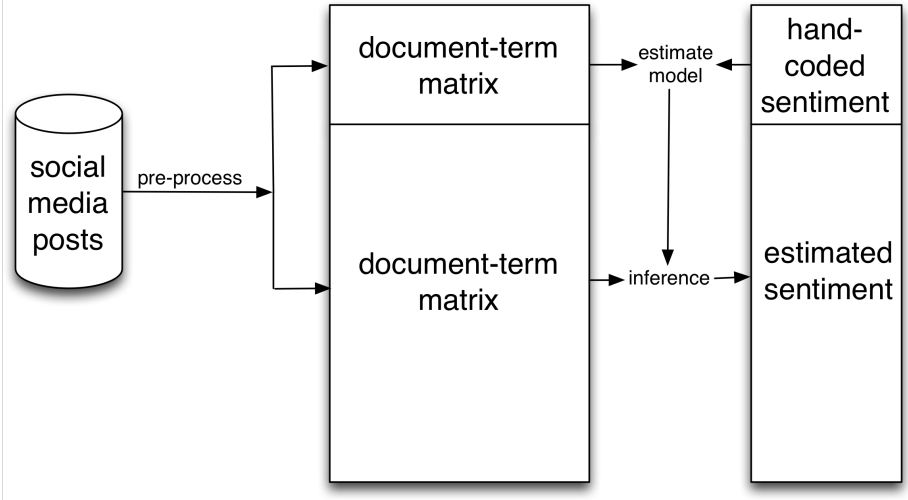
Neal Jean,[1,2*] Marshall Burke,[3,4,5*†] Michael Xie,[1] W. Matthew Davis,[4]
David B. Lobell,[3,4] Stefano Ermon[1]

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

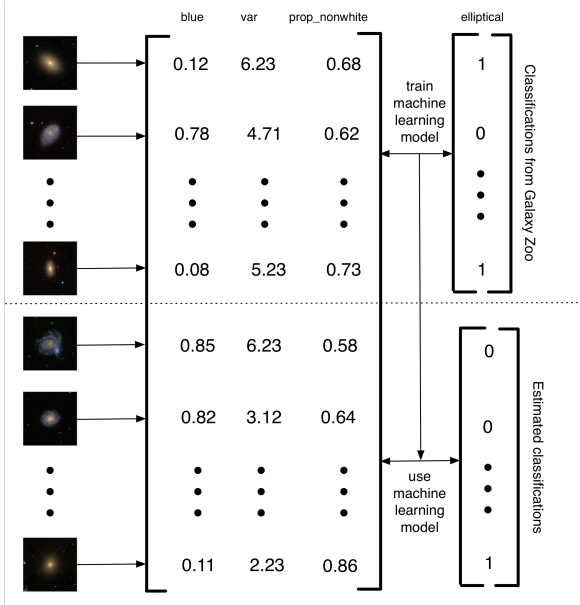
Supervised learning:
Lots of input-output pairs; goal is to develop a function that will predict the output from the input



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[1,2], Yoshua Bengio[3] & 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 Ermon[1]

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 Ermon[1]

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

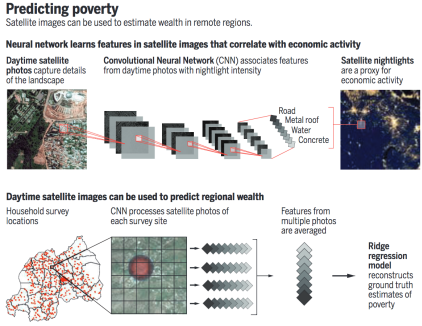
Live demo:
https://www.google.com/maps/place/Kigali,+Rwanda/@-1.9546259,30.0345059,26517m/data=!3m2!1e3!4b1!4m5!3m4!1s0x19dca4258ed8e797:0xf32b36a5411d0bc8!8m2!3d-1.9705786!4d30.1044288

But, most people had been using night lights

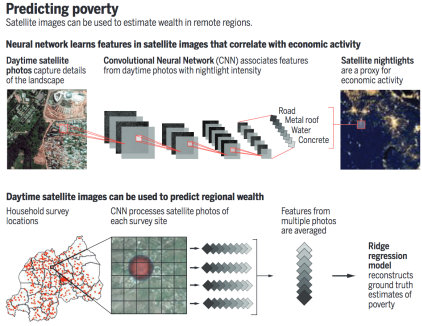
https://www.nasa.gov/multimedia/imagegallery/image_feature_2480.html

Prior research:
Nightlights + survey data to estimate wealth in places without surveys

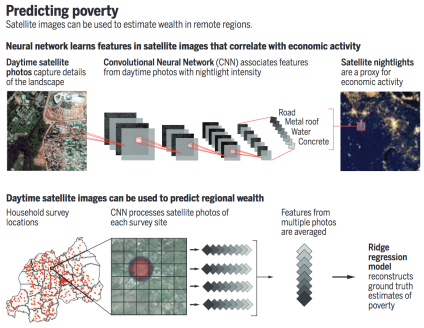
Jean et al. (2016):
Day pictures + Nightlights + survey data to estimate wealth in places without surveys



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

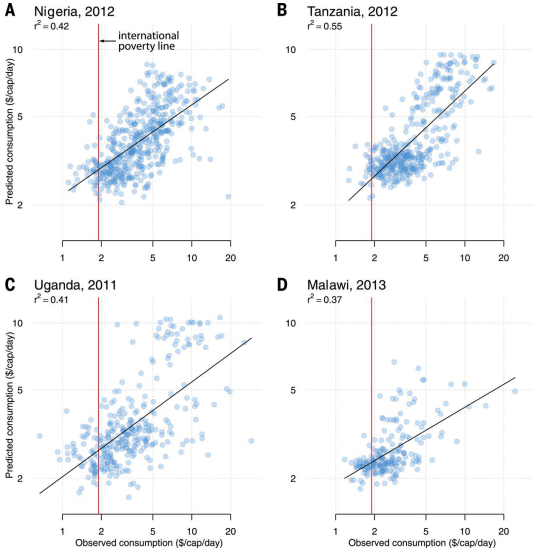


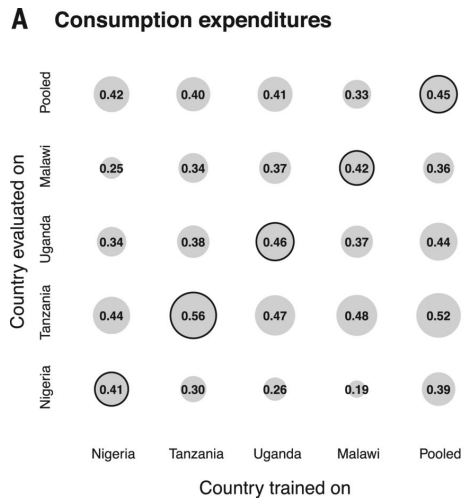
- Start with CNN pretrained on ImageNet (e.g. hampsters and weasels)
- Train CNN to predict nightlights from day pictures (lots of training data)

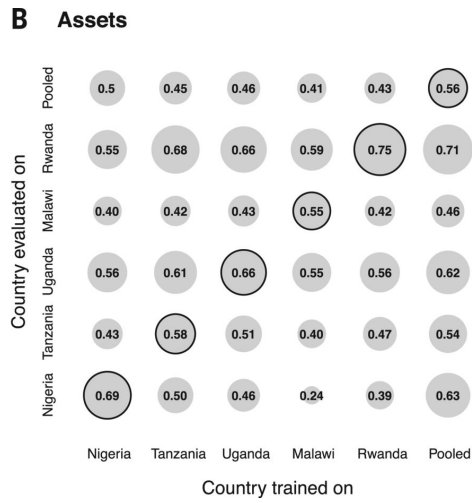


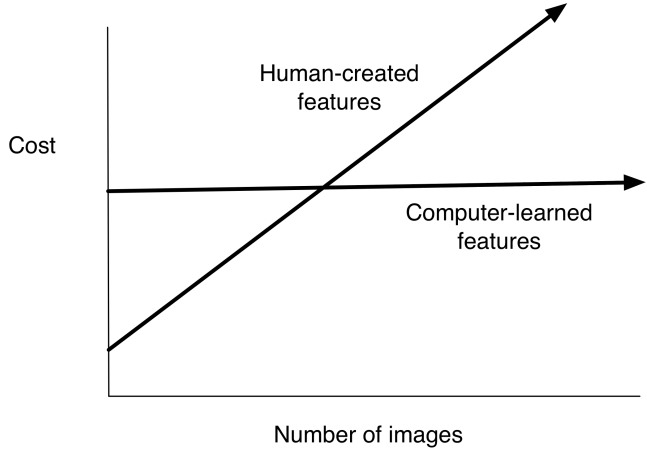
- Start with CNN pretrained on ImageNet (e.g. hampsters and weasels)
- Train CNN to predict nightlights from day pictures (lots of training data)
- Take features from CNN and train ridge regression to predict cluster mean survey response

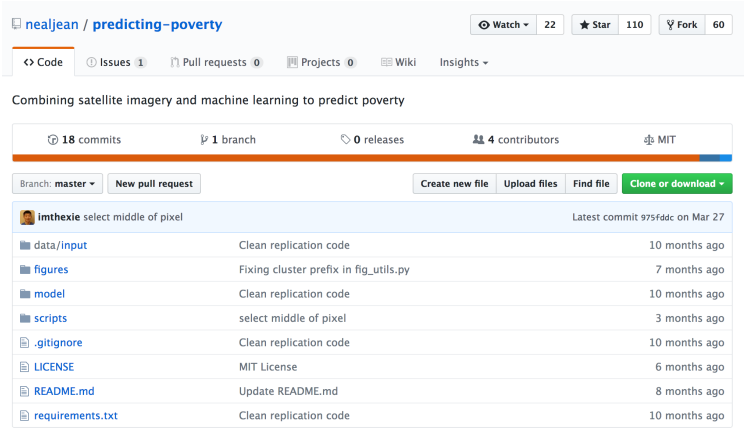
http://dx.doi.org/10.1126/science.aah5217


A
Nigeria, 2012
$r^2 = 0.42$
international poverty line
B
Tanzania, 2012
$r^2 = 0.55$
C
Uganda, 2011
$r^2 = 0.41$
D
Malawi, 2013
$r^2 = 0.37$
Predicted consumption ($/cap/day)
Observed consumption ($/cap/day)
1
2
5
10
20


A
Consumption expenditures
Country evaluated on
Pooled
Malawi
Uganda
Tanzania
Nigeria
0.42
0.40
0.41
0.33
0.45
0.25
0.34
0.37
0.42
0.36
0.34
0.38
0.46
0.37
0.44
0.44
0.56
0.47
0.48
0.52
0.41
0.30
0.26
0.19
0.39
Nigeria
Tanzania
Uganda
Malawi
Pooled
Country trained on


B
Assets
Country evaluated on
Pooled
Rwanda
Malawi
Uganda
Tanzania
Nigeria
0.5
0.45
0.46
0.41
0.43
0.56
0.55
0.68
0.66
0.59
0.75
0.71
0.40
0.42
0.43
0.55
0.42
0.46
0.56
0.61
0.66
0.55
0.56
0.62
0.43
0.58
0.51
0.40
0.47
0.54
0.69
0.50
0.46
0.24
0.39
0.63
Nigeria
Tanzania
Uganda
Malawi
Rwanda
Pooled
Country trained on


Human-created
features
Cost
Computer-learned
features
Number of images



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.