GR 6307
Public Economics and Development

4. Building State Capacity with Data & Technology

Michael Best

Spring 2018

Outline

Prediction Problems in Government

Examples of Data & Technology in Development

Outline

Prediction Problems in Government

- Kleinberg, Ludwig, Mullainathan & Obermeyer (AER PnP 2015) *Prediction Policy Problems*
- Kleinberg, Lakkarju, Leskovec, Ludwig & Mullainathan (WP 2017) *Human Decisions and Machine Predictions*

Kleinberg et al 2015: Prediction

- In empirical policy research, we often focus on *causal inference*. Policy choices depend on understanding the counterfactual–what would happen without the policy.
- But, there are many policy problems where causal inference is not central, or even necessary!
- Consider 2 toy examples:
  1. There is a drought. Should we invest in a rain dance to increase the chance of rain?
  2. It is cloudy. Should I take an umbrella to work?
- In both cases data are going to be useful. But they require different estimators:
  1. Do rain dances cause rain? *Causal inference*
  2. Is the chance of rain high enough to merit an umbrella? *Prediction inference*
- This paper: Policy prediction problems are everywhere, and machine learning can help us solve them more effectively.

Kleinberg et al 2015: Prediction and Causation

- Consider this framework for thinking about this.
- Let Y be an outcome (rain) that depends on X_0 (a policy choice) and other Xs.
- The policymaker must pick X_0 (umbrella, rain dance) to maximize known payoff $\pi(X_0, Y)$.
- Decision depends on

$$\frac{d\pi(X_0, Y)}{dX_0} = \frac{\partial}{\partial X_0} \underbrace{\pi(Y)}_{\text{prediction}} + \frac{\partial \pi}{\partial Y} \underbrace{\frac{\partial Y}{\partial X_0}}_{\text{causation}}$$

1. Prediction: We know the payoff, but we need to evaluate it at Y, so we need to predict what Y will be
2. Causality: How much will Y change if I change X_0?

- 2 things to note
  1. Prediction is useful when $\partial\pi/\partial X$ depends on Y (benefit of umbrella depends on rain)
  2. Only $\hat{Y}$ enters the decision, so we just need a low error estimate of $\hat{Y}$, not an unbiased or causal one

Kleinberg et al 2015: Machine Learning

- Standard empirical techniques aren't great for prediction because they focus on **unbiasedness**.
- e.g. Suppose you have 2 variables to predict y and you get OLS estimates $\hat{\beta}_1 = 1 \pm 0.001$ and $\hat{\beta}_2 = 4 \pm 10$. What's the best prediction? $x_1 + 4x_2$? or perhaps the unbiased estimator x_1 since $\hat{\beta}_2$ is noisy?
- General setup:
  - Suppose you have a dataset D of n points $(y_i, x_i) \sim G$
  - Use this data to pick a function $\hat{f} \in \mathcal{F}$ to predict the y value of a new data point $(y, x) \sim G$
  - Goal is to minimize a loss function $\mathcal{L}\left(y, \hat{f}\right) = \left(y - \hat{f}\left(x\right)\right)^2$

Kleinberg et al 2015: Machine Learning

- OLS minimizes *in-sample* error by choosing among linear functions $\mathcal{F}_{lin}$

$$\hat{f}_{OLS} = \underset{f_\beta \in \mathcal{F}_{lin}}{\arg\min} \sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2$$

- This does great in-sample (sum over the is in *this* dataset). But could do *arbitrarily badly* on a new dataset (out of sample)!!

$$\begin{aligned} MSE\left(x\right) &\equiv \mathbb{E}_D\left[\left(\hat{f}\left(x\right) - y\right)^2\right] \\ &= \underbrace{\mathbb{E}_D\left[\left(\hat{f}\left(x\right) - \mathbb{E}_D\left[\hat{y}_0\right]\right)^2\right]}_{\text{Variance}} + \underbrace{\left(\mathbb{E}_D\left[\hat{y}_0\right] - y\right)^2}_{\text{Bias}^2} \end{aligned}$$

Kleinberg et al 2015: Machine Learning

- So what's the alternative? ML techniques minimize

$$\hat{f}_{ML} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2 + \lambda R\left(f\right)$$

- $R\left(f\right)$ is a *regularizer* that penalizes choosing functions that create variance. Constructed so that the set of functions $\mathcal{F}_c = \{f | R\left(f\right) \leq c\}$ creates more variable predictions as c increases.
- For linear models, larger coefficients generate more variable predictions so a natural regularizer is $R\left(f_\beta\right) = \|\beta\|^d$ ($d = 1$=LASSO, $d = 2$=ridge)
- λ is the price at which we trade off variance and bias. OLS has infinite price for bias $1/\lambda = \infty$.

Kleinberg et al 2015: Machine Learning

- How should we pick λ? A key insight of machine learning is that you can ask the data to tell you:
- Split the data into f subsets (folds).
- For a set of potential λs estimate the algorithm on $f-1$ of the folds and then see which λ predicts best on the fth fold (the hold-out fold). This procedure is called (f-fold) *cross-validation*
- This is cool because they expand the set of predictors we can consider:
  - They allow for *wide* data (e.g. text data, internet activity data) where we have many more variables than data points.
  - Much more flexible functional forms
- **But NB, you will get a great $\hat{y}$, you don't get any guarantee that you have useful $\hat{\beta}$s**

Kleinberg et al 2015: Policy Examples

1. Who should get joint replacement surgery?
2. Which teacher will have the greatest value added?
3. How long will unemployment spells last?
4. How should health inspections be targeted?
5. Predicting at-risk youth
6. Which borrowers are credit-worthy?

Outline

Prediction Problems in Government

- Kleinberg, Ludwig, Mullainathan & Obermeyer (AER PnP 2015) *Prediction Policy Problems*
- Kleinberg, Lakkarju, Leskovec, Ludwig & Mullainathan (WP 2017) *Human Decisions and Machine Predictions*

Kleinberg et al 2017: Introduction

- Machine learning is all about prediction. This paper looks at what, at first glance, is an ideal application.
- After an arrest, judges decide whether defendants await trial in jail or at home. Law says this bail decision depends only on the judges prediction of whether the defendant wil reoffend or flee if released.
- Here use data on 758,027 defendands in NYC between 2008 and 2013 to predict probability of reoffending.
- *Can we use these predictions to understand and improve judges' decisions?*

1. Needs methods from *both* machine learning and microeconometrics.
2. *Omitted payoffs:* We can predict probability of reoffending, but what if there are other things judges care about?
3. *Selective labels:* We only see the crime outcomes of people who are released. What would those who were jailed have done if they had been released?

Kleinberg et al 2017: Data & Context

- Shortly after someone is arrested, there's a bail hearing. Judges can a) release; b) set a dollar bail; c) detain with no chance of bail.
- Judges asked to decide based on a prediction of whether the defendant would fail to appear in court or be re-arrested for a new crime.
- Data on all arrests in NYC between 11/2008 and 11/2013: 1,460,462 cases.
- 758,027 subject to a pre-trial release decision. Randomly sample 203,338 cases to keep in a "lock box": Not used for training the algorithm or writing drafts of the paper, will only be used for the final version.
- Working data is 554,689 cases.

Table 1: Summary Statistics

	Full Sample	Judge Releases	Judge Detains	P-value
Sample Size	554,689	408,283	146,406	
Release Rate	.7361	1.0000	0.00	
Outcomes				
Failure to Appear (FTA)	.1521	.1521		
Arrest (NCA)	.2581	.2581		
Violent Crime (NVCA)	.0372	.0372		
Murder, Rape, Robbery (NMRR)	.0187	.0187		
Defendant Characteristics				
Age	31.98	31.32	33.84	<.0001
Male	.8315	.8086	.8955	<.0001
White	.1273	.1407	.0897	<.0001
African American	.4884	.4578	.5737	<.0001
Hispanic	.3327	.3383	.3172	<.0001
Arrest County				
Brooklyn	.2901	.2889	.2937	.0006
Bronx	.2221	.2172	.2356	<.0001
Manhattan	.2507	.2398	.2813	<.0001
Queens	.1927	.2067	.1535	<.0001
Staten Island	.0440	.0471	.0356	<.0001

Arrest Charge				
Violent Crime				
Violent Felony	.1478	.1193	.2272	<.0001
Murder, Rape, Robbery	.0581	.0391	.1110	<.0001
Aggravated Assault	.0853	.0867	.0812	<.0001
Simple Assault	.2144	.2434	.1335	<.0001
Property Crime				
Burglary	.0206	.0125	.0433	<.0001
Larceny	.0738	.0659	.0959	<.0001
MV Theft	.0067	.0060	.0087	<.0001
Arson	.0006	.0003	.0014	<.0001
Fraud	.0696	.0763	.0507	<.0001
Other Crime				
Weapons	.0515	.0502	.0552	<.0001
Sex Offenses	.0089	.0086	.0096	.0009
Prostitution	.0139	.0161	.0078	<.0001
DUI	.0475	.0615	.0084	<.0001
Other	.1375	.1433	.1216	<.0001
Gun Charge	.0335	.0213	.0674	<.0001
Drug Crime				
Drug Felony	.1411	.1175	.2067	<.0001
Drug Misdemeanor	.1142	.1156	.1105	<.0001

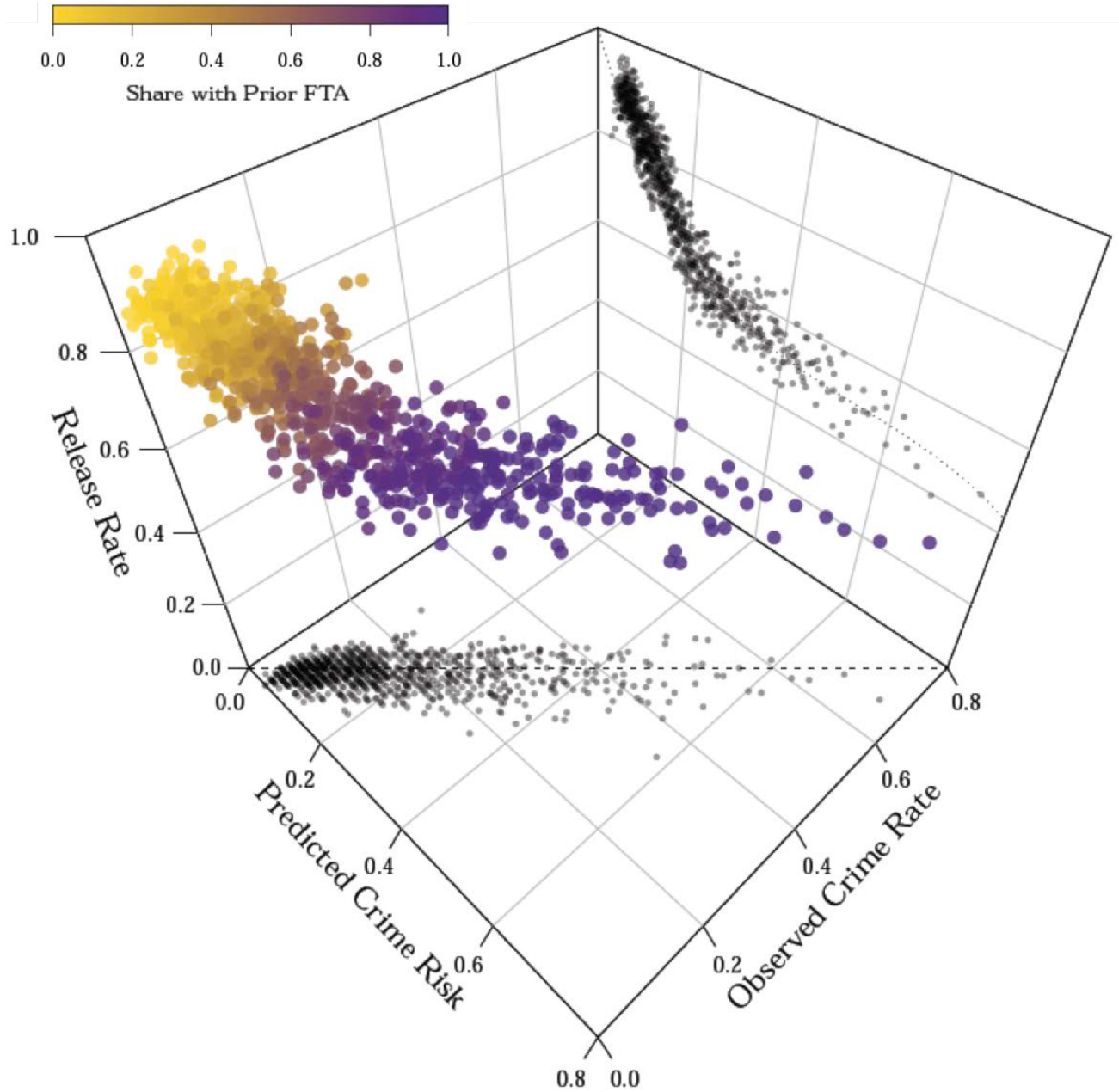
	Full Sample	Judge Releases	Judge Detains	P-value
Defendant Priors				
FTAs	2.093	1.305	4.288	<.0001
Felony Arrests	3.177	2.119	6.127	<.0001
Felony Convictions	.6157	.3879	1.251	<.0001
Misdemeanor Arrests	5.119	3.349	10.06	<.0001
Misdemeanor Convictions	3.122	1.562	7.473	<.0001
Violent Felony Arrests	1.017	.7084	1.879	<.0001
Violent Felony Convictions	.1521	.1007	.2955	<.0001
Drug Arrests	3.205	2.144	6.163	<.0001
Felony Drug Convictions	.2741	.1778	.5429	<.0001
Misdemeanor Drug Convictions	1.049	.5408	2.465	<.0001
Gun Arrests	.2194	.1678	.3632	<.0001
Gun Convictions	.0462	.0362	.0741	<.0001

Kleinberg et al 2017: Machine Learning

- Form predictions $\hat{y} = m(x)$ to minimize loss function $L(y, \hat{y})$
- Consider functions $m(x)$ that generate predicted probabilities in $[0, 1]$
- Use Bernoulli loss

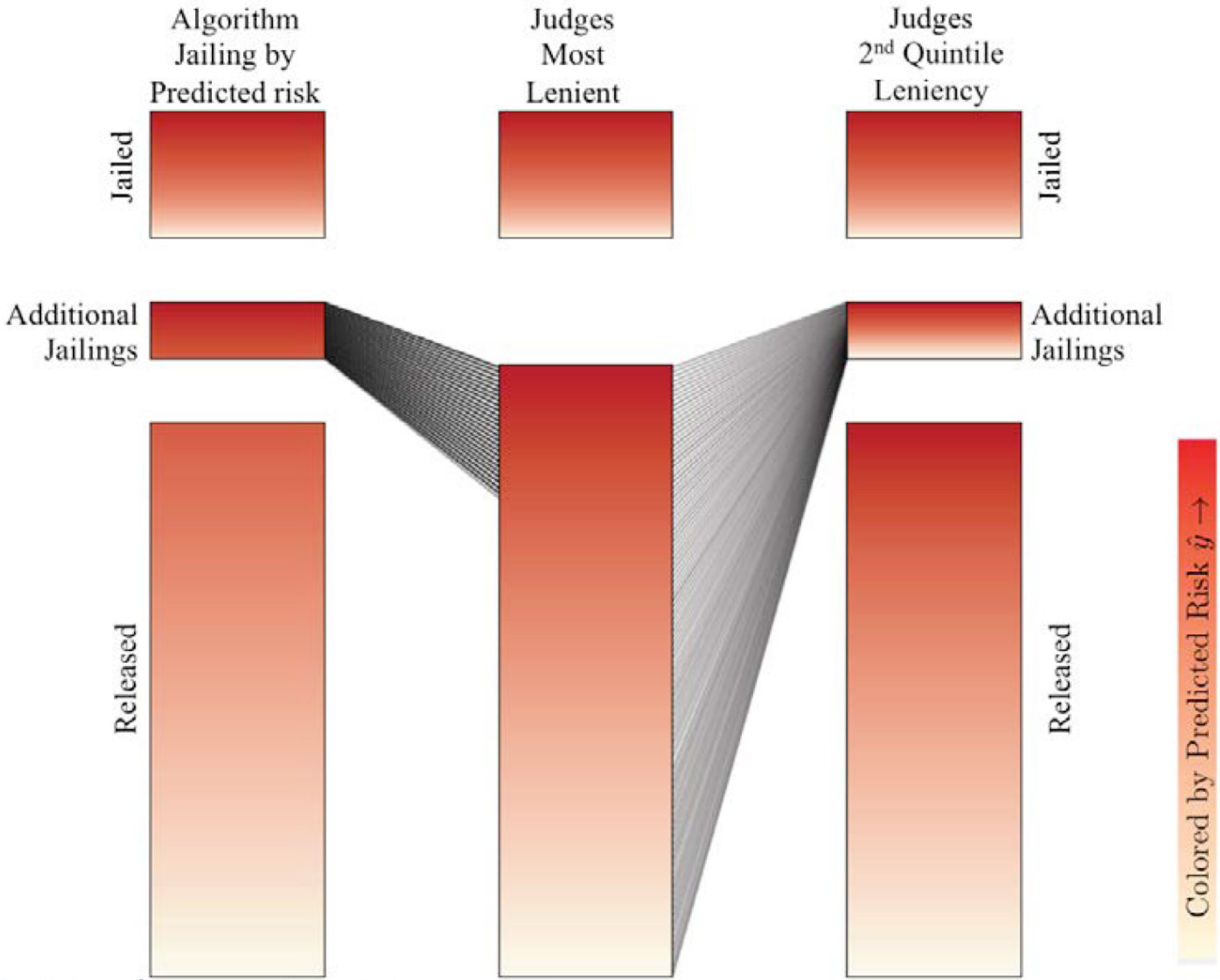
$$L(y_i, m(x_i)) = -[y_i \log(m(x_i)) + (1 - y_i) \log(1 - m(x_i))]$$

- For this paper, use gradient boosted decision trees to form $m(x_i)$. Average multiple trees built sequentially where each iteration up-weights observations that fit poorly in the sequence of trees up to that point.


0.0
0.2
0.4
0.6
0.8
1.0
Share with Prior FTA
1.0
0.8
0.6
0.4
0.2
0.0
Release Rate
0.0
0.2
0.4
0.6
0.8
Predicted Crime Risk
0.8
0.6
0.4
0.2
0.0
Observed Crime Rate

Kleinberg et al 2017: Misranking?

- The riskiest 1% of defendants have a predicted risk of 62.6%, but 48.5% of them are released, and reoffend 56.3% of the time!
- Suggests Judges are misranking the defendants. But could also be that the judges use an even higher threshold risk for detention.
- Look across judges of different leniencies. If it's the case that more lenient judges just have a higher threshold risk, then the predicted risk scores should be good at predicting which people will be jailed by a less lenient judge relative to a lenient judge.
- Identified by quasi-random assignment of cases to judges: It depends on who happens to be on duty in that bourough $\times$ court house $\times$ day


Algorithm
Jailing by
Predicted risk
Judges
Most
Lenient
Judges
2nd Quintile
Leniency
Jailed
Jailed
Additional
Jailings
Additional
Jailings
Released
Released
Colored by Predicted Risk $\hat{y} \rightarrow$

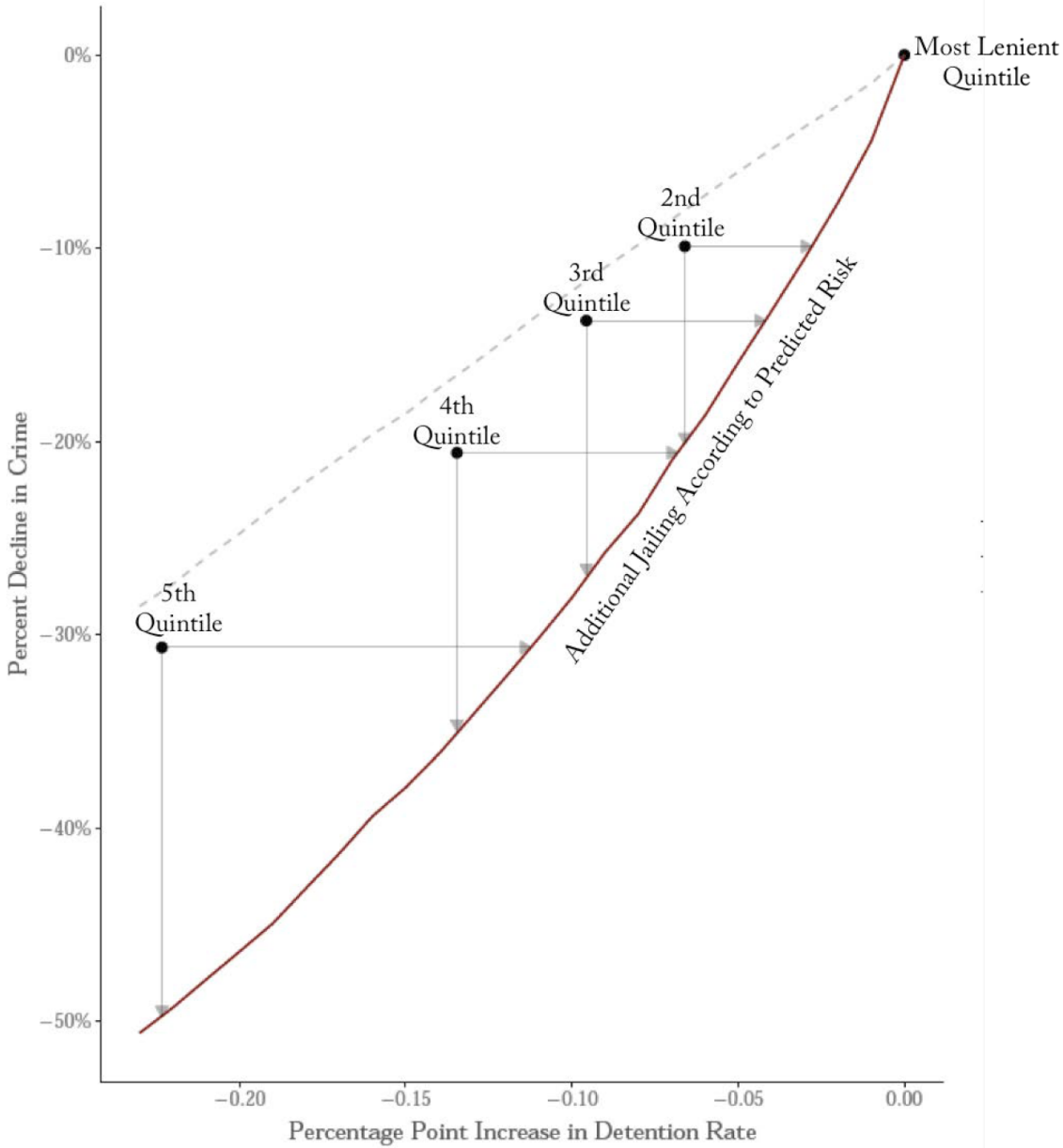


Figure 5: Comparing Impact of Detaining in order of Predicted Risk To What Judges Achieve

Kleinberg et al 2017: Alternative Policies

- So can we use the algorithm to improve release decisions?
- Consider 2 types of policies

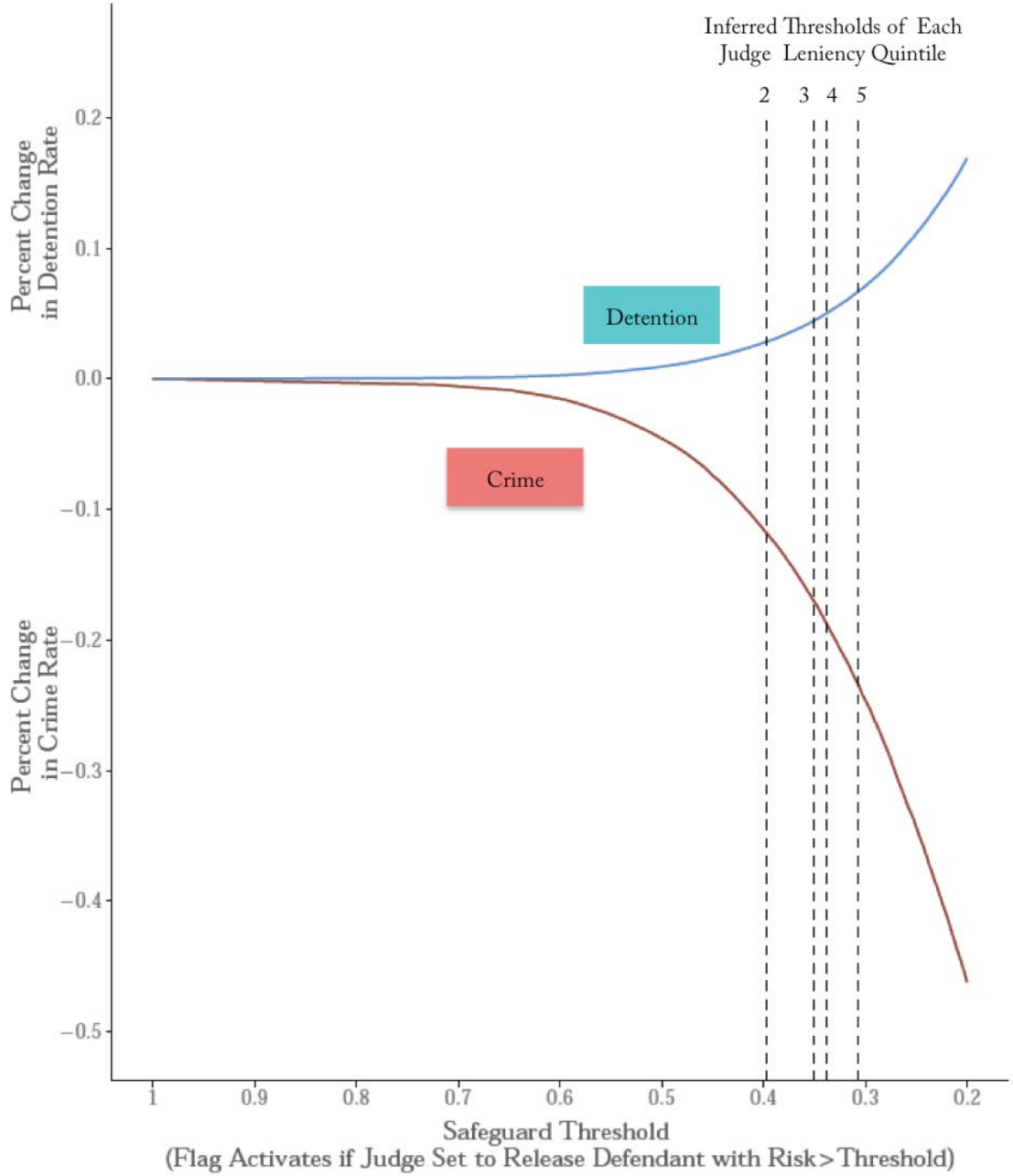
1. Contraction: A warning whenever the judge is about to release a high-risk defendant. Like a driver assist system that warns when the car does something potentially dangerous.
2. Reranking: Risk tool danks defendants and makes recommendations for all decisions. Like an auto-pilot that the judge could overrule. Improves both high- and low-risk decisions.

- Now we need a counterfactual: 2 issues arise

1. Compliance: Will the judge pay attention to the warnings / comply with the risk score?
2. Reranking involves releasing some of the jailed defendants. How much crime *would they have* committed? A missing labels problem.

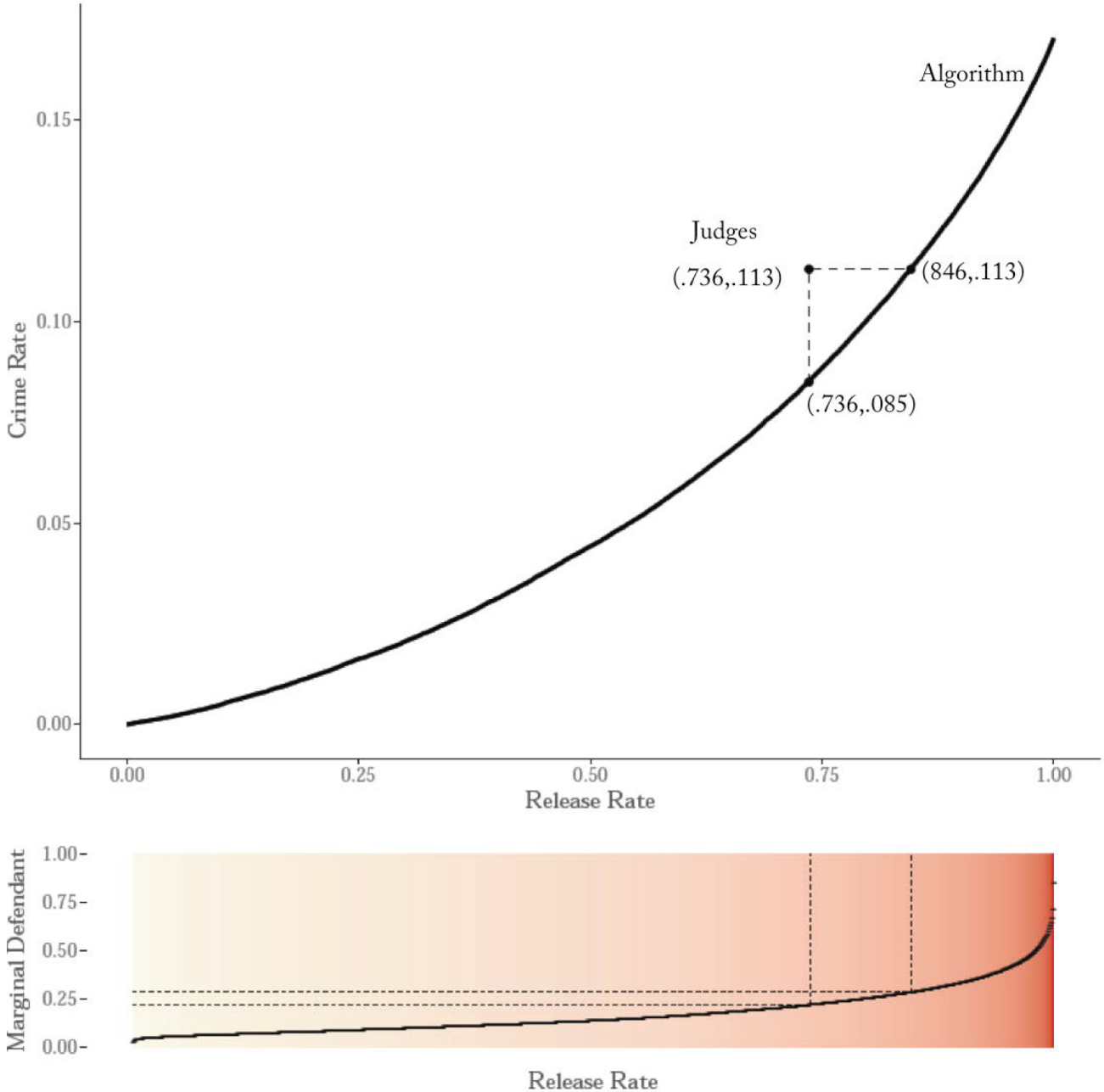
Kleinberg et al 2017: Contraction

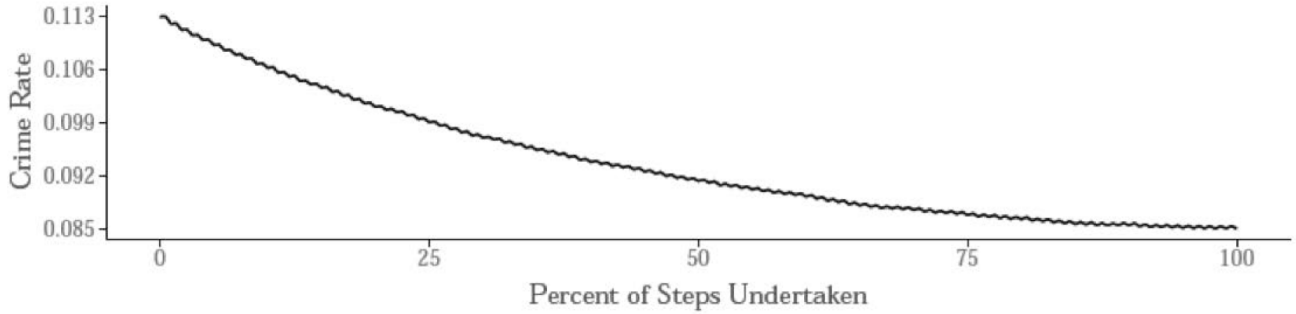
- What threshold risk score should trigger the warning?
- Low threshold averts more crime, but triggers more warnings, and increases jailings of people who would not reoffend.
- Implicitly, the judges have some threshold too that we can compare to.


Inferred Thresholds of Each Judge Leniency Quintile
2
3
4
5
Percent Change in Detention Rate
0.2
0.1
0.0
Detention
Crime
−0.1
−0.2
−0.3
−0.4
−0.5
Percent Change in Crime Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Safeguard Threshold
(Flag Activates if Judge Set to Release Defendant with Risk>Threshold)

Kleinberg et al 2017: Reranking

- Reranking creates a selective labels problem: We only see crime rates of those who are jailed.
- Approach here: Impute based on observables. Then do 2 bounding exercises
  1. Decompose algorithm's gains into
     - 1.1 jail a high risk defendant and release an *average* risk defendant
     - 1.2 release a low risk defendant and jail an *average* risk defendant. Selective labels problem only comes in here.
  2. Assume imputing underestimates by α and use a prediced risk of $\min\{1, \alpha\hat{y}\}$ and do robustness to α.


Algorithm
Judges
(.736,.113)
(846,.113)
(.736,.085)
Crime Rate
0.15
0.10
0.05
0.00
0.00
0.25
0.50
0.75
1.00
Release Rate
Marginal Defendant
1.00
0.75
0.50
0.25
0.00
Release Rate


Crime Rate
0.113
0.106
0.099
0.092
0.085
0
25
50
75
100
Percent of Steps Undertaken

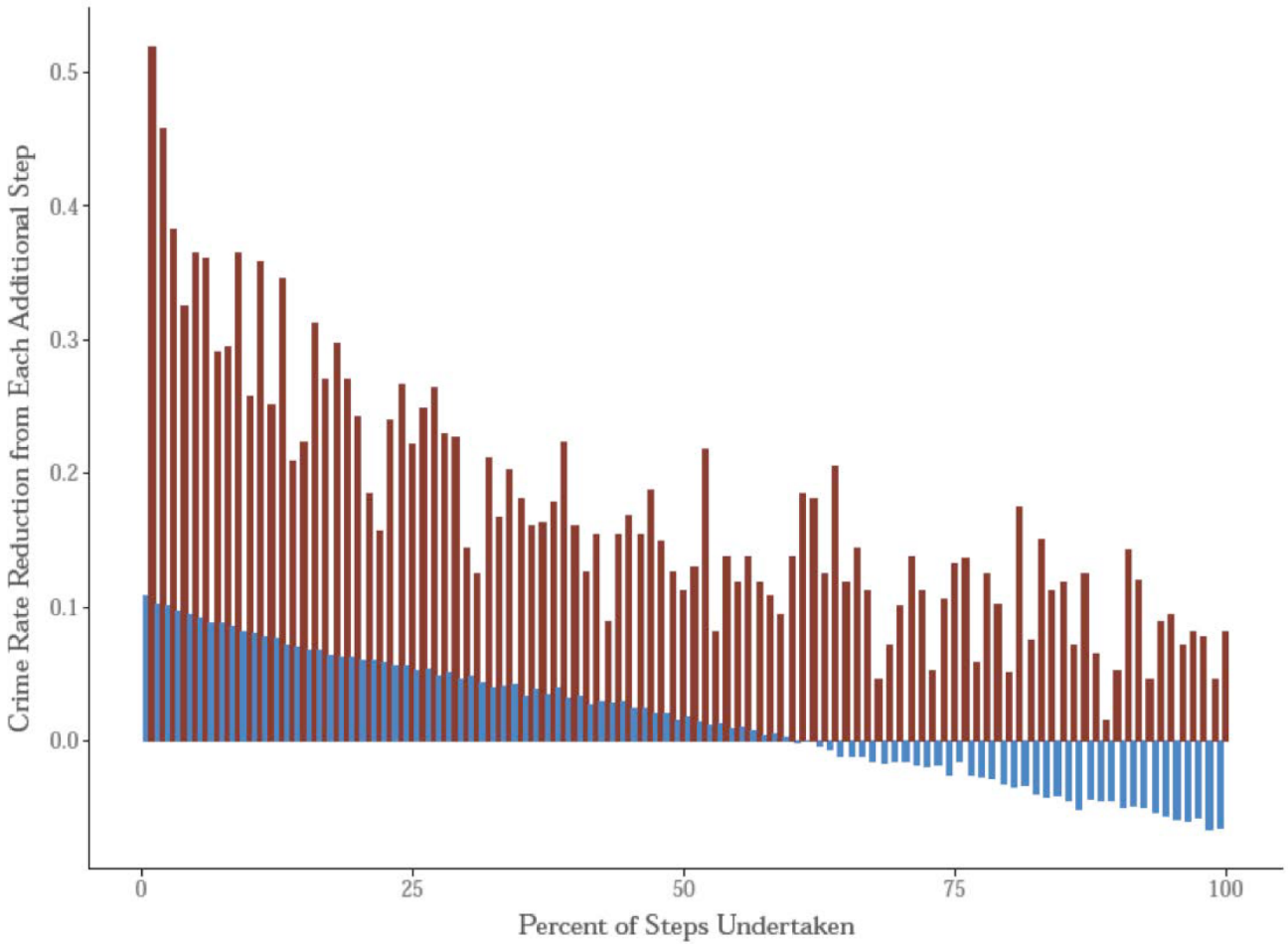

Crime Rate Reduction from Each Additional Step
0.5
0.4
0.3
0.2
0.1
0.0
0
25
50
75
100
Percent of Steps Undertaken

Table 5: Policy Simulation Under Different Assumptions

	Assume $y = min(1, \alpha\hat{y})$ for Additional Releases Beyond Most Lenient Judge Quintile's Release Rate						
	Value of α						
	1	1.25	1.5	2	3	$\cdots$	∞
Algorithm's Crime Rate	.0854	.0863	.0872	.0890	.0926		.1049
at Judge's Jail Rate	(.0008)	(.0008)	(.0008)	(.0009)	(.0009)		(.0009)
Percentage Reduction	-24.68%	-24.06%	-23.01%	-21.23%	-18.35%		-14.39%
Algorithm's Jail Rate	.1531	.1590	.1642	.1733	.1920		.2343
at Judge's Release Rate	(.0011)	(.0011)	(.0011)	(.0011)	(.0012)		(.0013)
Percentage Reduction	-41.85%	-40.13%	-38.37%	-34.87%	-29.36%		-18.51%

Outline

Prediction Problems in Government

Examples of Data & Technology in Development

Outline

Examples of Data & Technology in Development

- Abelson, Varshney & Sun (2014): *Targeting Direct Cash Transfers to the Extremely Poor*
- Blumenstock, Cadamuro & On (Science 2015) *Predicting Poverty and Wealth from Mobile Phone Metadata*
- Jean, Burke, Xie, Davis, Lobell & Ermon *Combining Satellite Imagery and Machine Learning to Predict Poverty*

Abelson et al 2014

- Work with GiveDirectly to see how data can be used to improve targeting.
- Use satellite data and roof material (thatch vs metal) to predict poverty.
  - Metal roof better: Mosquitos live in thatch, leak and collapse regularly.
  - Metal roof is expensive: $\$564$.
  - Good proxy for poverty
- Aggregate up to the village level to rank villages for GiveDirectly operations.

Abelson et al 2014

(a)

(b)

Figure 1: Homes in central east Africa with (a) metal and (b) thatched roofs.

Abelson et al 2014

Figure 2: Example of metal roof in center of satellite image.

Figure 3: Example of thatched roof in center of satellite image.

Abelson et al 2014

- Data has issues:
  - Satellite data is from google maps: Sometimes there are clouds, some pictures from the wet season, others from the dry season.
- Training data from GiveDirectly which does census in villages it works in and has roof type in it.
- Set up a crowdsourcing application (using Flask in Python)

Abelson et al 2014

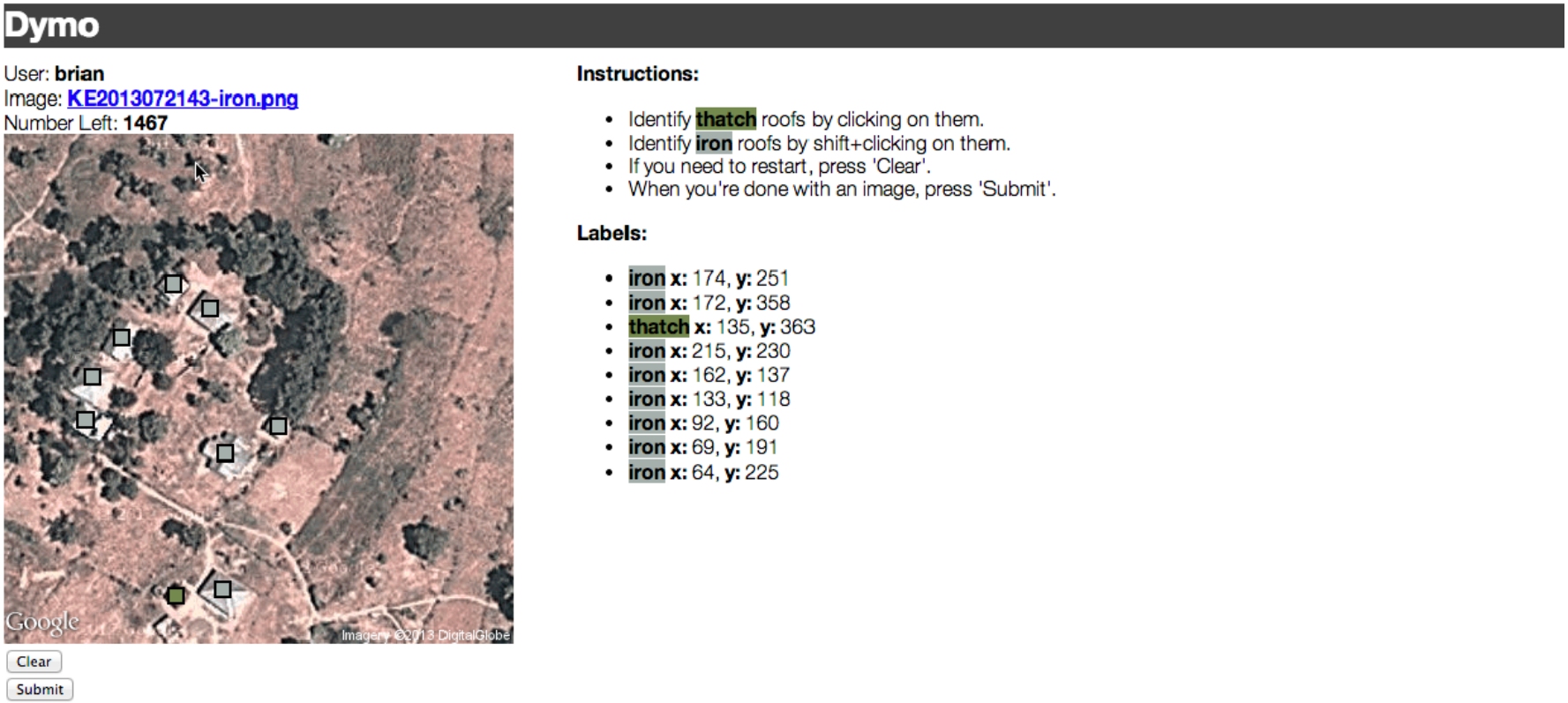


Figure 6: Screen shot of application deployed for crowdsourced labeling of roofs in satellite images.

Abelson et al 2014

- Train algorithms for two problems:
  1. How many roofs in each 400 x 400 pixel satellite image?
  2. What fraction of the roods are metal?
- Use random forests for this problem.

Abelson et al 2014

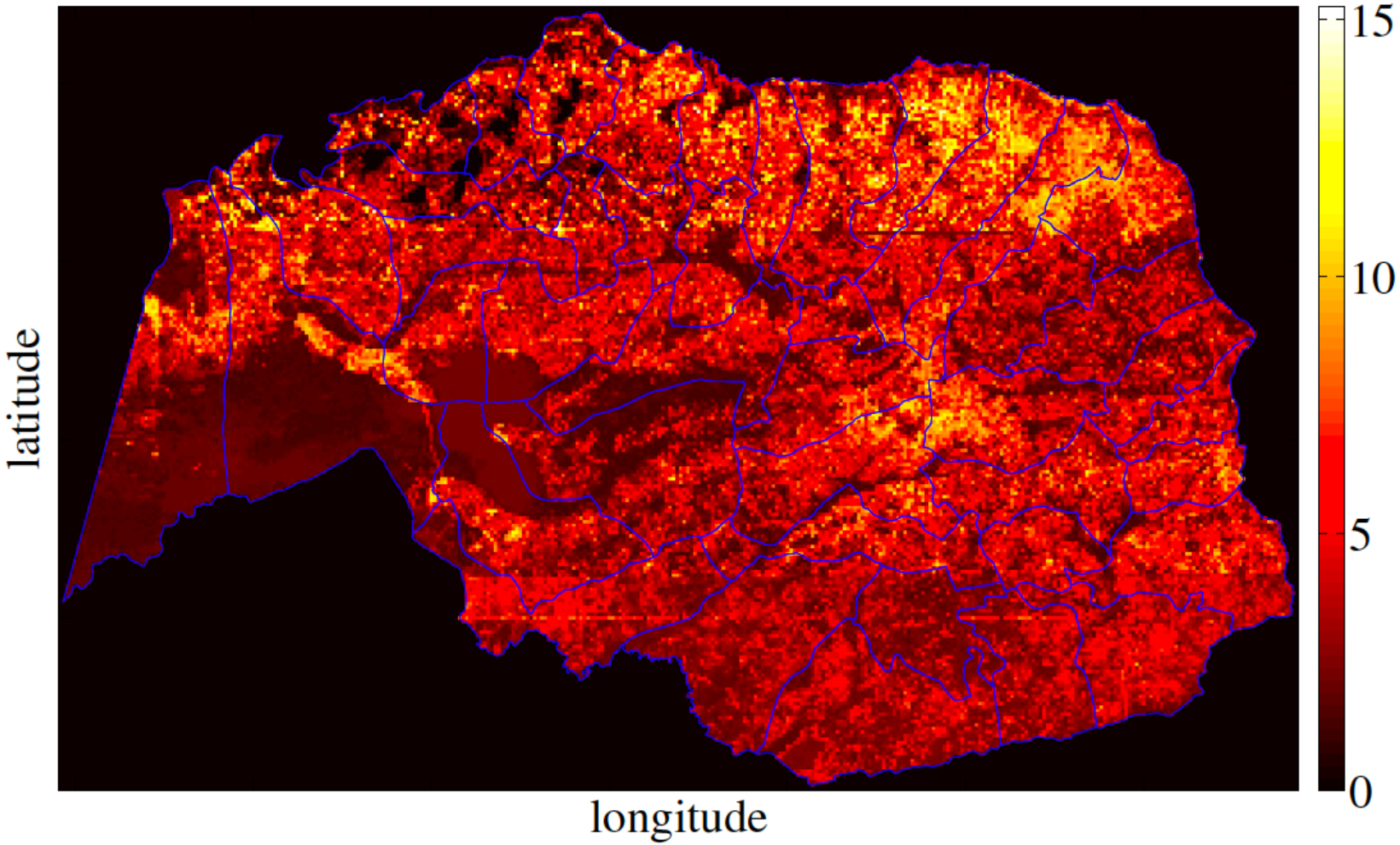


Figure 10: Heat map of number of total estimated roofs per 400×400 pixel image in the region of interest.

Abelson et al 2014

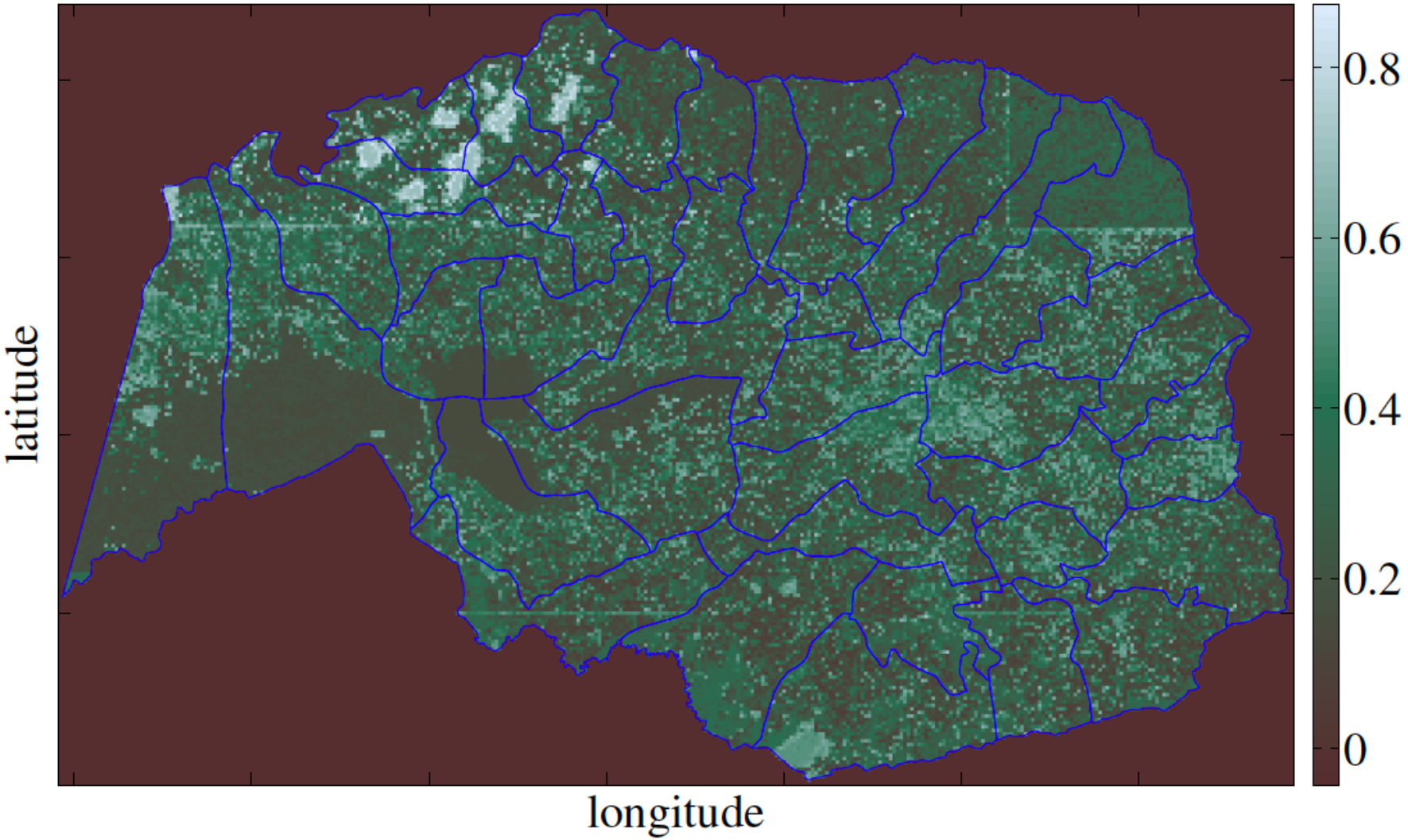


Figure 11: Heat map of proportion of roofs that are metal in the region of interest.

Abelson et al 2014

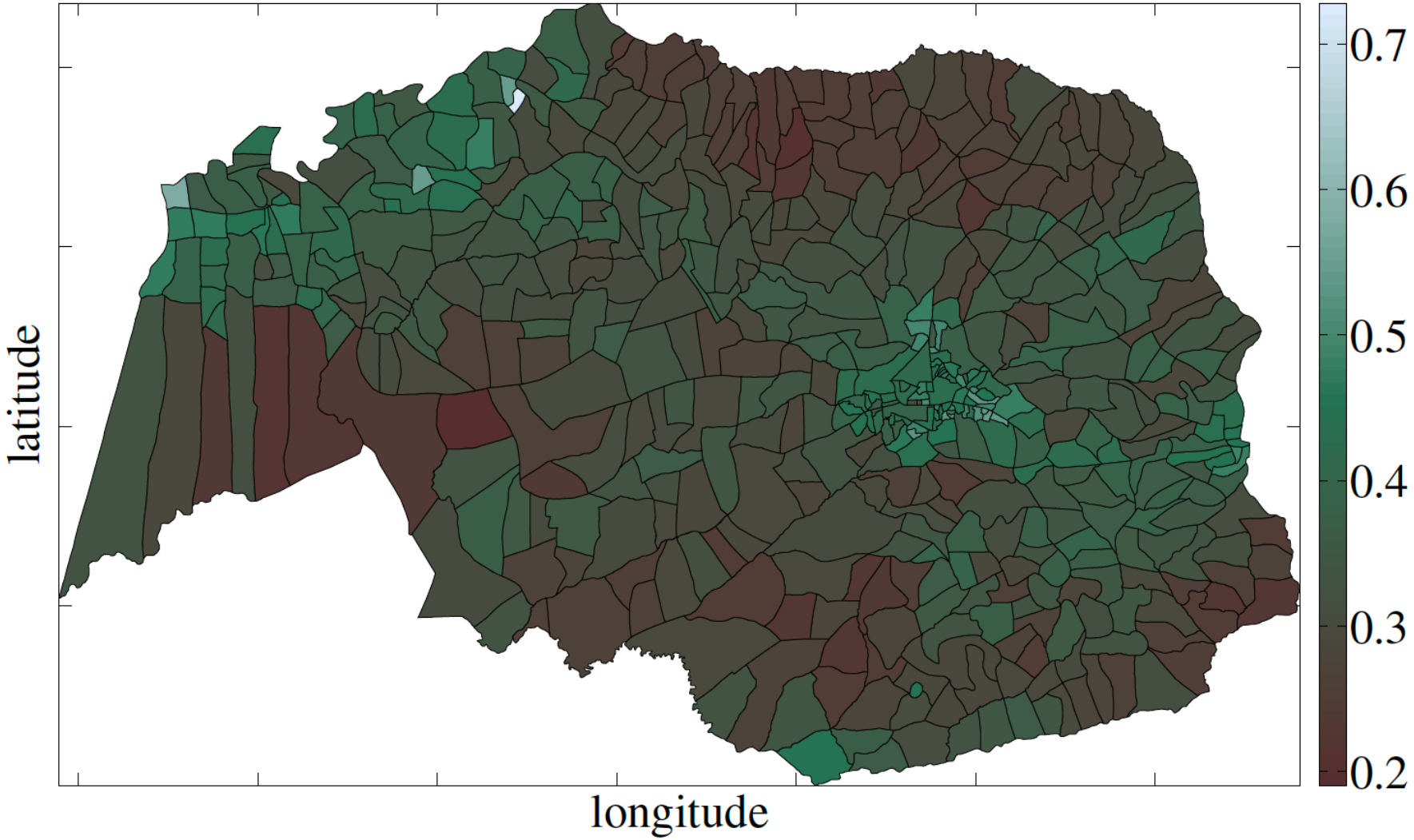


Figure 12: Estimated proportion of metal roofs in villages in the region of interest.

Outline

Examples of Data & Technology in Development

- Abelson, Varshney & Sun (2014): *Targeting Direct Cash Transfers to the Extremely Poor*
- Blumenstock, Cadamuro & On (Science 2015) *Predicting Poverty and Wealth from Mobile Phone Metadata*
- Jean, Burke, Xie, Davis, Lobell & Ermon *Combining Satellite Imagery and Machine Learning to Predict Poverty*

Blumenstock et al 2015

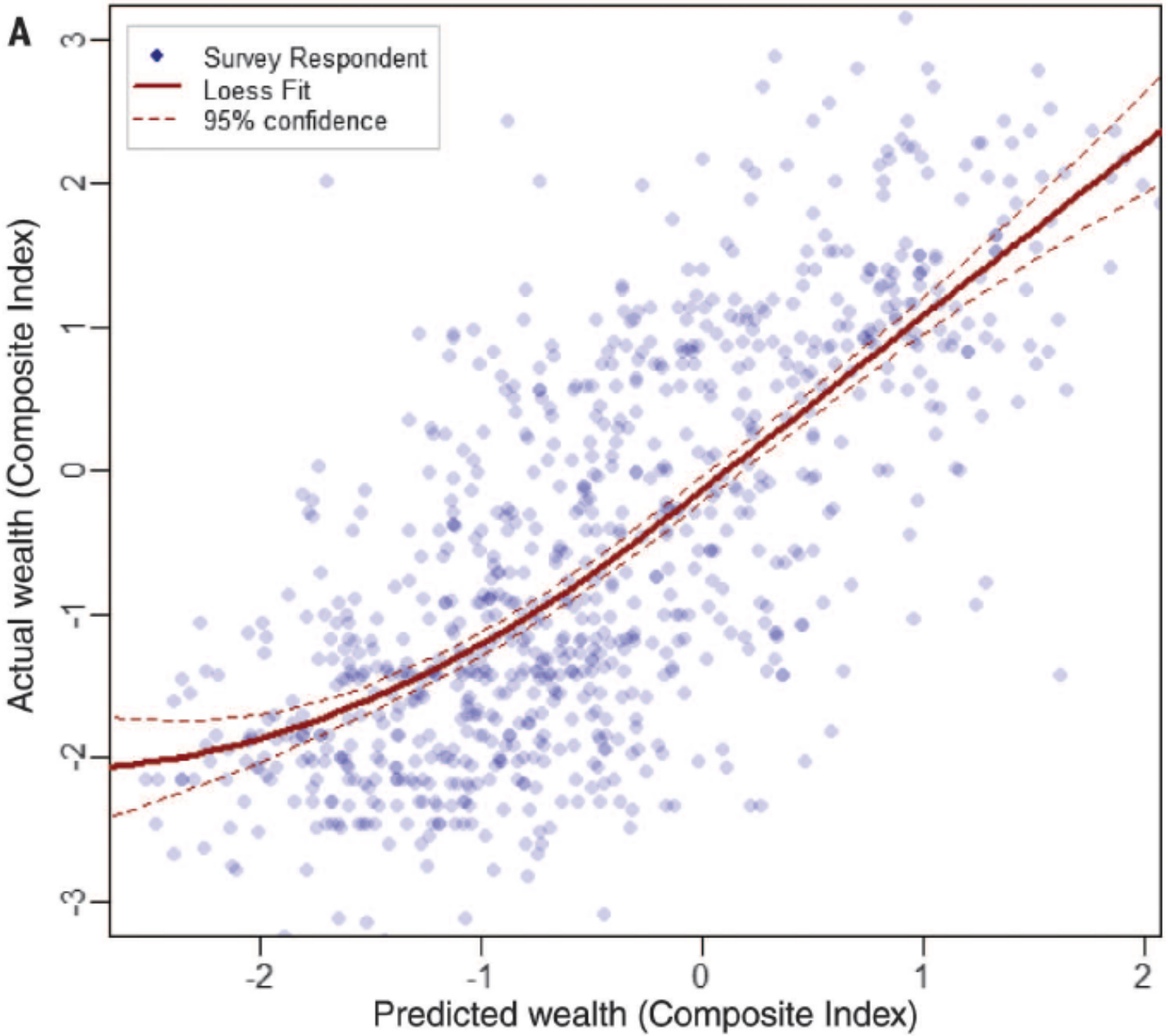
- In many countries, survey data is either completely missing, or unreliable (national stats shown to be off by as much as 50%)
- This paper will try to use metadata from mobile phone users to predict the user's socioeconomic characteristics.
  - Very fine-grained prediction
  - Usedul for applications that require individual-level information (targeting, policy interventions etc)
- Use data on billions of interactions over Rwanda's largest mobile phone network and a follow-up phone survey of 856 individual subscribers.

Blumenstock et al 2015

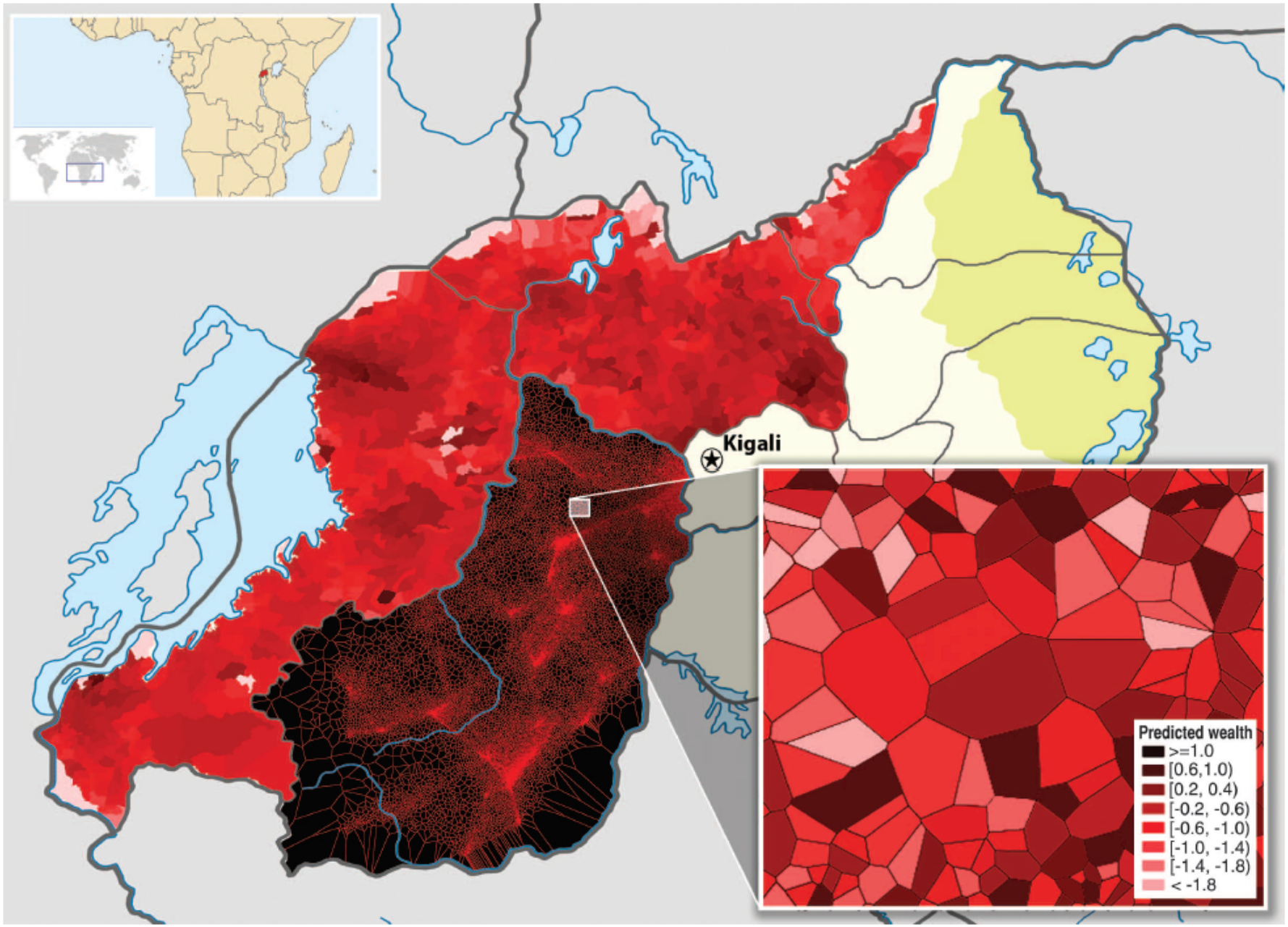
Table 1. Summary statistics for primary data sets. Phone survey data were collected by the authors in Kigali, in collaboration with the Kigali Institute of Science and Technology. Call detail records were collected by the primary mobile phone operator in Rwanda at the time of the phone survey. Demographic and Health Survey (DHS) data were collected by the Rwandan National Institute of Statistics. N/A, not applicable.

Summary statistic	Phone survey	Call detail records	DHS (2007)	DHS (2010)
Number of unique individuals	856	1.5 million	7377	12,792
Data collection period	July 2009	May 2008–May 2009	Dec. 2007–Apr. 2008	Sept. 2010–Mar. 2011
Number of questions in survey	75	N/A	1615	3396
Primary geographic units	30 districts	30 districts	30 districts	30 districts
Secondary geographic units	300 cell towers	300 cell towers	247 clusters	492 clusters

Blumenstock et al 2015



Blumenstock et al 2015



Blumenstock et al 2015

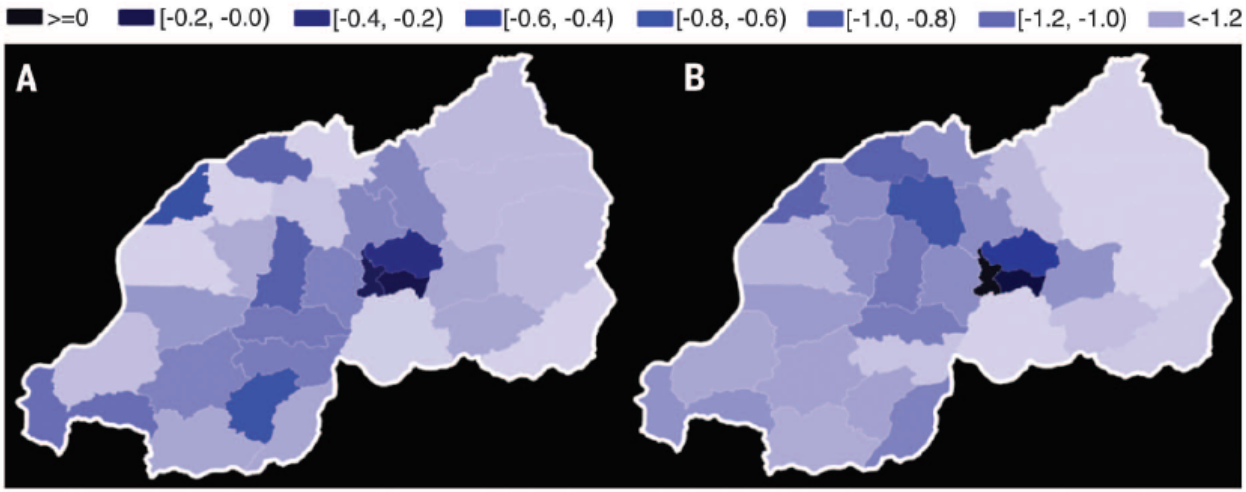
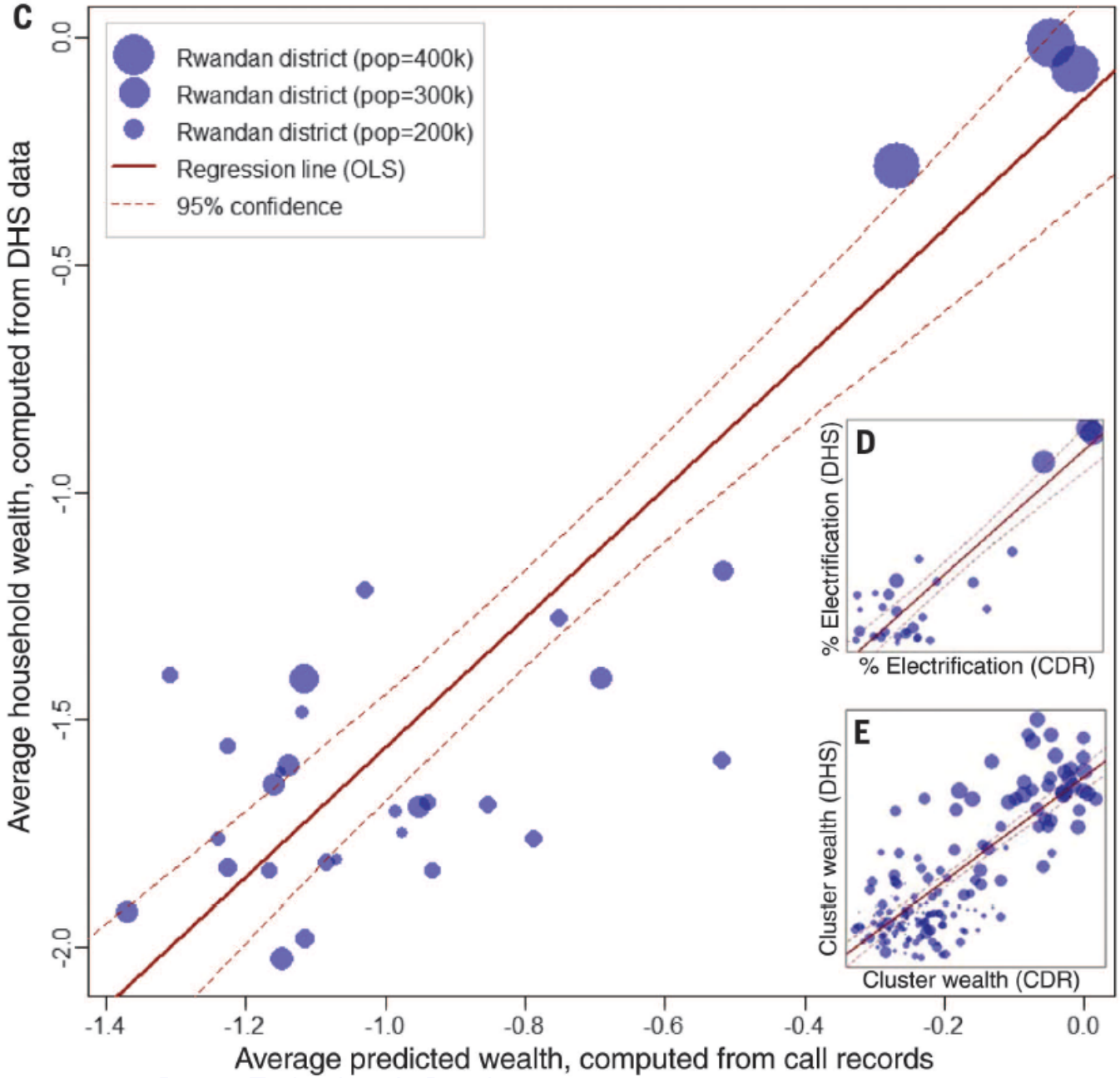


Fig. 3. Comparison of wealth predictions to government survey data. (**A**) Predicted composite wealth index (district average), computed from 2009 call data and aggregated by administrative district. (**B**) Actual composite wealth index (district average), as computed from a 2010 government DHS of 12,792 households. (**C**) Comparison of actual and predicted district wealth, for each of the 30 districts, with dots sized by population. (**D**) Comparison of actual and predicted rates of electrification, for each of the 30 districts. (**E**) Comparison of actual and predicted cluster wealth, for each of the 492 DHS clusters. CDR, call detail records.

Blumenstock et al 2015



Outline

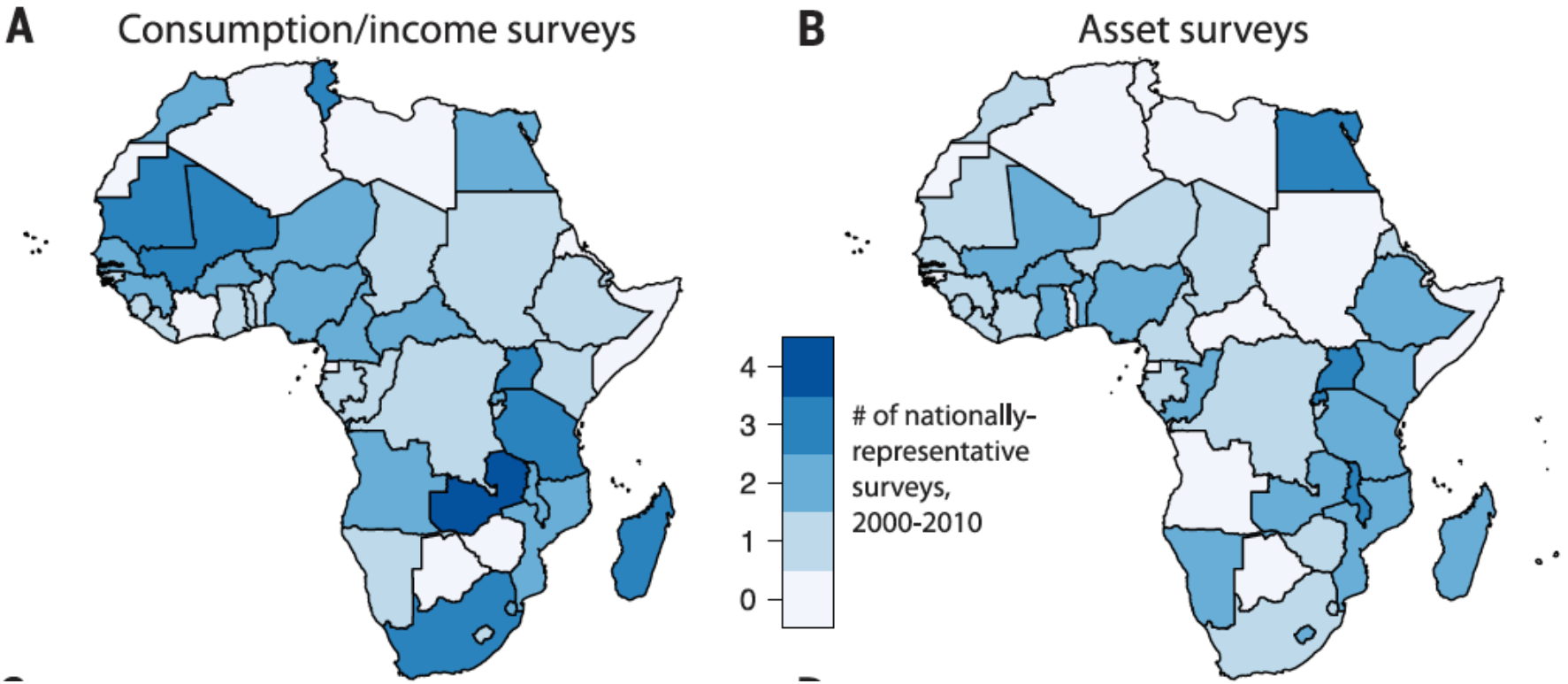
Examples of Data & Technology in Development

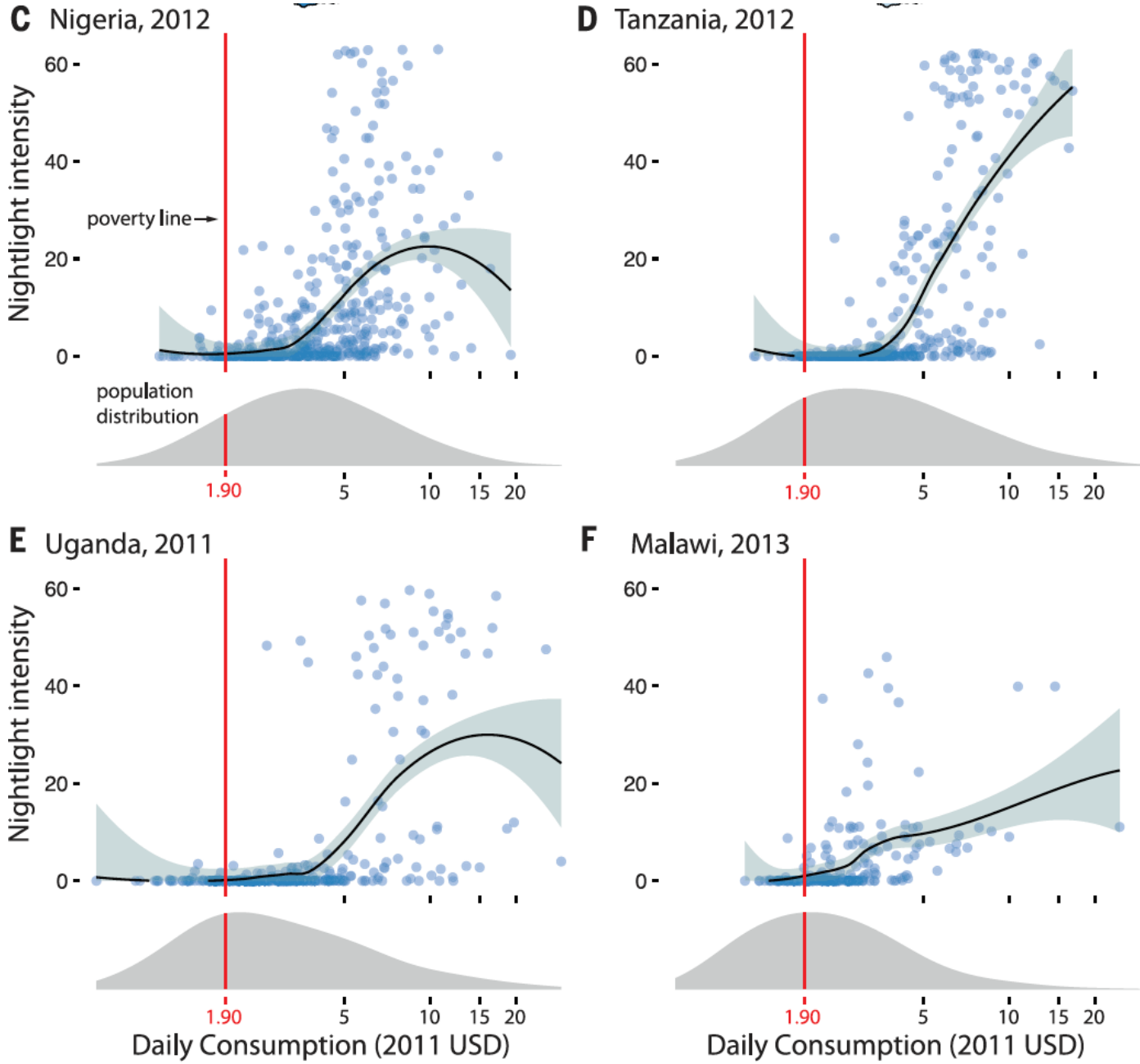
- Abelson, Varshney & Sun (2014): *Targeting Direct Cash Transfers to the Extremely Poor*
- Blumenstock, Cadamuro & On (Science 2015) *Predicting Poverty and Wealth from Mobile Phone Metadata*
- Jean, Burke, Xie, Davis, Lobell & Ermon *Combining Satellite Imagery and Machine Learning to Predict Poverty*

Jean et al 2016

- Focus on 5 African countries: Nigeria, Tanzania, Uganda, Malawi & Rwanda.
- One popular approach in data-scarce environments has been to use nightlights in satellite images.
- Combine satellite data with:
  - World Bank Living Standards Measurement Study (LSMS) surveys: measures expenditure
  - Demographic and Health Surveys (DHS): measures welath.
- Proceed in 3 steps
  1. Train a Convolutional Neural Network (CNN) on ImageNet to identify low-level features of images common to many image classification tasks
  2. Fine-tune the CNN by getting it to predict nightlights using daytime satellite images. Reduce the dimensionality of the daytime pictures to the features that tend to predict nightlights: a coarse measure of wellbeing
  3. Use ridge regession of survey data on cluster-level image features to predict expenditure and wealth.

Jean et al 2016




C Nigeria, 2012
D Tanzania, 2012
E Uganda, 2011
F Malawi, 2013
Nightlight intensity
poverty line
population distribution
1.90
5
10
15
20
0
20
40
60
Daily Consumption (2011 USD)

Jean et al 2016

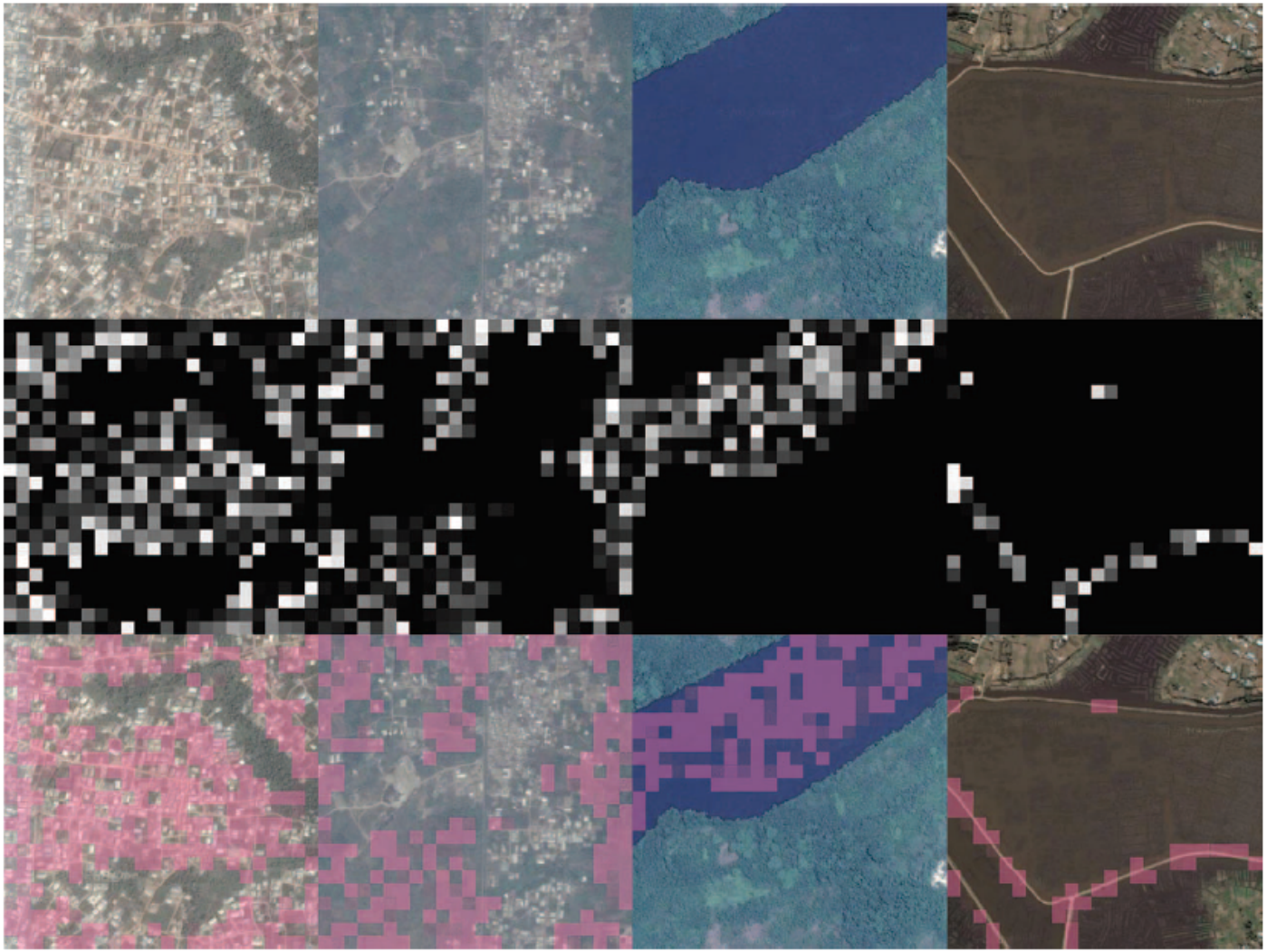
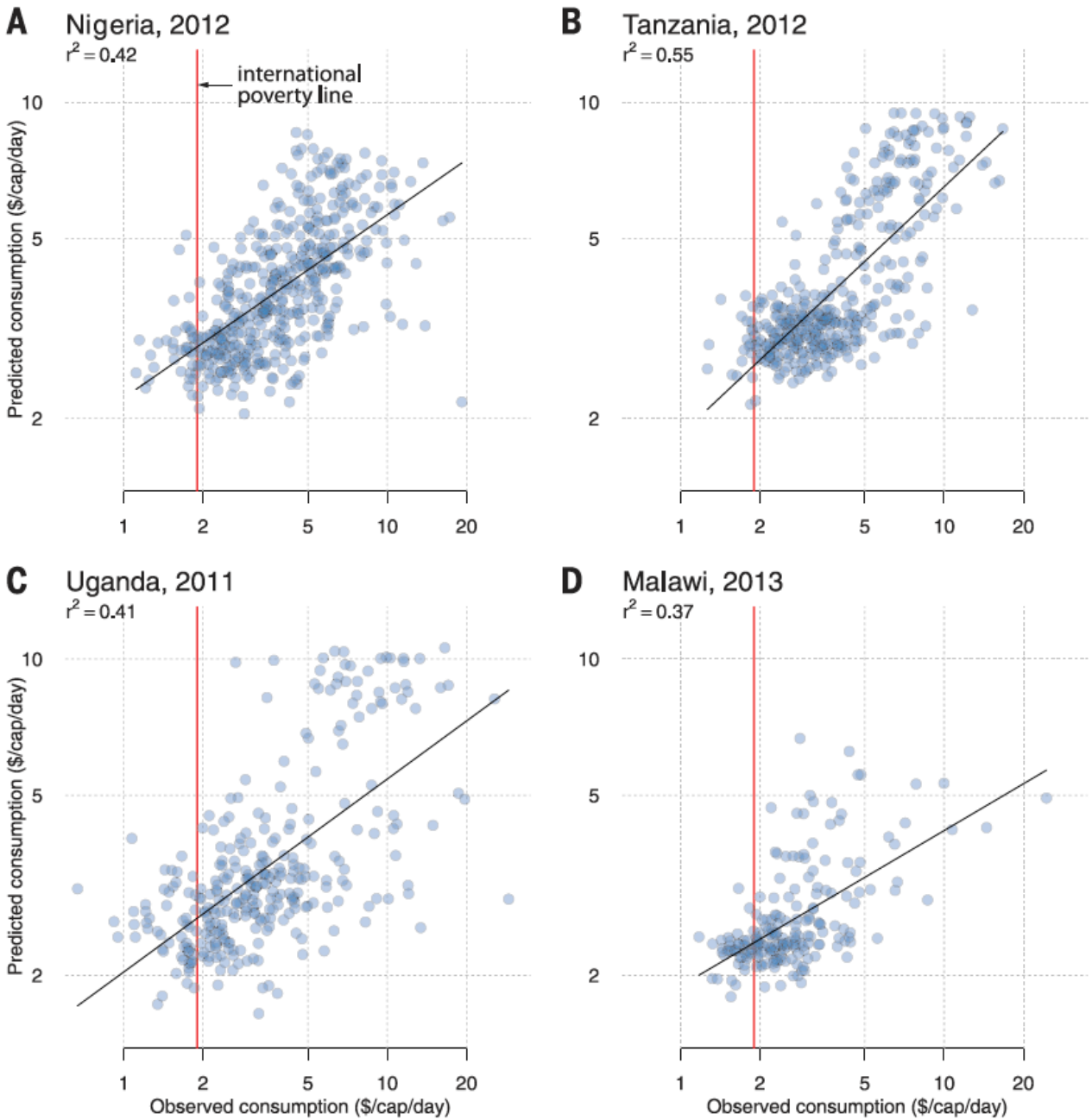


Fig. 2. Visualization of features. By column: Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, nonurban areas, water, and roads) in the convolutional neural network model used for extracting features. Each filter "highlights" the parts of the image that activate it, shown in pink. By row: Original daytime satellite images from Google Static Maps, filter activation maps, and overlay of activation maps onto original images

Jean et al 2016



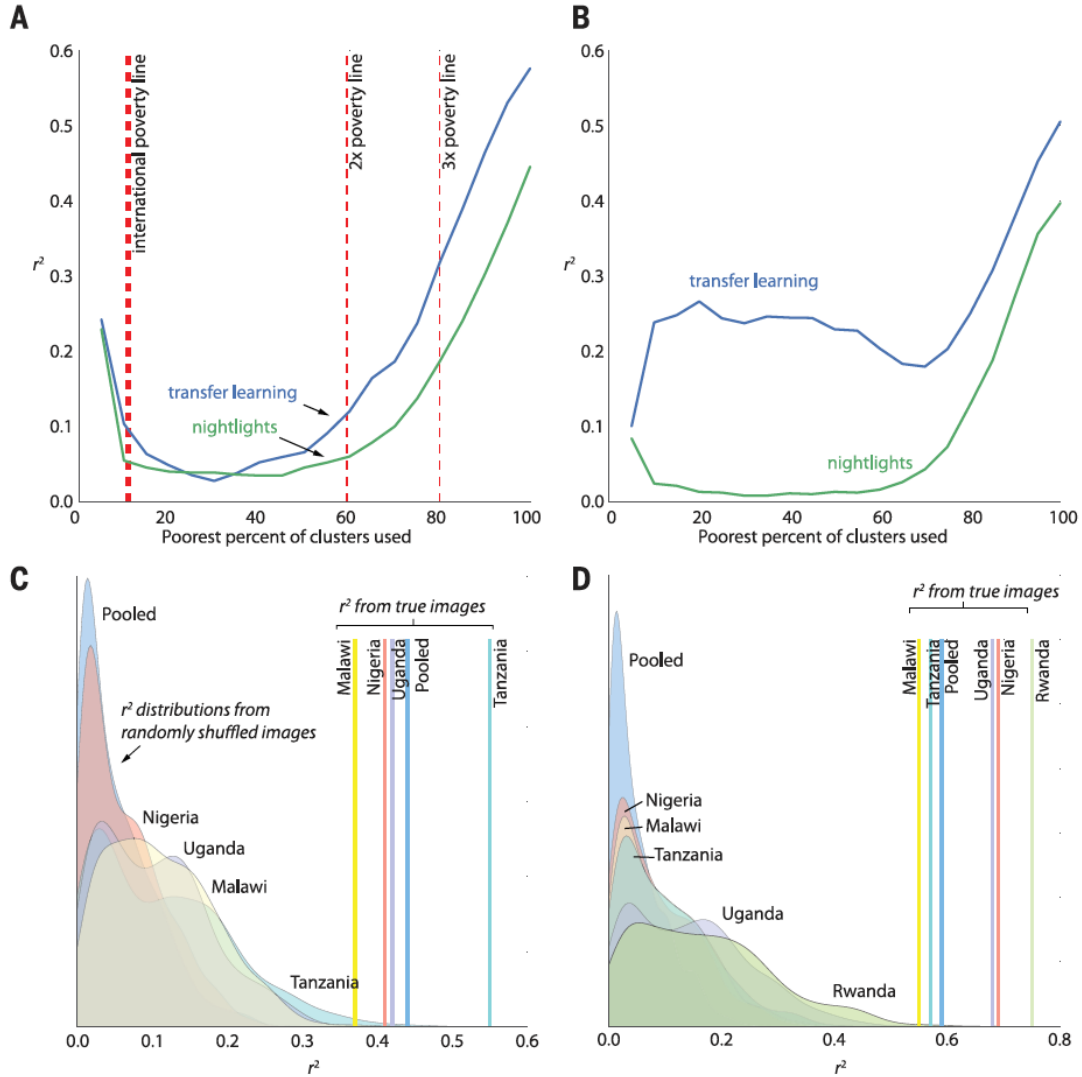


Fig. 4. Evaluation of model performance. (**A**) Performance of transfer learning model relative to nightlights for estimating consumption, using pooled observations across the four LSMS countries. Trials were run separately for increasing percentages of the available clusters (e.g., *x*-axis value of 40 indicates that all clusters below 40th percentile in consumption were included). Vertical red lines indicate various multiples of the international poverty line. Image features reduced to 100 dimensions using principal component analysis. (**B**) Same as (A), but for assets. (**C**) Comparison of r^2 of models trained on correctly assigned images in each country (vertical lines) to the distribution of r^2 values obtained from trials in which the model was trained on randomly shuffled images (1000 trials per country). (**D**) Same as (C), but for assets. Cross-validated r^2 values are reported in all panels.

Papers

- Banerjee et al ID cards
- Muralidharan et al