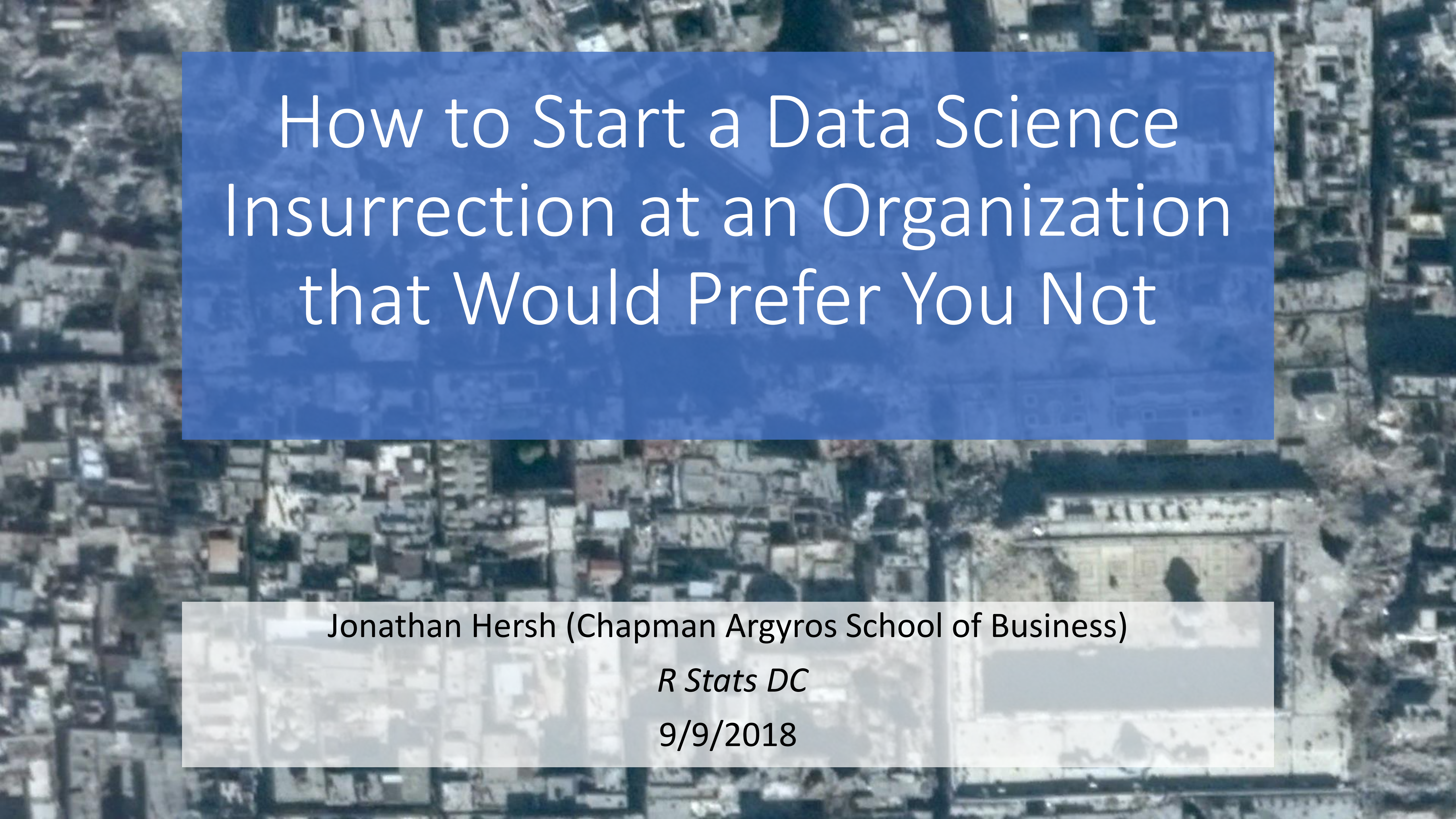


How to Start a Data Science Insurrection at an Organization that Would Prefer You Not

Jonathan Hersh (Chapman Argyros School of Business)

R Stats DC

9/9/2018

(Four Tips for Implementing Machine Learning Projects that Scale within Organizations)

Jonathan Hersh (Chapman Argyros School of Business)

R Stats DC

9/9/2018

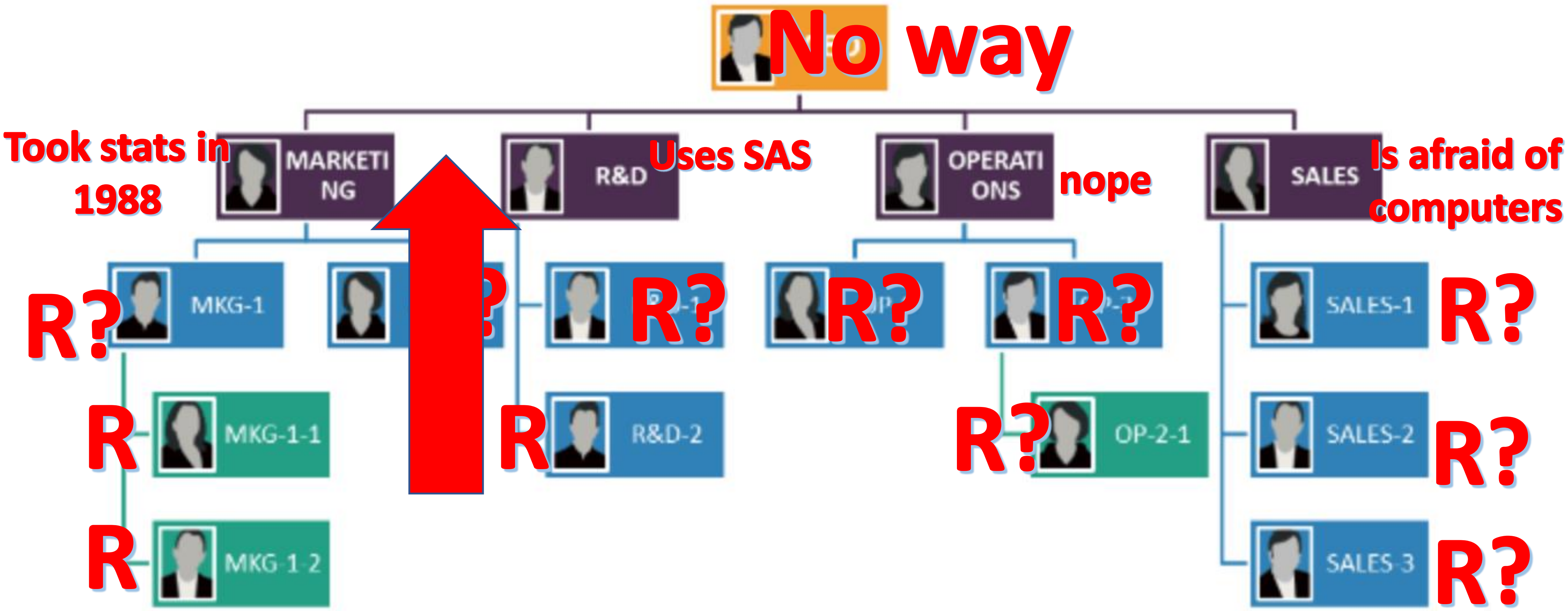
imgflip.com

When you think of machine learning

Not necessarily these organizations

THE WORLD BANK
United States Census Bureau
cfpb
Consumer Financial Protection Bureau
IDB
Inter-American Develo
UNITED STATES FEDERAL RESERVE SYSTEM
INTERNATIONAL MONETARY FUND

Who actually understands machine learning?



No way

Took stats in 1988

Uses SAS

nope

Is afraid of computers

R?

R

R

?

R?

R

R?

R?

R?

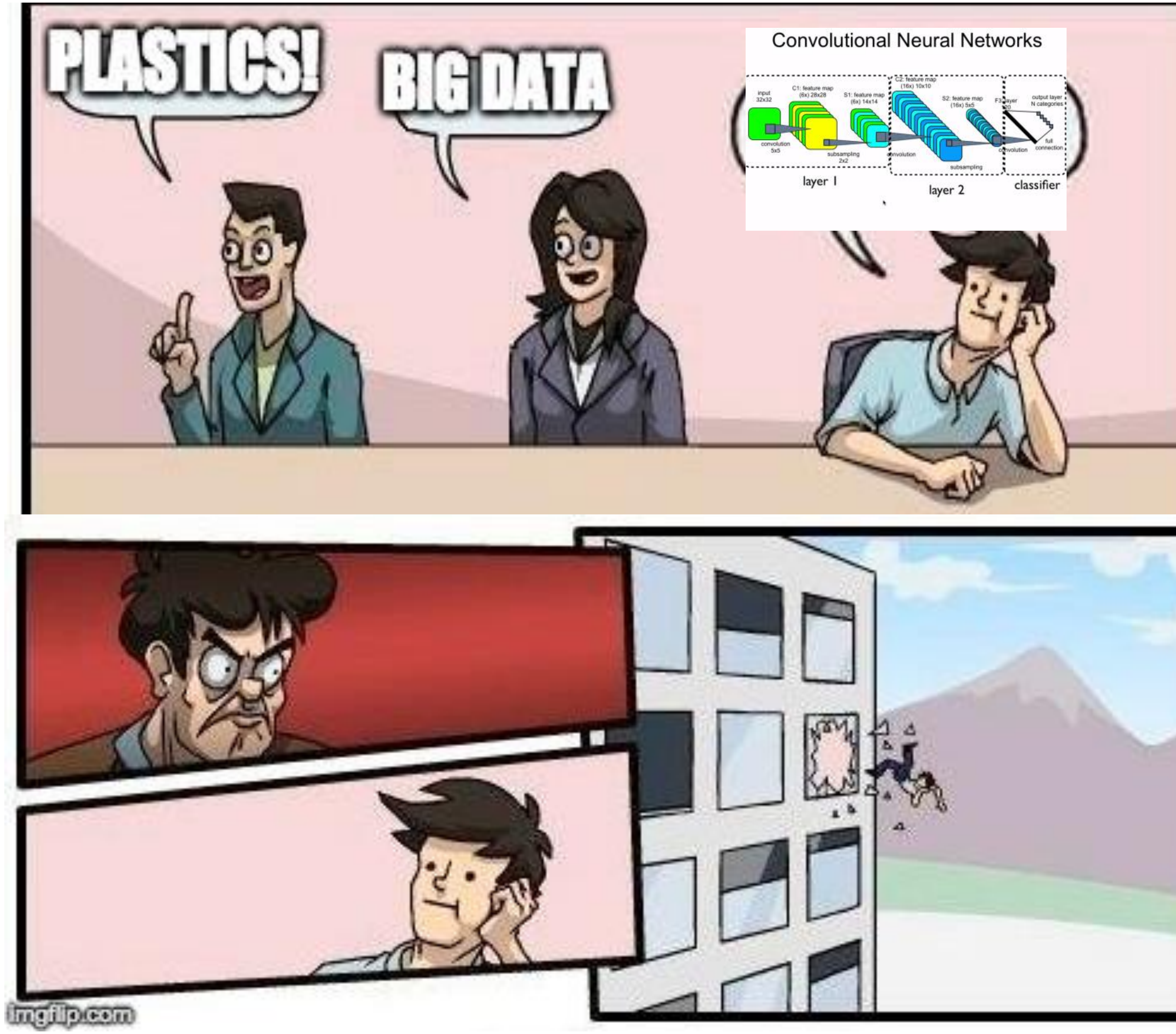
R?

R?

R?

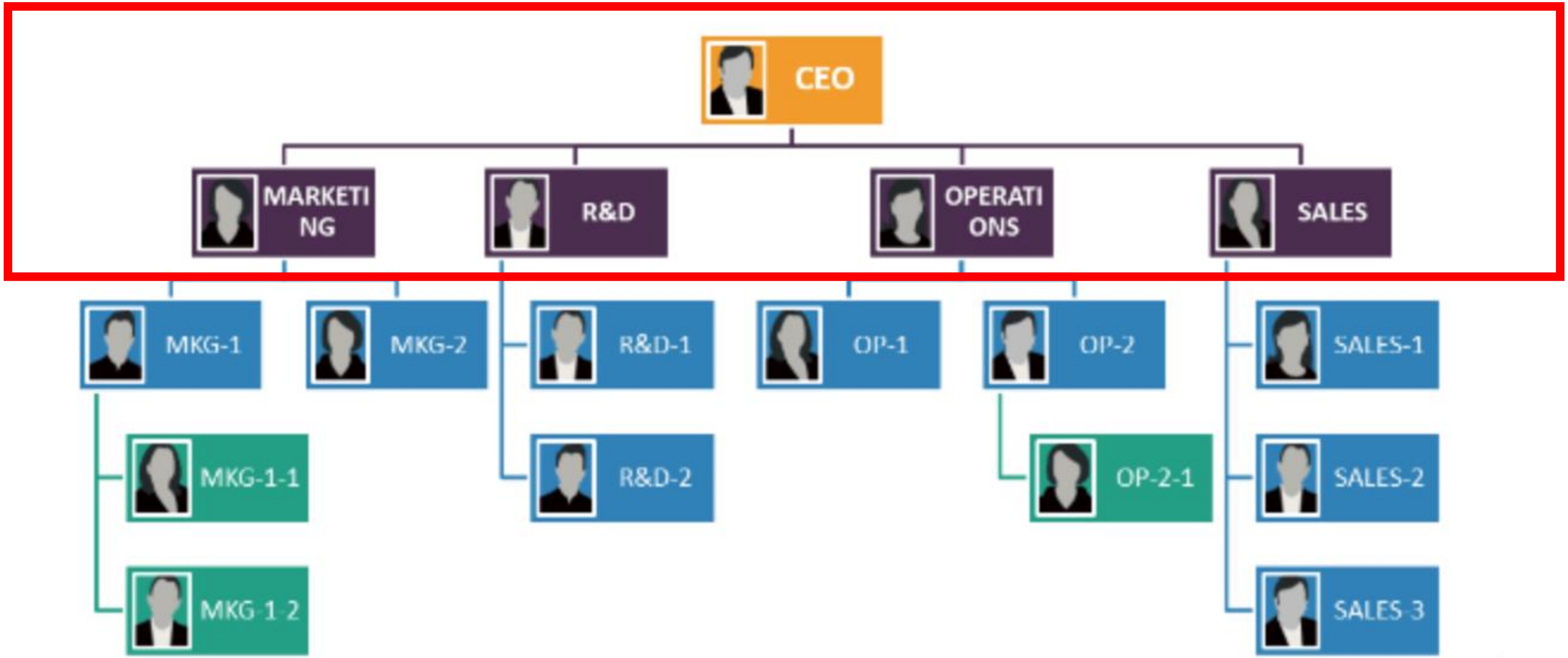
Tip 1:

Explain how your model works in language your boss's boss can understand

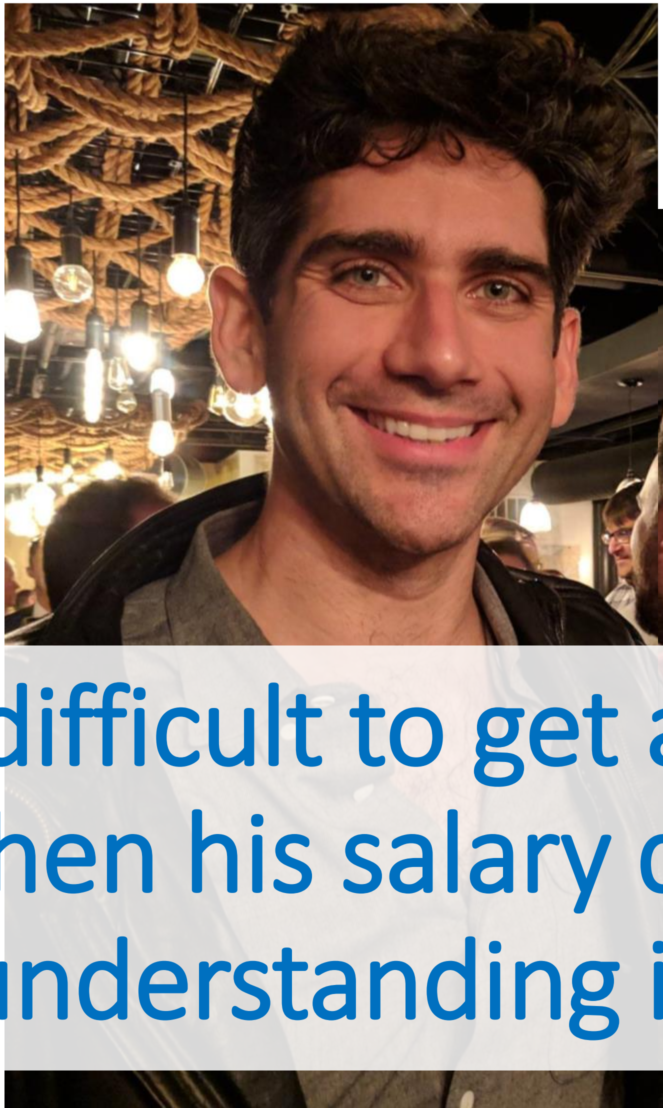

PLASTICS!
BIG DATA
Convolutional Neural Networks
input 32x32
C1: feature map (6x) 28x28
S1: feature map (6x) 14x14
convolution 5x5
subsampling 2x2
C2: feature map (16x) 10x10
S2: feature map (16x) 5x5
convolution
subsampling
F3: layer 120
output layer N categories
convolution
full connection
layer 1
layer 2
classifier
imgflip.com

Use *their* language to explain what you do

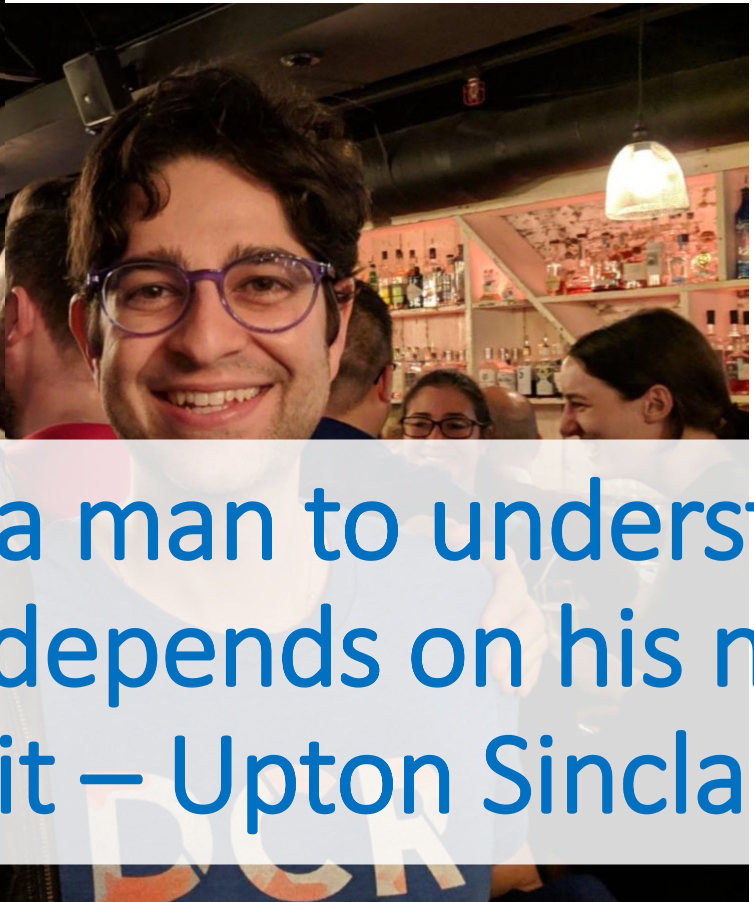
How did these people get to the top?



You
(new cool tools)



You boss
(old new tools)



It’s difficult to get a man to understand when his salary depends on his not understanding it – Upton Sinclair

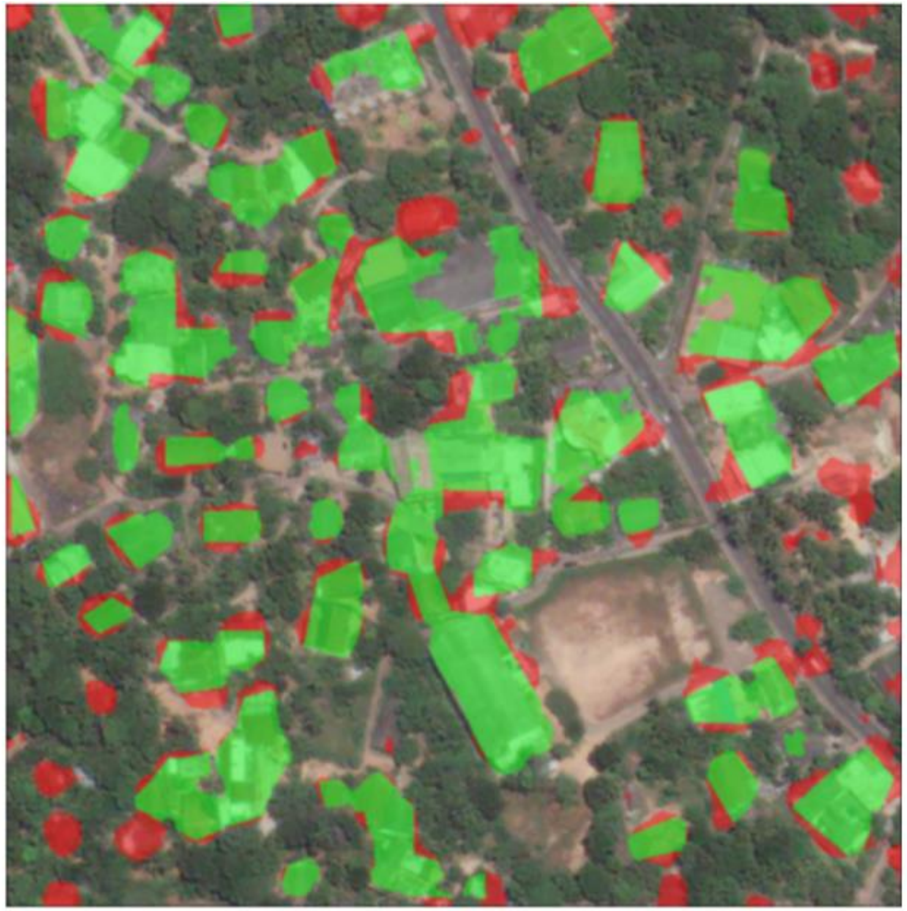
Tip 2:

Motivate machine learning projects by showing how current methods are insufficient

Previous Research (Engstrom, Hersh, Newhouse, 2017) Using Intermediate Features to Estimate Poverty in Sri Lanka

Satellite Image

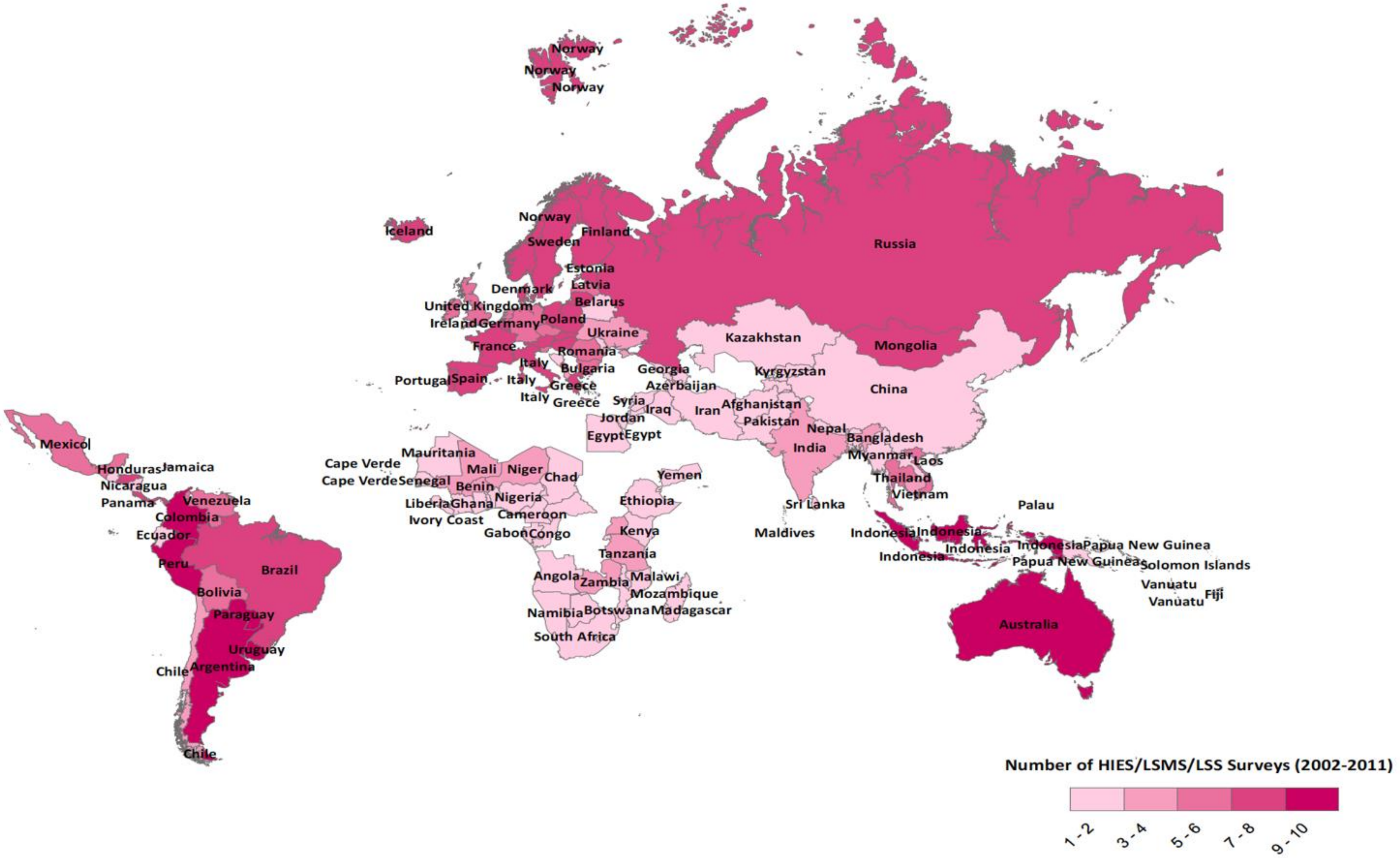
Buildings



How Most Data is Collected

57 Countries Have Zero or One Poverty Estimate 2002-2011

Number of Poverty Data Points, 2002 - 2011

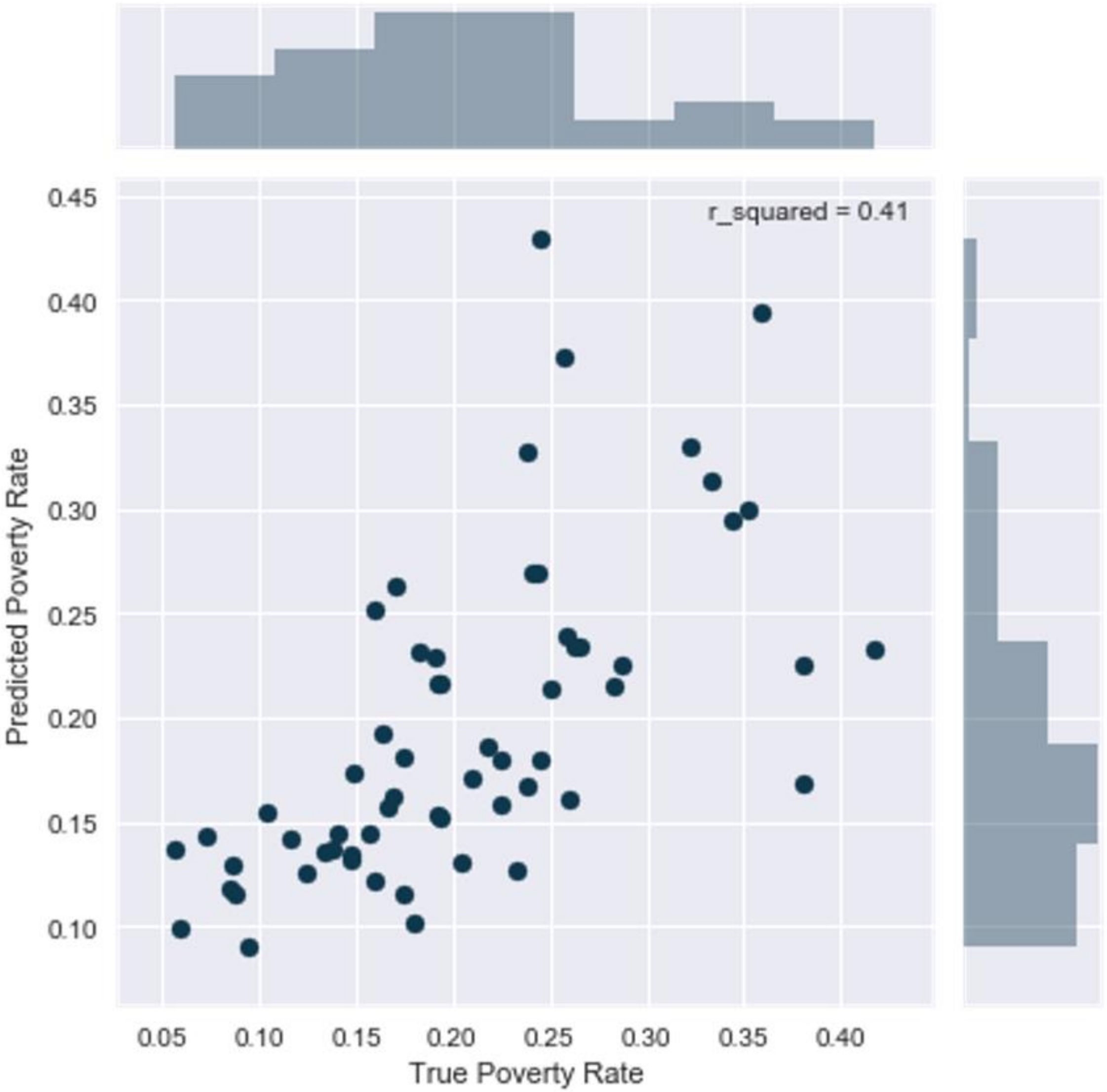


Tip 3:

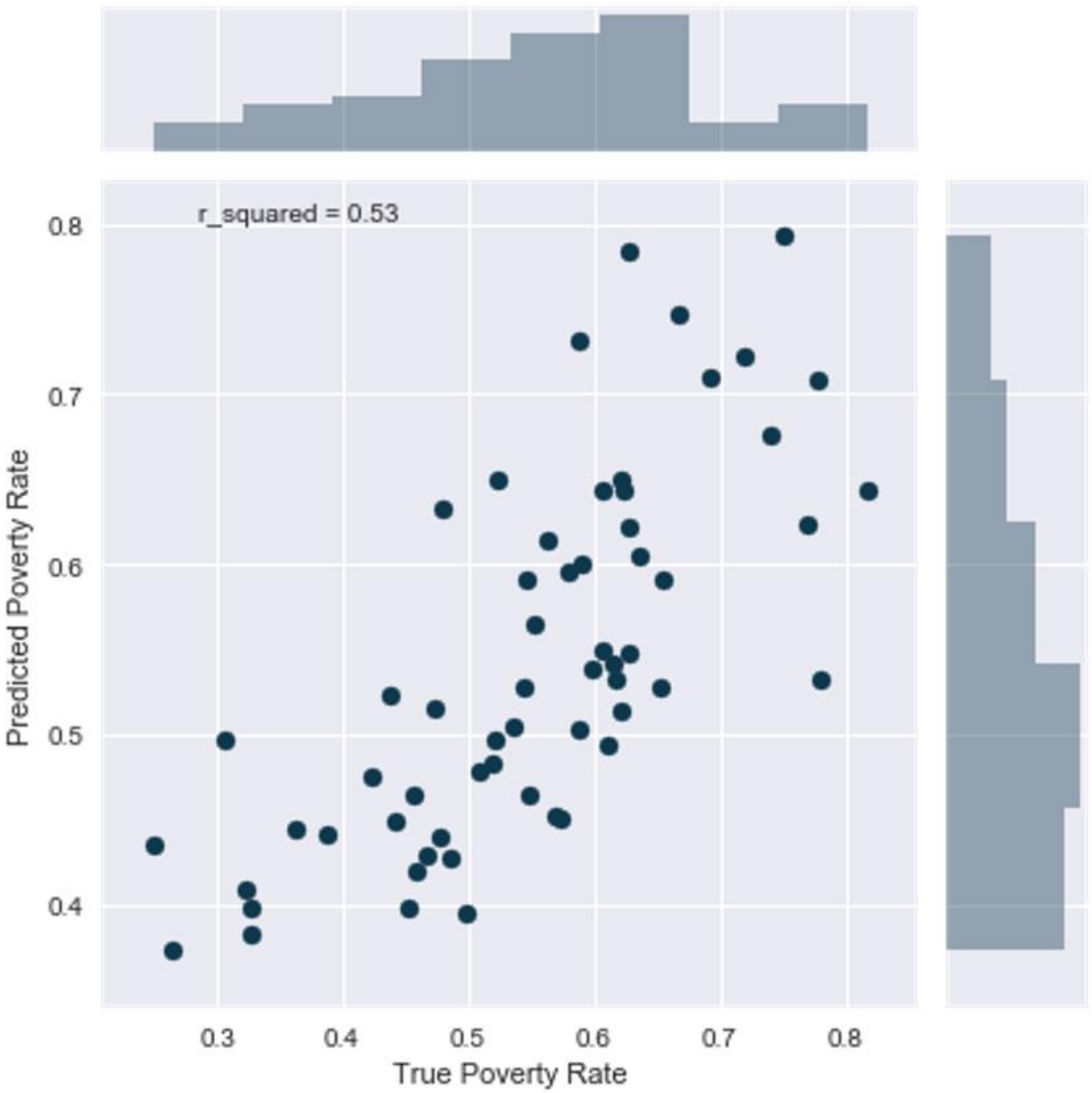
Be honest about the limitations of your methods

Urban Municipalities: Predicted vs True

Severe Poverty



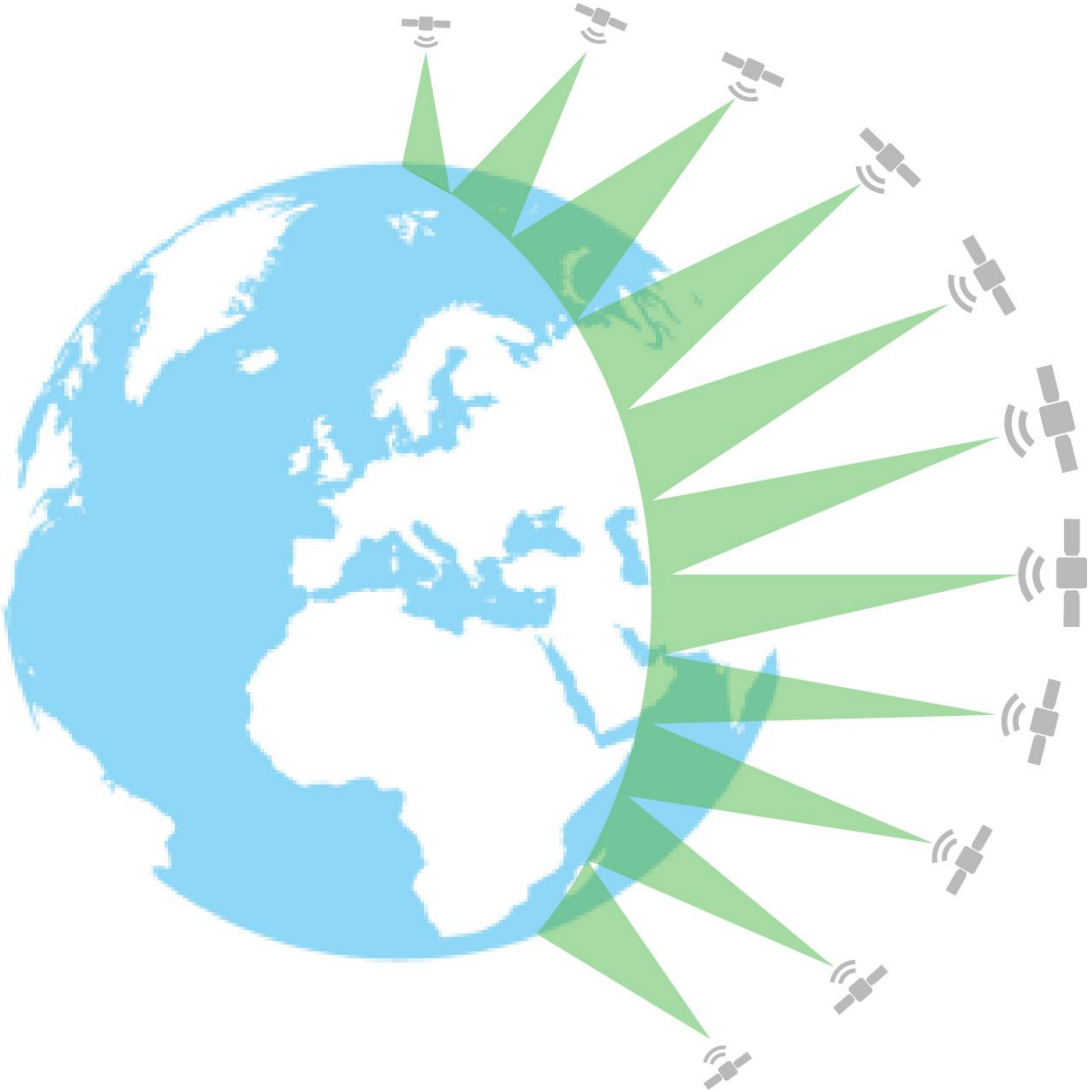
Moderate Poverty

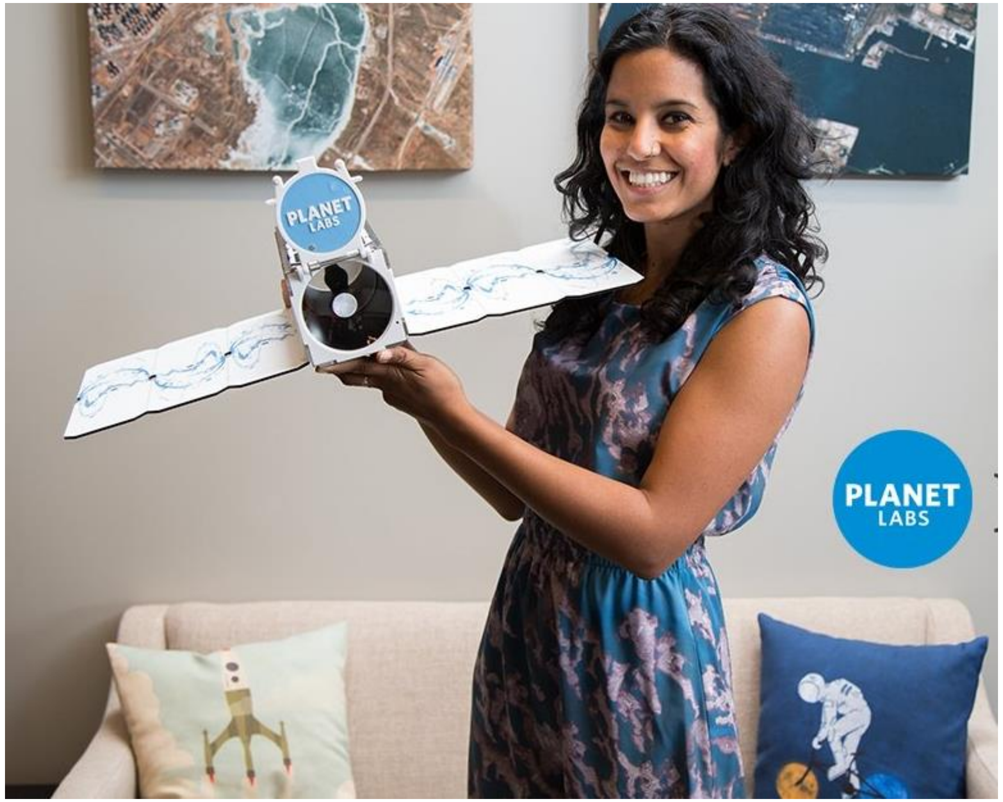


Tip 4:

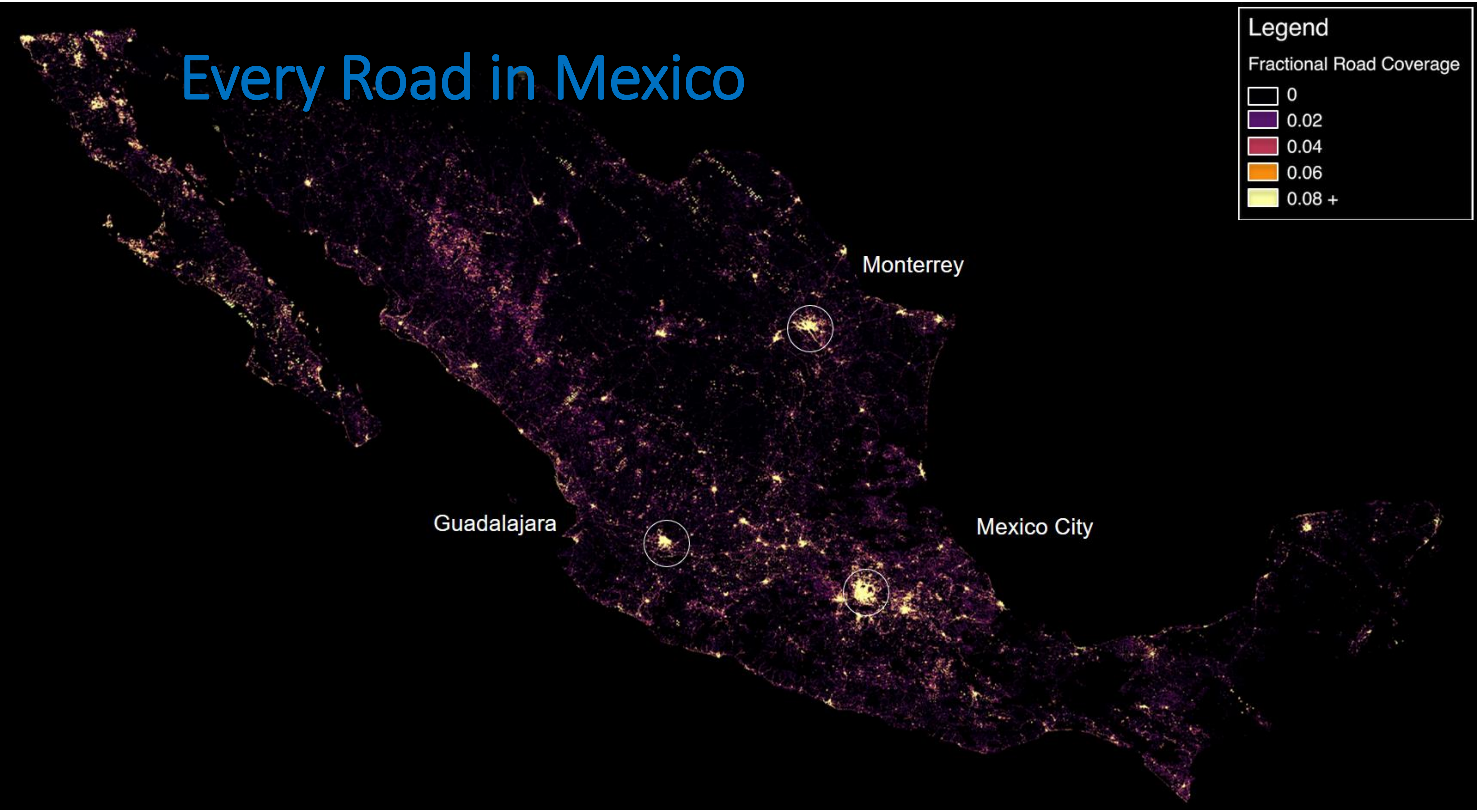
But always show the upside if it works out

“Micro-Satellites” ~ Daily Revisit Rate





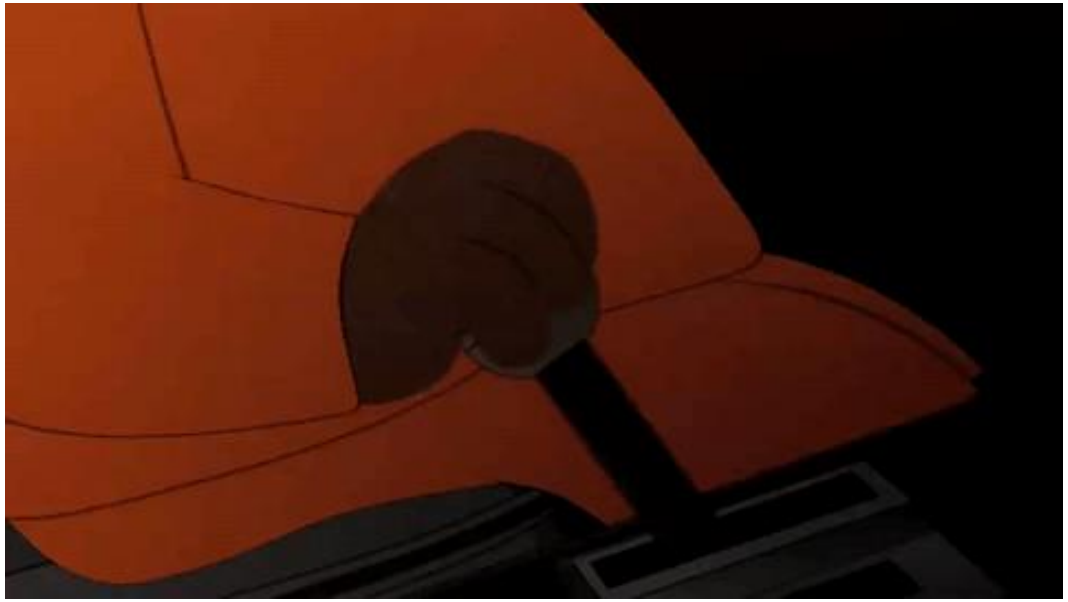
Every Road in Mexico



Shifting gears: do mobile phones cause car accidents?

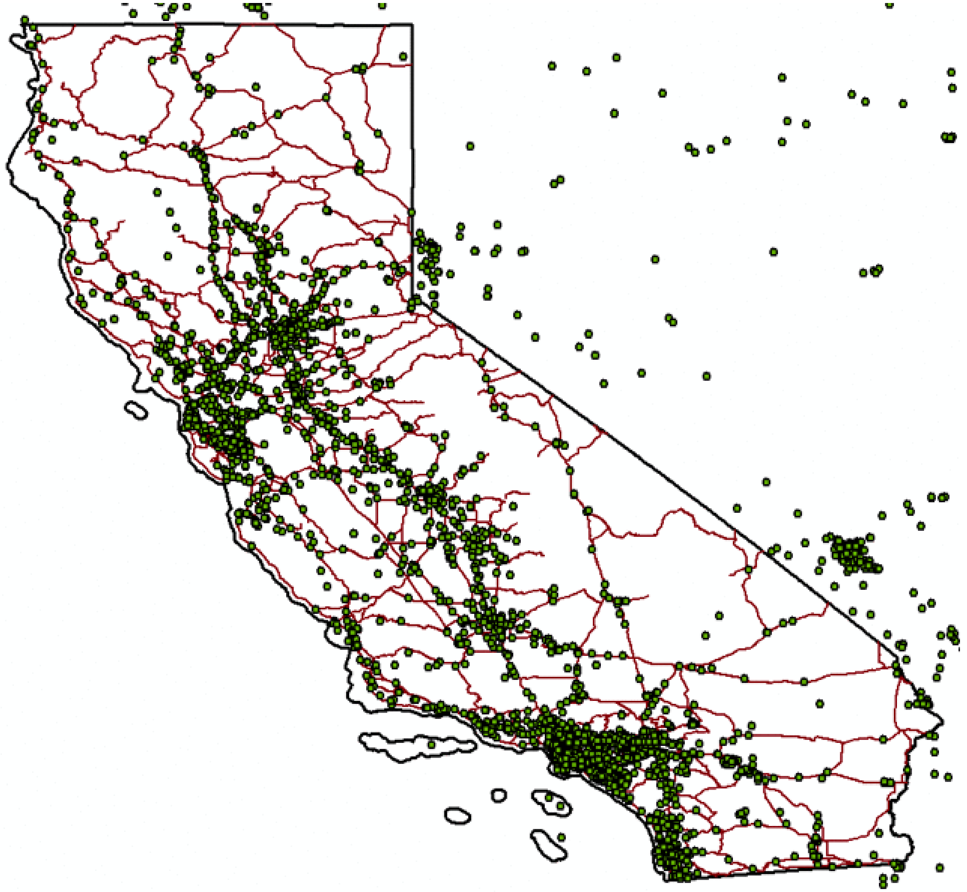
- Research Question: do internet enabled mobile phones increase traffic accidents?

- Joint work with Matt and Bree Lang at UC Riverside – "Car Accidents and 3G Coverage: New Evidence Using Cell Phone Towers"

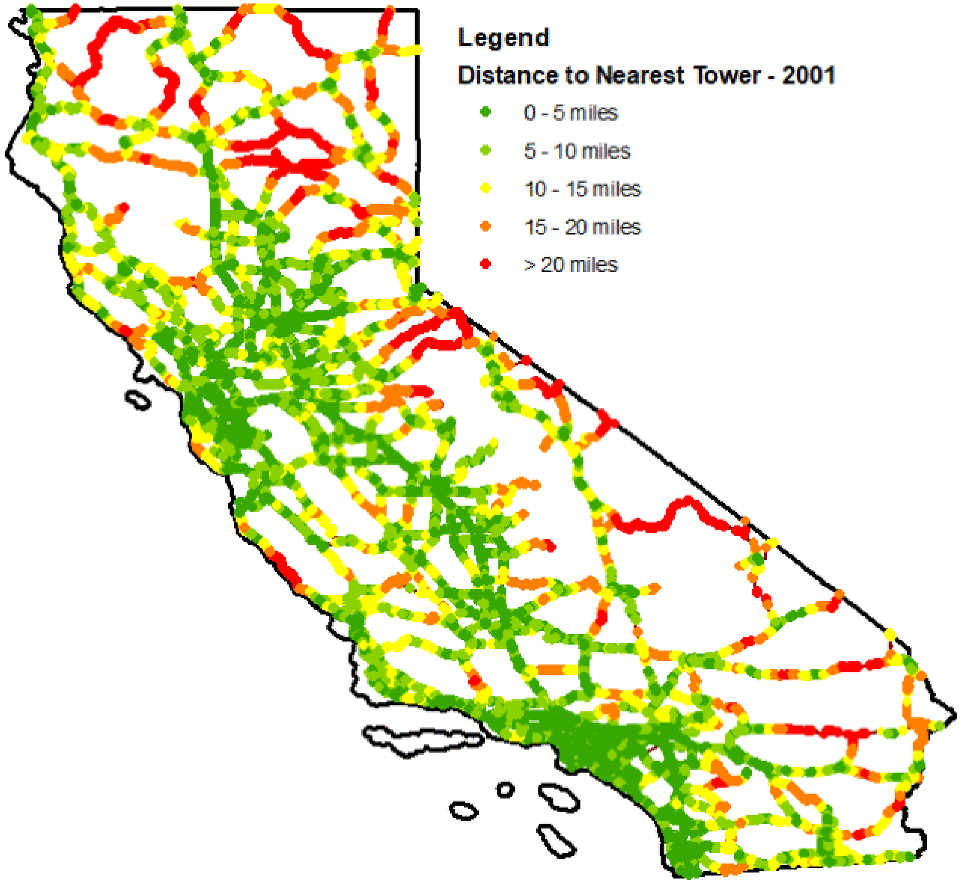


Event study: growth of 3G

Cell Tower Growth
2001 - 2011

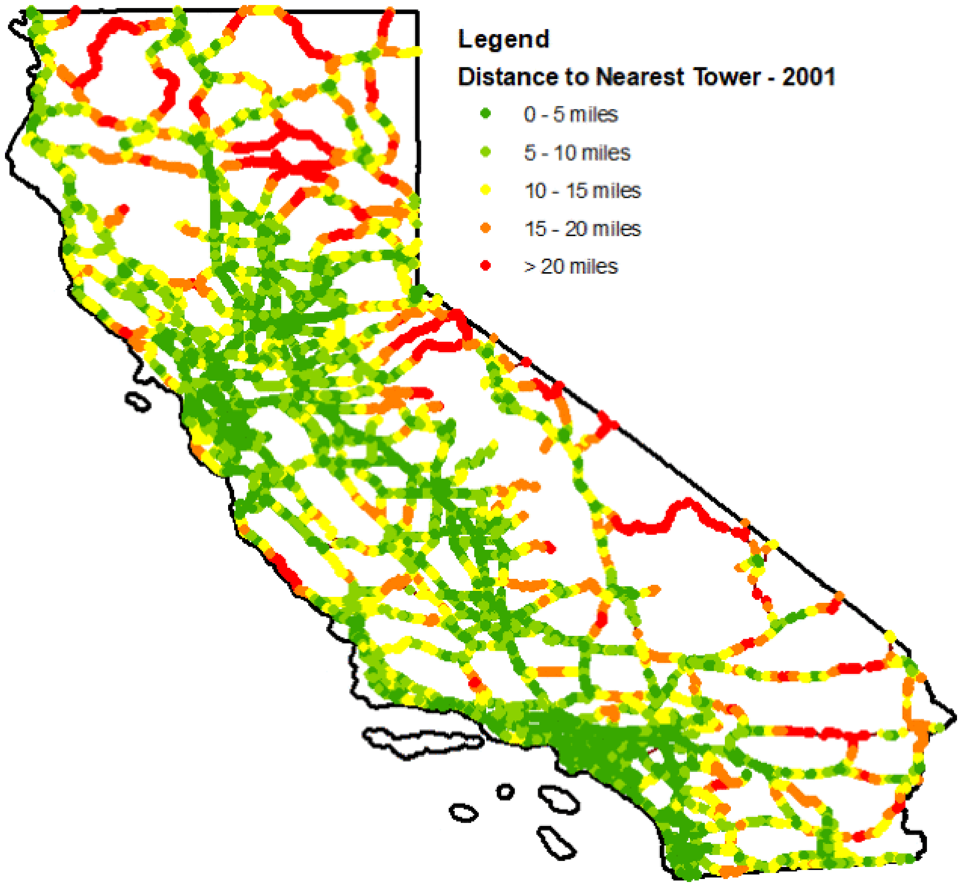


Road distance to
nearest tower
2001 - 2011

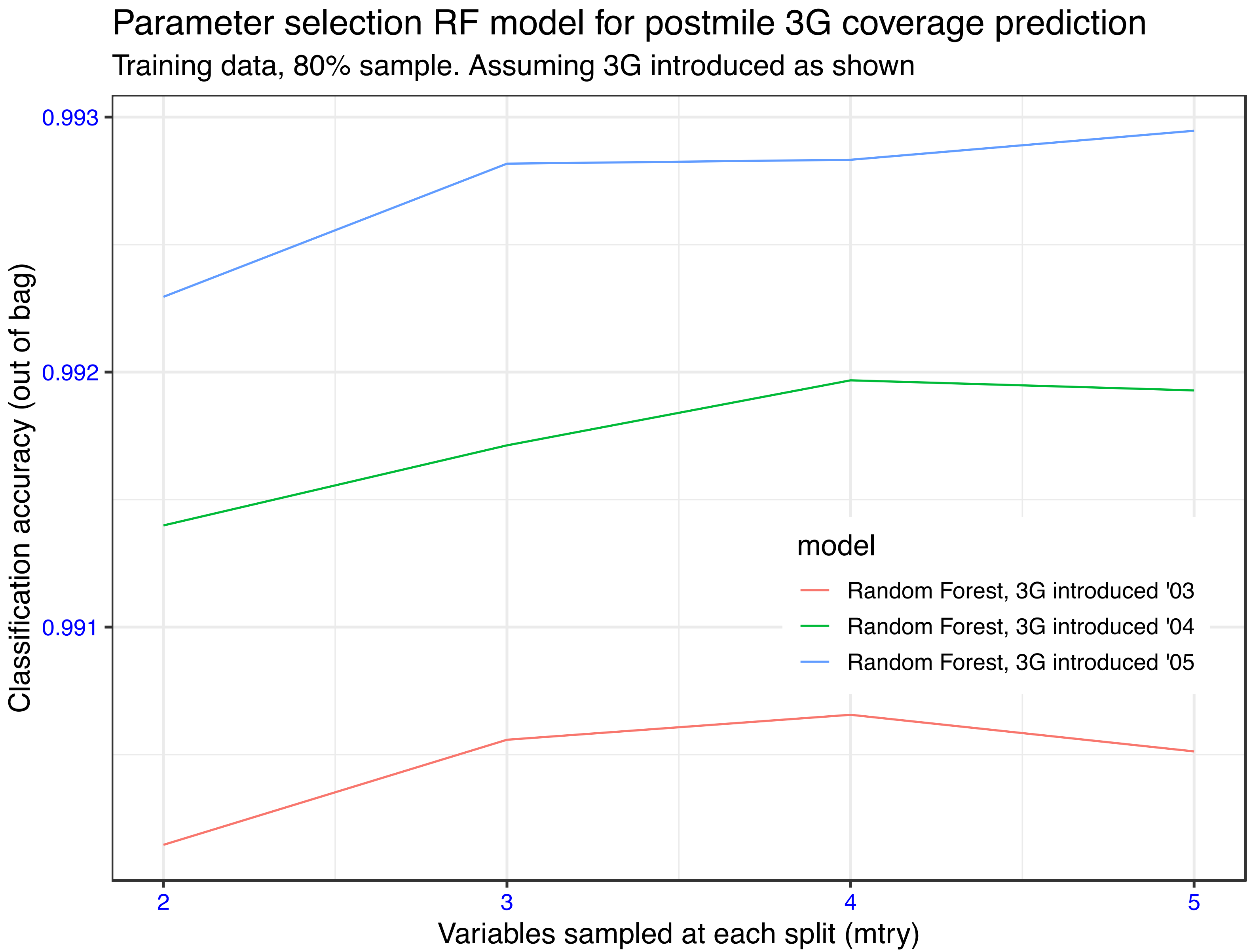


Problem: Only only know if a road has 3G access in 2016

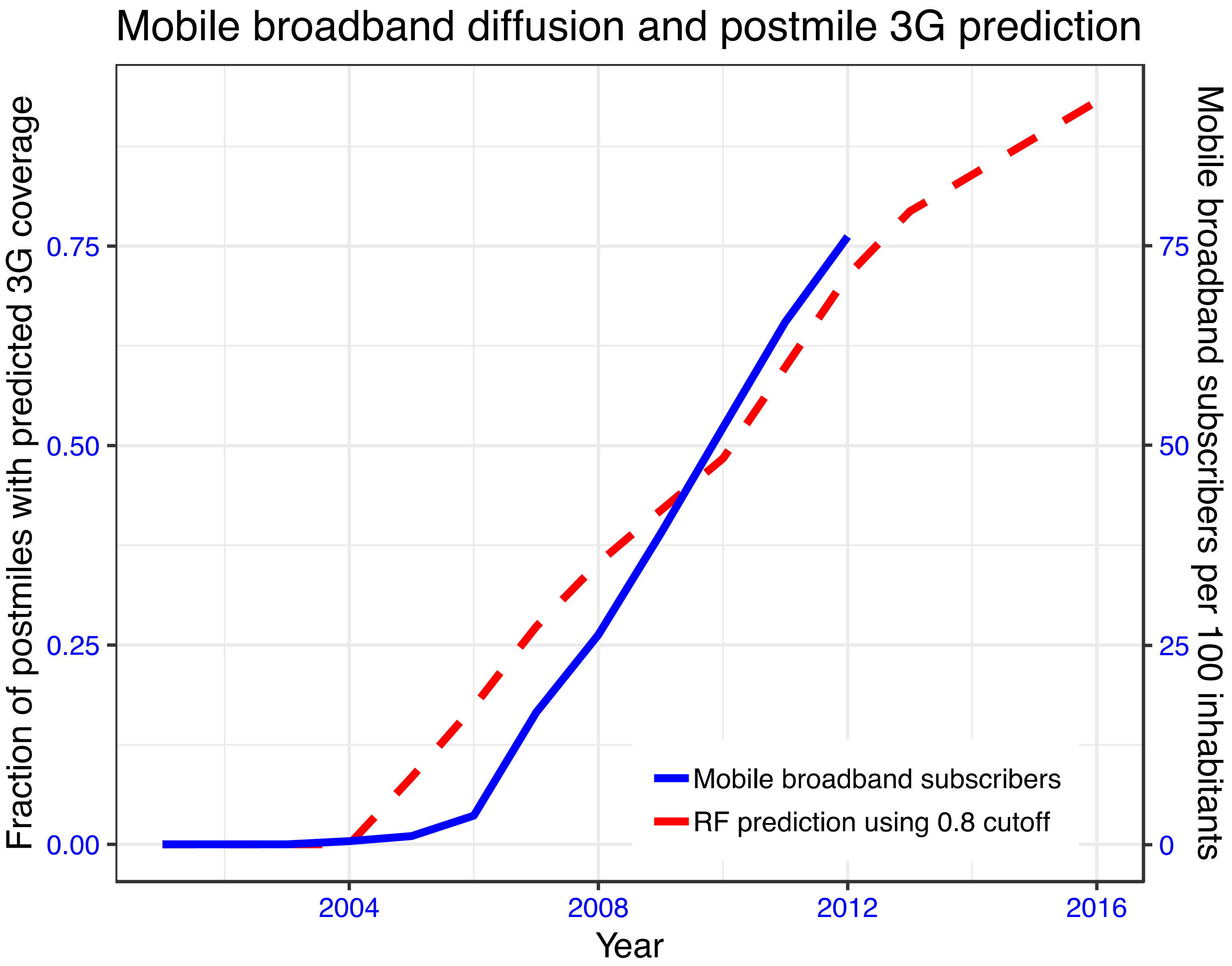
- But: we know closest tower characteristics 2001 – 2016
- Solution: build random forest model to predict 3G coverage 2001 – 2016 based on closest tower characteristics



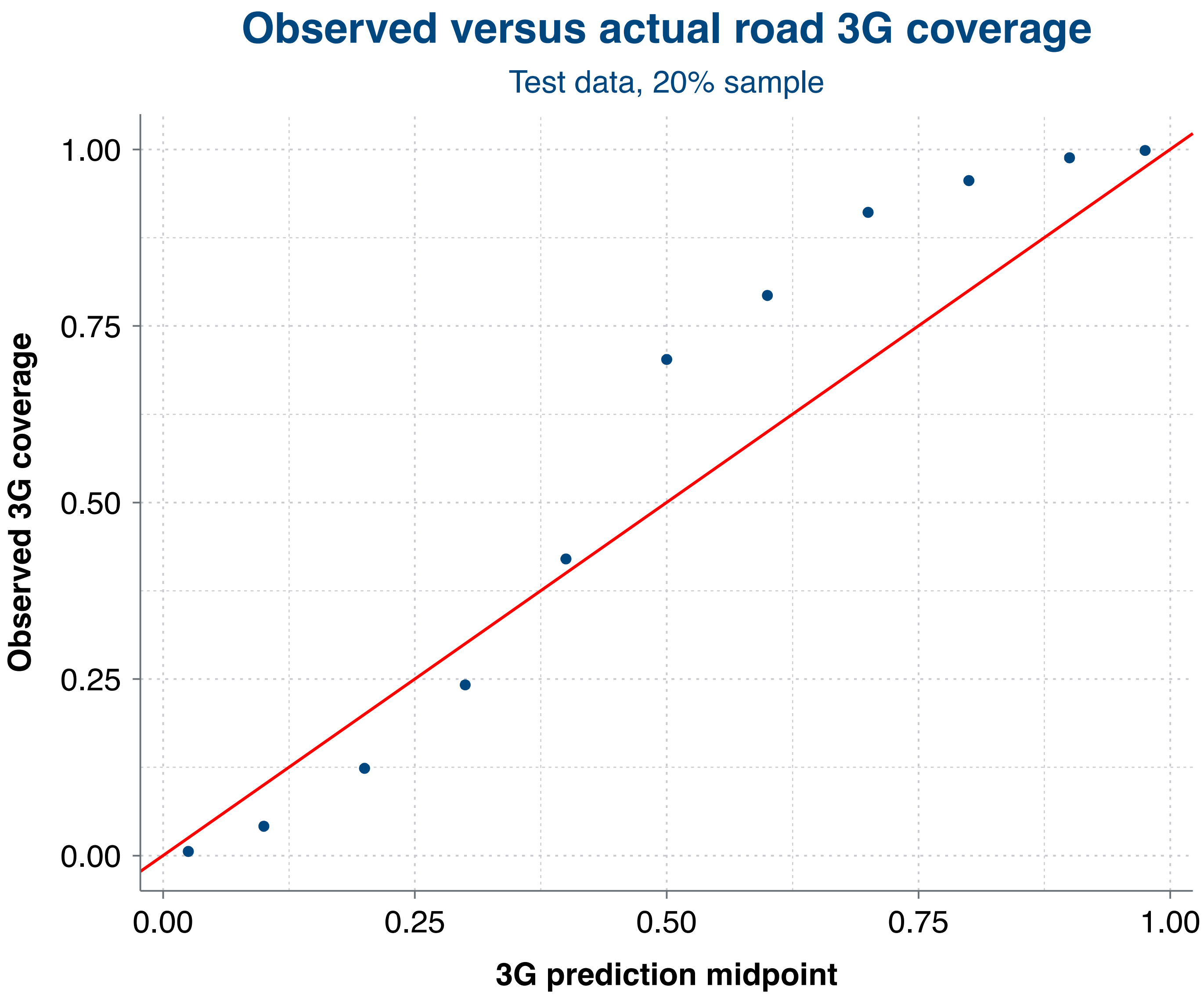
Cross-validate and select optimal 'mtry'



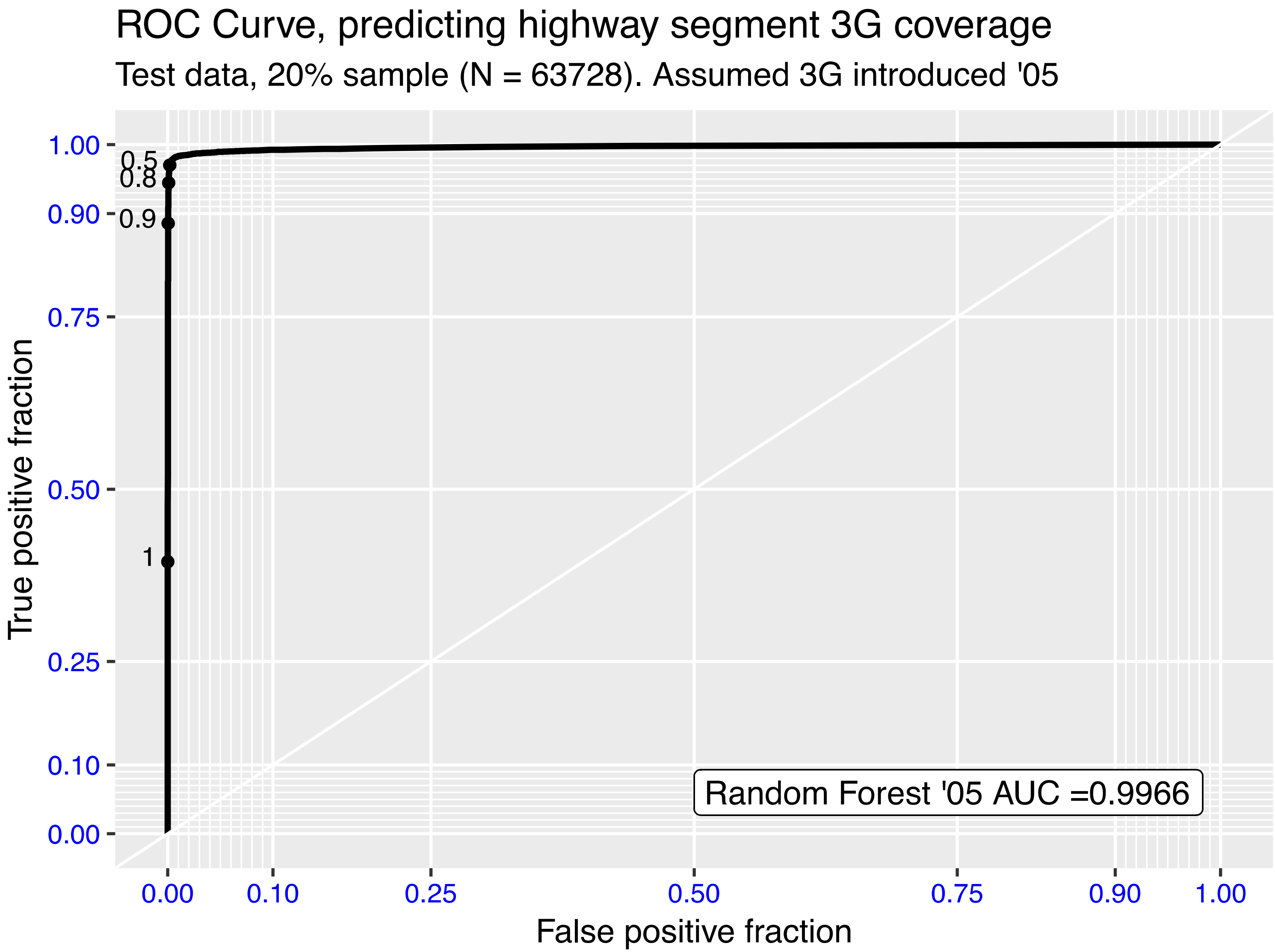
Predicted growth in 3G coverage matches observed coverage



Model Calibration Plot



Test Set: ROC Accuracy Plot



ROC Curve, predicting highway segment 3G coverage

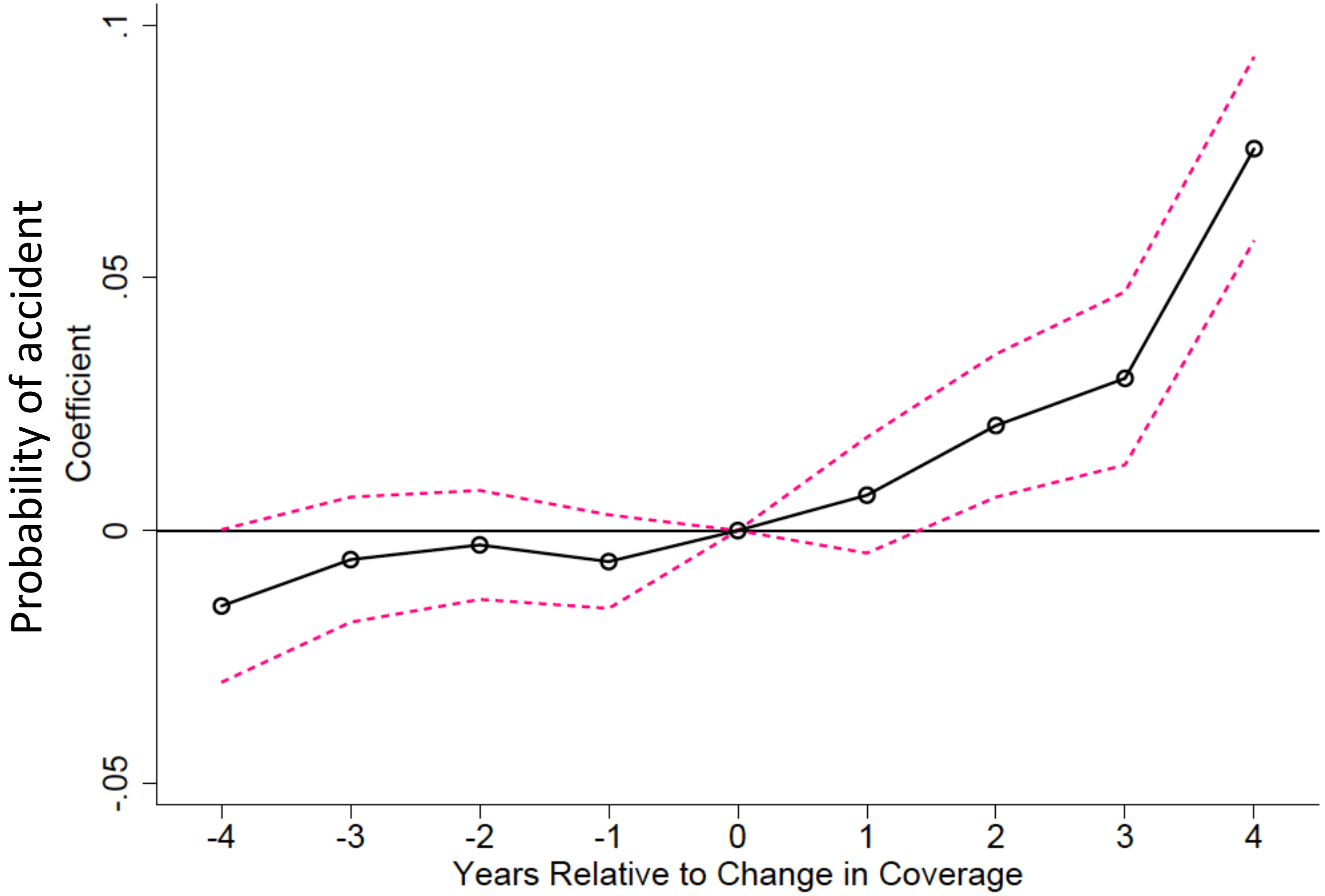
Test data, 20% sample (N = 63728). Assumed 3G introduced '05

Event Study Fixed Effect Poisson Model

$$\mathrm{E}[Accident_{it} |\, X_{it}\ ] = \exp\left(\sum_{k=-4}^{-1} \theta_k S_{it+k} + \sum_{j=1}^{4+} \theta_j S_{jt+j} + \ln(Road\ Traffic_{it}) + \gamma_i + \tau_t + v_{it}\right)$$

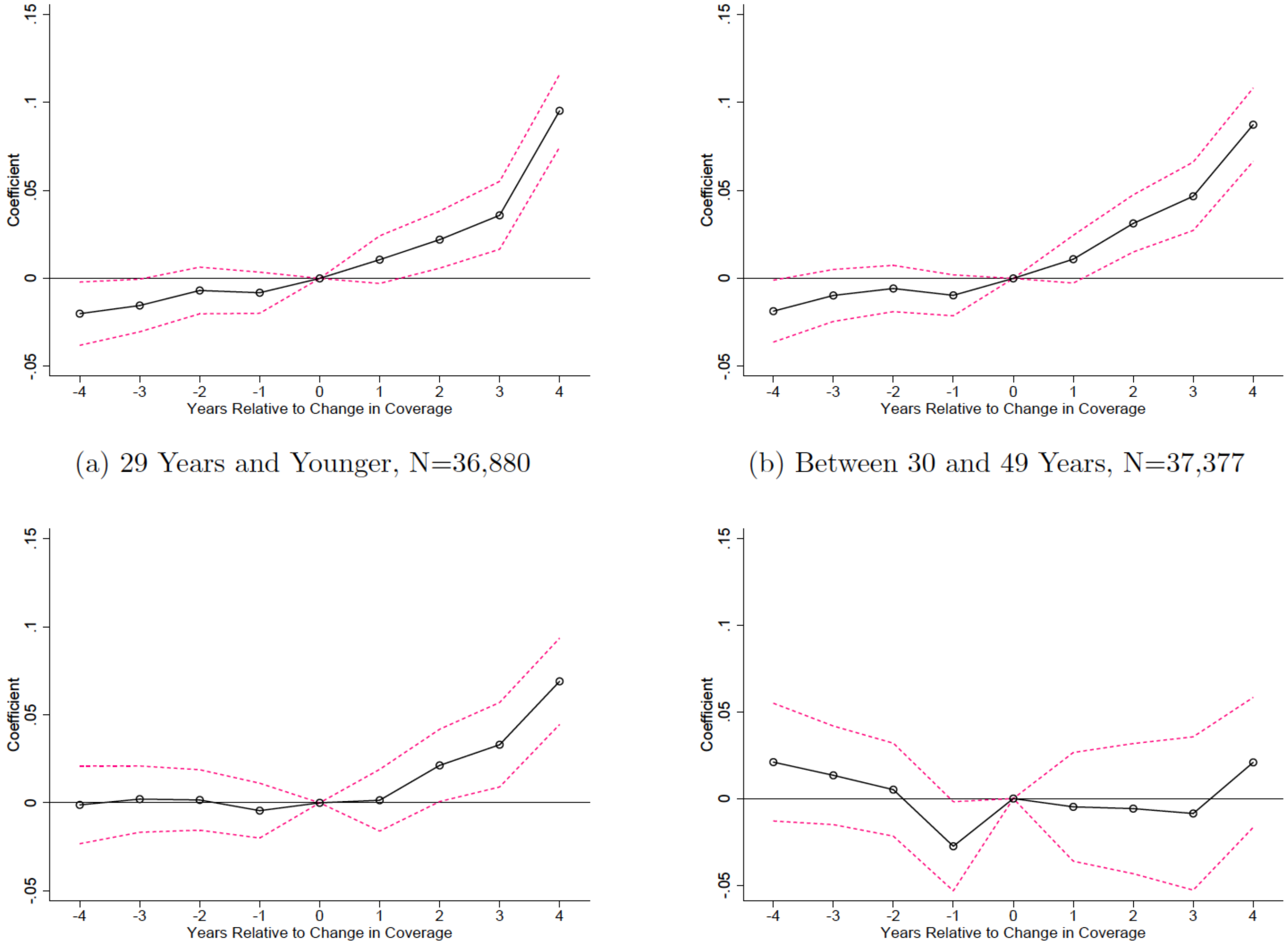
- θ_j impact of mobile internet coverage on accident probability
- S_{jt} when predicted 3G coverage = 1
- γ_i road segment fixed effect
- τ_t time fixed effect
- v_{it} error

Event study: impact of predict 3G road coverage on traffic accident probability



(c) $\geq$0.80 Threshold, N=40,877

Younger Drivers More Affected by 3G road Access



(a) 29 Years and Younger, N=36,880

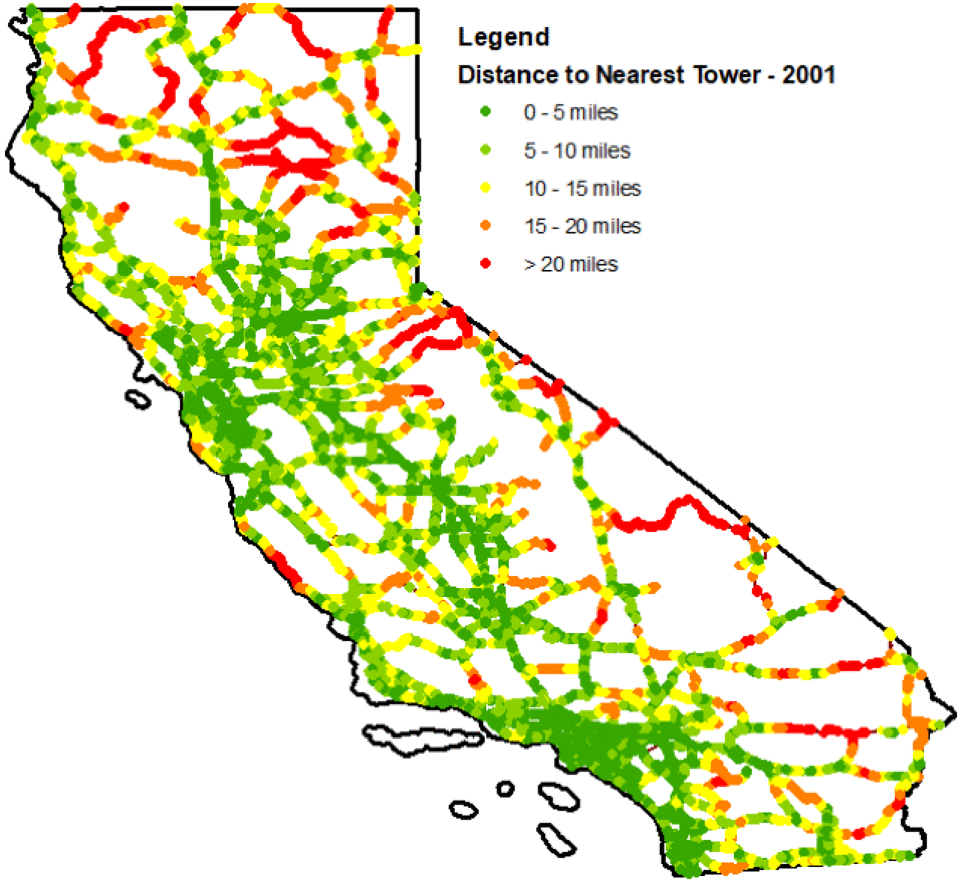
(b) Between 30 and 49 Years, N=37,377

(c) Between 50 and 64 Years, N=33,857

(d) 65 Years and Older, N=27,494

Conclusion

- Accident rates increase 1.1 percent a road gets access to 3G coverage
- Internet connected mobile phones cause 3,305 accidents per year in California
- Further evidence you can embed machine learning in causal inference models



Comments/suggestions appreciated!

Jonathan Hersh

Argyros School of Business, Chapman University

hersh@chapman.edu

(Please talk to me if you’re interested in teaching machine learning @ Chapman)