KAIST Summer Session 2018

Module 1. Research Design for Data Analytics

Two Paradigms of Analytics

KAIST College of Business

Jiyong Park

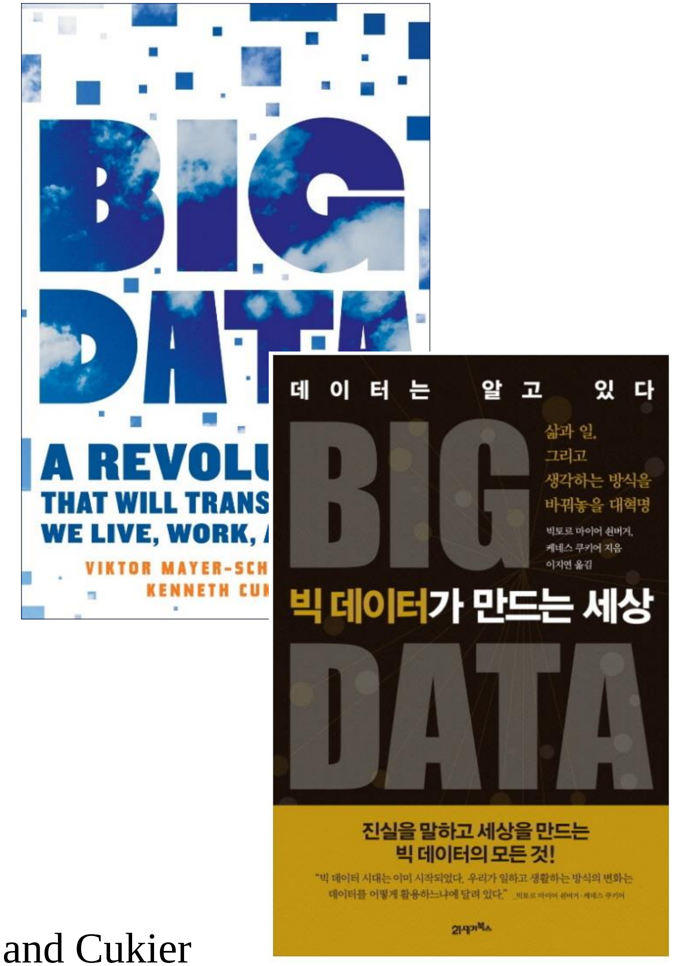
25 June, 2018

Two Paradigms of Analytics
: Causal Inference and Prediction

“Correlation is Enough in the Era of Big Data”

“인간은 원인을 찾도록 길들여져 있었다. 하지만 분명한 것은 이제 사회가 인과성에 대한 그동안의 집착을 일부 포기하고 단순한 상관성에 만족해야 할 것이라는 점이다. 즉 이유는 모른 채 결론만 알게 되는 것이다. 이것은 수백 년간 이어져온 관행을 뒤집는 일이며, 우리는 의사결정 방식이나 현실에 대한 이해 방식을 기초적인 부분부터 다시 생각해야 할지도 모른다.

많은 경우에 우리는 그 정도면 충분하다. *빅데이터에서 중요한 것은 결론이지 이유가 아니다.* 어떤 현상의 원인을 알아야 할 필요는 없다. 우리는 데이터 스스로 진실을 드러내게 하면 된다.”



<Big Data (빅데이터가 만드는 세상)>, Mayer-Schonberger and Cukier

“Causality Lies at the Heart of Our Understanding”

“예측 알고리즘은 상관관계를 찾아내는 데 초자연적인 능력을 발휘할지 모르지만, 그 특성과 현상이 생기는 근본적인 원인에는 무관심하다.

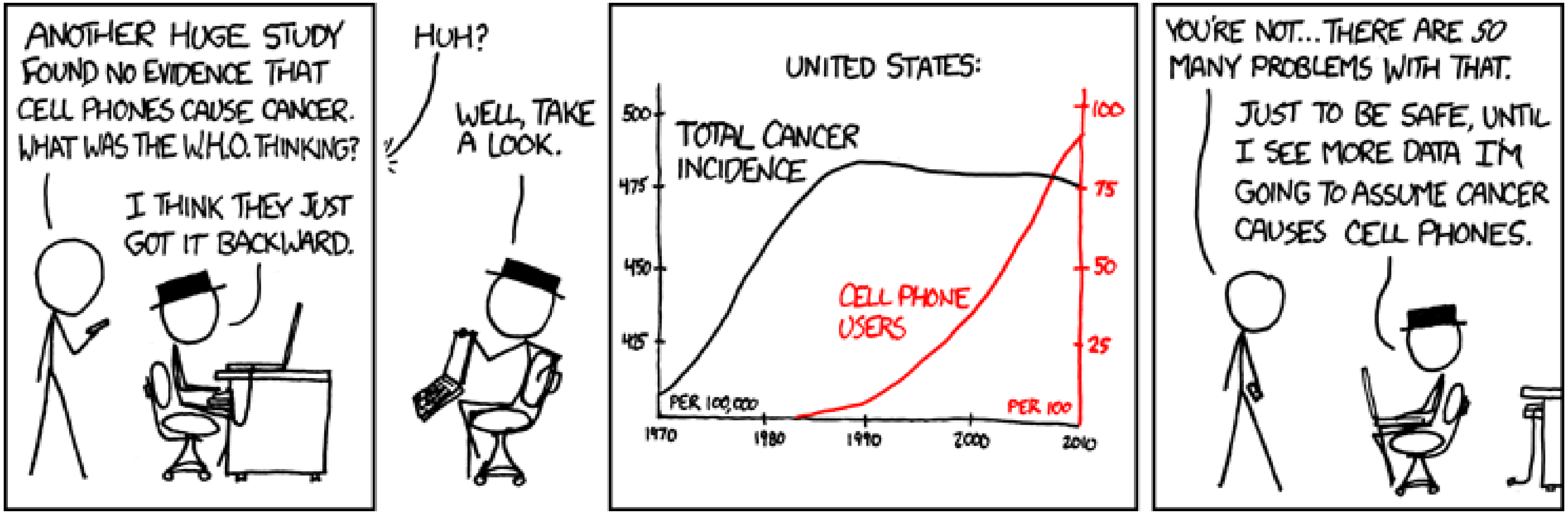
하지만, 인간이 가진 이해력의 범위를 확대하면서 궁극적으로 ***우리는 지식 추구 활동에 의미를 부여하는 것은 인과관계의 해독이다.*** 이는 세상이 돌아가는 원리를 세심하게 풀어헤치는 것이다.”

<The Glass Cage (유리감옥)>, Nicholas Carr

생각하지 않는 사람들
인터넷이 우리의 뇌 구조를 바꾸고 있다
니콜라스 카 지음 | 최지향
The Sha
기술이 발전할수록 인간은 왜 무능해지나
THE GLASS CAGE
니콜라스 카 지음
이진원 옮김
생각을 통제하는 거대한 힘
유리감옥
세계적 디지털 사상가 니콜라스 카
《생각하지 않는 사람들》 이후 4년 만의 신작!
전 세계 동시 출간
한국경제신문

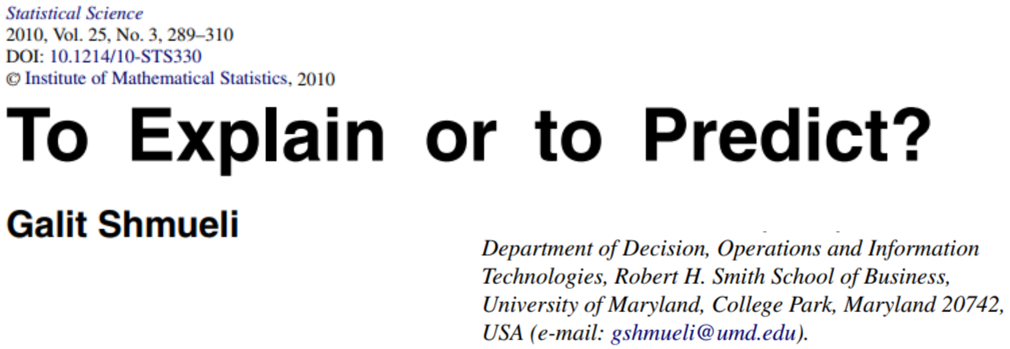
Correlation versus Causation

- Indeed, it is the timeless conundrum, “correlation or causation?”


ANOTHER HUGE STUDY FOUND NO EVIDENCE THAT CELL PHONES CAUSE CANCER. WHAT WAS THE W.H.O. THINKING?
I THINK THEY JUST GOT IT BACKWARD.
HUH?
WELL, TAKE A LOOK.
UNITED STATES:
TOTAL CANCER INCIDENCE
CELL PHONE USERS
500
475
450
425
PER 100,000
100
75
50
25
PER 100
1970
1980
1990
2000
2010
YOU'RE NOT... THERE ARE SO MANY PROBLEMS WITH THAT.
JUST TO BE SAFE, UNTIL I SEE MORE DATA I'M GOING TO ASSUME CANCER CAUSES CELL PHONES.

Source: https://xkcd.com/925/

To Explain or To Predict?



Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330
© Institute of Mathematical Statistics, 2010

To Explain or to Predict?

Galit Shmueli

Department of Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742, USA (e-mail: gshmueli@umd.edu).

- Four major disparities between explanation and prediction (Shmueli 2010)
  - ~~Theory – Data~~
  - **Causation – Association (Correlation)**
  - **Bias – Variance**
  - Retrospective (*in-sample*) – Prospective (*out-of-sample*)

Shmueli, G., 2010. To Explain or to Predict?. *Statistical Science*, 25(3), pp.289-310.

Bias versus Variance

- Bias – Variance decomposition

$$True\ Relationship: Y = f(x) + \varepsilon$$

$$Empirical\ Model:\ \ Y = \hat{f}(x)$$

$$Expected\ Estimation\ Error = E\left[\left(Y - \hat{f}(x)\right)^2\right]$$

$$= Var[\varepsilon] + \underline{\left\{E\left[\hat{f}(x)\right] - f(x)\right\}^2} + \underline{E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right]}$$

$$= Var[\varepsilon] + Bias^2 + Var\left[\hat{f}(x)\right]$$

- Explanatory modeling focuses on minimizing bias to obtain the most accurate representation of the underlying true relationship.
- Predictive modeling seeks to minimize the combination of bias and estimation variance, occasionally sacrificing theoretical accuracy for improved empirical precision.

Shmueli, G., 2010. To Explain or to Predict?. *Statistical Science*, 25(3), pp.289-310.

What is Causal Inference?

- Causal inference is to connect one process (*cause*) with another process or state (*effect*) where the first is partly responsible for the second, and the second is partly dependent on the first.

- Goal of causal inference: (In-sample) **Unbiased estimates**
  - ➢ But, it is quite challenging in observational studies.

- Potential outcome framework
  - ➢ Ideal causality

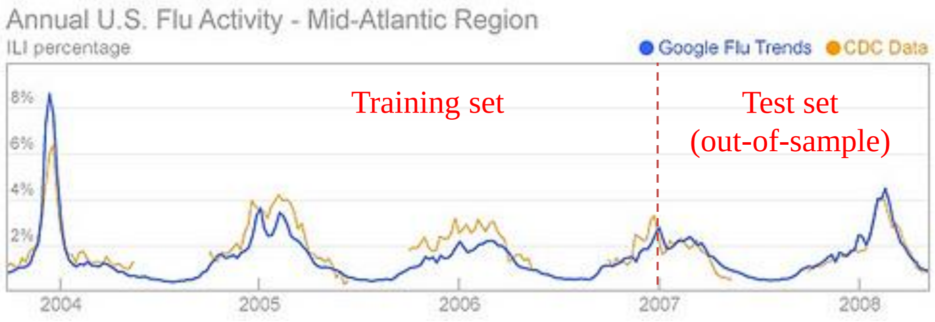
    = *(Outcome for treated if treated) – (Outcome for treated if not treated)*
  - ➢ But, in reality…

    = *(Outcome for treated if treated) – (Outcome for untreated if not treated)*

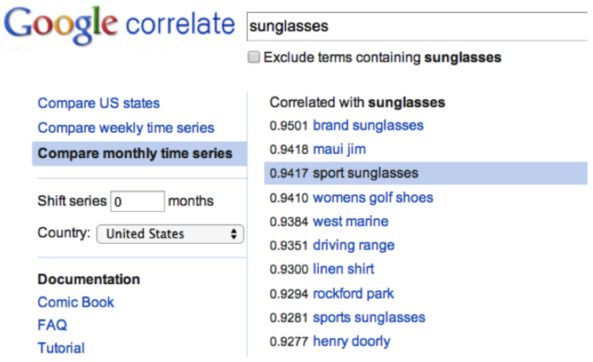
What is Prediction?

- Prediction is to state about the future event or the unknowns, based upon experience or data.

- Goal of prediction: (Out-of-sample) **High predictive power**
  - ➢ Prediction does not require causation. Rather, it exploits a range of correlations as much as researchers can (*reducing bias*), while avoiding overfitting (*minimizing variance*).



Google's research on prediction



sunglasses

□ Exclude terms containing **sunglasses**

Compare US states
Compare weekly time series
Compare monthly time series

Shift series 0 months
Country: United States

Documentation
Comic Book
FAQ
Tutorial

Correlated with **sunglasses**

0.9501	brand sunglasses
0.9418	maui jim
0.9417	sport sunglasses
0.9410	womens golf shoes
0.9384	west marine
0.9351	driving range
0.9300	linen shirt
0.9294	rockford park
0.9281	sports sunglasses
0.9277	henry doorly

Google's follow-up service

Different Goals Need Different Methodologies

- Example: Explaining review helpfulness (Ghose and Ipeirotis 2011)
  - “The explanatory study that we described above revealed what factors influence the helpfulness.” (p. 1507)

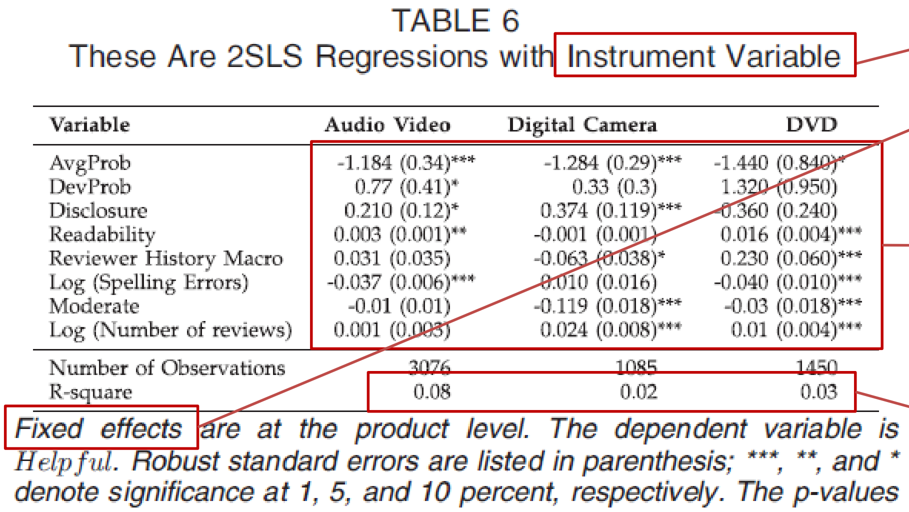


TABLE 6
These Are 2SLS Regressions with Instrument Variable

Variable	Audio Video	Digital Camera	DVD
AvgProb	-1.184 (0.34)***	-1.284 (0.29)***	-1.440 (0.840)*
DevProb	0.77 (0.41)*	0.33 (0.3)	1.320 (0.950)
Disclosure	0.210 (0.12)*	0.374 (0.119)***	0.360 (0.240)
Readability	0.003 (0.001)**	-0.001 (0.001)	0.016 (0.004)***
Reviewer History Macro	0.031 (0.035)	-0.063 (0.038)*	0.230 (0.060)***
Log (Spelling Errors)	-0.037 (0.006)***	0.010 (0.016)	-0.040 (0.010)***
Moderate	-0.01 (0.01)	-0.119 (0.018)***	-0.03 (0.018)***
Log (Number of reviews)	0.001 (0.003)	0.024 (0.008)***	0.01 (0.004)***
Number of Observations	3076	1085	1450
R-square	0.08	0.02	0.03

*Fixed effects are at the product level. The dependent variable is $Helpful$. Robust standard errors are listed in parenthesis; ***, **, and * denote significance at 1, 5, and 10 percent, respectively. The p-values*

For unbiased and consistent estimation

A explanatory study seeks to investigate statistically significant effect of factors of interest.

Low explanatory power (not good for out-of-sample prediction)

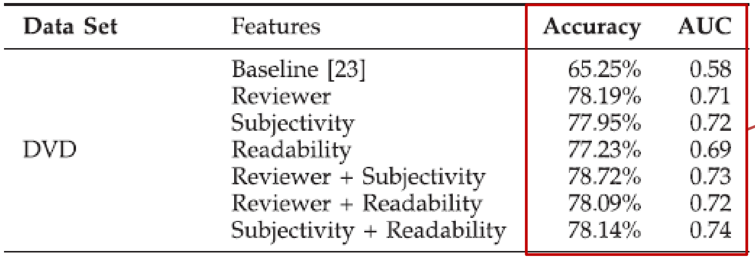
- In predicting for out-of-sample (e.g., products not included in training), can we include any fixed effects?

Ghose, A. and Ipeirotis, P.G., 2011. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10), pp.1498-1512.

Different Goals Need Different Methodologies

- Example: Predicting review helpfulness (Ghose and Ipeirotis 2011)
  - ➢ "The main goal now is not to explain which factors affect helpfulness and impact, but to examine whether, given an existing review, how well can we predict the helpfulness and economic impact of an unseen review." (p. 1507)

TABLE 8
Accuracy and Area under the ROC Curve for the Helpfulness Classifiers



Data Set	Features	Accuracy	AUC
DVD	Baseline [23]	65.25%	0.58
	Reviewer	78.19%	0.71
	Subjectivity	77.95%	0.72
	Readability	77.23%	0.69
	Reviewer + Subjectivity	78.72%	0.73
	Reviewer + Readability	78.09%	0.72
	Subjectivity + Readability	78.14%	0.74

The primary goal of prediction is to maximize the predictive power.

"Our evaluation results are based on stratified 10-fold cross validation and we use as valuation metrics the classification accuracy and the area under the ROC curve." (p. 1508)

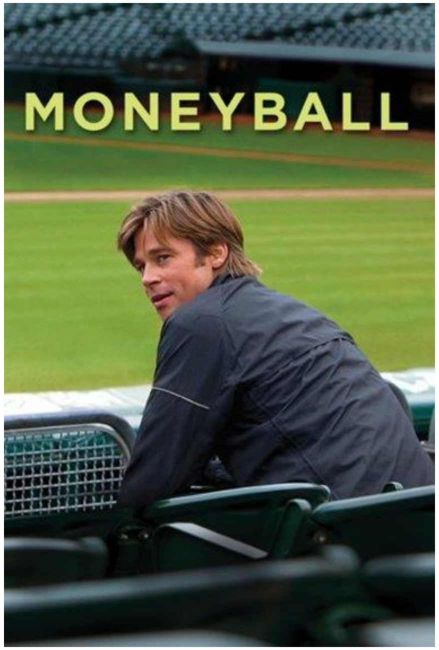
  - ➢ "While it is hard, at this point, to claim causality… it is definitely possible to show a strong correlation between the two." (p. 1509)

Ghose, A. and Ipeirotis, P.G., 2011. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10), pp.1498-1512.

Input-Output Framework

Lesson from Moneyball

- Moneyball (2011)
  - A film about the general manager, Billy Beane, of the Oakland Athletics baseball team



1. What is the problem? (https://www.youtube.com/watch?v=pWgyy_rlmag)

2. Moneyball – data science (https://www.youtube.com/watch?v=zjU7R-6ShdE)

3. Science vs scouts (https://www.youtube.com/watch?v=DtumWOsgFXc)

Lesson from Moneyball

- Moneyball (2011)

Baseball scouts

Data scientists

Suppose you are the general manager (in the movie, Brad Pitt stars as Billy Beane).

Would you hire only either scouts or scientists for the championship, or both of them?

Causal Inference versus Prediction

- Indeed, they play different roles with different goals.

Baseball scouts

“He’s got a great swing. Natural swing.”

“Look, we’re gonna find 25 guys, put them through player development, teach them how to play Oakland A baseball.”

Data scientists

“They’re still asking the wrong questions.”

“The goal shouldn’t be to buy players, what you want to buy is wins. To buy wins, you buy runs.”

Causal Inference versus Prediction

- Causal inference and prediction are necessary for the team management.

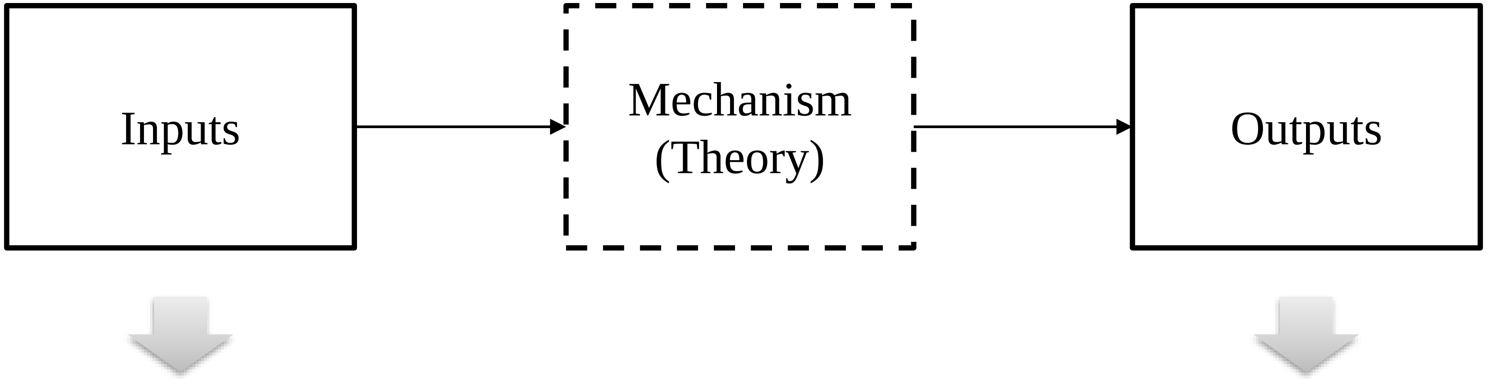
In the farm teams or training camps,
we need to know the causal factors
to teach and cultivate players' ability.

In the games, the only thing of importance
is to increase the probability of winning.

Lesson from Moneyball

WE'RE ASKING THE WRONG QUESTIONS. DO YOU BELIEVE IN THIS THING?
YES.

Input-Output Framework



If you are interested in contemplating some intervention in the *inputs*, identification strategy for causal inference would be the right tool.

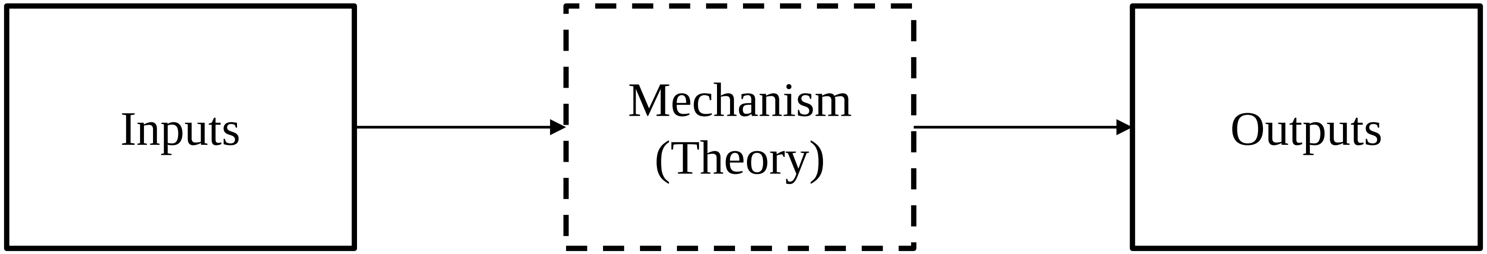
→ **Intervention-oriented research**

If you are interested in obtaining the best (or precise) *output*, predictive analytics would be the right tool.

→ **Solution-oriented research**

Same Phenomenon, Different Questions

- Example: Safety inspection (Athey 2017)


Inputs

Mechanism (Theory)

Outputs

“Should we step up the safety inspections?”

“What is the best way to allocate food-safety inspectors?”

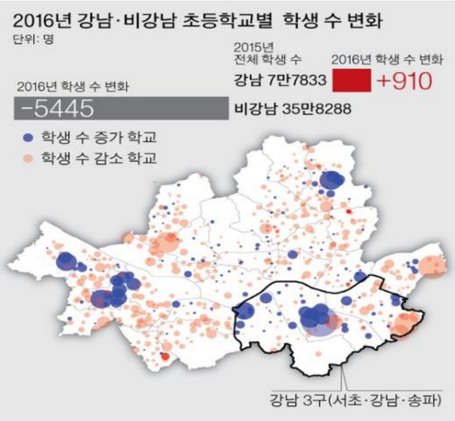
Athey, S., 2017. Beyond Prediction: Using Big Data for Policy Problems. *Science*, 355(6324), pp.483-485.

Examples of Causal Inference and Prediction
: A Case of Income Level

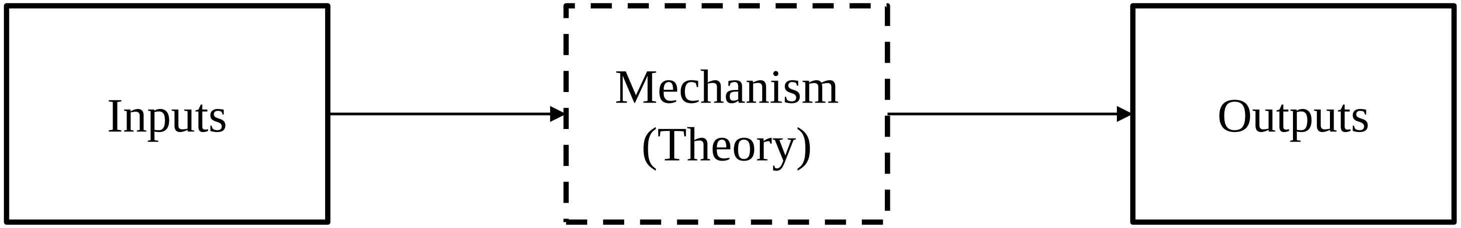
A Researcher Raises a Question…

- “Are neighborhoods important for the children’s future?”

맹모삼천지교



- Input-output framework



Quality of neighborhoods

Children’s future income

What Identification Strategy Does

- Ideal approach is to randomly assign where children live and observe their future income… *(is it possible?)*

- Chetty and Hendren (2018) use a quasi-experiment design of family moves.

family move →

13 years old

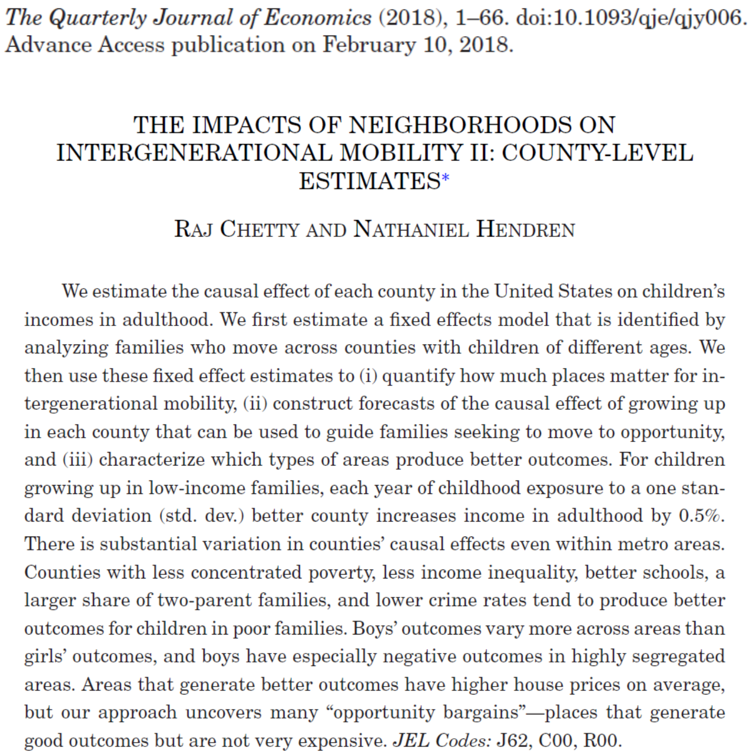
11 years old

Comparing the future income between siblings to identify an effect of “2 years exposure” to new town

Chetty, R. and Hendren, N., 2018. The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. *Quarterly Journal of Economics*, forthcoming

How Can We Utilize the Findings of This Research?

- Academic application
  - ➢ Subsequent research on which factors are correlated to (causal) social mobility

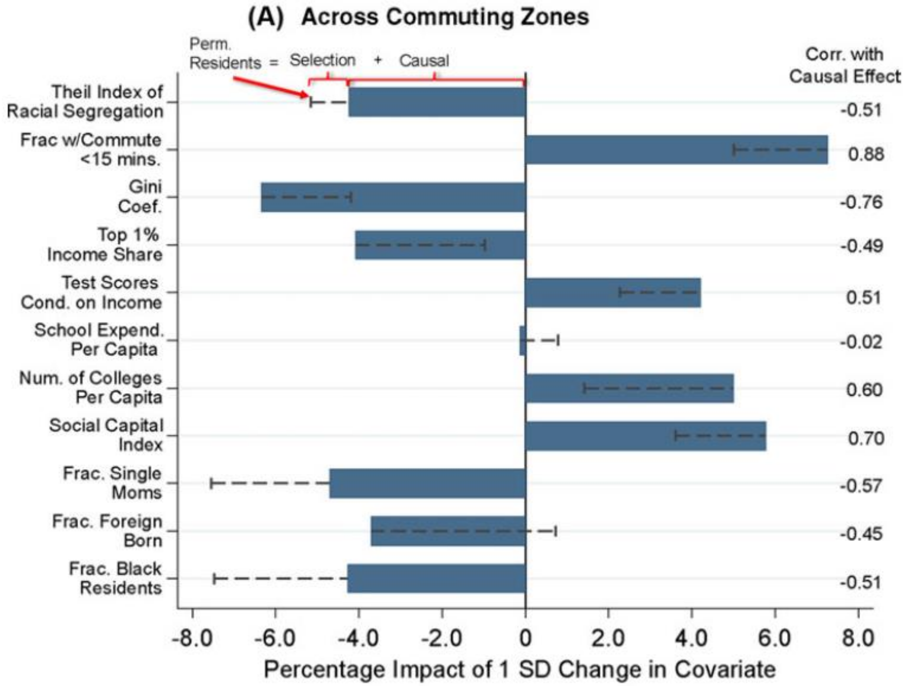


The Quarterly Journal of Economics (2018), 1–66. doi:10.1093/qje/qjy006.
Advance Access publication on February 10, 2018.

THE IMPACTS OF NEIGHBORHOODS ON INTERGENERATIONAL MOBILITY II: COUNTY-LEVEL ESTIMATES*

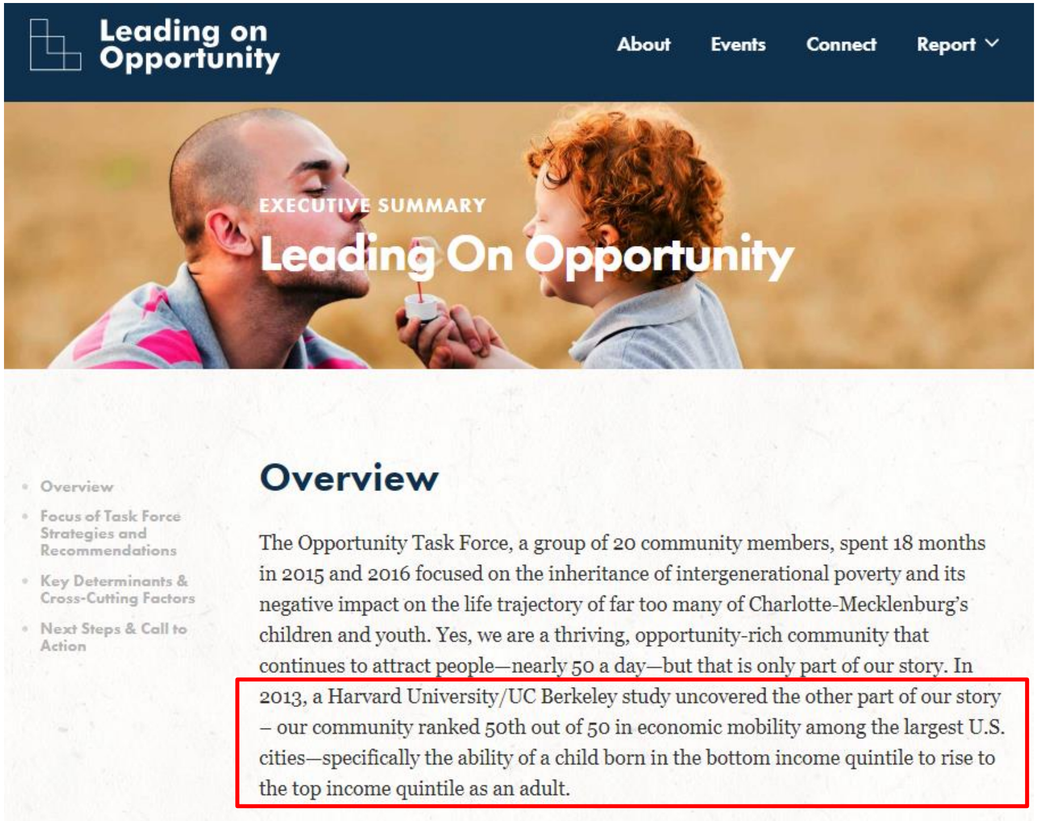
RAJ CHETTY AND NATHANIEL HENDREN

We estimate the causal effect of each county in the United States on children's incomes in adulthood. We first estimate a fixed effects model that is identified by analyzing families who move across counties with children of different ages. We then use these fixed effect estimates to (i) quantify how much places matter for intergenerational mobility, (ii) construct forecasts of the causal effect of growing up in each county that can be used to guide families seeking to move to opportunity, and (iii) characterize which types of areas produce better outcomes. For children growing up in low-income families, each year of childhood exposure to a one standard deviation (std. dev.) better county increases income in adulthood by 0.5%. There is substantial variation in counties' causal effects even within metro areas. Counties with less concentrated poverty, less income inequality, better schools, a larger share of two-parent families, and lower crime rates tend to produce better outcomes for children in poor families. Boys' outcomes vary more across areas than girls' outcomes, and boys have especially negative outcomes in highly segregated areas. Areas that generate better outcomes have higher house prices on average, but our approach uncovers many "opportunity bargains"—places that generate good outcomes but are not very expensive. *JEL Codes:* J62, C00, R00.



How Can We Utilize the Findings of This Research?

- Practical application
  - ➢ Community task force to improve the quality of neighborhoods



Leading on Opportunity
About
Events
Connect
Report
EXECUTIVE SUMMARY
Leading On Opportunity
Overview
Focus of Task Force Strategies and Recommendations
Key Determinants & Cross-Cutting Factors
Next Steps & Call to Action

Overview

The Opportunity Task Force, a group of 20 community members, spent 18 months in 2015 and 2016 focused on the inheritance of intergenerational poverty and its negative impact on the life trajectory of far too many of Charlotte-Mecklenburg's children and youth. Yes, we are a thriving, opportunity-rich community that continues to attract people—nearly 50 a day—but that is only part of our story. In 2013, a Harvard University/UC Berkeley study uncovered the other part of our story – our community ranked 50th out of 50 in economic mobility among the largest U.S. cities—specifically the ability of a child born in the bottom income quintile to rise to the top income quintile as an adult.

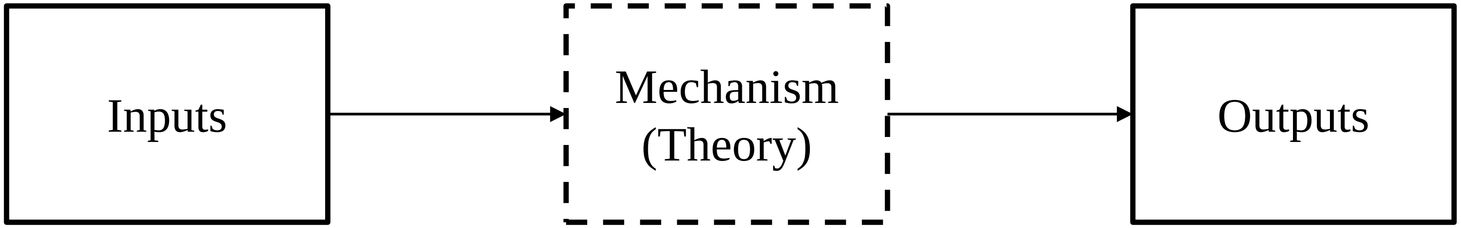
https://leadingonopportunity.org/introduction/executive-summary/

A Researcher Raises Another Question…

- “Where should we provide food aid in low-income countries in Africa?”
  - Unfortunately, there is no local level statistics especially in rural areas in Africa.

UNICEF

- Input-output framework



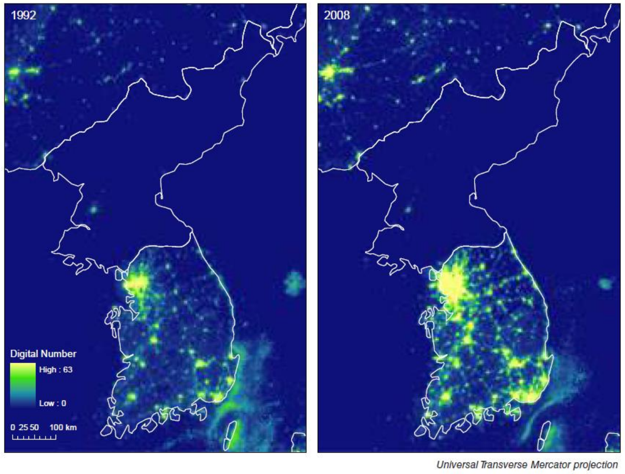
??

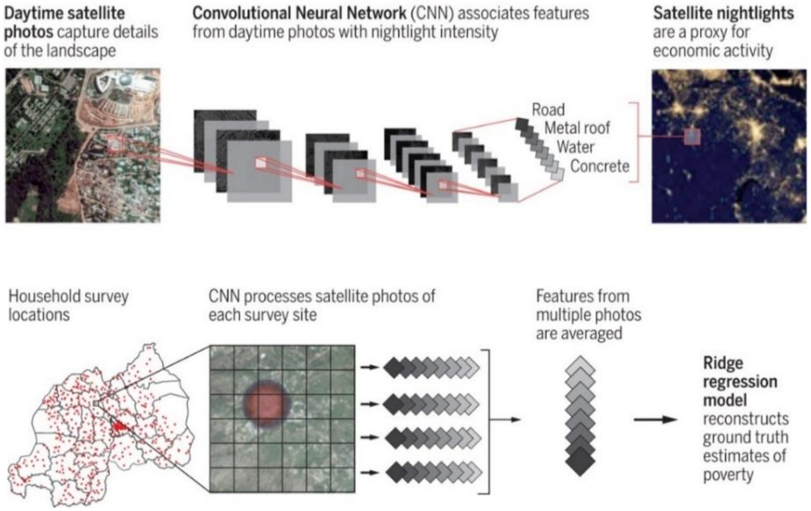
Local poverty level

What Predictive Analytics Does

- Ideal approach is to walk around the whole Africa and investigate the poverty level… *(is it possible?)*

- Jean et al. (2016) employ a deep learning algorithm (CNN) to **predict** the income level from satellite images.

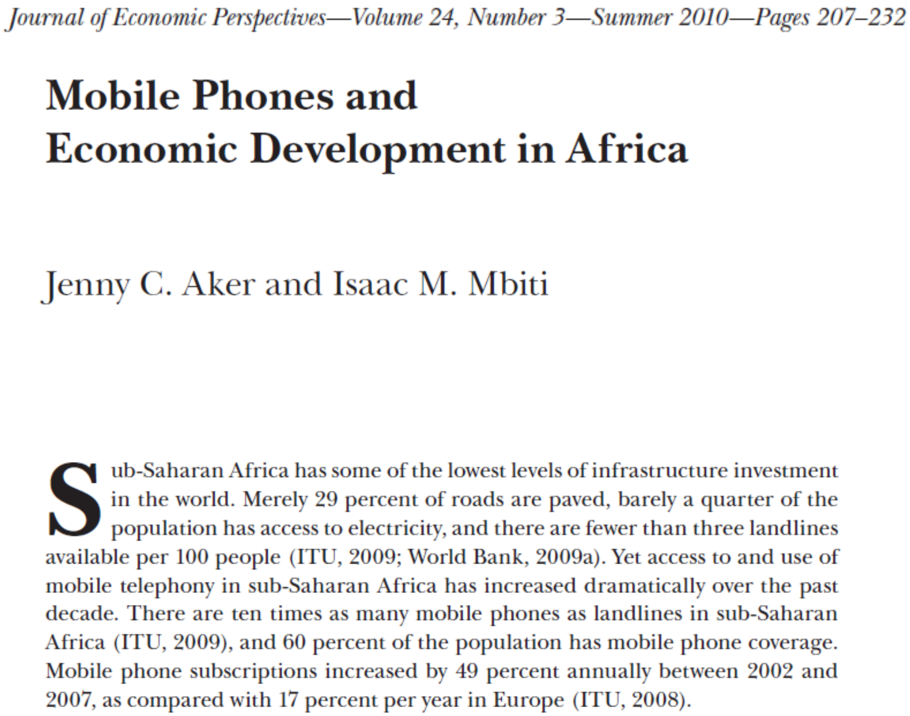



Daytime satellite photos capture details of the landscape
Convolutional Neural Network (CNN) associates features from daytime photos with nightlight intensity
Satellite nightlights are a proxy for economic activity
Road
Metal roof
Water
Concrete
Household survey locations
CNN processes satellite photos of each survey site
Features from multiple photos are averaged
Ridge regression model reconstructs ground truth estimates of poverty

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B. and Ermon, S., 2016. Combining Satellite Imagery and Machine Learning to Predict Poverty. *Science*, 353(6301), pp.790-794.

How Can We Utilize the Findings of This Research?

- Academic application
  - ➢ New indigent for new empirical research on the effect of mobile phones on economic development in Africa



Journal of Economic Perspectives—Volume 24, Number 3—Summer 2010—Pages 207–232

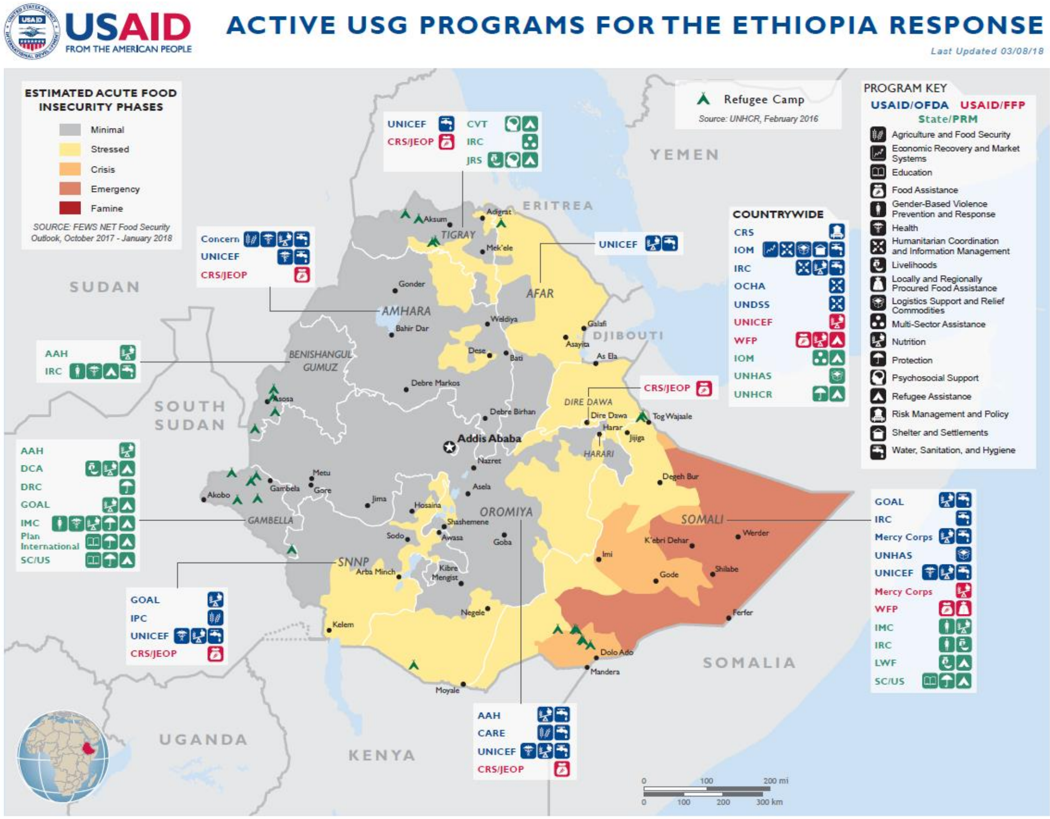
Mobile Phones and Economic Development in Africa

Jenny C. Aker and Isaac M. Mbiti

Sub-Saharan Africa has some of the lowest levels of infrastructure investment in the world. Merely 29 percent of roads are paved, barely a quarter of the population has access to electricity, and there are fewer than three landlines available per 100 people (ITU, 2009; World Bank, 2009a). Yet access to and use of mobile telephony in sub-Saharan Africa has increased dramatically over the past decade. There are ten times as many mobile phones as landlines in sub-Saharan Africa (ITU, 2009), and 60 percent of the population has mobile phone coverage. Mobile phone subscriptions increased by 49 percent annually between 2002 and 2007, as compared with 17 percent per year in Europe (ITU, 2008).

How Can We Utilize the Findings of This Research?

- Practical application
  - ➢ Allocation of foreign aid in Africa

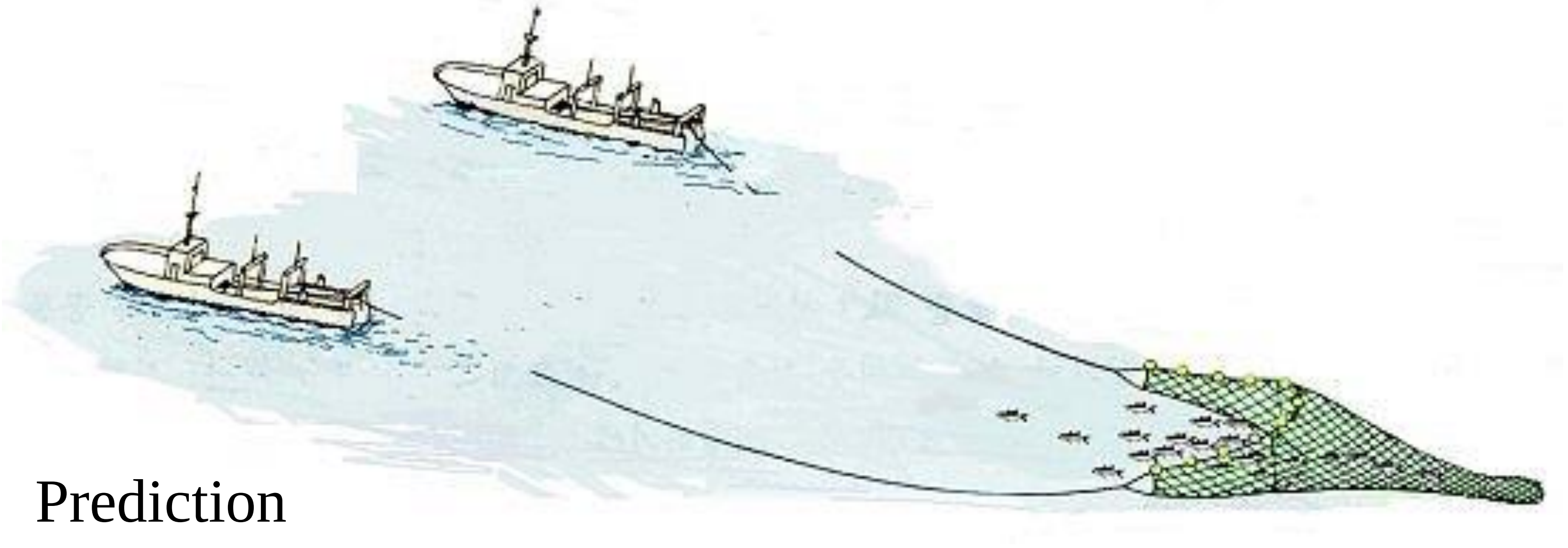


https://www.usaid.gov/crisis/ethiopia

Conclusion

Two Paradigms of Analytics

Causal Inference



Prediction

Applied Analytics

End of Document