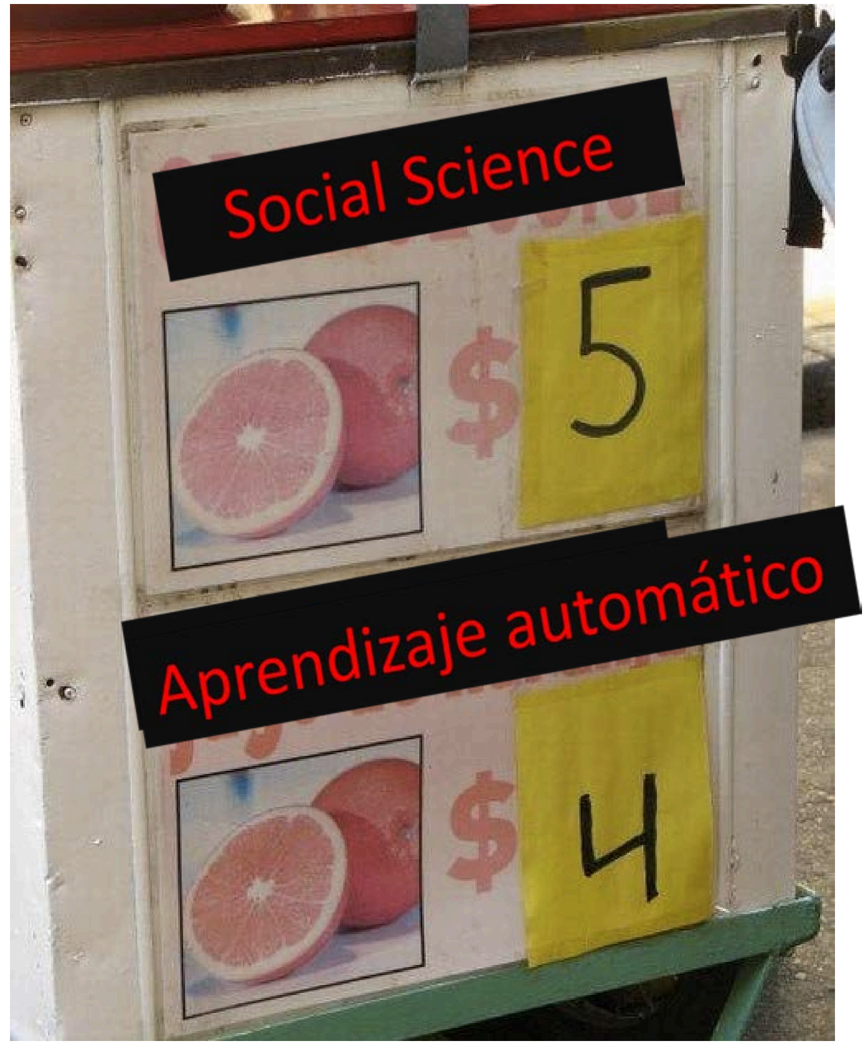
Bugbears or Legitimate Threats? (Social) Scientists' Criticisms of Machine Learning

Sendhil Mullainathan

Harvard University

This is a Poorly Titled Talk

Arbitrage



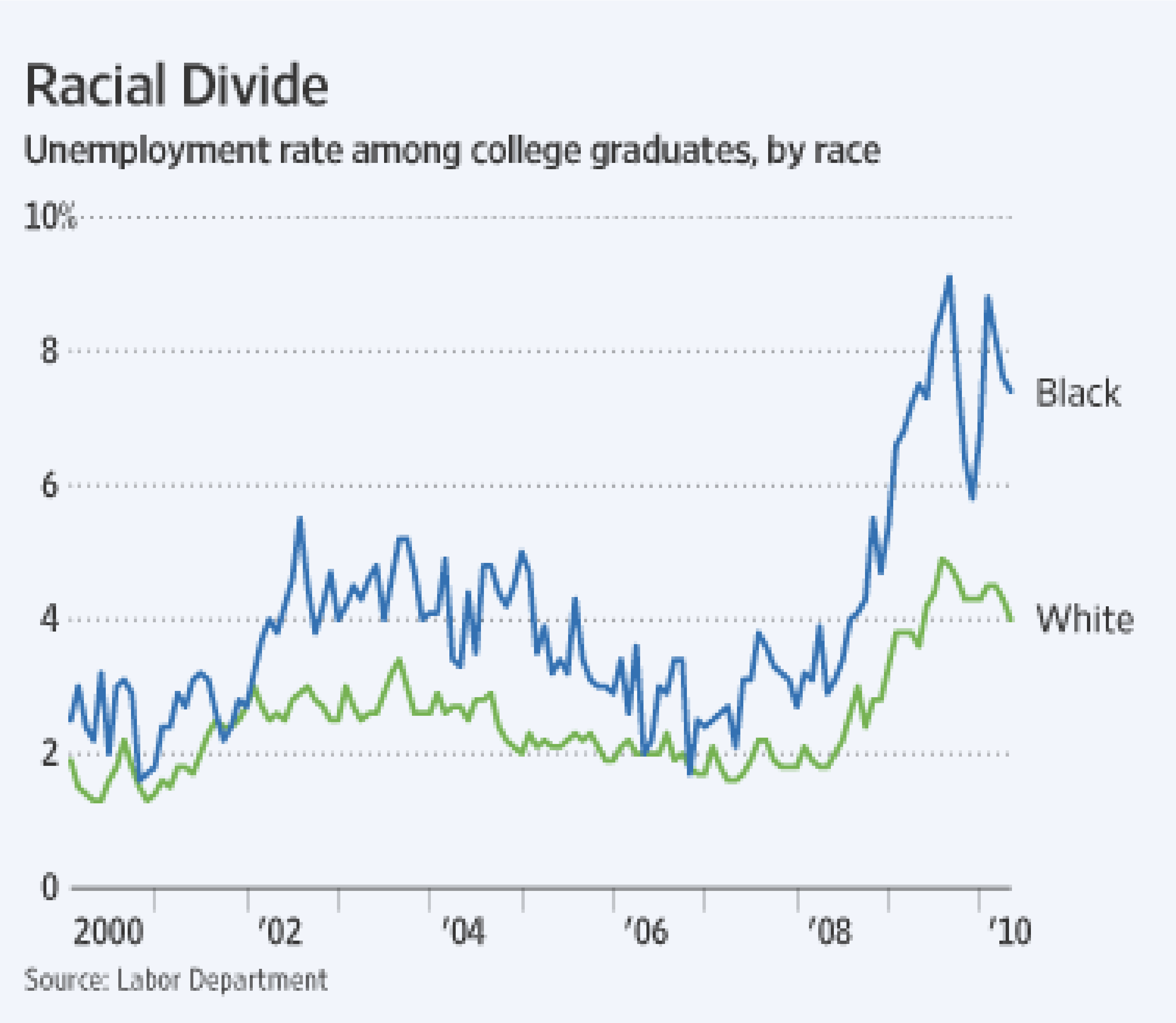
Outline of Talk

- Some past papers of mine
- Barrier 1: Predicting “versus” Theory Testing
- Barrier 2: Correlation versus Causation
- How I would redo some old papers

Outline of Talk

- Some past papers of mine
- Barrier 1: Predicting “versus” Theory Testing
- Barrier 2: Correlation versus Causation
- How I would redo some old papers

20
GREATEST HITS


Racial Divide
Unemployment rate among college graduates, by race
10%
8
6
4
2
0
2000
'02
'04
'06
'08
'10
Black
White
Source: Labor Department

Brendan Miller

Jamal Jones

JOHN DOE

Full Address • City, State, ZIP • Phone Number • E-mail

OBJECTIVE: Design apparel print for an innovative retail company

EDUCATION:

UNIVERSITY OF MINNESOTA — City, State
College of Design — May 2011

- Bachelor of Science in Graphic Design
- Cumulative GPA 3.93, Dean's List
- Twin cities Iron Range Scholarship

WORK EXPERIENCE:

AMERICAN EAGLE — City, State
Sales Associate — July 2009 - present

- Collaborated with the store merchandiser creating displays to attract clientele
- Use my trend awareness to assist customers in their shopping experience
- Thoroughly scan every piece of merchandise for inventory control
- Process shipment to increase my product knowledge

PLANET BEACH — City, State
Spa Consultant — Aug. 2008 - present

- Sell retail and memberships to meet company sales goals
- Build organizational skills by single handedly running all operating procedures
- Communicate with clients to fulfill their wants and needs
- Attend promotional events to market our services
- Handle cash and deposits during opening and closing
- Received employee of the month award twice

HEARTBREAKER — City, State
Sales Associate — May 2008 – Aug. 2008

- Stocked sales floor with fast fashion inventory
- Marked down items allowing me to see unsuccessful merchandise in a retail market
- Offered advice and assistance to each guest

VICTORIA'S SECRET — City, State
Fashion Representative — Jan. 2006 – Feb. 2009

- Applied my leadership skills by assisting in the training of coworkers
- Set up mannequins and displays in order to entice future customers
- Provided superior customer service by helping with consumer decisions
- Took seasonal inventory

VOLUNTEER EXPERIENCE:

TARGET CORPORATION — City, State
Brand Ambassador — August 2009

- Represented Periscope Marketing and Target Inc. at a college event
- Engaged University of Minnesota freshman in the Target brand experience

Call Back Rates

9.65%

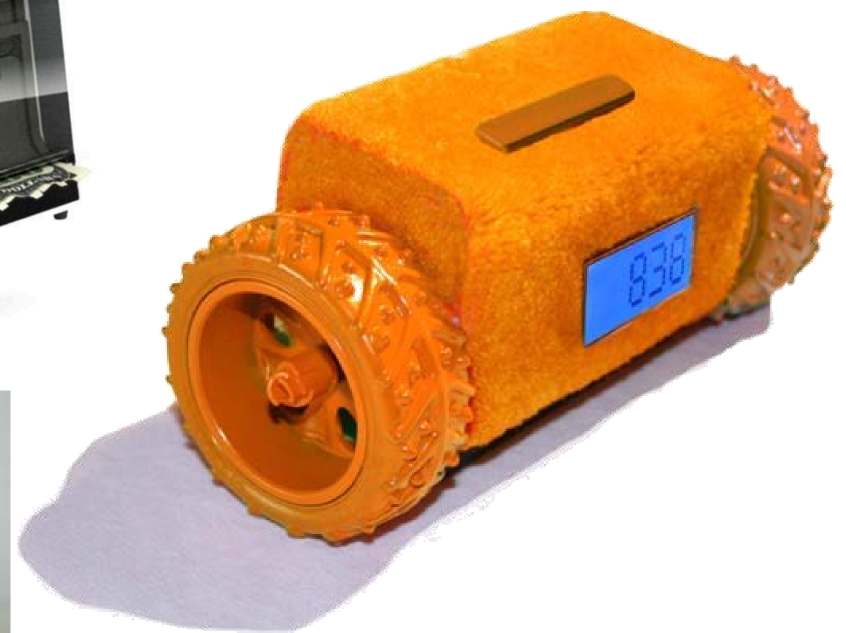
6.45%

TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

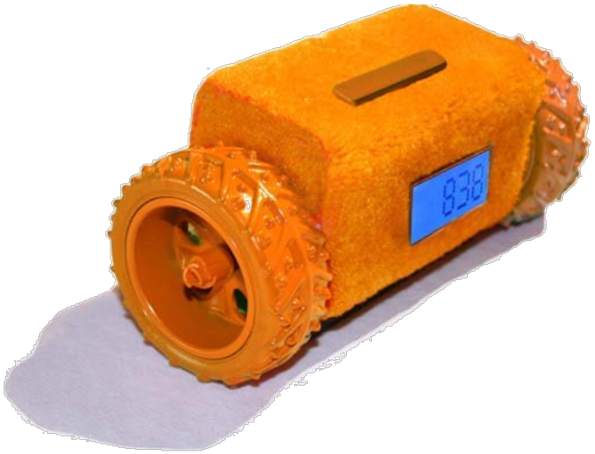
	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (p-value)
Sample:				
All sent resumes	9.65 [2,435]	6.45 [2,435]	1.50	3.20 (0.0000)
Chicago	8.06 [1,352]	5.40 [1,352]	1.49	2.66 (0.0057)
Boston	11.63 [1,083]	7.76 [1,083]	1.50	4.05 (0.0023)
Females	9.89 [1,860]	6.63 [1,886]	1.49	3.26 (0.0003)
Females in administrative jobs	10.46 [1,358]	6.55 [1,359]	1.60	3.91 (0.0003)
Females in sales jobs	8.37 [502]	6.83 [527]	1.22	1.54 (0.3523)
Males	8.87 [575]	5.83 [549]	1.52	3.04 (0.0513)

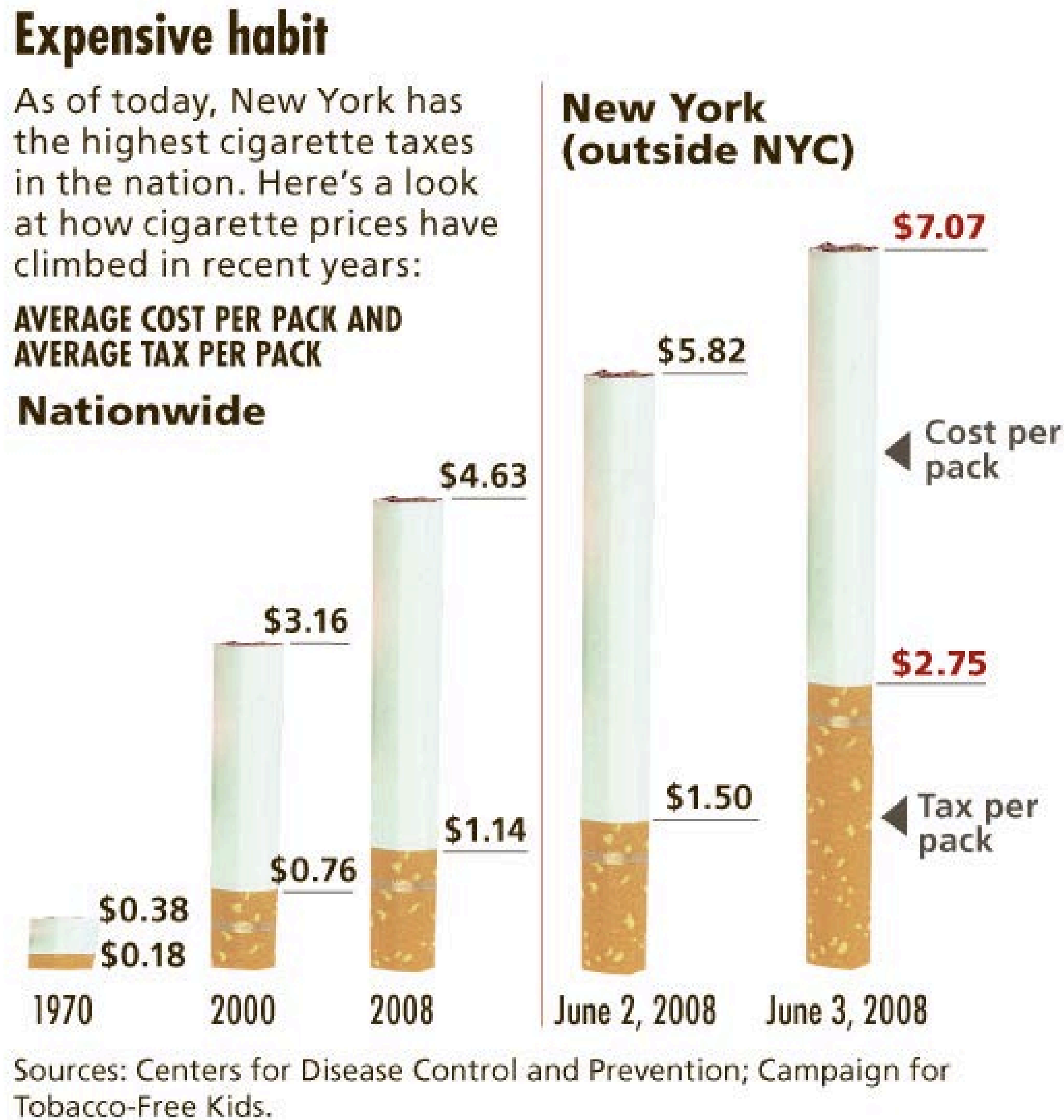
Bertrand and Mullainathan


24 HOUR
FITNESS



An alarm clock for people who have trouble getting out of bed


838


Expensive habit
As of today, New York has the highest cigarette taxes in the nation. Here's a look at how cigarette prices have climbed in recent years:
AVERAGE COST PER PACK AND AVERAGE TAX PER PACK
Nationwide
New York (outside NYC)
$0.38
$0.18
1970
$3.16
$0.76
2000
$4.63
$1.14
2008
$5.82
$1.50
June 2, 2008
$7.07
$2.75
June 3, 2008
Cost per pack
Tax per pack
Sources: Centers for Disease Control and Prevention; Campaign for Tobacco-Free Kids.

Cigarette Taxes Make Smokers Happier

Table 3: Distinguishing Impacts of Tax By Propensity to Smoke

	Very Happy	Somewhat Happy	Unhappy
Tax Rate - High Propensity	**-0.005**	**0.050**	**-0.055**
	(0.042)	**(0.045)**	**(0.029)**
Tax Rate - Low Propensity	**-0.005**	**0.003**	**0.011**
	(0.040)	**(0.040)**	**(0.017)**

Tax Rate	-0.027	-0.005	0.032
	(.033)	(.034)	(.020)
Propensity to Smoke	-0.069	-0.014	0.075
	(.038)	(.040)	(.026)
Propensity to Smoke * Tax Rate	**0.047**	**0.109**	**-0.156**
	(.078)	**(.070)**	**(.045)**

Gruber and Mullainathan

Common Themes

Theory Testing not Predicting

- Does race affect hiring?
  - NOT: What predicts hiring?
- Impact of commitment on smoker happiness
  - NOT: What predicts (smoker) happiness?

Causation not correlation

- Randomly assign name
  - NOT: Residual effect of race
- Exogenous tax variation
  - NOT: Direct effect of tax
  - NOT: quitting on happiness

Outline of Talk

- Some past papers of mine
- Barrier 1: Predicting “versus” Theory Testing
- Barrier 2: Correlation versus Causation
- How I would redo some old papers

Theory Testing

- What does it mean to test a theory?
- Is it any different than a simple hypothesis test? `

Kleinberg, Mullainathan, Tan and Zimmerman

A Fictional Example

- Anachronistic 19th century health researcher
  - Mind-body connection: pessimism theory

- How to test?

- Does room-mate health matter?

Sets Up An Experiment

- Randomly assigns roommates

Sets Up An Experiment

- Randomly assigns roommates
- Wants to control for other theories
  - Doctor quality

Sets Up An Experiment

- Randomly assigns roommates
- Wants to control for other theories
  - Doctor quality
  - Ensures roommate assignment does not lead to correlated doctor assignment

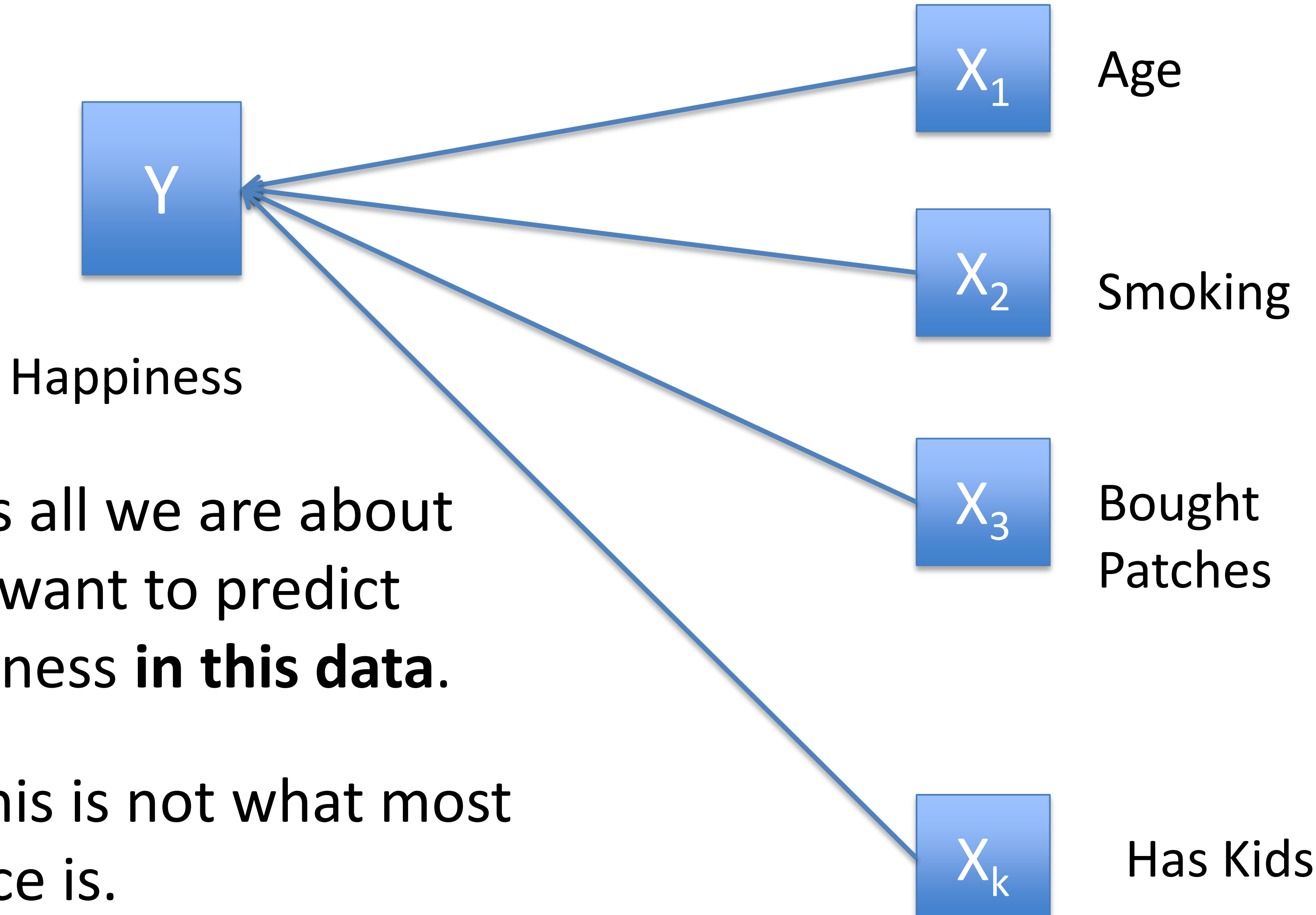
Pessimism

- Roommate health still matters!
- Concludes support for her theory
- But over time new data comes out
  - Someone notices that health of ward-mate matters
    - Even if you don’t ever see or or talk to ward-mate
  - Someone else had data on instrument/hand washing practices and find it matters
  - . . . .
- Germ theory eventually rises

What goes wrong?

- This was a good hypothesis test
  - Empirical relation is true: room-mate health does matter
- This was a less good theory test (pessimism theory)
  - Structural statement: Pessimism is not the reason
- Most science: theory testing not just hypothesis testing
- Requires a model of scientific theorizing

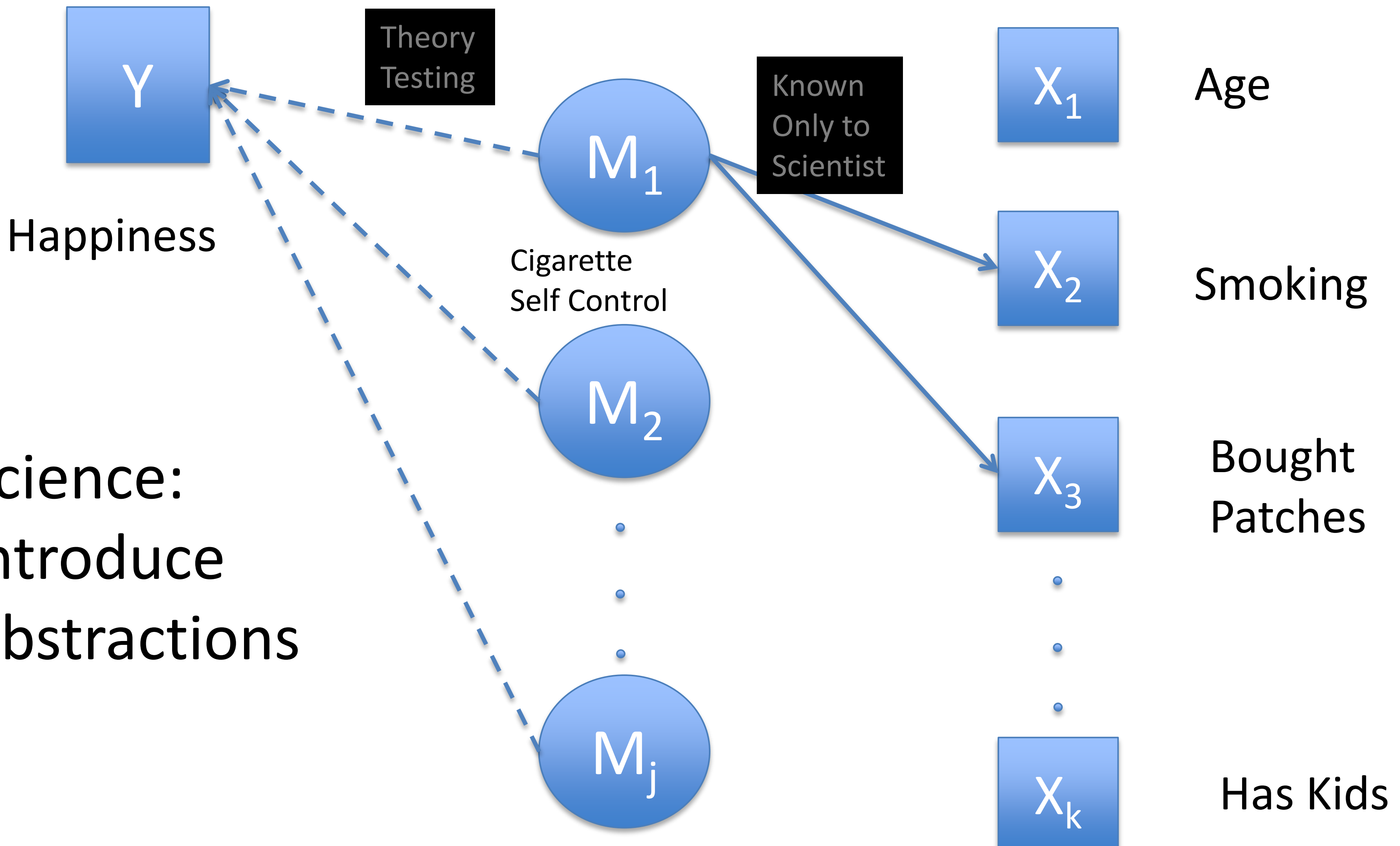
Atheoretical



This is all we are about
If we want to predict
happiness **in this data**.

But this is not what most
science is.

Modeling Modeling



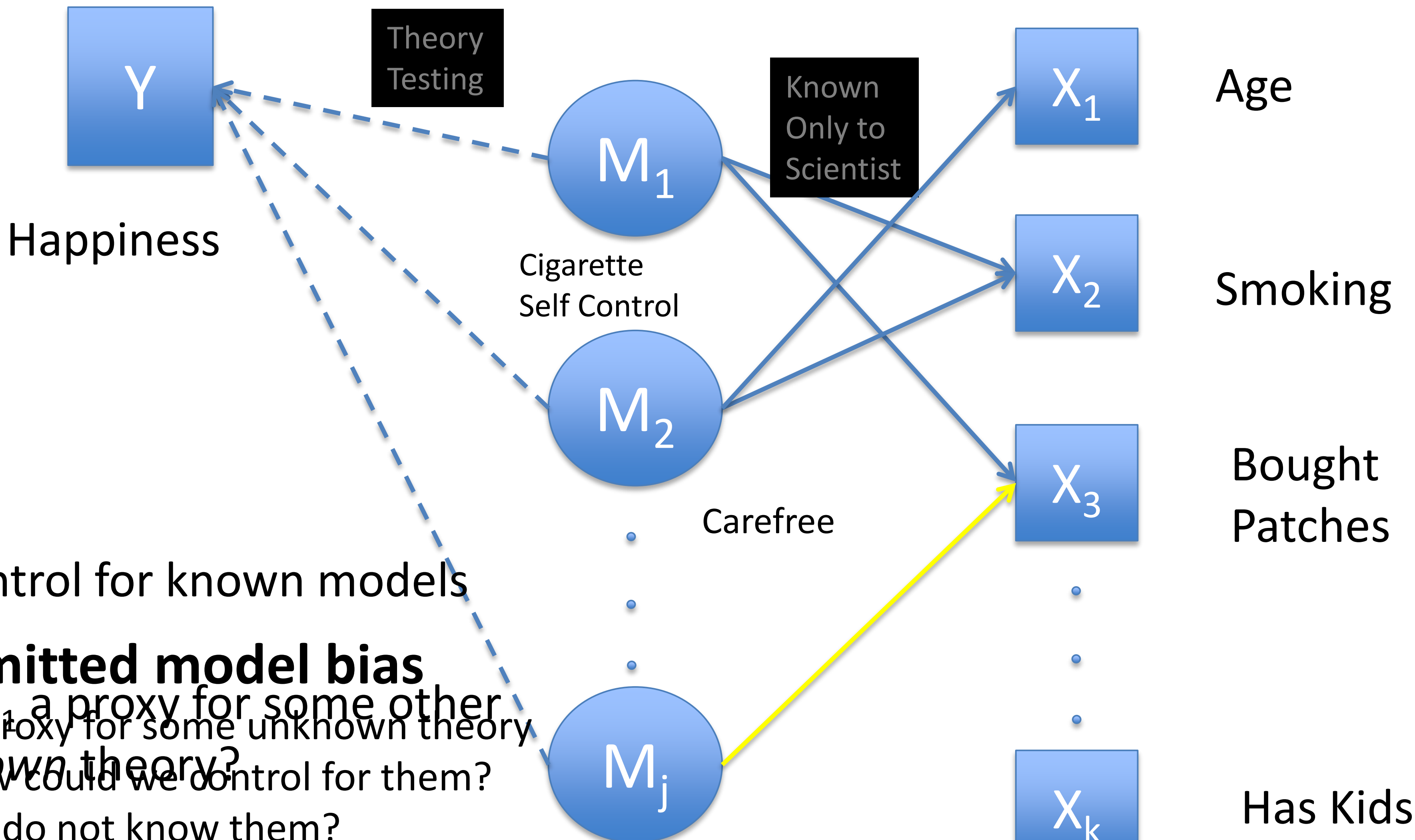
Science: Introduce Abstractions

Models

- Models allow generalization
  - Can map how X -> M in new contexts
  - Belief that M-> Y_0 implies M-> Y_i for some other *I*
    - Self control for smoking cigarettes -> for smoking weed

- Note:
  - Models are in scientists heads
  - Their structure extends past any one data set or Y
  - Latent variables analysis cannot extract them with *one* data set

Deduction



Control for known models

Omitted model bias

Is X_1 a proxy for some other *known* theory?

X_0 proxy for some unknown theory

How could we control for them?

We do not know them?

Control for those variables

Induction

1. Identify all variables *S* related to M_0
2. Predict Y using full variable set: Performance P^*
3. Predict Y without *S*: Performance P_{-S}
4. Inductive test: M_0 valid if

$$P^* > P_{-S}$$

- Key insight: do not curate inclusion.
  - Curate exclusion
- Note: Machine learning techniques are what allow induction
  - Regularization allows high dimensional data analysis

What does Induction Do?

- Controls for *all models covered* by X
  - Both known and unknown

- Suggests theory testing only as powerful as *diversity* of the data

- Does not induct NEW theories
  - Interpretability an issue but not the only issue

Prediction

Maximize predictive fit

Minimal curation of included features

- To make regularization easier

Induction

Maximize power of test

Minimal curation of included features

Maximal curation of **excluded** features in test

- Those related to theory to be tested

Deduction

Maximize power of test

Maximal curation of included features

- Control for known alternative theories

Example

- Prospect Theory:
  - Losses loom larger than gains

- Key test: Disposition Effect
  - Stocks in the loss domain (today price – purchase price) should be less likely to be sold

Deductive Test

Table 2: Odean statistics

	Balanced Sample
Proportion Gains Realized	0.536
Proportion Losses Realized	0.452
Difference	0.084
t-statistic	19.987***

Creating a Feature Set

- Four functions

Gain	Quartile	Max	Min
$p_{end} > p_{start}$	$p_t \in Q_k(\text{range})$	$p_t > \max(range)$	$p_t < \min(range)$

- Ranges

$A(i, j)$, where $0 \leq i < 10, 0 \leq j < 10$. These domains define a broad range of prices from the distant past around the buying action to the recent past close to time t. They are also commonly associated with the disposition effect.

$B(i, j)$, where $0 \leq i < 5, i + 1 \leq j < 5$. These define recent price movements.

$B(i, j)$, where $i = 0$, $j \in \{20, 40, \ldots, 200\}$. These define medium term to long-term price movements relative to t.

Deductive Test

Table 3: Inductive tests using linear regression.

feature sets	improvement in mean squared error	accuracy
$Gain(A(0,0))$ only	0.001673(***)	0.542156(***)

Inductive Test

Table 3: Inductive tests using linear regression.

feature sets	improvement in mean squared error	accuracy
$Gain(A(0,0))$ only	0.001673(***)	0.542156(***)
all features	0.018338	0.605383

Inductive Test

Table 3: Inductive tests using linear regression.

feature sets	improvement in mean squared error	accuracy
$Gain(A(0,0))$ only	0.001673(***)	0.542156(***)
all features	0.018338	0.605383
remove $Gain(A(0,0))$	0.018320	0.605480

Inductive Test

Table 3: Inductive tests using linear regression.

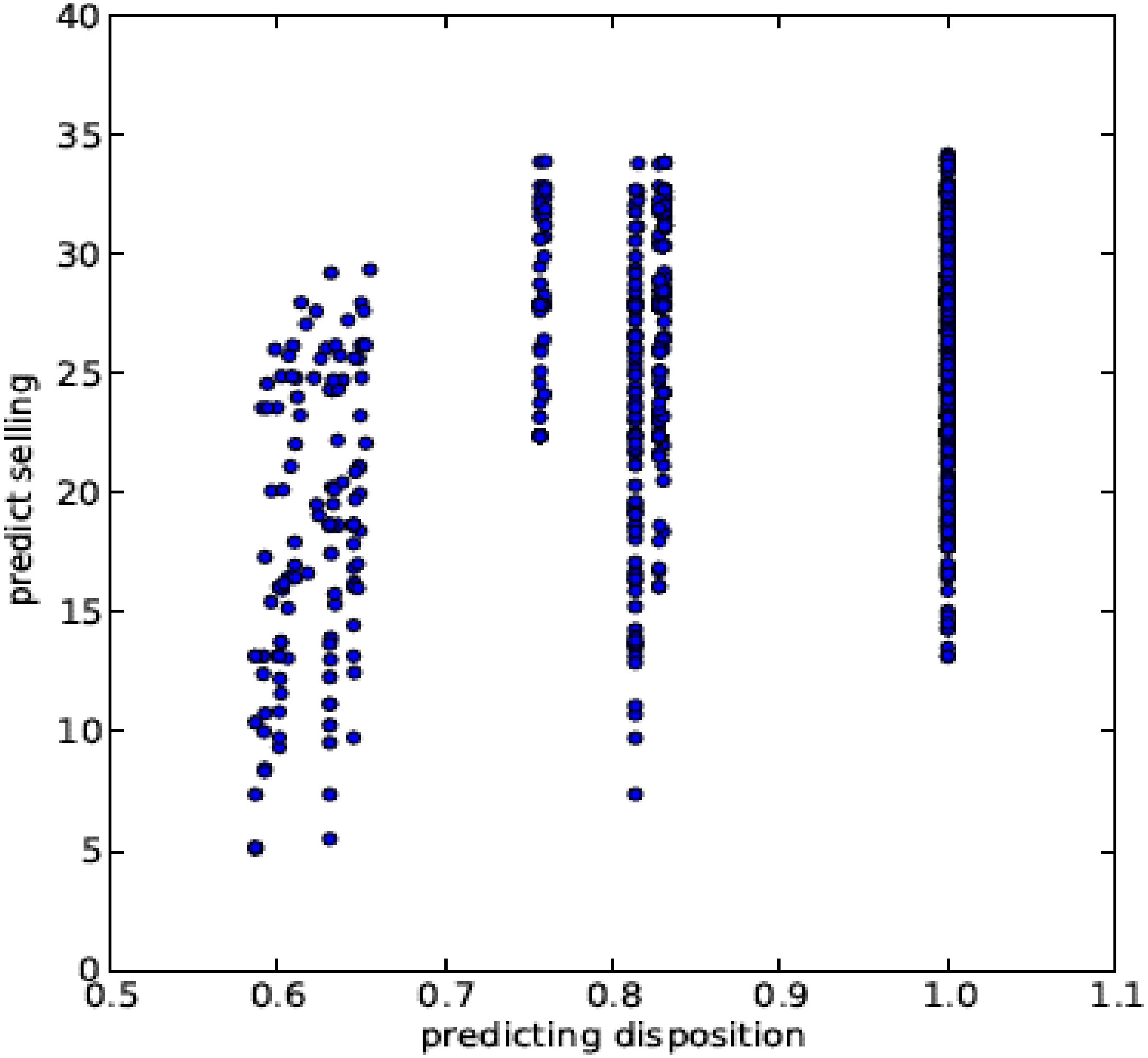
feature sets	improvement in mean squared error	accuracy
$Gain(A(0,0))$ only	0.001673(***)	0.542156(***)
all features	0.018338	0.605383
remove $Gain(A(0,0))$	0.018320	0.605480
remove $Gain(A(i,j)), 0 \leq i < 3, 0 \leq j < 3$	0.018284	0.604957
remove all $Gain(A(i,j))$	0.018216	0.604733

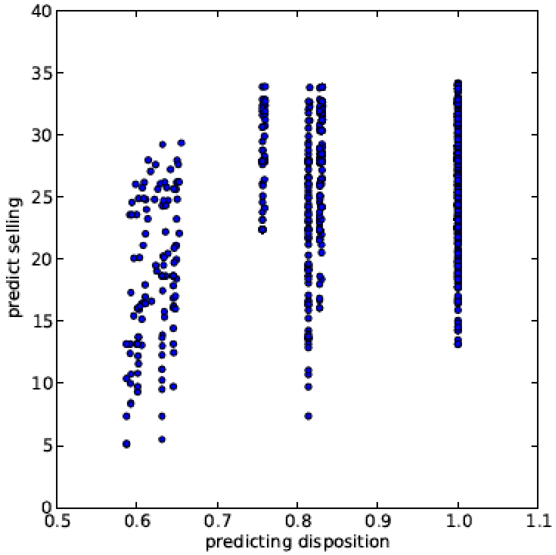
Inductive Test

Table 3: Inductive tests using linear regression.

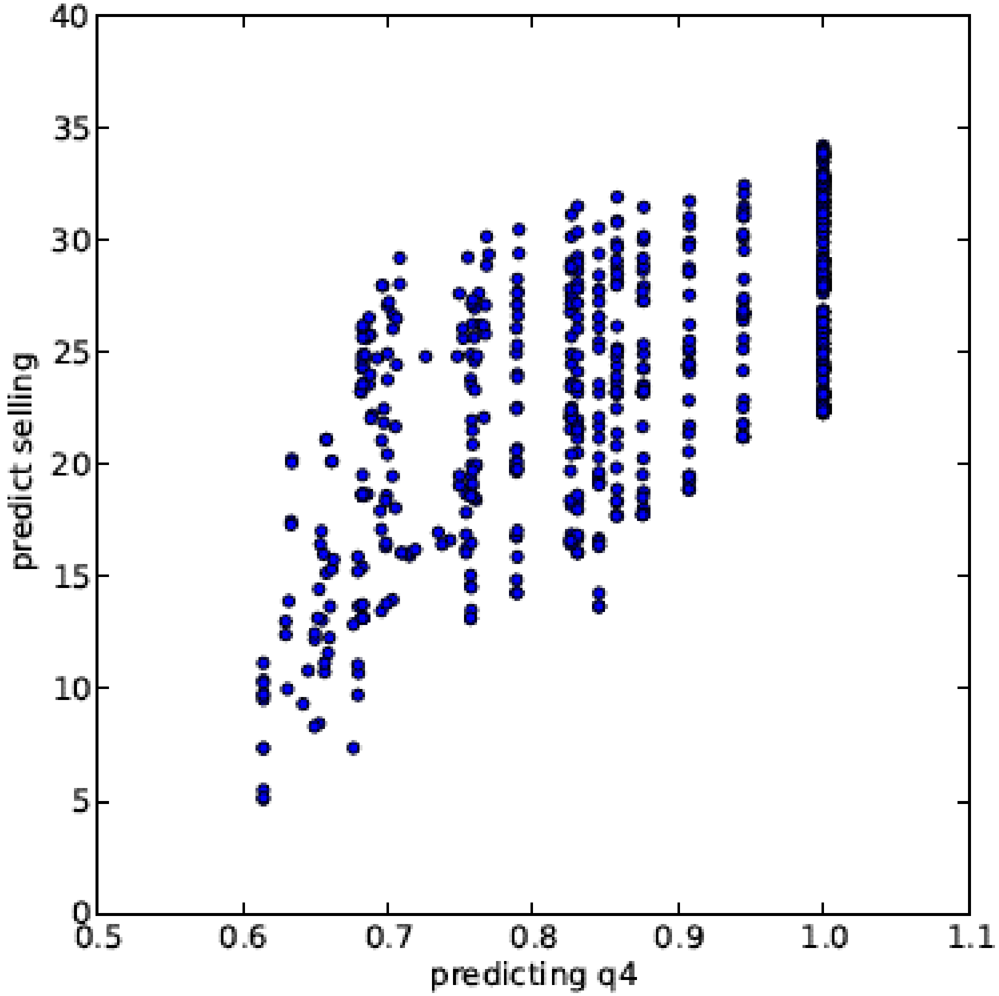
feature sets	improvement in mean squared error	accuracy
$Gain(A(0,0))$ only	0.001673(***)	0.542156(***)
all features	0.018338	0.605383
remove $Gain(A(0,0))$	0.018320	0.605480
remove $Gain(A(i,j)), 0 \leq i < 3, 0 \leq j < 3$	0.018284	0.604957
remove all $Gain(A(i,j))$	0.018216	0.604733
remove all quartile	0.014989(***)	0.589089(***)

The Clone Problem





(a) Reward in the game vs. predicting disposition



(b) Reward in the game vs. predicting q4

Table 11: Top patterns for trend and quartile features

Pattern			Prediction	Average reward
Δp_t	Δp_{t-1}	Quartile		
Up	Up	4	Sell	12.062
Down	Down	4	Sell	4.223
Down	Down	1	Sell	4.12
Down	Up	4	Sell	4.007
Down	Down	3	Sell	1.398
Down	Down	2	Sell	0.347
Up	Down	1	Hold	3.94
Up	Down	2	Hold	3.219
Up	Down	3	Hold	3.031
Up	Up	1	Hold	2.581
Up	Down	4	Hold	2.164
Up	Up	2	Hold	2.034

What is Needed

- More work to help us *test* structure provided by theories
  - Expansions of induction
  - Other methods?
- Note:
  - Currently we *use* theories to structure predictions
  - But testing theories different than using them

Outline of Talk

- Some past papers of mine
- Barrier 1: Predicting “versus” Theory Testing
- Barrier 2: Correlation versus Causation
- How I would redo some old papers

Policy

- Interested in taking an action (T—treatment). Should we or should we not?

- Core issue here is usually causal effect of T
  - The unknown: what will outcome Y be without treatment

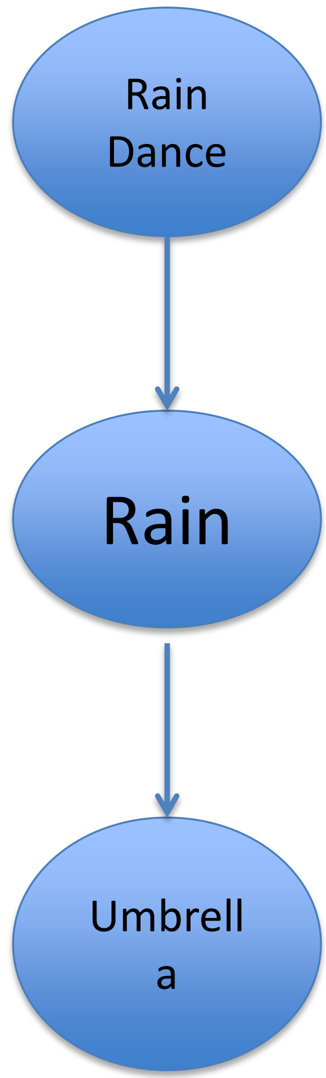
- Pretty far from machine learning

Two Important Policy Problems

- Rain Dances

- Umbrellas

A Very Complex Graphical model



Upstream Decisions
Causal Inference

Downstream Decisions
Predictions

Causality for Policy

- We focus on causal inference because that’s where the lamp shines

- But many policy problems are prediction problems

Example

- Defendant comes before judge
  – Judge must decide whether to release or not (bail)

- Defendant when out on bail can behave badly:
  – Fail to appear at case
  – Commit a crime

- Judge release based on *predicted* defendant misbehavior while out on bail

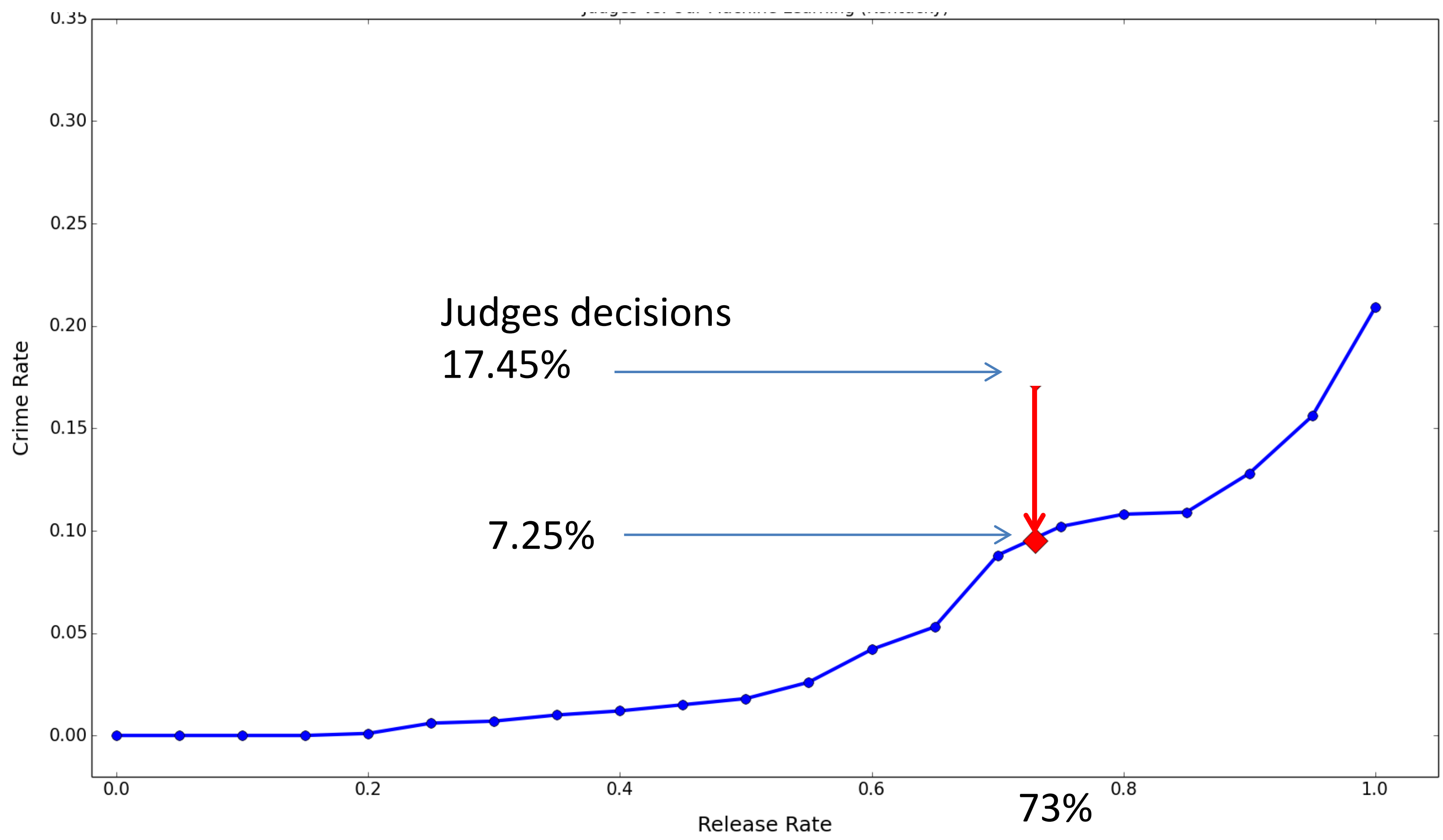
Important Policy Problem

- Each year police make over 12 million arrests

- Release vs. detain high stakes
  – Pre-trial detention spells avg. 2-3 months (can be up to 9-12 months)
  – Nearly 750,000 people in jails in US
  – Consequential for jobs, families as well as crime

Lakkaraju et. al.

Crime Rate Prediction



Notes: Standard errors too small to display on graph

Lakkaraju et. al.

Causality Lessons

1. Even for policy causality not always necessary

Causal Identification

- Difference – in – Differences
  - Smoking tax changes
  - Many policy changes use this paper

- Instrumental variable

- Regression Discontinuity

- Random assignment

Causal Identification

- **Difference – in – Differences**
  - Smoking tax changes
  - Many policy changes use this paper

- Instrumental variable

- Regression Discontinuity

- Random assignment

Table 5: Robustness Checks

Panel A: US Data					
Tax	0.032	0.033	0.036	0.070	0.015
	(.020)	(.020)	(.022)	(.021)	(.022)
Propensity to Smoke	0.075	-0.006	0.011	0.073	-0.190
	(.026)	(.036)	(.059)	(.025)	(.025)
Propensity to Smoke*Tax	**-0.156**	**-0.152**	**-0.167**	**-0.152**	**-0.104**
	(.045)	**(.049)**	**(.046)**	**(.042)**	**(.077)**
Panel B: Canadian Data					
Tax	0.000	0.000	0.010	0.018	0.003
	(.011)	(.011)	(.009)	(.016)	(.015)
Propensity to Smoke	0.096	0.072	0.180	0.097	0.096
	(.040)	(.061)	(.061)	(.040)	(.051)
Propensity to Smoke*Tax	**-0.048**	**-0.048**	**-0.082**	**-0.048**	**-.057**
	(.020)	**(.021)**	**(.026)**	**(.020)**	**(.031)**
Demographic Controls	Yes	Yes	Yes	Yes	Yes
State Dummies	Yes	Yes	Yes	Yes	Yes
Year Dummies	Yes	Yes	Yes	Yes	Yes
Propensity to Smoke*Unemployment Rate	No	Yes	No	No	No
State Dummies*Trend	No	No	Yes	No	No
Propensity to Smoke*Trend	No	No	No	Yes	No
State Dummies*Propensity to Smoke	No	No	No	No	Yes

Table 6: "Effect" of Other Taxes

Panel A: US Data

	Beer Tax	Gas Tax	Sales Tax	Total Revenues
Cigarette Tax	0.038	0.035	0.033	0.029
	(.024)	(.020)	(.020)	(.019)
Other Tax	-0.017	-0.001	0.003	-0.004
	(.008)	(.001)	(.004)	(.023)
Propensity to Smoke	0.055	0.060	0.060	0.125
	(.031)	(.048)	(.033)	(.038)
Propensity to Smoke*Cigarette Tax	**-0.181**	**-0.162**	**-0.159**	**-0.144**
	(.055)	**(.043)**	**(.045)**	**(.043)**
Propensity to Smoke*OtherTax	**0.034**	**0.001**	**0.003**	**-0.037**
	(.014)	**(.003)**	**(.006)**	**(.021)**

Can Improve on D-in-D

- Choose control variables using a prediction model
  – Controlling for confounds = predicting the residual

- Replace “by hand” robustness checks with “machine” robustness

Can Improve on Other Strategies

- Instrumental Variables
  - Choice of exact instrument – prediction problem
- Regression discontinuity
  - Choice of control set
- Propensity score matching
  - Predict treatment assignment

Causality Lessons

1. Even for policy causality not always necessary

2. Many causal identification strategies can be improved by machine learning

What is Needed

- Working on machine learning issues specific to policy contexts
  - More explicit integration of the policy *decision* into the prediction framework

- Integration of machine learning “technology” with causal inference “technology”

Outline of Talk

- Some past papers of mine
- Barrier 1: Predicting “versus” Theory Testing
- Barrier 2: Correlation versus Causation
- How I would redo some old papers

Discrimination

1. Complement experiment:
   – Is race predictive with “machine learning controls”?

2. Massively increase scale of experiment

3. Understand heterogeneity of treatment

Discrimination

TABLE 4—AVERAGE CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES AND RESUME QUALITY

	Panel A: Subjective Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (*p*-value)
White names	8.50	10.79	1.27	2.29
	[1,212]	[1,223]		(0.0557)
African-American names	6.19	6.70	1.08	0.51
	[1,212]	[1,223]		(0.6084)
	Panel B: Predicted Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (*p*- value)
White names	7.18	13.60	1.89	6.42
	[822]	[816]		(0.0000)
African-American names	5.37	8.60	1.60	3.23
	[819]	[814]		(0.0104)

Cigarette Smokers

1. Much better data
   - Happiness from twitter, instagram, facebook
   - Smoking could be inferred directly

1. Better casual inference
   - Machine learning for robustness checks

2. Inductive hypothesis testing

Conclusion

There’s a lot of profits in the orange juice market