Statistics, Machine Learning, and Public Policy Analysis

Propensity score analysis of observational data

Greg Ridgeway

http://www.i-pensieri.com/gregr

RAND Statistics Group, Santa Monica, CA

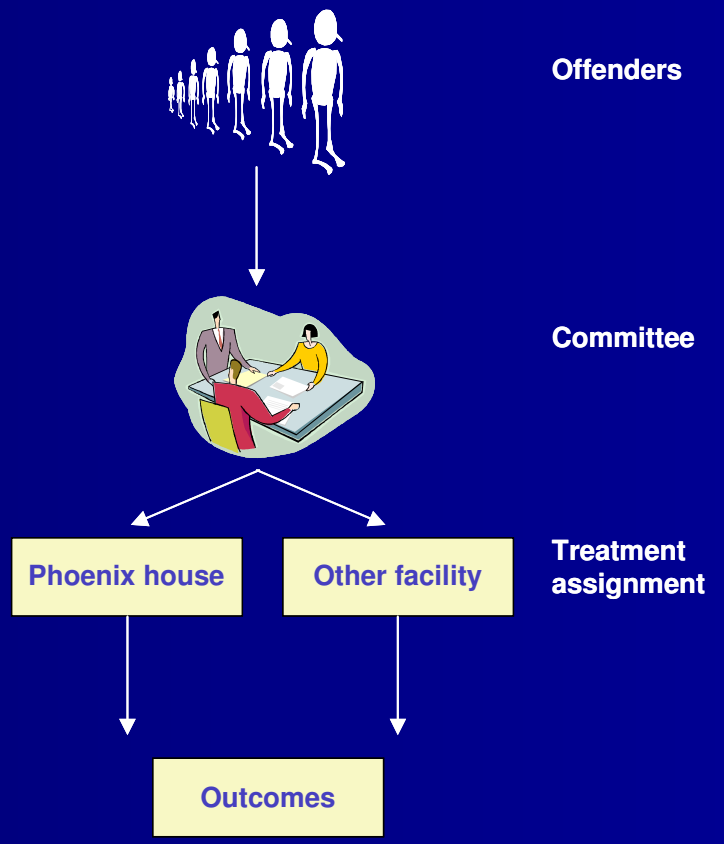
Problems I face at RAND

Assessing public policy almost always asks "what would have happened if..."

- youths sent to residential drug treatment had been sent to alternative programs
- officers treated drivers that they stopped equitably regardless of race
- military reservists were offered a DoD subsidized health plan

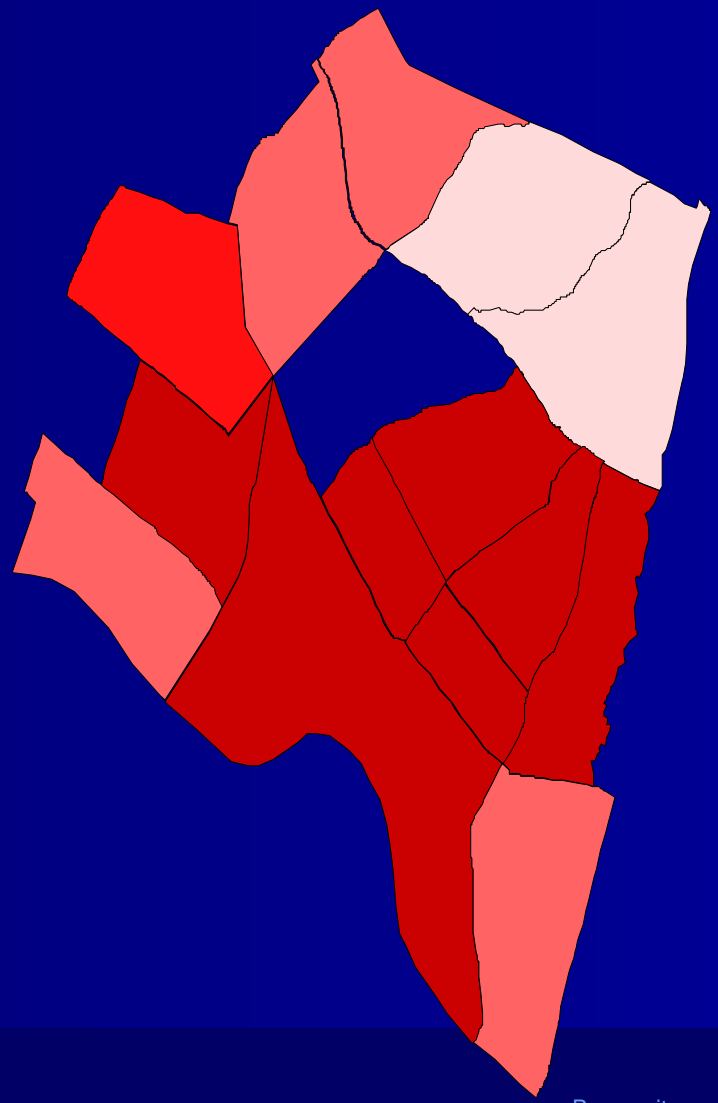
Example: Phoenix house

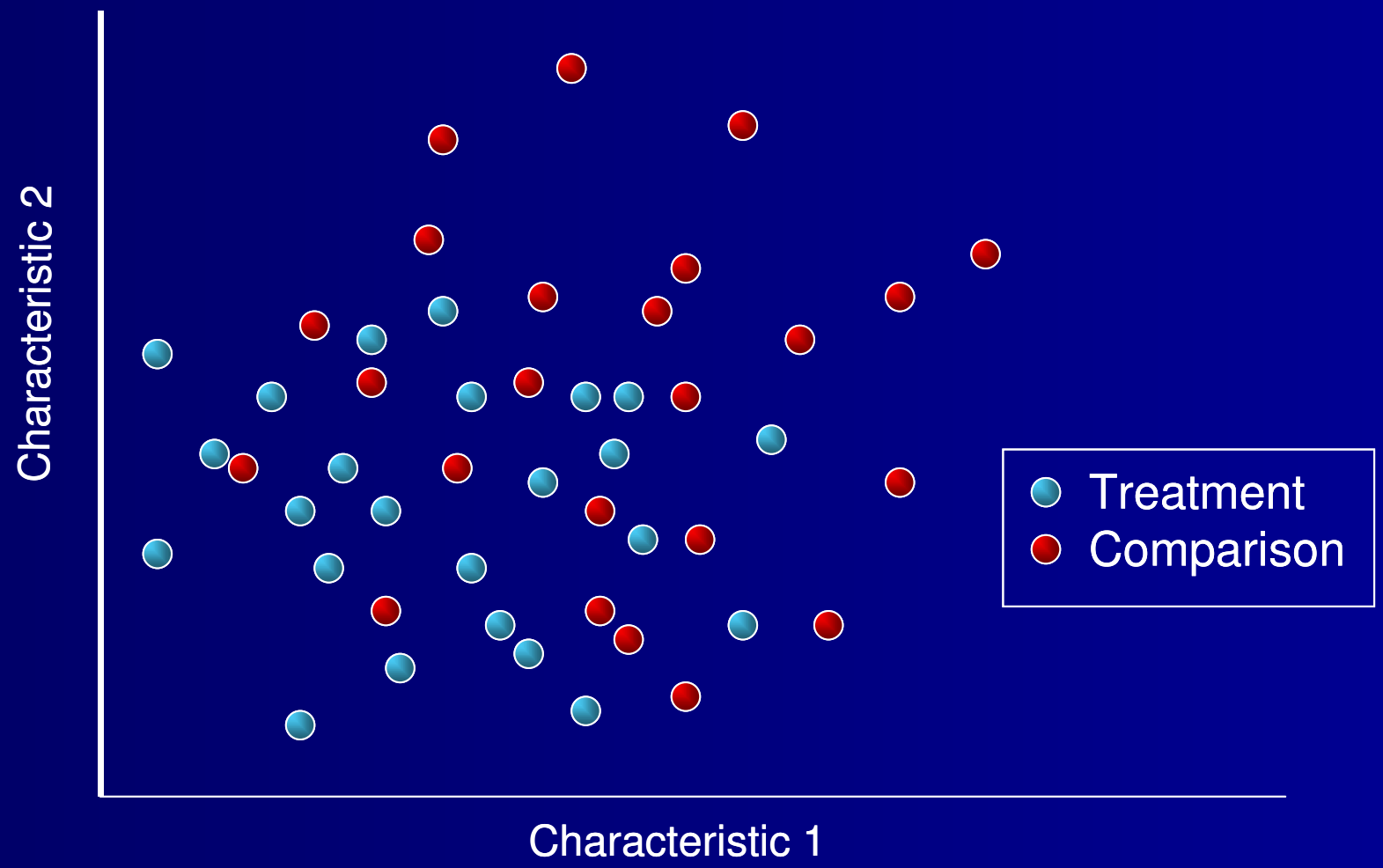
- The treatment assignments are non-random
- Youths in the treatment have no violent criminal history, moderate drug use
- A direct comparison ignores baseline differences

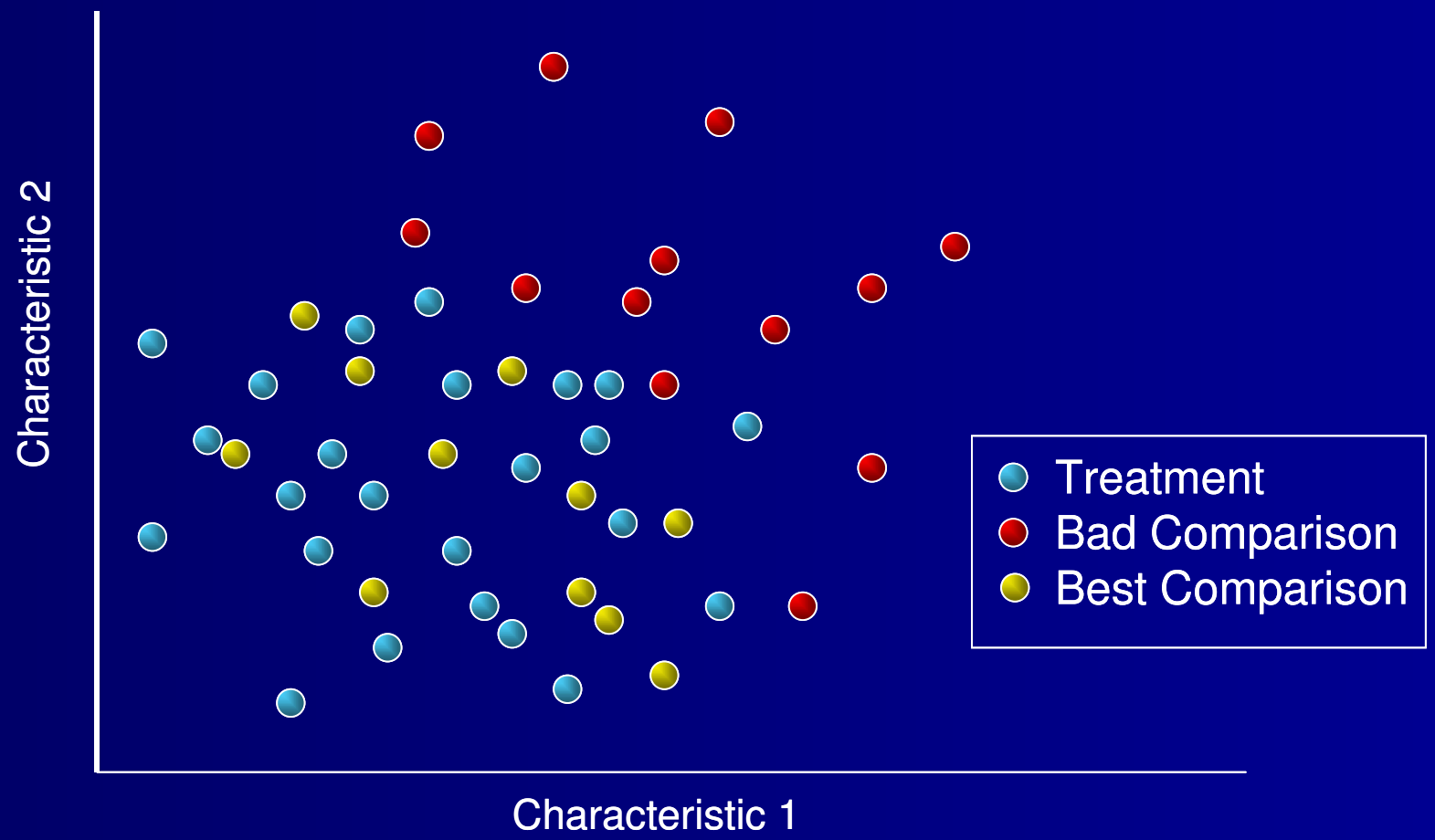


Example: Racially biased policing

- Drivers of different races may traverse different streets
- Policing practices can vary by neighborhood, crime patterns vary
- Direct comparisons of black drivers to white drivers ignore these differences




Characteristic 2
Characteristic 1
Treatment
Comparison


Characteristic 2
Characteristic 1
Treatment
Bad Comparison
Best Comparison

Adjusting for $\mathbf{x}$

- We must "adjust for" or "control for" $\mathbf{x}$, commonly interpreted as

$$y_{obs} = \beta_0 + \beta_1 T + \beta_2 x_1 + \ldots + \beta_{d+1} x_d + \epsilon$$

- Requires a lot of "expert knowledge" for selecting $\mathbf{x}$
- Challenging to diagnose
- When the two groups have little overlap in terms of $\mathbf{x}$, the model assumptions completely drive the result. This situation is difficult to detect

Causal estimation

- Each individual has a control outcome, y_0, and a treatment outcome, y_1

Average treatment effect of the treated

$$= \mathrm{E}(y_1|T=1) - \mathrm{E}(y_0|T=1)$$

$$\mathrm{E}(y_1|T=1) \approx \frac{\sum_{i \in T} y_{1i}}{N_T}$$

Causal estimation

$$\mathrm{E}(y_0|T=1) \;=\; \iint y_0 f(y_0, \mathbf{x}|T=1)\, d\mathbf{x}\, dy_0$$

Causal estimation

$$\begin{aligned} \mathrm{E}(y_0|T=1) &= \iint y_0 f(y_0, \mathbf{x}|T=1)\, d\mathbf{x}\, dy_0 \\ &= \iint y_0 \frac{f(y_0, \mathbf{x}|T=1)}{f(y_0, \mathbf{x}|T=0)} f(y_0, \mathbf{x}|T=0)\, d\mathbf{x}\, dy_0 \end{aligned}$$

- Apply Bayes Theorem to $f(y_0, \mathbf{x}|T)$.

Causal estimation

$$\mathrm{E}(y_0|T=1) =$$
$$\iint y_0 \frac{f(T=1|y_0,\mathbf{x})}{f(T=0|y_0,\mathbf{x})} \frac{f(y_0,\mathbf{x})}{f(y_0,\mathbf{x})} \frac{f(T=0)}{f(T=1)} f(y_0,\mathbf{x}|T=0)\, d\mathbf{x}\, dy_0$$

- Assume $f(T|y_0,\mathbf{x}) = f(T|\mathbf{x})$
- This the strong ignorability assumption. If $\mathbf{x}$ contains all the information used in assigning treatments, then this assumption holds.

Causal estimation

$$\mathrm{E}(y_0|T=1) = \frac{f(T=0)}{f(T=1)} \iint y_0 \frac{p(\mathbf{x})}{1-p(\mathbf{x})} f(y_0, \mathbf{x}|T=0)\, d\mathbf{x}\, dy_0$$

$$\mathrm{E}(y_0|T=1) \approx \frac{\sum_{i \in C} w_i y_{0i}}{\sum_{i \in C} w_i}$$

Balance on $\mathbf{x}$

- Even if the causal interpretation or strong ignorability is suspect, weighting comparison subjects with $p/(1-p)$ matches the joint distributions of $\mathbf{x}$

$$f(\mathbf{x}|T=1) \propto \frac{p(\mathbf{x})}{1-p(\mathbf{x})} f(\mathbf{x}|T=0)$$

Summary of the method

$$\mathrm{E}(y_1|T=1) \approx \frac{\sum_{i=1}^{N} t_i y_{1i}}{N_T}$$

$$\mathrm{E}(y_0|T=1) \approx \frac{\sum_{i=1}^{N} w_i(1-t_i)y_{0i}}{\sum_{i=1}^{N} w_i(1-t_i)}$$

- $w_i = \frac{p_i}{1-p_i}$, and p_i is the probability that subject i goes to the treatment group
- Derivation requires that treatment assignments depend only on $\mathbf{x}$
- For years machine learning has focused on estimating $p(\mathbf{x})$ when $\mathbf{x}$ is high-dimensional

Logistic log-likelihood

- Let $p(\mathbf{x}) = 1/\left(1 + e^{-F(\mathbf{x})}\right)$
- Find $F(\mathbf{x})$ to maximize

$$\ell(F) = \mathrm{E}_{t,\mathbf{x}}\, tF(\mathbf{x}) - \log\left(1 + e^{F(\mathbf{x})}\right)$$

Gradient boosting

- Initialize $F(\mathbf{x}) = 0$
- Find a $g(\mathbf{x})$ such that $F(\mathbf{x}) + \lambda g(\mathbf{x})$ has a larger log-likelihood than $F(\mathbf{x})$
- The $g(\mathbf{x})$ offering the greatest local improvement in the log-likelihood is

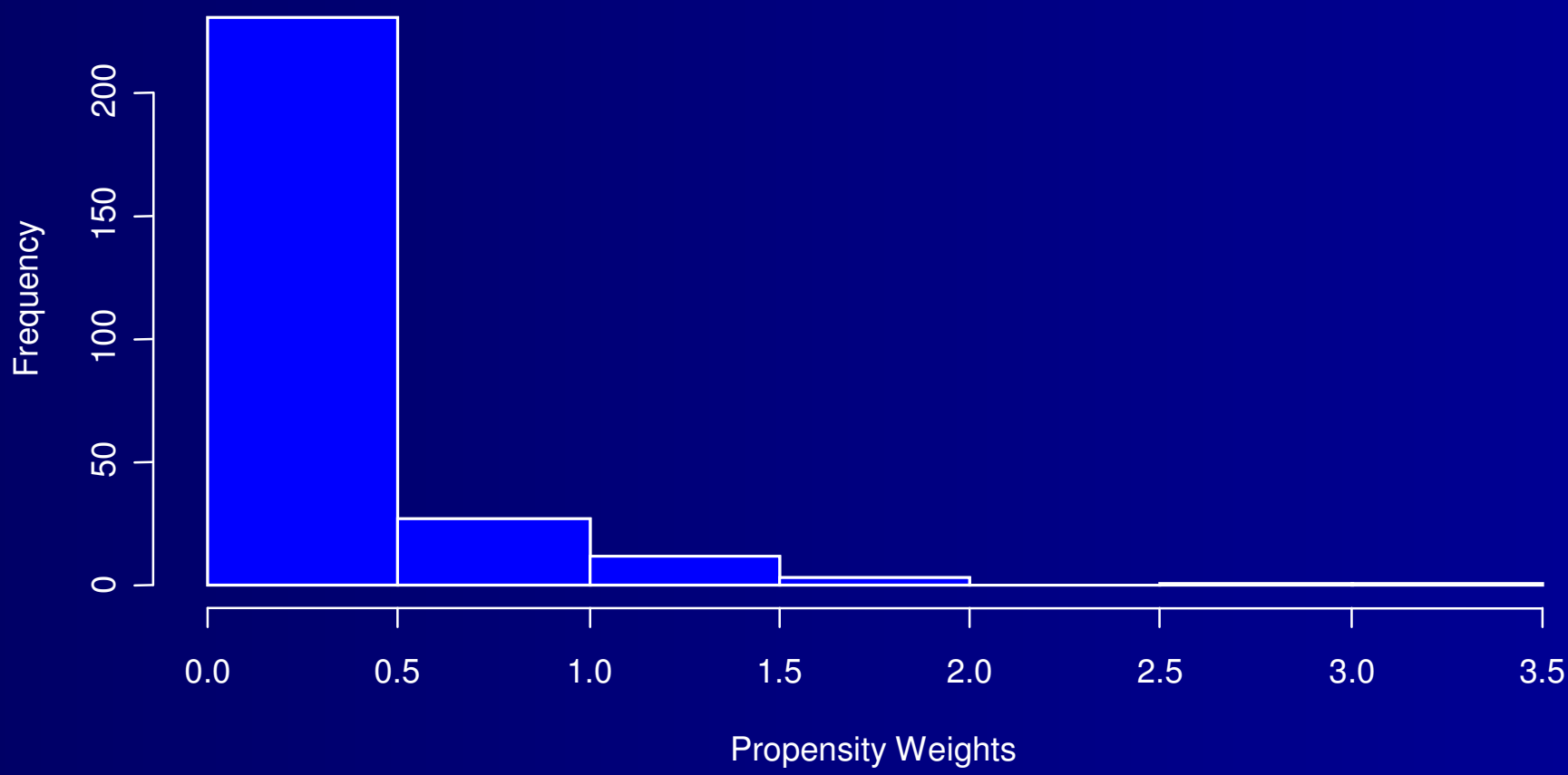
$$g(\mathbf{x}) = \mathrm{E}\left[t - \frac{1}{1 + e^{-F(\mathbf{x})}} | \mathbf{x}\right]$$

- We will use regression trees to estimate $\mathrm{E}\left[t - p(\mathbf{x}) | \mathbf{x}\right]$

Advantages

1. Excellent estimation of $p(\mathbf{x})$
2. The resulting model handles continuous, nominal, ordinal, and missing x's
3. Invariant to one-to-one transformations of the x's
4. Model higher interaction terms with more complex regression trees
5. Implemented in R in the `gbm` library

Observed control group weights



$$\mathrm{ESS} = (\sum w_i)^2 / \sum w_i^2$$

Balance of subject features

Variable	treatment mean	weighted control mean	unweighted control mean	effect size weighted	effect size unweighted
Treatment motivation	2.52	2.22	1.35	0.23	0.89
Environmental risk	30.61	31.09	28.94	-0.05	0.17
Substance use	7.61	6.94	4.59	0.16	0.69
Complex behavior	12.84	13.00	12.11	-0.02	0.09
Age	15.82	15.76	15.31	0.07	0.56
⋮				⋮	⋮
ESS	175	107.5	274		
Average |ES|				**0.107**	**0.307**

Results: Phoenix house

	Unweighted	GBM	Logit, 0.05	Logit, 0.20
Estimated Treatment Effect (confidence interval)				
Marijuana	-11.8	-5.9	-1.9	-5.2
	(-19.7, -3.8)	(-16.2, 4.3)	(-12.7, 8.8)	(-24.4, 14.1)
Alcohol	-1.2	2.8	1.5	3.1
	(-5.5, 3.0)	(-3.6, 9.3)	(-10.2, 13.3)	(-10.5, 16.7)

Results: Phoenix house

	Unweighted	GBM	Logit, 0.05	Logit, 0.20
Estimated Treatment Effect (confidence interval)				
Marijuana	-11.8	-5.9	-1.9	-5.2
	(-19.7, -3.8)	(-16.2, 4.3)	(-12.7, 8.8)	(-24.4, 14.1)
Alcohol	-1.2	2.8	1.5	3.1
	(-5.5, 3.0)	(-3.6, 9.3)	(-10.2, 13.3)	(-10.5, 16.7)
Measures of model fit				
Deviance	NA	466.4	539.2	511.4
ASAM	0.31	0.11	0.14	0.20
SE, Marijuana	4.0	5.2	6.6	11.8
SE, Alcohol	2.2	3.3	7.2	8.3

Balance of driver features

	% Black drivers N=3,703	% Non-black drivers (weighted) ESS=2,089	% Non-black drivers (unweighted) N=3,033
Region			
A	31%		27%
B	32%		14%
C	1%		3%
D	11%		21%
E	9%		8%
F	3%		6%
G	14%		21%

Balance of driver features

	% Black drivers N=3,703	% Non-black drivers (weighted) ESS=2,089	% Non-black drivers (unweighted) N=3,033
Region			
A	31%	29%	27%
B	32%	30%	14%
C	1%	1%	3%
D	11%	13%	21%
E	9%	9%	8%
F	3%	3%	6%
G	14%	15%	21%

Balance of driver features

	% Black drivers N=3,703	% Non-black drivers (weighted) ESS=2,089	% Non-black drivers (unweighted) N=3,033
Time			
12am-4am	16%	13%	7%
4am-8am	4%	4%	4%
8am-12pm	17%	17%	21%
12pm-4pm	20%	23%	28%
4pm-8pm	24%	25%	26%
8pm-12am	20%	18%	13%
Age			
Under 18	3%	3%	3%
18-29	47%	45%	38%
30-39	22%	25%	26%
40+	28%	27%	33%

Stop outcomes

	Black drivers	Non-black drivers (weighted)
Citation rate	68%	72%
	(66.6%, 69.7%)	(70.3%, 74.5%)
0-9 minutes	47%	53%
	(45.4%, 48.6%)	(51.0%, 56.1%)
Pat search	2.7%	2.6%
	(2.1%, 3.2%)	(1.8%, 3.4%)
Consent search	2.2%	1.6%
	(1.7%, 2.7%)	(0.9%, 2.2%)
Probable cause	3.2%	1.4%
	(2.6%, 3.9%)	(0.8%, 2.0%)

Remaining questions

- The bias/variance tradeoff is difficult to optimize. Aggressively trying to balance on subject features costs power
- Subject features associated with group assignment but not outcomes can greatly increase variance without offering any reduction in bias
- Detecting insufficient overlap between the groups is fairly easy using ESS or histograms of estimated propensity scores
- Sensitivity to the strong ignorability assumption...

Assessing sensitivity

- Assume that w_i is off by a factor $a_i \in [1/G, G]$ due to an unobserved factor
- Assume this unobserved factor is strongly associated with the outcome
- Find a_i's that maximize and another set of a_i's that minimize the estimated treatment effect.

G	Maximum	Minimum
1.24	0.00	-11.32
2.00	13.78	-20.58
3.00	23.19	-26.52
4.00	28.06	-29.87

Conclusions

- Statistics, machine learning, and policy analysis find a happy marriage in propensity score studies
- Statistics pins down the analytical question
- Machine learning balances the groups by accurately assessing the propensity score
- Policy analysis inspects the groups for balance on the essential features and interprets differences