Potential Outcomes: From Matching to Propensity Scores

Dr. Jordi Mur-Petit
datascience.barcelona

UNIVERSITAT DE BARCELONA

Tecniospring INDUSTRY

ACCIÓ
Generalitat de Catalunya
Government of Catalonia

Intended Learning Outcomes

- Potential Outcomes framework
- The fundamental problem of CI
- Dealing with the CIA and Exchangeability
  - Matching
  - Propensity scores

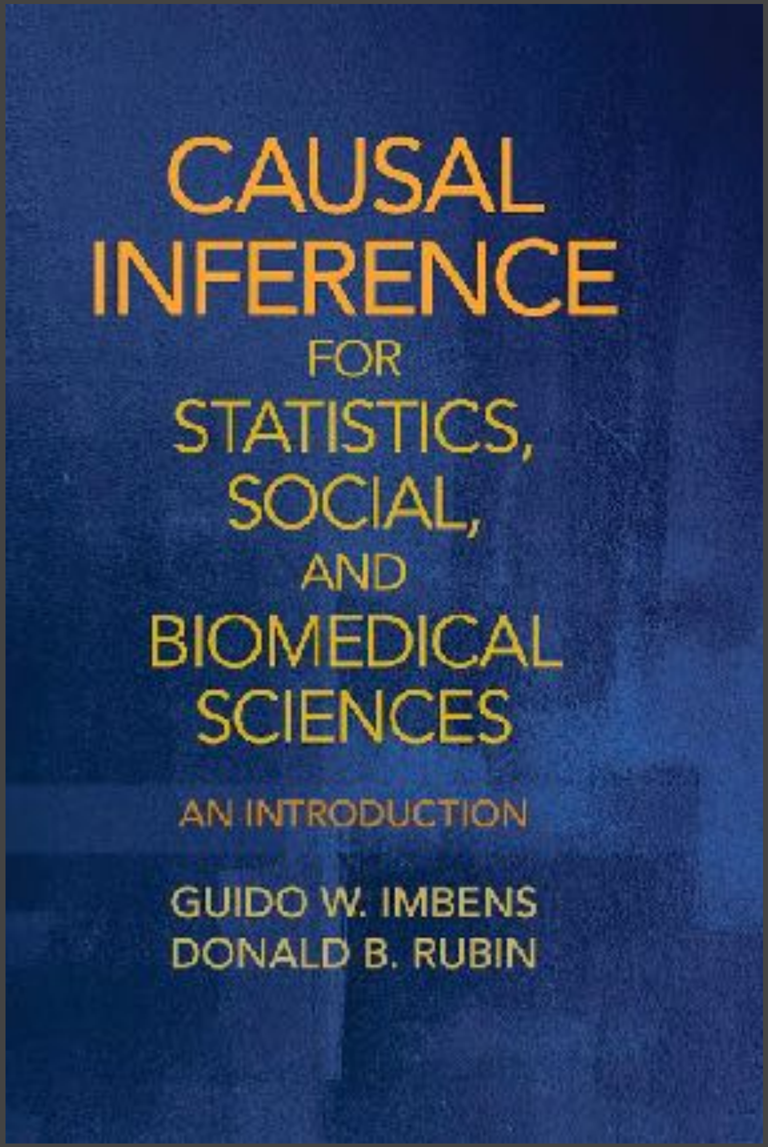
Potential Outcomes framework
a.k.a. (Neyman-)Rubin causal model

- 1923 - Jerzy **Neyman** - first idea, limited to RCTs

- 1974 - Donald **Rubin** - extension to observational studies

- 1994 - **Imbens & Angrist** - application to economics (instrumental variables)

- Today - Applied throughout medicine, economics, social sciences



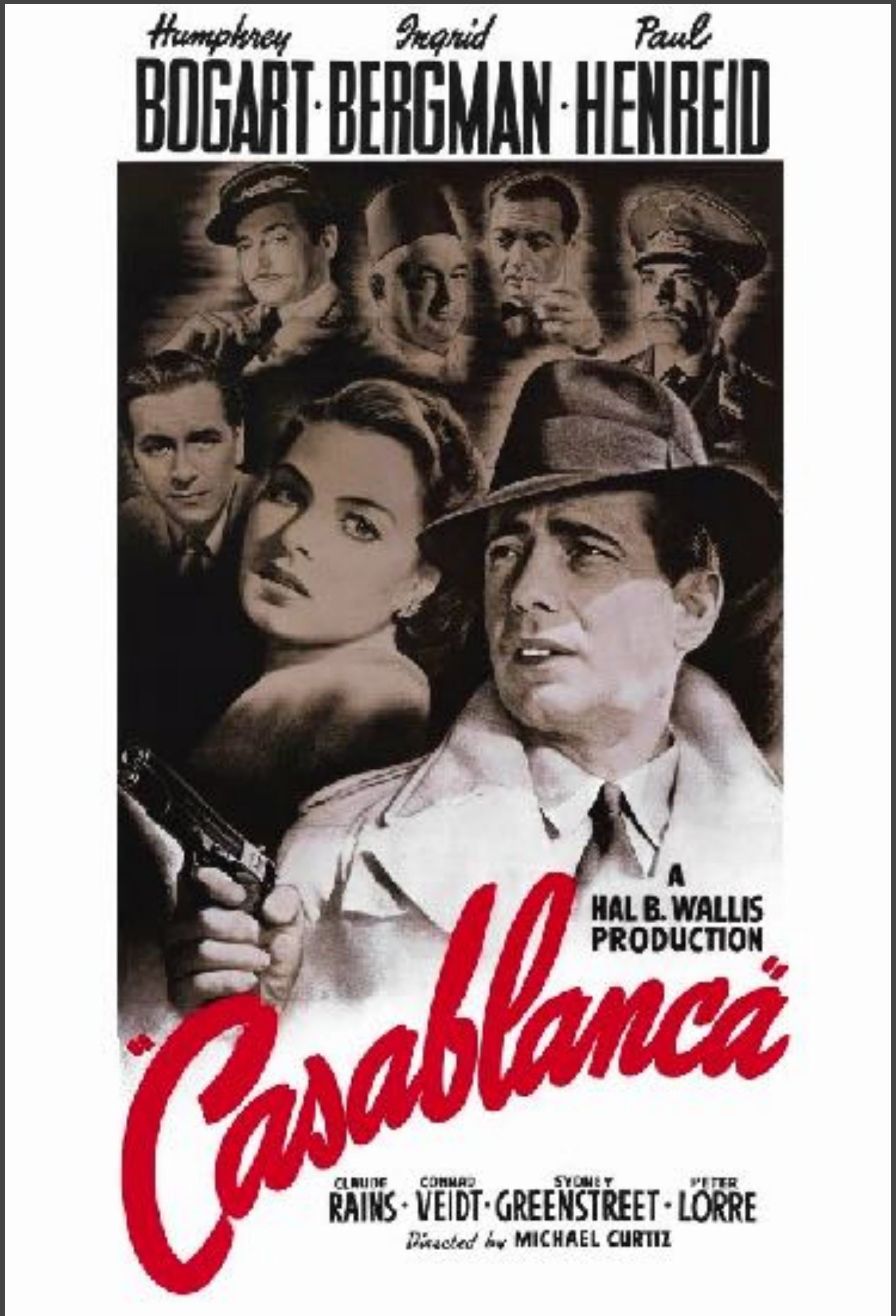
CAUSAL INFERENCE FOR STATISTICS, SOCIAL, AND BIOMEDICAL SCIENCES

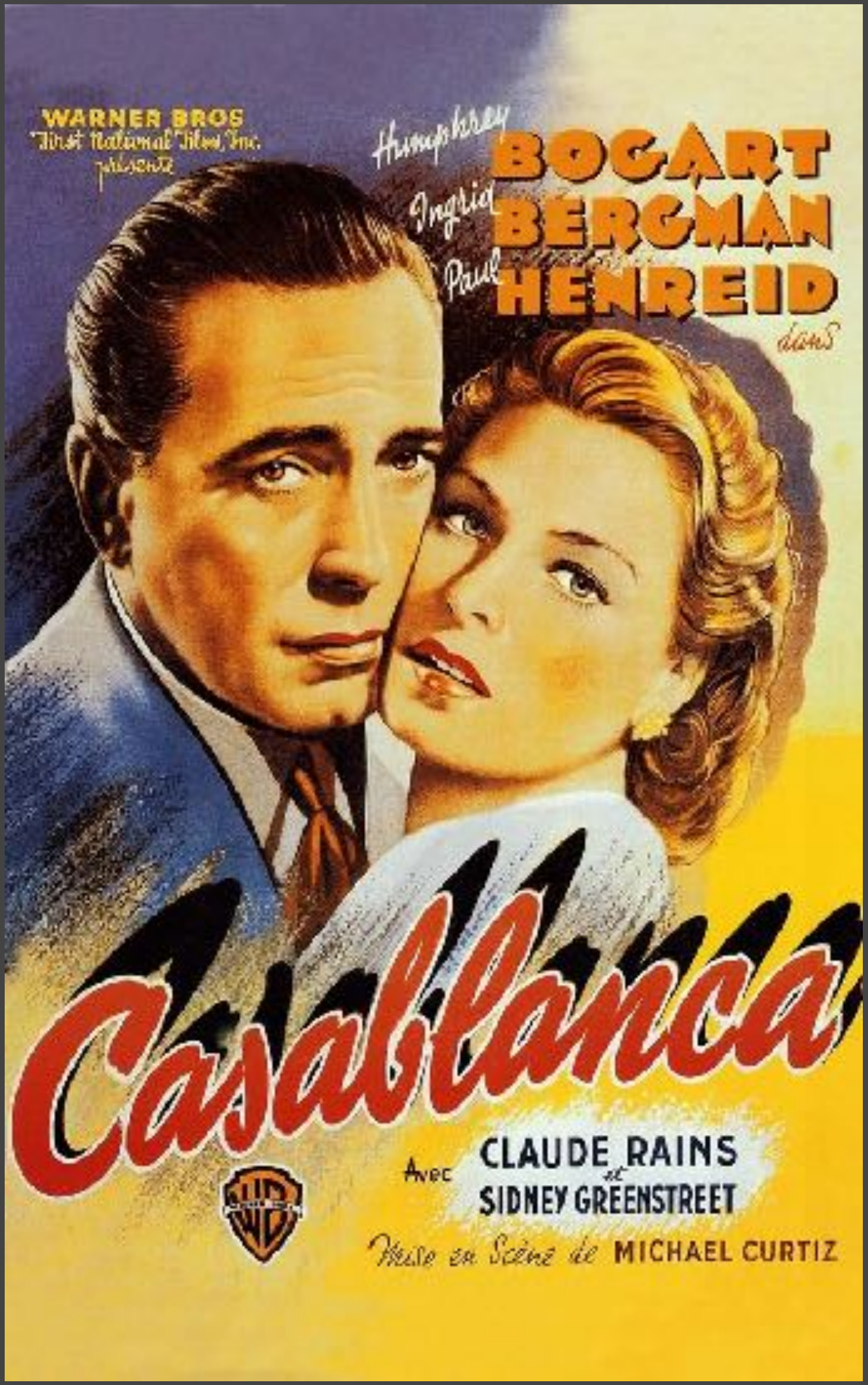
AN INTRODUCTION

GUIDO W. IMBENS
DONALD B. RUBIN

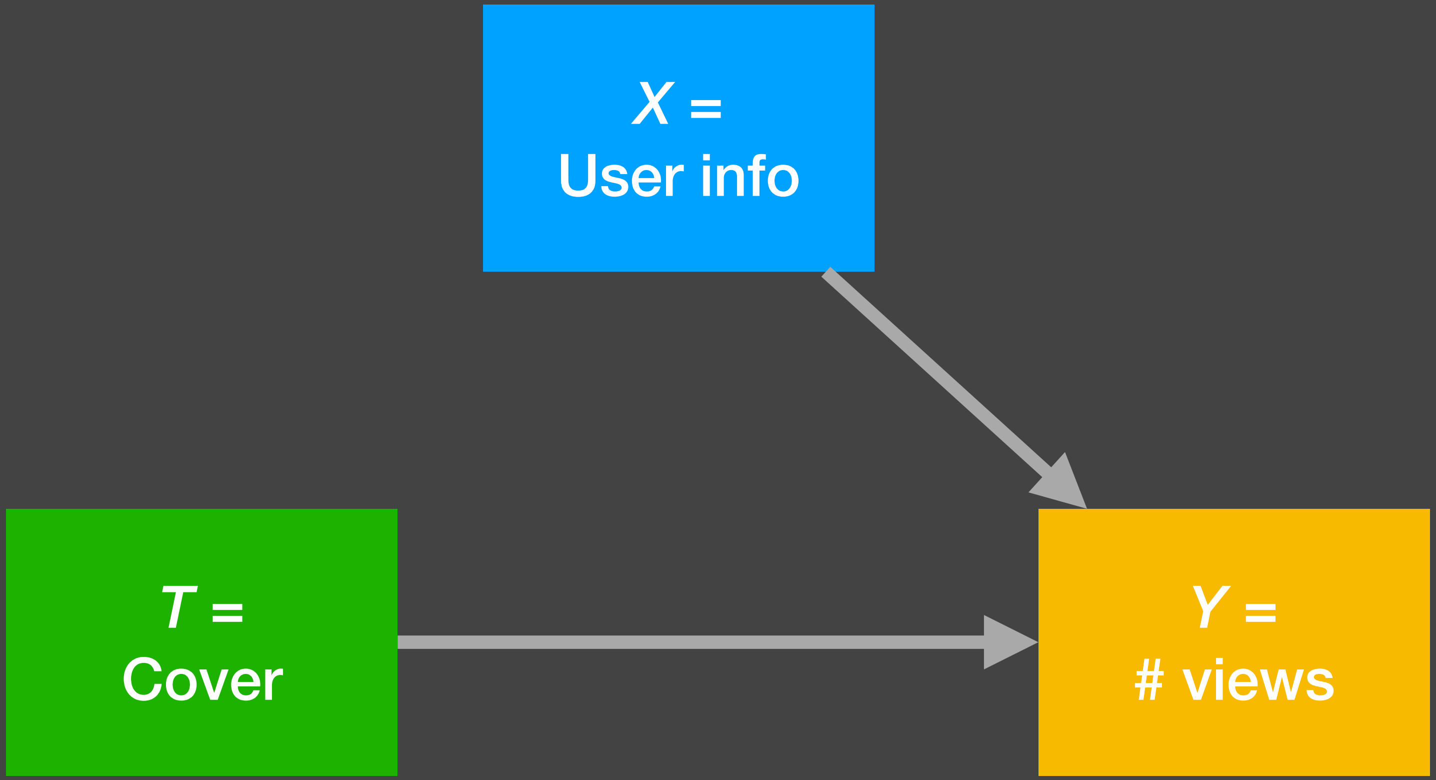
J. Neyman: «Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes», Master's Thesis (1923)
D. Rubin: «Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies». J. Educ. Psychol. 66, 688–701 (1974)
G.W. Imbens & J.D. Angrist: «Identification and estimation of local average treatment effects», Econometrica 61, 467-476 (1994).

Which cover would *make you* watch this movie?


Humphrey Ingrid Paul
BOGART · BERGMAN · HENREID
A HAL B. WALLIS PRODUCTION
"Casablanca"
CLAUDE RAINS · CONRAD VEIDT · SYDNEY GREENSTREET · PETER LORRE
Directed by MICHAEL CURTIZ


WARNER BROS
"First National" Films, Inc.
présente
Humphrey BOGART
Ingrid BERGMAN
Paul HENREID
dans
Casablanca
Avec CLAUDE RAINS et SIDNEY GREENSTREET
Mise en Scène de MICHAEL CURTIZ

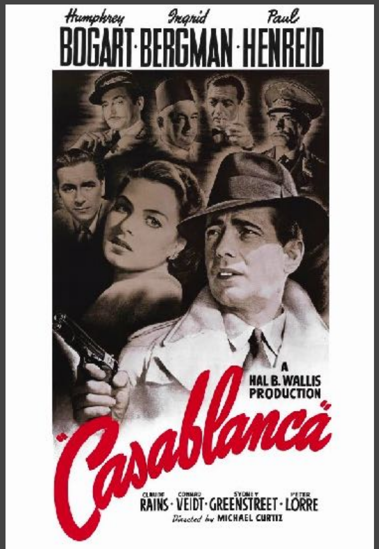
Which cover would *make you* watch this movie?



X =
User info

T =
Cover

Y =
# views


Humphrey Ingrid Paul
BOGART·BERGMAN·HENREID
A HAL B. WALLIS PRODUCTION
Casablanca
CLAUDE RAINS · CONRAD VEIDT · SYDNEY GREENSTREET · PETER LORRE
Directed by MICHAEL CURTIZ

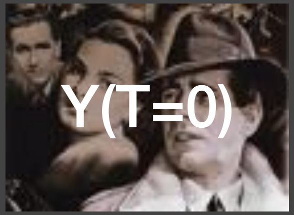
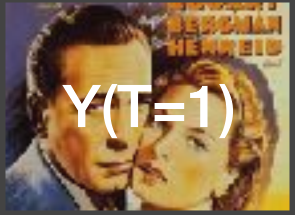
WARNER BROS "First National Pictures, Inc." présente
Humphrey BOGART
Ingrid BERGMAN
Paul HENREID
dans
Casablanca
Avec CLAUDE RAINS
SIDNEY GREENSTREET
Mise en Scène de MICHAEL CURTIZ

Which cover would *make you* watch this movie?

User features *X*				Treatment	Outcome
Age	Gender	Movies viewed	…	Cover	Watched?
55	F	…	…		Y
27	M	…	…		N
33	…	…	…		…

Looks like a prediction or classification problem: $Y = f(X, T)$

The fundamental problem

User features *X*				Treatment	Outcome	Potential Outcomes		Causal effect
Age	Gender	Movies viewed	…	Cover	Watched?	 Y(T=0)	 Y(T=1)	Y(1)-Y(0)
55	F	…	…		Y	Y	?	?
27	M	…	…		N	?	N	?
33	…	…	…		…	?	Y	?
					$E[Y] =$	4.1%	5.9%	(+1.8%) *«observational»*

Looks like a prediction or classification problem: $Y = f(X, T)$

But actually we have two populations — are they «equivalent» («exchangeable»)?

Are the underlying populations similar across *X*? Are there confounding features?

When can we mix data?

- **Intuition:** check if the two populations (*control* and *treatment*) are **exchangeable (~ i.i.d.)**

- **Formally:**
  Conditional Independence Assumption (CIA):
  «Assignment to *Treatment* or *Control* group has been at random [w.r.t. observed features]»

User features *X*				Treatment
Age	Gender	Movies viewed	…	Cover
55	F	…	…	
…	…	…	…	
…	…	…	…	
27	M	…	…	
33	…	…	…	

When can we mix data?

- **Intuition:** check if the two populations (*control* and *treatment*) are **exchangeable (~ i.i.d.)**

- **Formally:**
  Conditional Independence Assumption (CIA):
  «Assignment to *Treatment* or *Control* group has been at random [w.r.t. observed features]»

- In Randomized Controlled Trials (RCTs), validity of the CIA is assessed by checking e.g. averages of relevant features (age, sex…) → «Table 1»

- Unlikely to be satisfied in observational data.

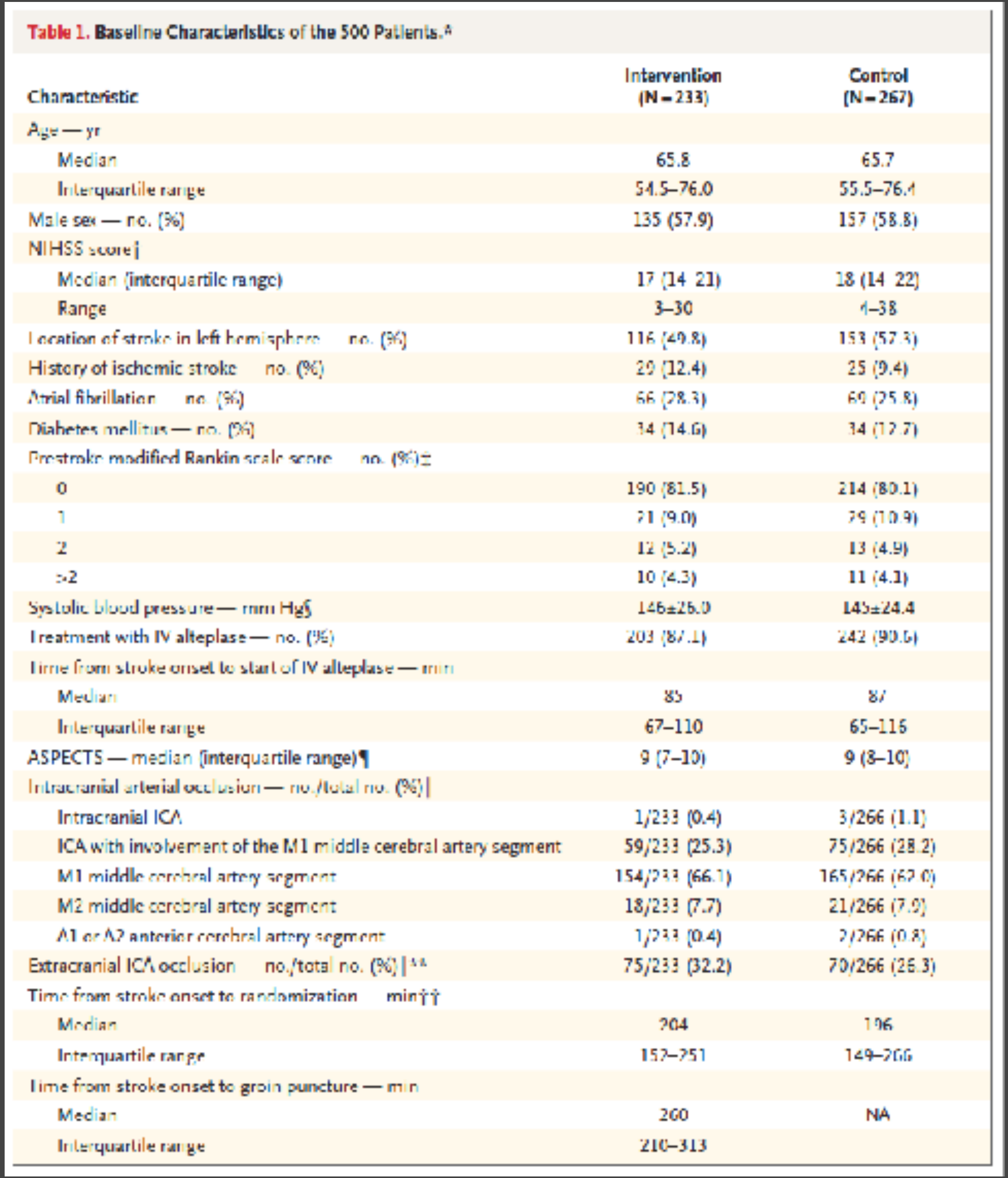


Table 1. Baseline Characteristics of the 500 Patients.[*]

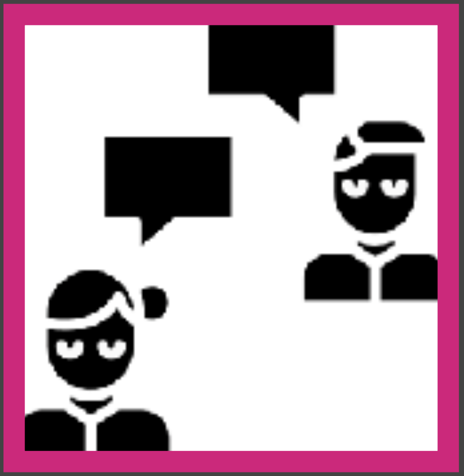
Characteristic	Intervention (N=233)	Control (N=267)
Age — yr		
Median	65.8	65.7
Interquartile range	54.5–76.0	55.5–76.4
Male sex — no. (%)	135 (57.9)	157 (58.8)
NIHSS score[†]		
Median (interquartile range)	17 (14–21)	18 (14–22)
Range	3–30	4–38
Location of stroke in left hemisphere — no. (%)	116 (49.8)	153 (57.3)
History of ischemic stroke — no. (%)	29 (12.4)	25 (9.4)
Atrial fibrillation — no. (%)	66 (28.3)	69 (25.8)
Diabetes mellitus — no. (%)	34 (14.6)	34 (12.7)
Prestroke modified Rankin scale score — no. (%)[‡]		
0	190 (81.5)	214 (80.1)
1	21 (9.0)	29 (10.9)
2	12 (5.2)	13 (4.9)
>2	10 (4.3)	11 (4.1)
Systolic blood pressure — mm Hg[§]	146±26.0	145±24.4
Treatment with IV alteplase — no. (%)	203 (87.1)	242 (90.6)
Time from stroke onset to start of IV alteplase — min		
Median	85	87
Interquartile range	67–110	65–116
ASPECTS — median (interquartile range)[¶]	9 (7–10)	9 (8–10)
Intracranial arterial occlusion — no./total no. (%)[‖]		
Intracranial ICA	1/233 (0.4)	3/266 (1.1)
ICA with involvement of the M1 middle cerebral artery segment	59/233 (25.3)	75/266 (28.2)
M1 middle cerebral artery segment	154/233 (66.1)	165/266 (62.0)
M2 middle cerebral artery segment	18/233 (7.7)	21/266 (7.9)
A1 or A2 anterior cerebral artery segment	1/233 (0.4)	2/266 (0.8)
Extracranial ICA occlusion — no./total no. (%)[‖**]	75/233 (32.2)	70/266 (26.3)
Time from stroke onset to randomization — min[††]		
Median	204	196
Interquartile range	152–251	149–266
Time from stroke onset to groin puncture — min		
Median	260	NA
Interquartile range	210–313	

Berkhemer et al., NEJM (2015)

The recommender’s view

A new user logs in…
What cover do we show them?

User features *X*				Treatment	Outcome
Age	Gender	Movies viewed	…	Cover	Watched?
55	F	…	…		Y
27	M	…	…		N
33	…	…	…		…
44	M	…	…	?	…



Matching

User features *X*				Treatment	Outcome
Age	Gender	Movies viewed	…	Cover	Watched?
55	F	…	…		Y
27	M	…	…		N
33	…	…	…		…
44	M	…	…		N
44	M	…	…	?	?

Intuition:

- See if you already met a similar case and apply what you learned

Matching

User features *X*				Treatment	Outcome
Age	Gender	Movies viewed	…	Cover	Watched?
55	F	…	…		Y
27	M	…	…		N
33	…	…	…		…
44	M	…	…		N
44	M	…	…	?	?

Intuition:

- See if you already met a similar case and apply what you learned

Drawbacks:

- Need to search whole dataset — potentially slow
- No guarantee to find a match!
- Curse of dimensionality — things get worse the more you know about your users [larger dim(X)]

Propensity Scores

User features *X*				Treatment	Outcome
Age	Gender	Movies viewed	…	Cover	Watched?
55	F	…	…		Y
27	M	…	…		N
33	…	…	…		…
44	M	…	…		N
44	M	…	…	?	?

Goals:

- Improve robustness by relying on more than N=1 observations.
- Exploit what we know about outcomes in C and T groups.
- Avoid curse of dimensionality
- Find objective way to define «distance» between units.

Idea:

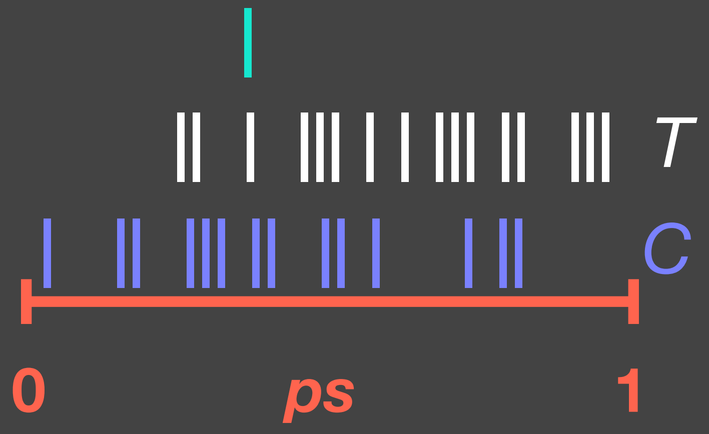
- Reduce information in X to a single number:
  What is the probability that a user with features X=x was in the treatment group?

$$ps = P(T|X)$$

- P(T|X) model tries to capture how biases cropped up in the assignment to Treatment group in the real world.

Propensity Score Stratification

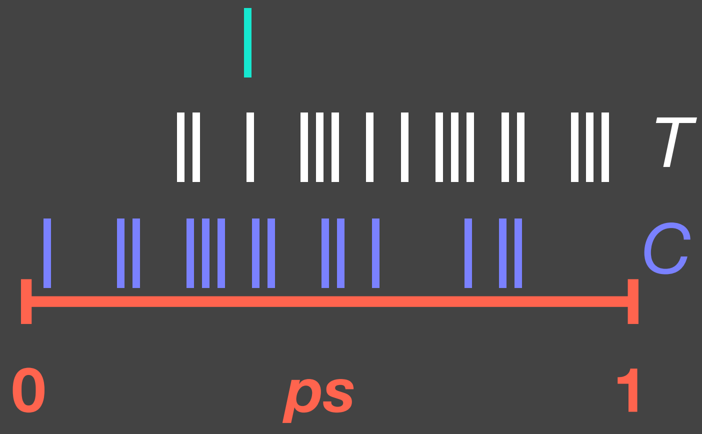
User features *X*				Treatment	PS	Outcome
Age	Gender	Movies viewed	…	Cover	P(T|X)	Watched?
55	F	…	…		0.8	Y
27	M	…	…		0.95	N
33	…	…	…		0.65	…
44	M	…	…		0.45	N
44	M	…	…	?	0.45	?



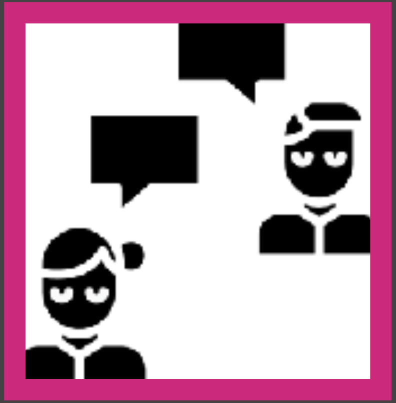
Recall: P(T|X) model tries to capture how biases cropped up in the assignment to Treatment group in the real world. Contains no info on outcomes.

Propensity Score Stratification

User features *X*				Treatment	PS	Outcome
Age	Gender	Movies viewed	…	Cover	P(T|X)	Watched?
55	F	…	…		0.8	Y
27	M	…	…		0.95	N
33	…	…	…		0.65	…
44	M	…	…		0.45	N
44	M	…	…	?	0.45	?



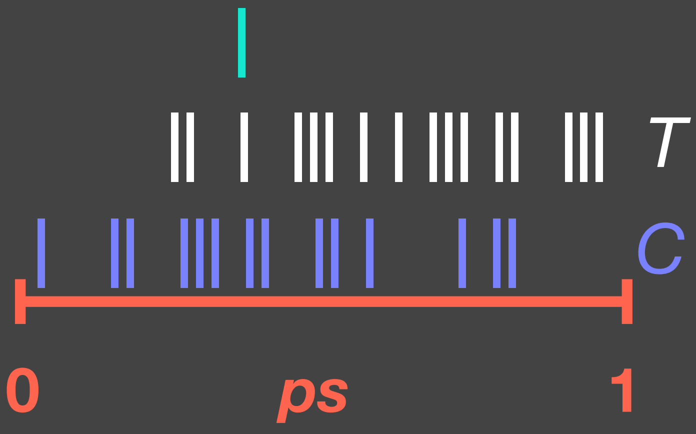
How would this look if assignment had been purely at random (CIA)?



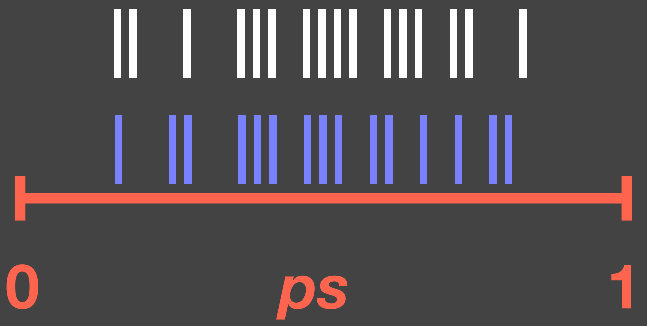
Recall: P(T|X) model tries to capture how biases cropped up in the assignment to Treatment group in the real world. Contains no info on outcomes.

Propensity Score Stratification

User features *X*				Treatment	PS	Outcome
Age	Gender	Movies viewed	…	Cover	P(T|X)	Watched?
55	F	…	…		0.8	Y
27	M	…	…		0.95	N
33	…	…	…		0.65	…
44	M	…	…		0.45	N
44	M	…	…	?	0.45	?

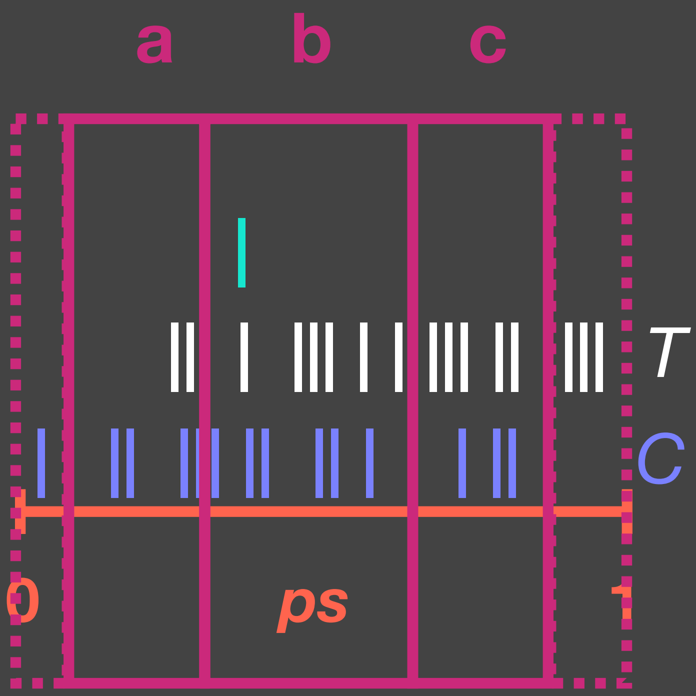


How would this look if assignment had been purely at random (CIA)?



Propensity Score Stratification

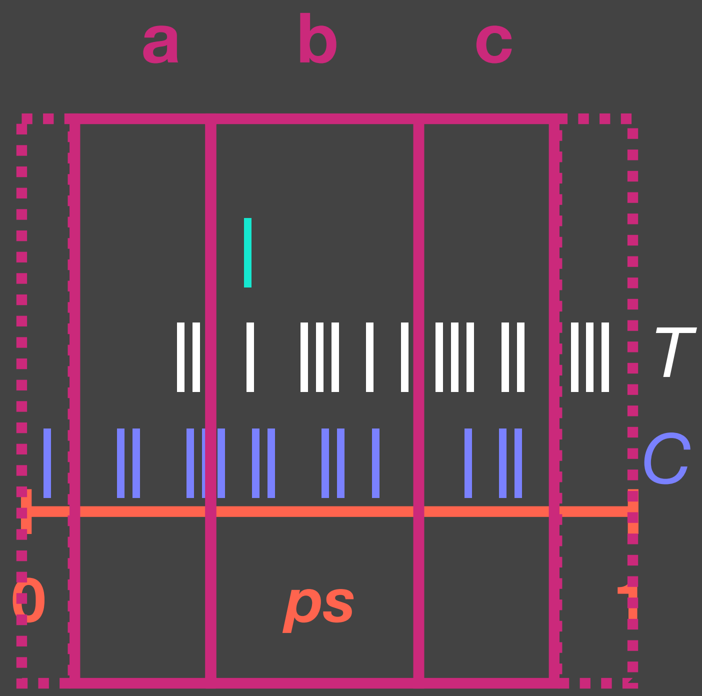
User features *X*				Treatment	PS	Outcome
Age	Gender	Movies viewed	…	Cover	P(T|X)	Watched?
55	F	…	…		0.8	Y
~~27~~	~~M~~	~~…~~	~~…~~		~~0.95~~	~~N~~
33	…	…	…		0.65	…
44	M	…	…		0.45	N
44	M	…	…	?	0.45	?



- Drop outliers in ps-space, not in X-space.
- Adjust bin width to satisfy «exchangeability» within the bin.

Propensity Score Stratification

	a	b	c
ps	0.1-0.3	0.3-0.6	0.6-0.85
# obs.	100	200	150
Age	36(4)	48(3)	55(5)
Gender	M(52%)	F(50.1%)	F(55%)
ATE	-2%	-0.5%	5.4%



- Adjust bin width to satisfy «exchangeability» within the bin.
- Then extract causal effect by bin.
- Allows us to design group-targeted actions → customer segmentation.

Propensity Score Stratification

	a	b	c
ps	0.1-0.3	0.3-0.6	0.6-0.85
# obs.	100	200	150
Age	36(4)	48(3)	55(5)
Gender	M(52%)	F(50.1%)	F(55%)
ATE	-2%	-0.5%	5.4%

Overall causal ATE:

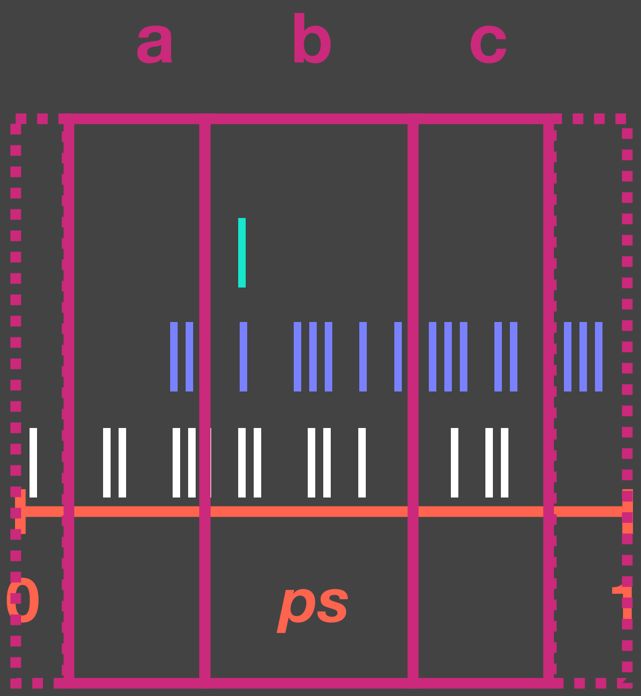
$$\frac{100}{450}(-0.02) + \frac{200}{450}(-0.005) + \frac{150}{450}(0.054) = +1.1\,\%$$

Recall «Observational» ATE: +1.8%

- Adjust bin width to satisfy «exchangeability» within the bin.
- Then extract causal effect by bin.
- Allows us to design group-targeted actions → customer segmentation.
- Causal estimate of population-wide ATE.

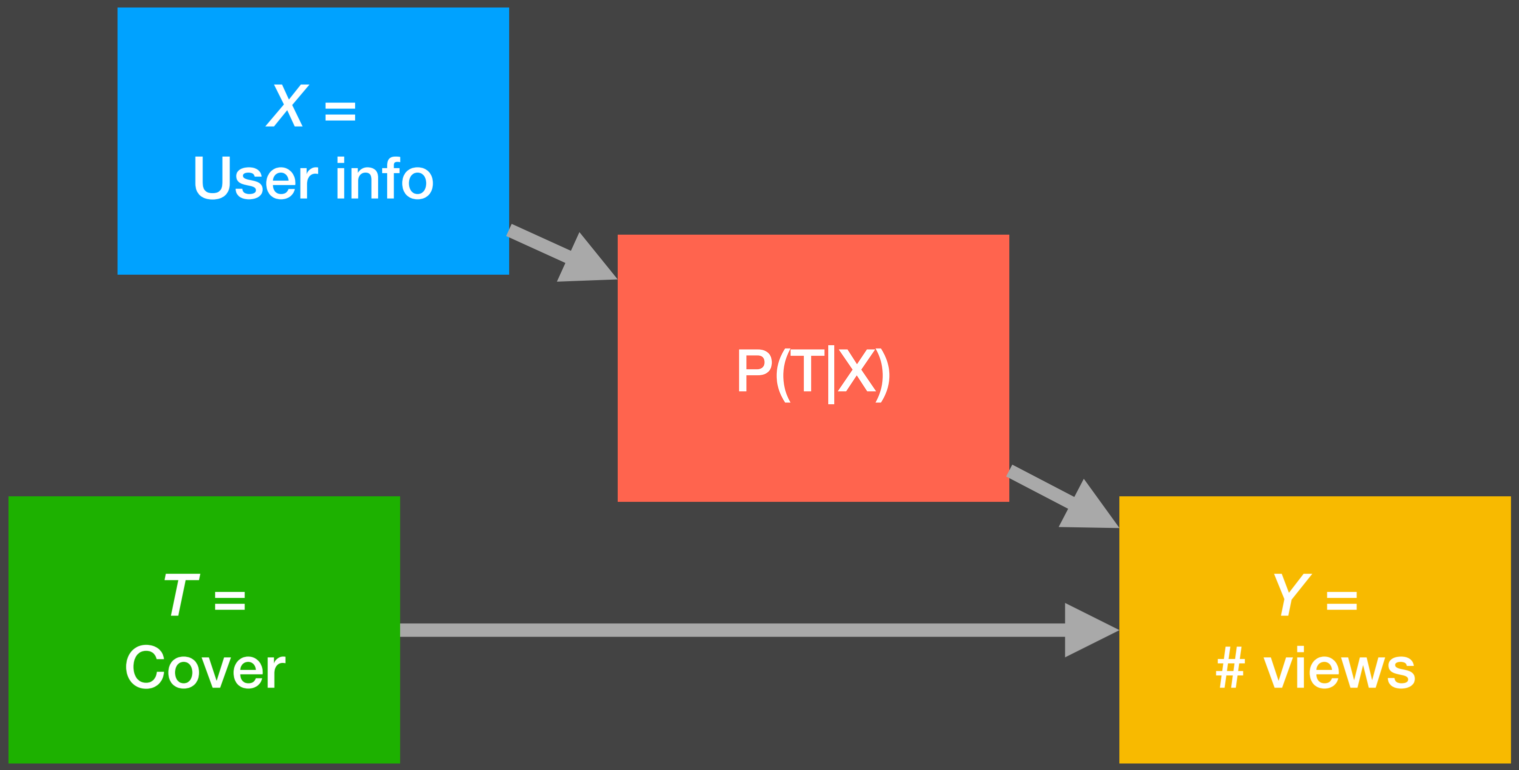
Connection with DAGs?

User features *X*				Treatment	PS	Outcome
Age	Gender	Movies viewed	…	Cover	P(T|X)	Watched?
55	F	…	…		0.8	Y
27	M	…	…		0.2	N



A good DAG underlying your model to estimate P(T|X) will allow you to:

- Not use irrelevant features
- Avoid biases
- Reduce uncertainty



X = User info

P(T|X)

T = Cover

Y = # views

Causal inference: Best practices

- **Always follow the four steps: Model, Identify, Estimate, Refute.**

  Refute is the most important step.

> «Try to prove yourself wrong»
> —W.D. Phillips, Nobel prize in Physics 1997

- **Aim for simplicity.**

  If your analysis is too complicated, it is most likely wrong.

- **Try at least two methods with different assumptions.**

  Higher confidence in estimate if both methods agree.

- **Remember the order for validity of estimates obtained: Randomization, Natural experiments, Conditioning.**

  Consider observational methods as strong hints (but they can be misleading)

Adapted from Amit Sharma
(Microsoft Research, *DoWhy*'s lead developer)

Now turn to the Notebook pscore_oil_wells_analyse.ipynb

Charles Addams, «Skier», The New Yorker, 13 Jan 1940