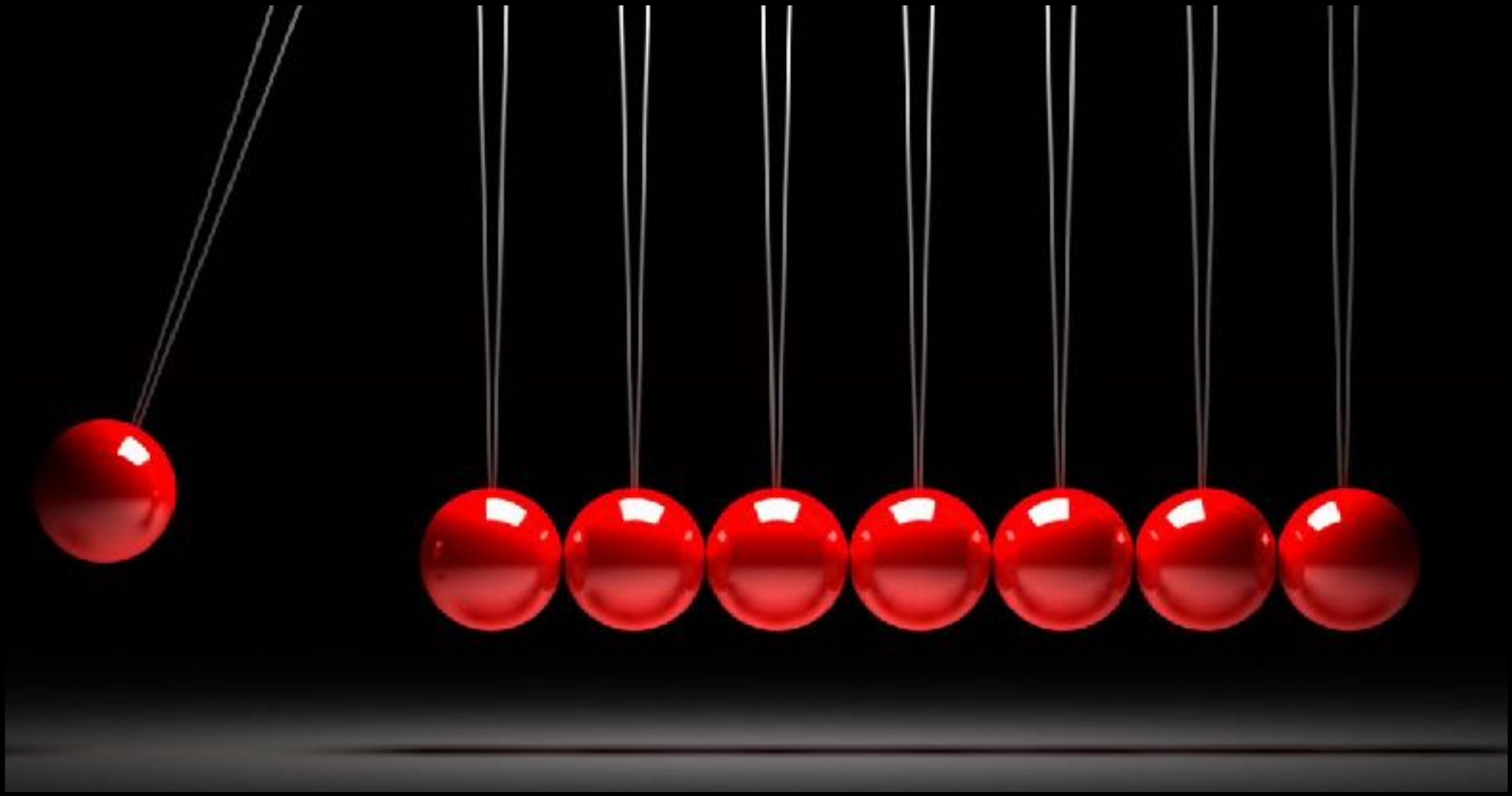
Causal Data Science workshop



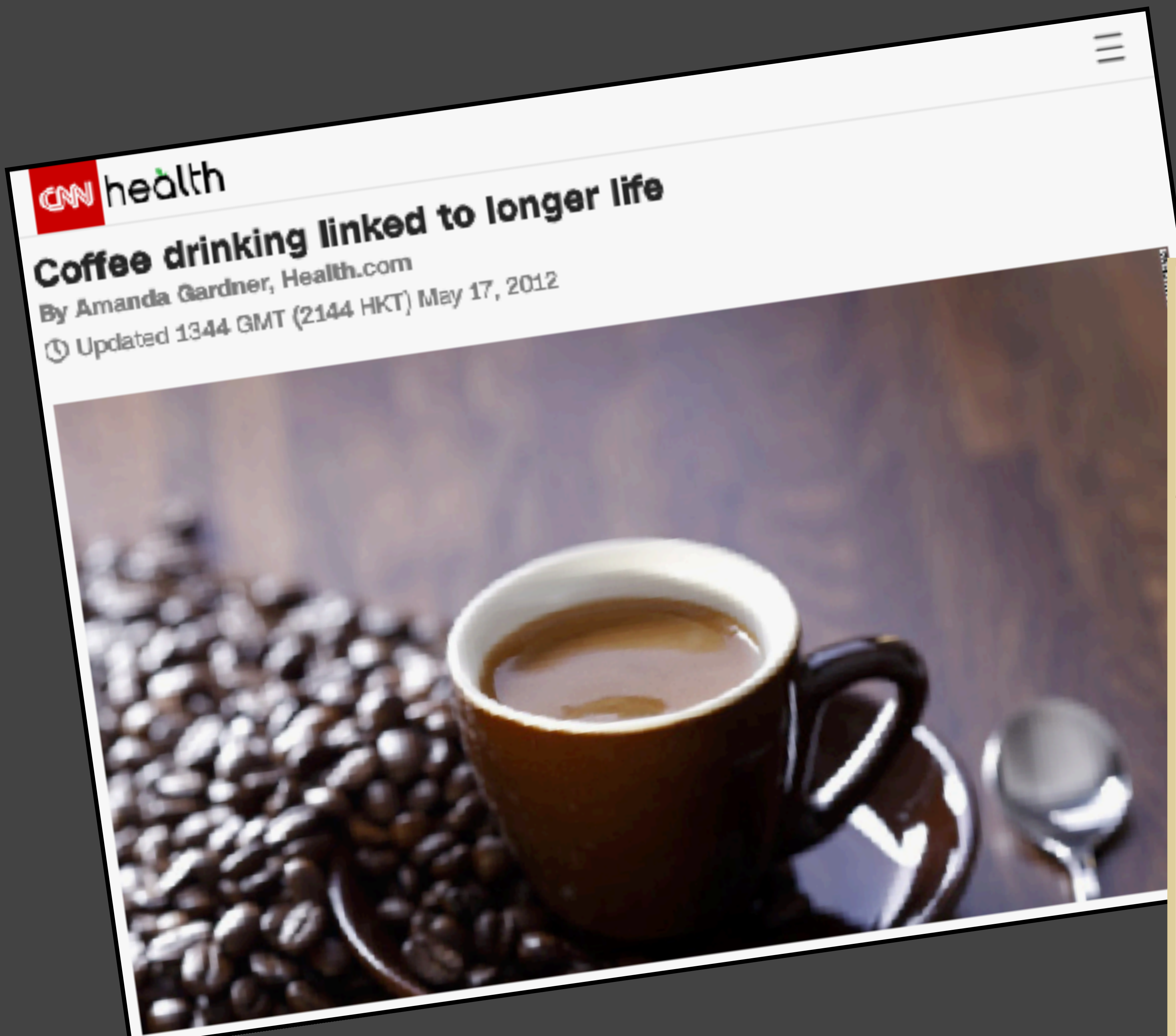
Jordi Mur-Petit, PhD
jordi.mur.work@gmail.com

UNIVERSITAT DE BARCELONA
Tecniospring INDUSTRY
ACCIÓ
Generalitat de Catalunya
Government of Catalonia

Schedule

- **14:00 - 14:45** Crash course on Causal Inference
    - Causality - what for?
    - The fundamental problem of causality
    - Potential Outcomes formalism: some practical methods (running example)
- **14:45 - 15:30** Hands-on work on Notebook

Causality: what for?



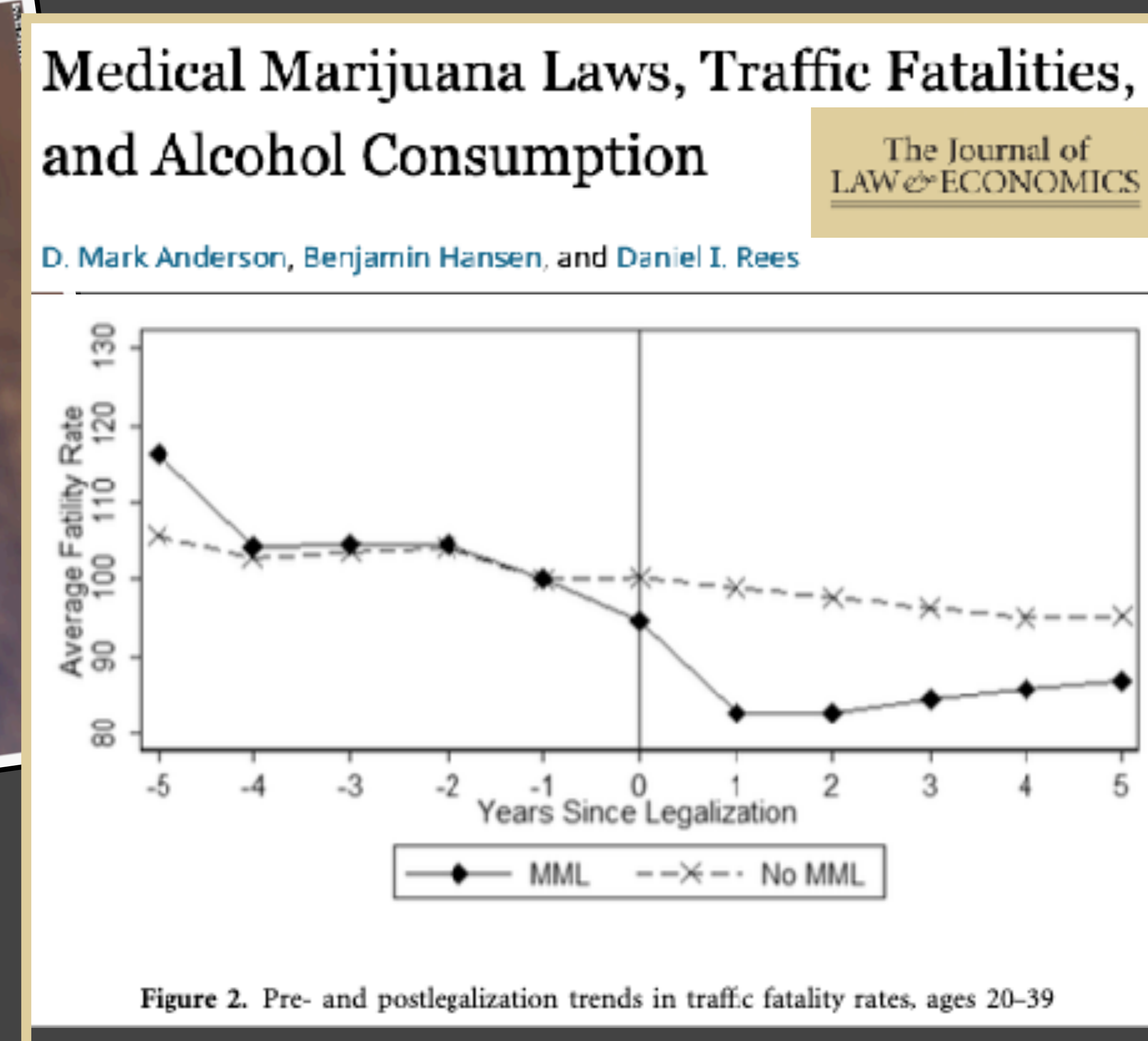
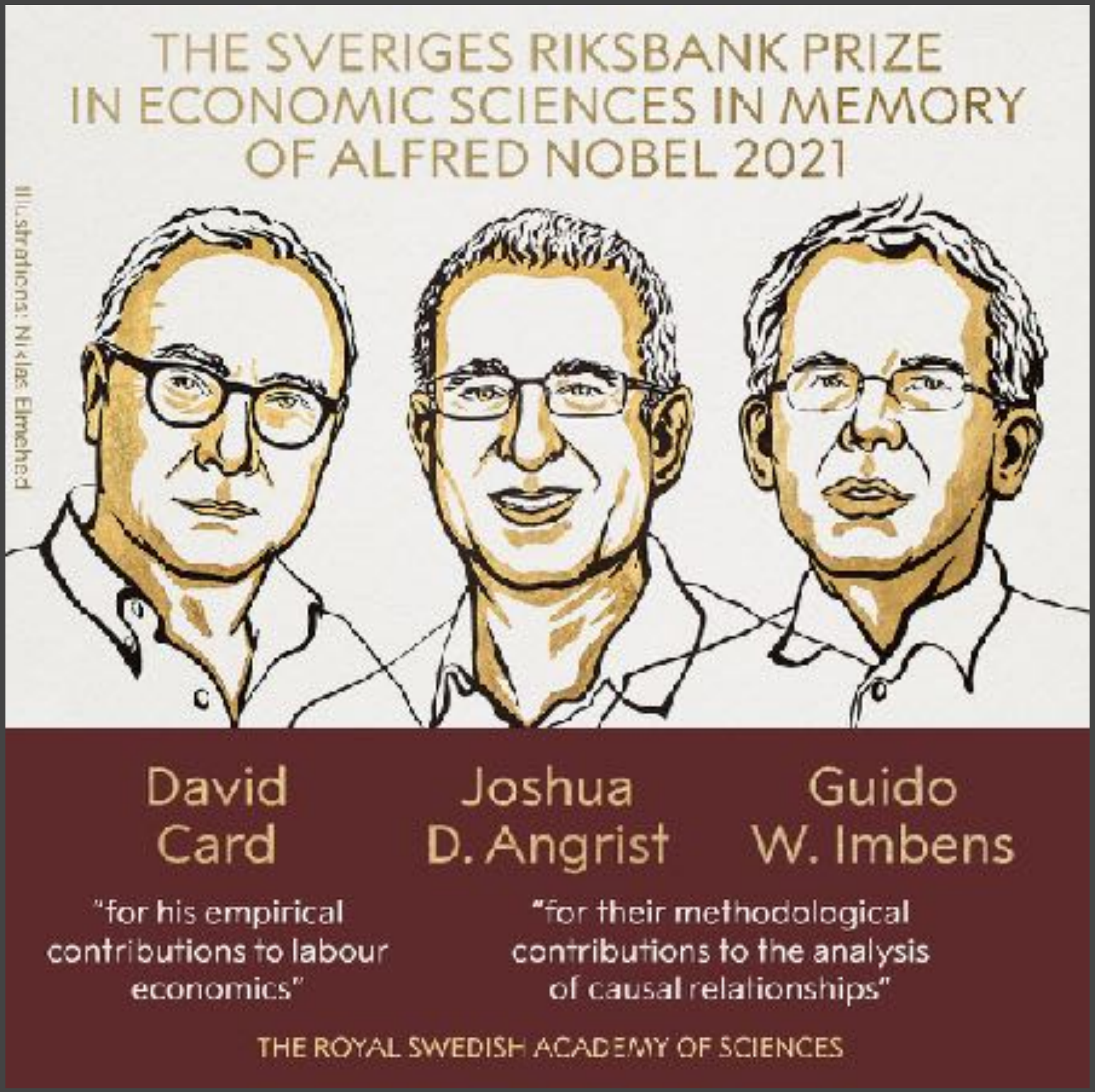


Figure 2. Pre- and postlegalization trends in traffic fatality rates, ages 20–39

Causality: what for?



Causality: what for?

Health:
- Dietary guidelines: calories, alcohol, red meat…
- Exercise: 10,000 steps…

Econ / politics:
- Legalize marijuana: medical, recreational uses?
- Best interventions against unemployment?

Business:
- Uplift — Best promo to get increase revenue?
- Recommenders — Item most likely to make customer come back?
- Customer retention — Best action to retain customers?
- …

3 kinds of queries

Pearl's «Ladder of Causation»

Causality

«Imagining»
Counterfactuals
What if...?

«Doing»
Interventions

Traditional Statistics/ML

«Seeing»
Associations, predictions

Beyond textbook ML

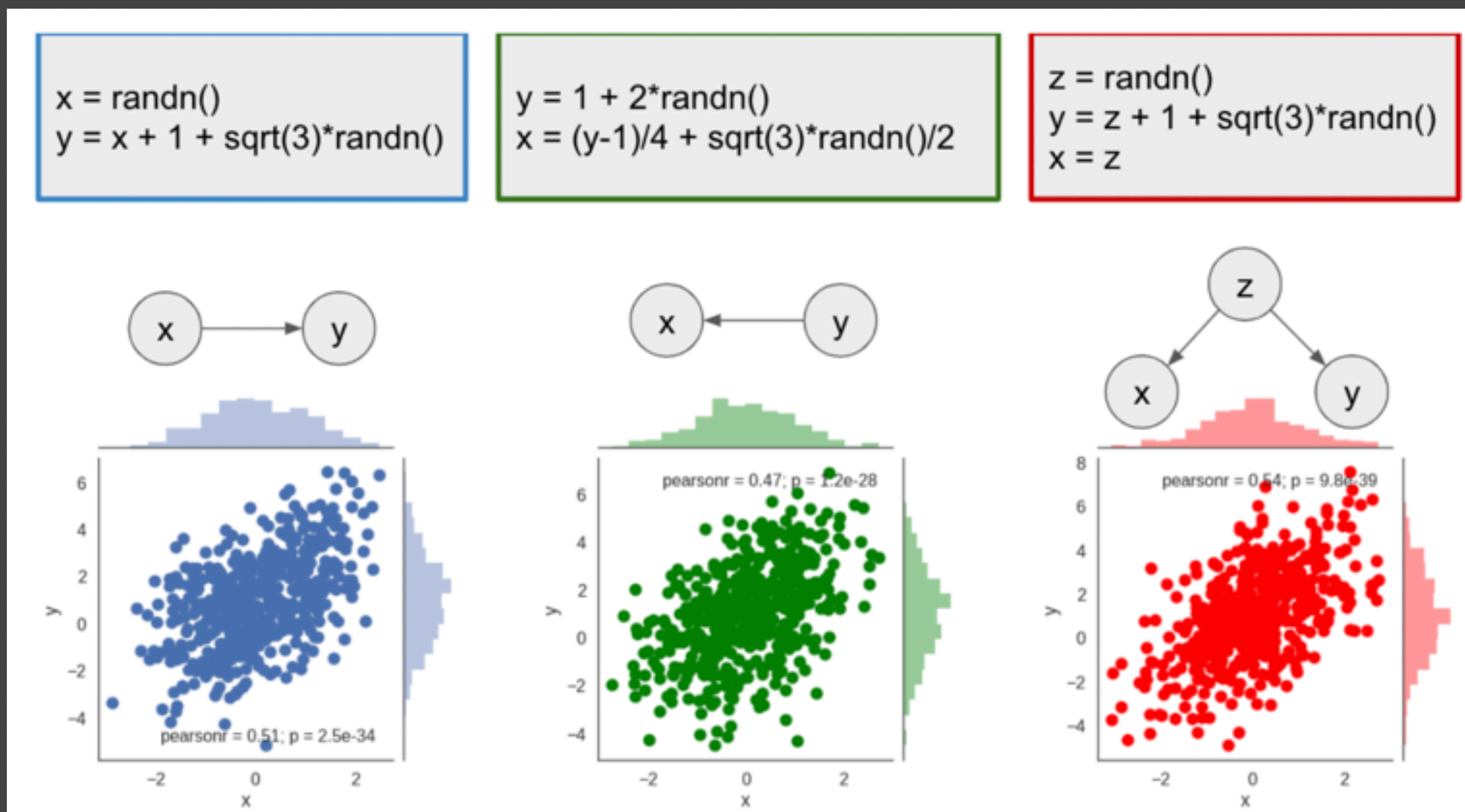
It depends on the needs:

- **I want to minimize the Empirical Risk** $R(f) = \frac{1}{n}\sum_{i=1}^{n} L[f_\theta(\mathbf{x}_i), y_i]$
- **I want to maximize robustness against changes in *p*(X, Y).**
- **I want to maximize robustness against adversarial attacks.**
- **I want to be able to explain my predictions.**
- **I want to measure and mitigate unwanted biases (discrimination).**
- **I want to use the prediction to inform a decision (intervention/treatment) that can change *p*(X).**
- **Etc.**

All these considerations involve causal thinking.

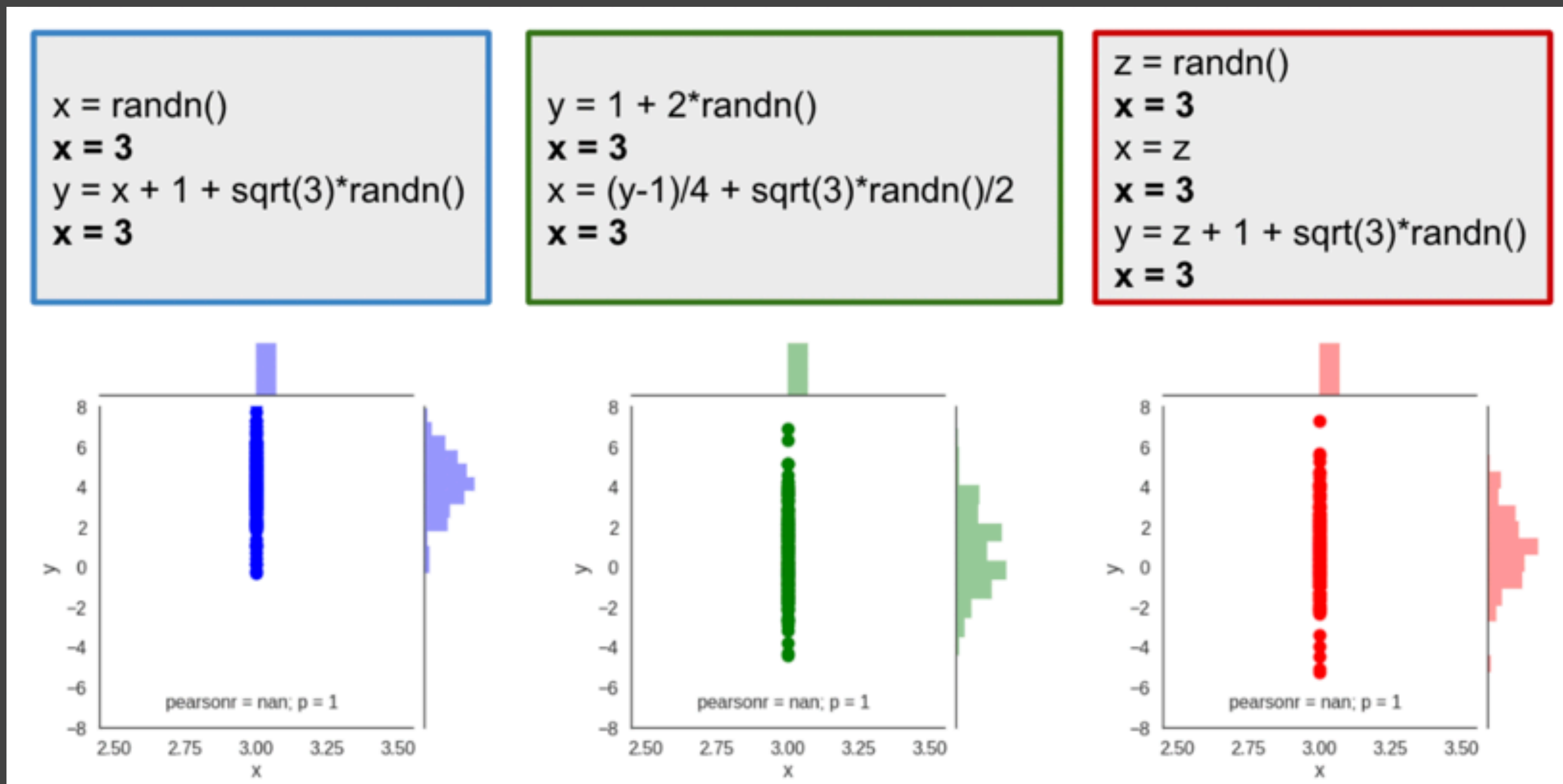
What’s an intervention?

Plots represent some observed PDF *P*(y,x)



What’s an intervention?

Plots represent the interventional PDF *P*(y, *do*(x=3)).



We will want to find ways to estimate PDFs like *P*(y | *do*(x), z) from known PDFs like *P*(x,y,z), *P*(y|x,z), etc.

This is called **identification** and there’s a set of graphical rules to do it.

Defining «Causality»

Causality («counterfactual» definition):

Imagine two worlds, identical in every way up until the point where a “treatment” occurs in one (*factual*) world but not the (*counterfactual*) other.

Any subsequent difference in a property *Y* between the two worlds is then, logically, a *consequence* of this treatment.

Counterfactual value/Potential outcome: hypothetical value of a variable under a treatment that did *not* occur.

$$T = \{0, 1\} \rightarrow Y_0 \equiv Y(0), \quad Y_1 \equiv Y(1)$$

Causal effect of a treatment: difference between the observed outcome after the intervention and its counterfactual value:

«*average treatment effect*»: $ATE = Y(1) - Y(0)$

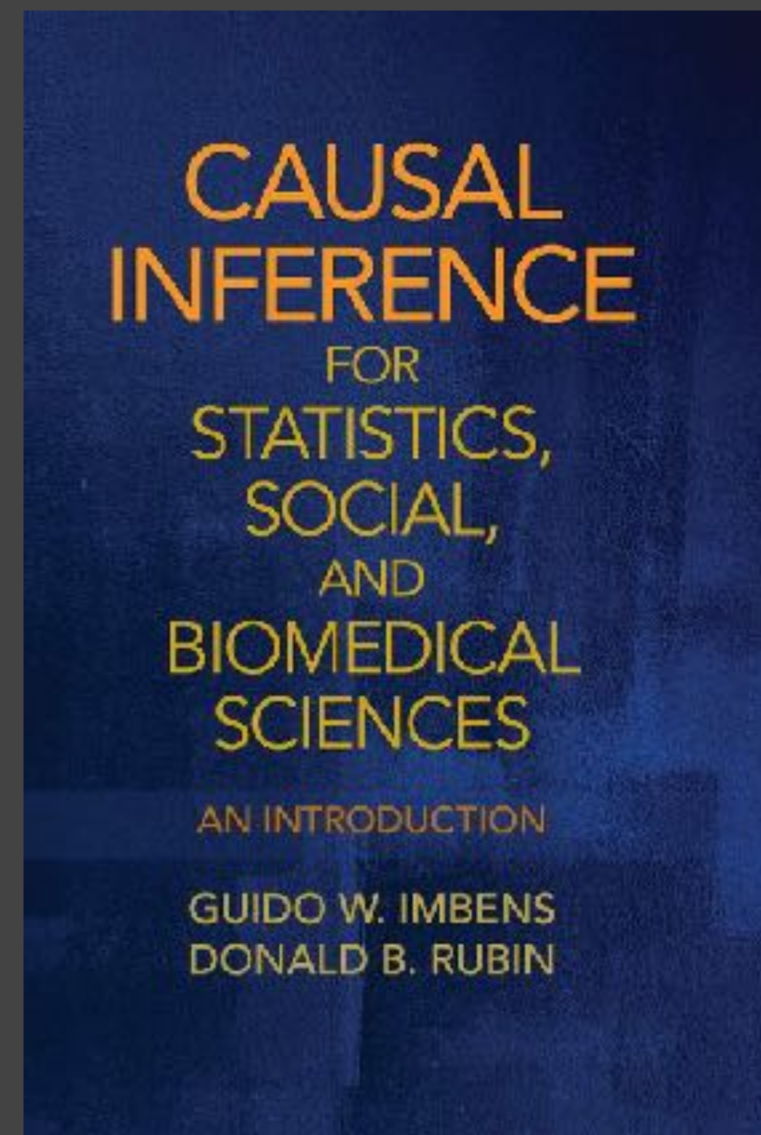
Potential Outcomes framework
a.k.a. (Neyman-)Rubin causal model

- 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



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).

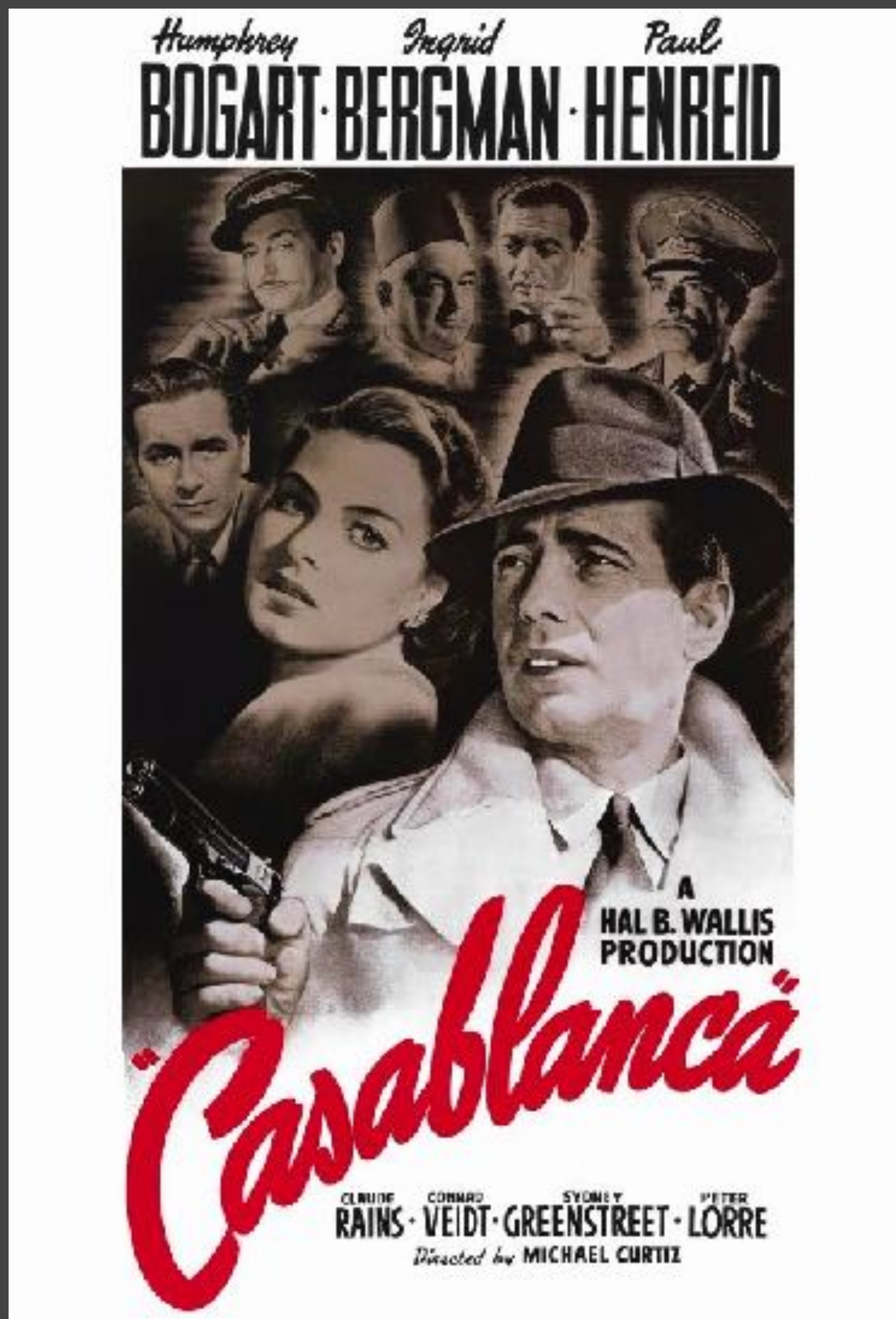
Potential Outcomes framework
a.k.a. (Neyman-)Rubin causal model

Key take-away message:

You can extract cause-effect information not only from experimental data (RCT, A/B test), but also from observational (real world) data.

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?

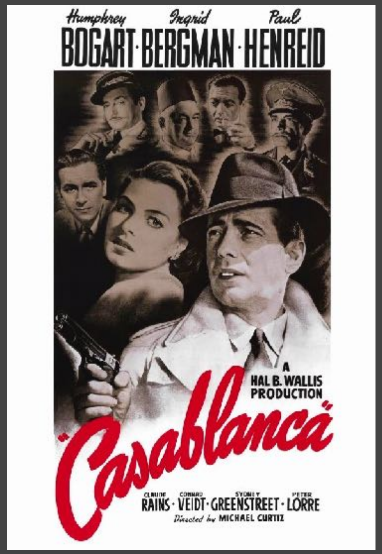


Which cover would *make you* watch this movie?

X =
User info

T =
Cover

Y =
views



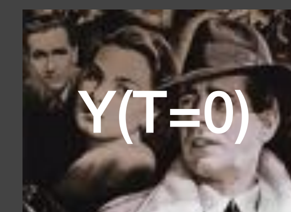
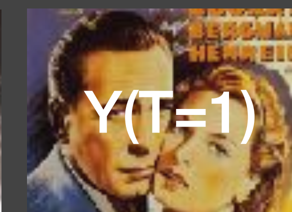
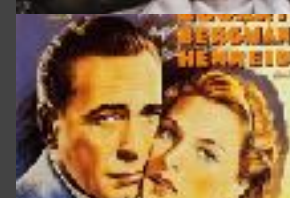
This drawing is called a DAG — directed acyclic graph

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 «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?

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

Berkhemer et al., NEJM (2015)

Table 1. Baseline Characteristics of the 500 Patients.*

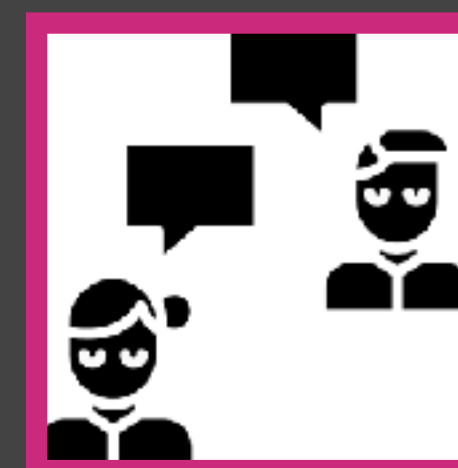
Characteristic	Intervention (N=233)	Control (N=267)
Age — yr		
Median	65.8	65.7
Interquartile range	54.5–76.0	55.5–76.4
Male sex — no. (%)	135 (57.9)	157 (58.8)
NIHSS score†		
Median (interquartile range)	17 (14–21)	18 (14–22)
Range	3–30	4–38
Location of stroke in left hemisphere — no. (%)	116 (49.8)	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)

- Unlikely to be satisfied in observational data — **But there's a way out!**

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	…	…	?	…



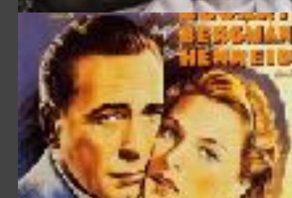
Matching

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	…	…	?	?

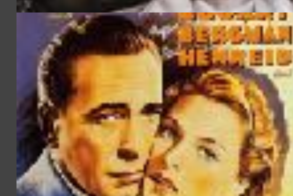
Intuition:

- 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:

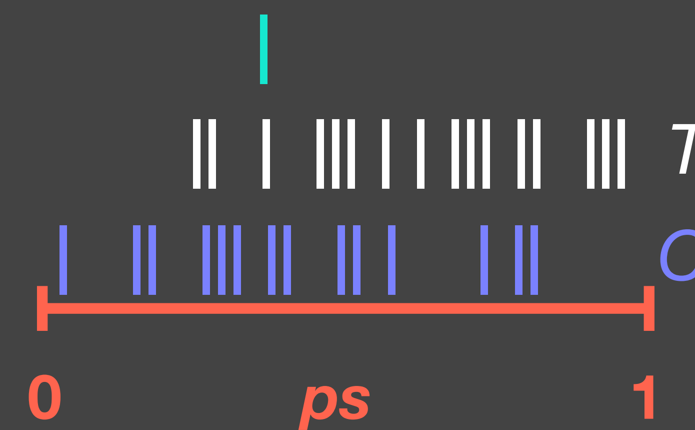
- Reduce information in X to a single number:
  What is the probability that a user with features X=x was in the treatment group?

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

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

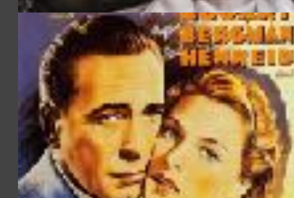
Propensity Score Stratification

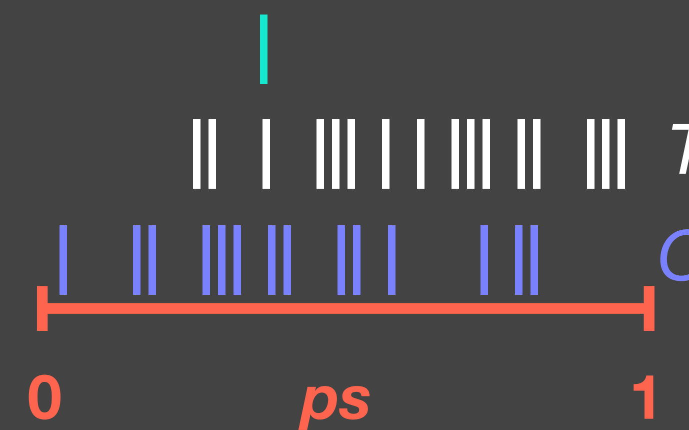
User features *X*				Treatment	PS	Outcome
Age	Gender	Movies viewed	…	Cover	P(T|X)	Watched?
55	F	…	…		0.8	Y
27	M	…	…		0.95	N
33	…	…	…		0.65	…
44	M	…	…		0.45	N
44	M	…	…	?	0.45	?



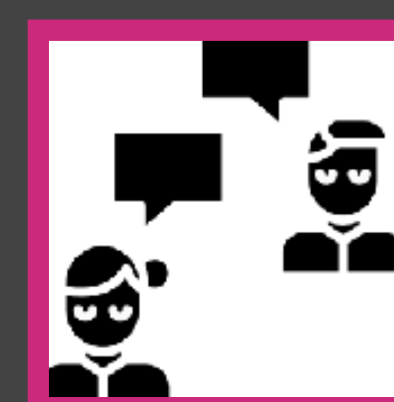
Remember: P(T|X) model tries to capture how biases cropped up in the assignment to Treatment group in the real world. Contains no info on outcomes.

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	?

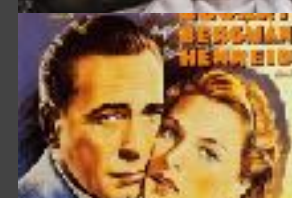


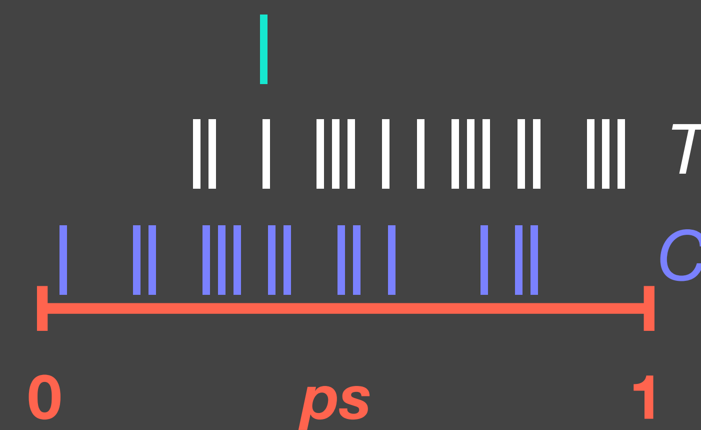
How would this look if assignment had been purely at random?



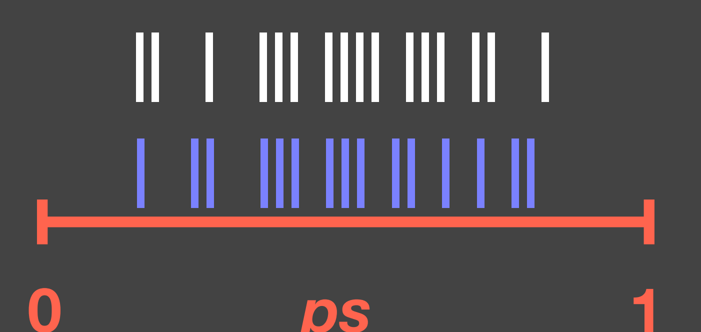
Remember: P(T|X) model tries to capture how biases cropped up in the assignment to Treatment group in the real world. Contains no info on outcomes.

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	?

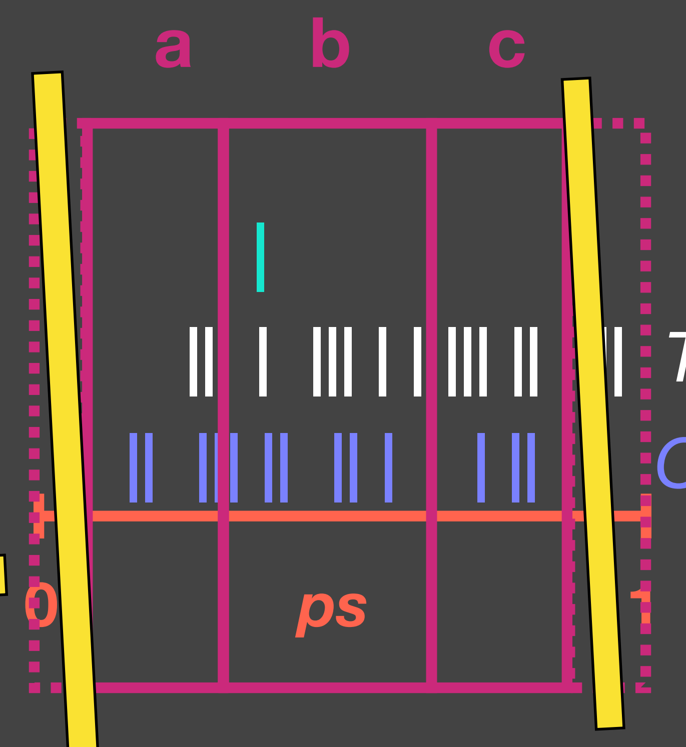


How would this look if assignment had been purely at random (CIA)?



Propensity Score Stratification

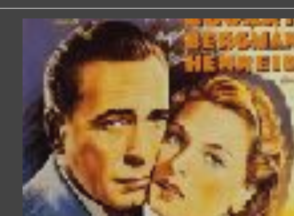
User features X				Treatment	PS	Outcome
Age	Gender	Movies viewed	…	Cover	P(T|X)	Watched?
55	F	…	…		0.8	Y
27	M				0.95	N
33	…	…	…		0.65	…
44	M	…	…		0.45	N
44	M	…	…	?	0.45	?

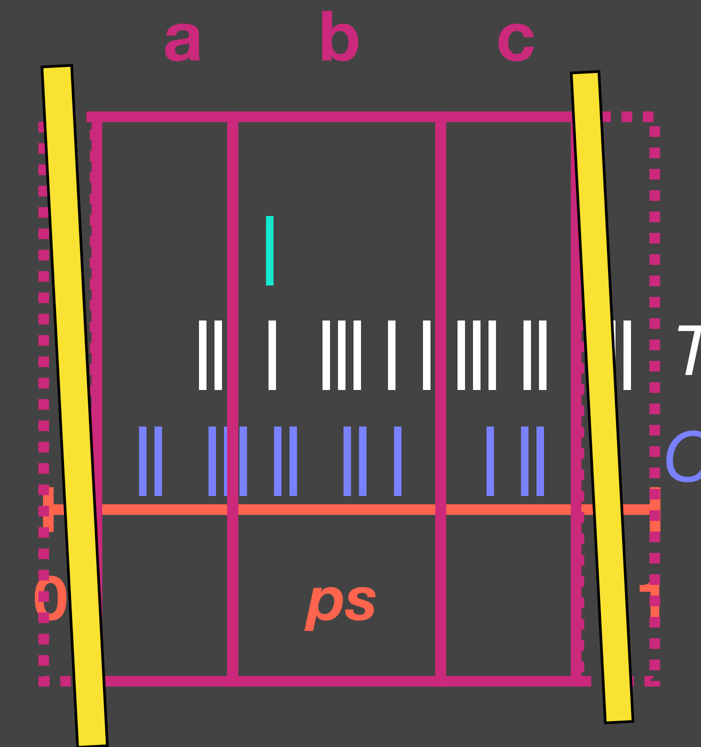


- Drop outliers in ps-space, not in X-space.
- Adjust bin width to satisfy «exchangeability» within the bin.

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%

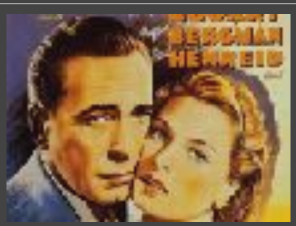




- Adjust bin width to satisfy «exchangeability» within the bin.
- Then extract causal effect by bin.
- Allows us to design group-targeted actions → customer segmentation.

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.

Causal inference: Best practices

- **Always follow the four steps:**
  1. **Model:** draw a DAG relating the flow of information through variables; *domain knowledge* enters here.
  2. **Identify:** relate desired interventional PDF to observable ones, e.g., $P[y \mid do(x)] = \sum_z P(y, x, z)P(z)$.
  3. **Estimate:** *ML* enters here: get best numerical estimates based on (2), e.g., fitting non-parametric models to $P(y,x,z)$.
  4. **Refute:** «try to prove yourself wrong».

- **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 estimates if different methods agree.

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

Good news! There are several Python libraries for Causal Inference!

- DoWhy, by Microsoft
- CausalML, by Uber

See a partial listing with many more libraries at
https://github.com/rguo12/awesome-causality-algorithms

**Now turn to Jupyter notebook
and we'll put all this into practice with DoWhy!**