Machine learning and causal inference:

Toward new clinical evidence?

Bénédicte Colnet, former Ph.D. student at Inria (Soda & PreMeDICaL teams)

Pyladies 🐍 Paris, Botify, September 27th 2023

Julie Josse
Missing values & causal inference

Gaël Varoquaux
ML & co-founder of scikit-learn

Erwan Scornet
Random forest & missing values

Inria

Inserm

I have to tell you something

I have to tell you something

I mostly used R during the past three years

I have to tell you something

I mostly used R during the past three years

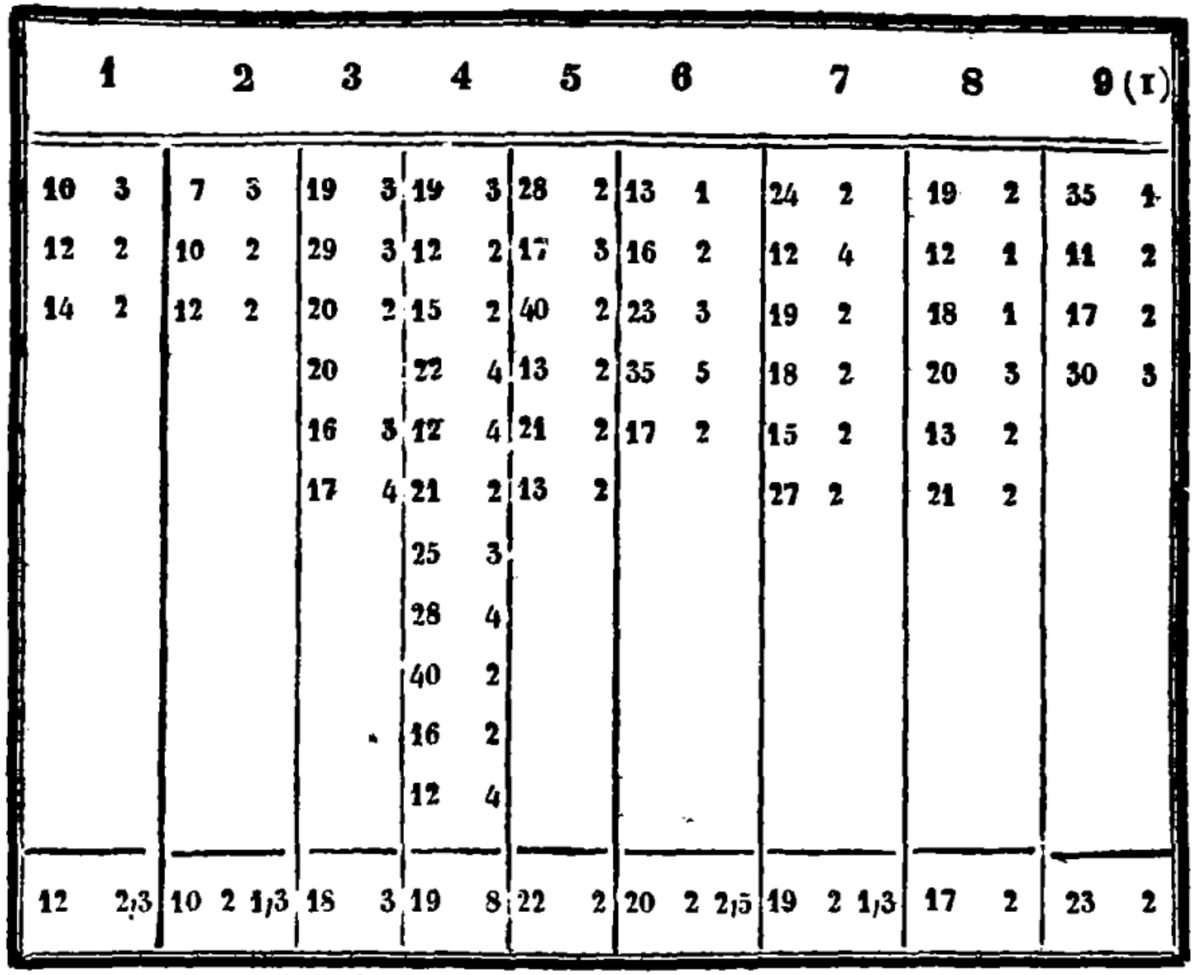
Why?

1. Collaborations with clinicians and medical doctors

2. Causal inference community mostly uses R

Evidence based medicine

The promise of big data



1		2		3		4		5		6		7		8		9 (1)	
16	3	7	3	19	3	19	3	28	2	13	1	24	2	19	2	35	1
12	2	10	2	29	3	12	2	17	3	16	2	12	4	12	1	11	2
14	2	12	2	20	2	15	2	40	2	23	3	19	2	18	1	17	2
				20		22	4	13	2	35	5	18	2	20	3	30	3
				16	3	17	4	21	2	17	2	15	2	13	2		
				17	4	21	2	13	2			27	2	21	2		
						25	3										
						28	4										
						40	2										
						16	2										
						12	4										
12	2,3	10	2 1/3	18	3	19	8	22	2	20	2 2,5	19	2 1/3	17	2	23	2

Source: Pierre Charles Alexandre Louis's experiment on bloodletting (1835)
— Original research work is made available by the French National Library (BnF)

A brief history of modern medical evidence: the ever increasing role of data and statistics

James Lind's scorbut experiment

1747

A brief history of modern medical evidence: the ever increasing role of data and statistics

James Lind's scorbut experiment

William Farr — *General Register Office*

1747 — 1828 — 1837 — 1854 — 1912

P.C.A. Louis's experiments on bloodletting

John Snow's discovery on cholera

Janet Lane-Clayton pioneered the use of cohort studies and case control studies (benefit of breast feeding versus cow milk)

A brief history of modern medical evidence: the ever increasing role of data and statistics

James Lind's scorbut experiment

William Farr — *General Register Office*

1747 — 1828 — 1837 — 1854 — 1912 — 1948

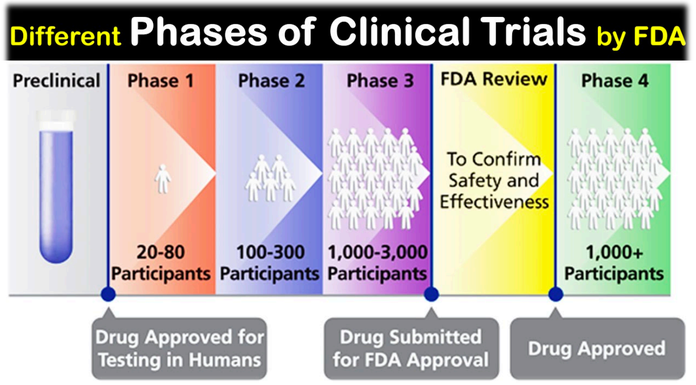
Streptomycin trial for pulmonary tuberculosis

So-called evidence based medicine's era

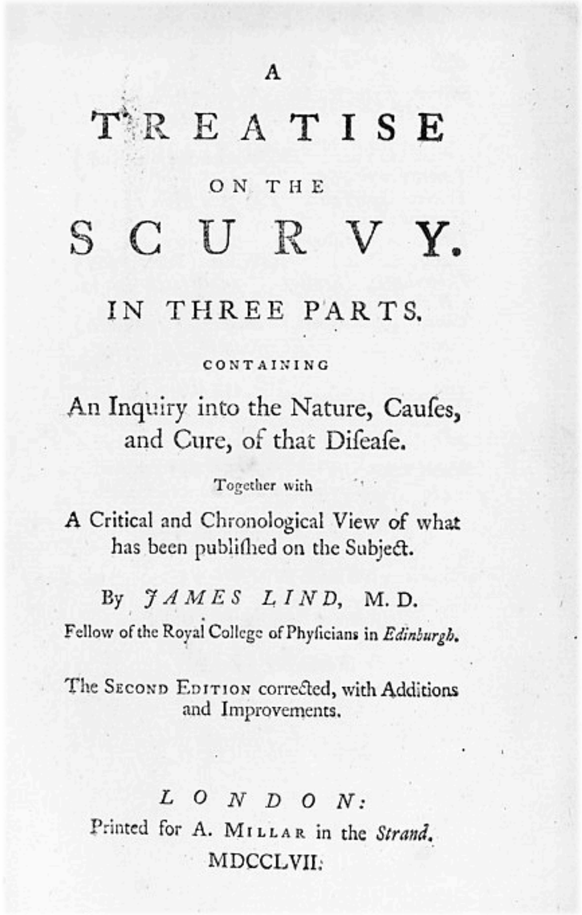
P.C.A. Louis's experiments on bloodletting

John Snow's discovery on cholera

Janet Lane-Clayton pioneered the use of cohort studies and case control studies (benefit of breast feeding versus cow milk)



A longstanding presence of Randomized Controlled Trials (RCTs) … now being the gold-standard



A

TREATISE

ON THE

SCURVY.

IN THREE PARTS.

CONTAINING

An Inquiry into the Nature, Cauſes, and Cure, of that Diſeaſe.

Together with

A Critical and Chronological View of what has been publiſhed on the Subject.

By *JAMES LIND*, M. D.

Fellow of the Royal College of Phyſicians in *Edinburgh*.

The SECOND EDITION corrected, with Additions and Improvements.

LONDON:

Printed for A. MILLAR in the *Strand*.

MDCCLVII.

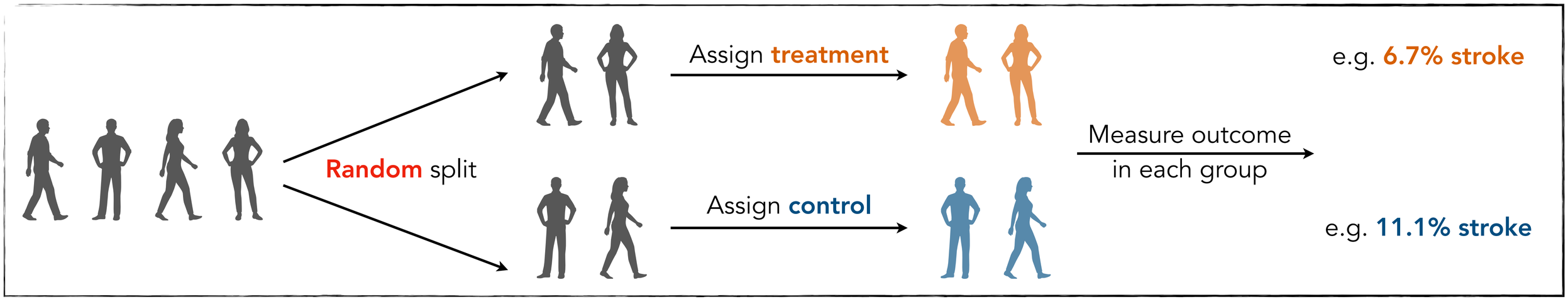
James Lind experiment on scorbut in **1757**
Source: Wikipedia

Drug Trials Snapshot	Active Ingredient	Date of FDA Approval	What is it Approved For
CABENUVA	cabotegravir and rilpivirine	January 20, 2021	Treatment of HIV-1 infection.
LUPKYNIS	voclosporin	January 22, 2021	Treatment of lupus nephritis
VERQUVO	vericiguat	January 19, 2021	Treatment of chronic heart failure
GEMTESA	vibegron	December 23, 2020	Treatment of symptoms of overactive bladder
EBANGA	ansuvimab-zykl	December 21, 2020	Treatment of Zaire ebolavirus infection
ORGOVYX	relugolix	December 18, 2020	Treatment of advanced prostate cancer

Recently approved drugs by the Food and Drug Administration (FDA), all with their corresponding RCT snapshot and information.
Source: www.fda.gov - **2022**

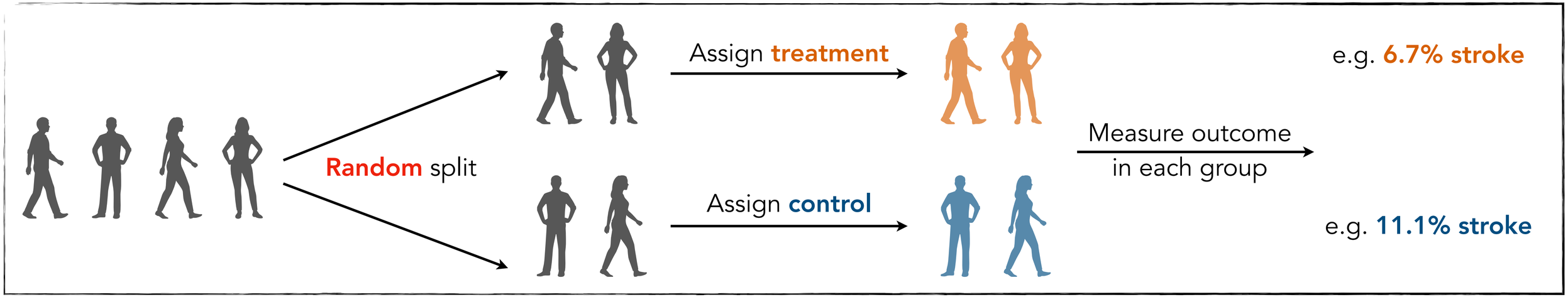
Randomized Controlled Trials (RCTs) as the current gold standard

Principle



Randomized Controlled Trials (RCTs) as the current gold standard

Principle



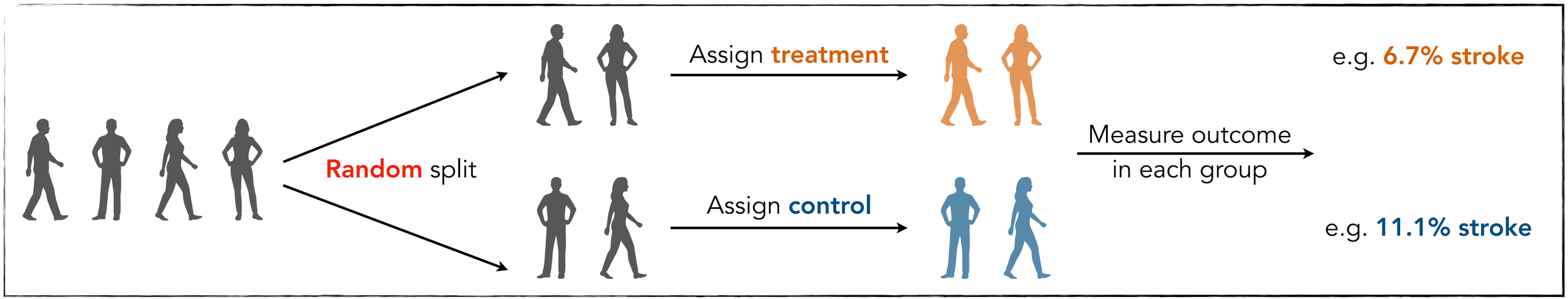
In practice : the CRASH-3 trial investigating Tranexamic Acid effect on brain injured related death

Results Between July 20, 2012, and Jan 31, 2019, we randomly allocated 12 737 patients with TBI to receive tranexamic acid (6406 [50·3%] or placebo [6331 [49·7%], of whom 9202 (72·2%) patients were treated within 3 h of injury. Among patients treated within 3 h of injury, the risk of head injury-related death was 18·5% in the tranexamic acid group versus 19·8% in the placebo group (855 *vs* 892 events; risk ratio [RR] 0·94 [95% CI 0·86–1·02]).

Source: Screenshot from the Lancet (CRASH-3 main report)

Randomized Controlled Trials (RCTs) as the current gold standard

Principle



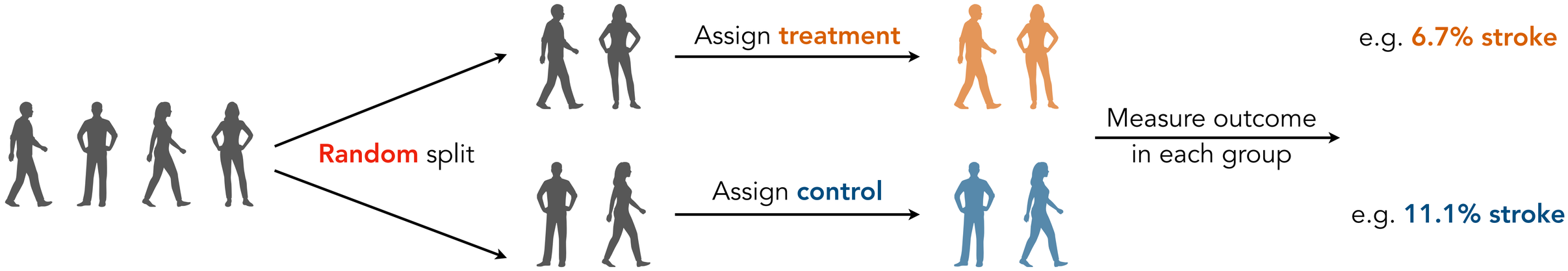
When it comes to the code

```
t.test(vector.goup.A, vector.group.B)
```

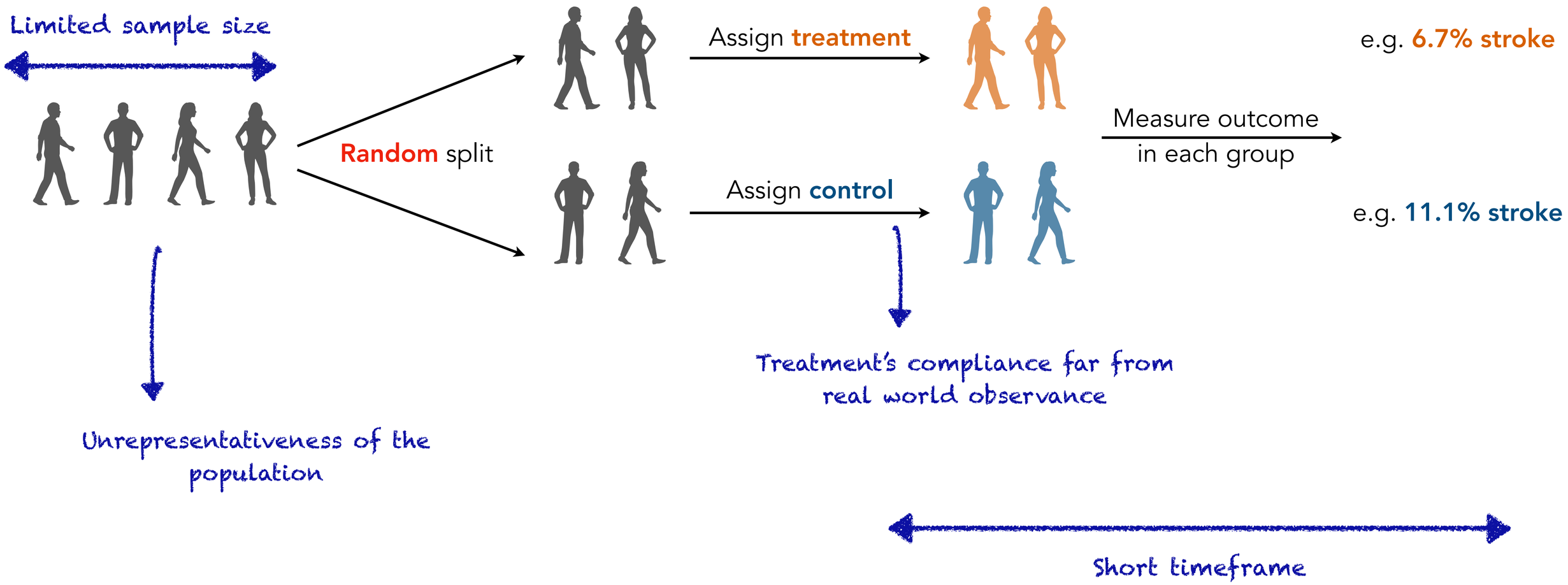
```
t.test(control, sample2)
```

```
##
##     Welch Two Sample t-test
##
## data:  control and sample2
## t = -4.6694, df = 140.62, p-value = 0.00000698
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.400420 -0.567307
## sample estimates:
## mean of x mean of y
##  7.111882  8.095745
```

The limited scope of RCTs is increasingly under scrutiny



The limited scope of RCTs is increasingly under scrutiny



The promise of detailed and larger observational or *real world* data sets

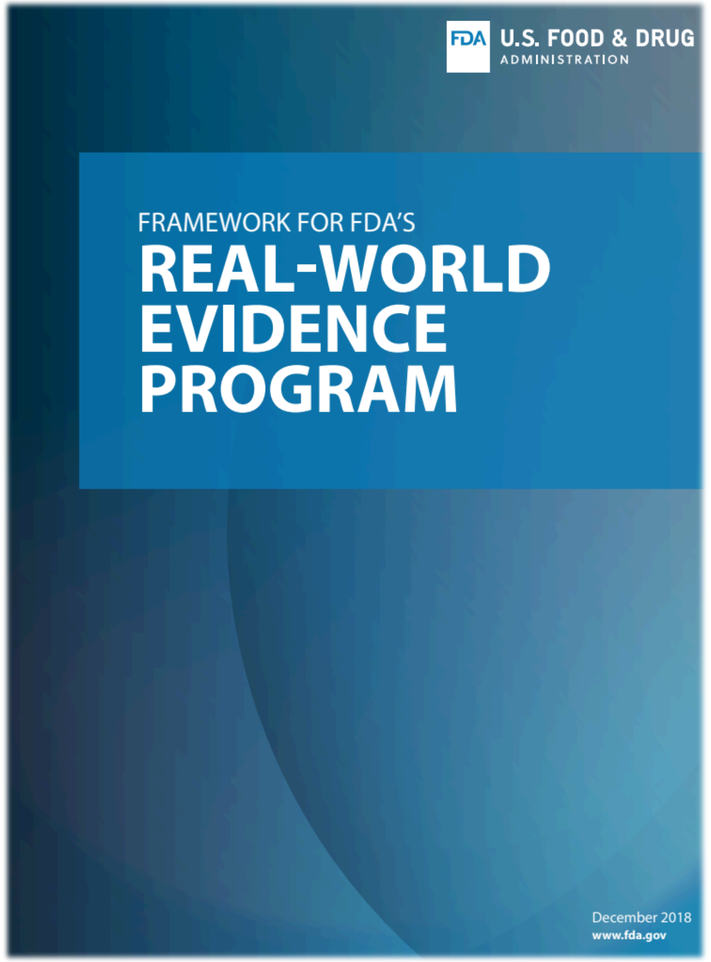
Estimate the efficacy in real-world conditions

- Using large cohorts like hospital data bases

- To emulate a target trial[1] leveraging observed confounding variables

- Solving both representativity and effective treatment given

🎁 *Large sample enabling more personalization (i.e stratified effects)*



Source: FDA's website

(1) Hernán and Robins, Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available, *Am J Epidemiol*, 2016

The example of a large French national cohort — The Traumabase

- 30,000 patients of unique size and granularity in Europe (~9,000 suffering from TBI)
- But randomisation does not hold, e.g. severe trauma are more likely to be treated

Among control	Among treated
16% dead	**38% dead**

← Confusion problem

The example of a large French national cohort — The Traumabase

- 30,000 patients of unique size and granularity in Europe (~9,000 suffering from TBI)
- But randomisation does not hold, e.g. severe trauma are more likely to be treated

Among control	Among treated
16% dead	**38% dead**

← Confusion problem

After adjustment on confounding covariates (Glasgow score, age, blood pressure, …), the null hypothesis of no effect can not be rejected[2].

Is there a paradox?

CRASH-3 key results

The risk of head injury-related death reduced with tranexamic acid in patients with mild-to-moderate head injury (RR 0·78 [95% CI 0·64–0·95]) but not in patients with severe head injury (0·99 [95% CI 0·91–1·07]

(2) Mayer et al., Doubly robust treatment effect estimation with missing attributes, *Annals of Applied Statistics* 2019

Machine Learning

- Works well with big data
- Non-parametric tools
- Seeking for predictions

- Goal: estimate

$$\mu := \mathbb{E}\left[Y \mid A = 1\right]$$

=> All that matters is prediction

Machine Learning

- Works well with big data
- Non-parametric tools
- Seeking for predictions

- Goal: estimate

$$\mu := \mathbb{E}\left[Y \mid A = 1\right]$$

=> All that matters is prediction

Causal inference

- Usually rather small data
- Linear or parametric model
- Willing to answer causal questions

- Goal ?

=> All that matters is inference

Machine Learning

- Works well with big data
- Non-parametric tools
- Seeking for predictions

- Goal: estimate

$$\mu := \mathbb{E}\left[Y \mid A = 1\right]$$

=> All that matters is prediction

Causal inference

- Usually rather small data
- Linear or parametric model
- Willing to answer causal questions

- Goal ?

=> All that matters is inference

Example of causal questions :

Effect of reducing car traffic on air pollution?
Is there an effect of financial incentives on teacher performance?
Do job training programs raise average future income?

What if?

Machine Learning

- Works well with big data
- Non-parametric tools
- Seeking for predictions

- Goal: estimate

$$\mu := \mathbb{E}\left[Y \mid A = 1\right]$$

=> All that matters is prediction

Causal inference

- Usually rather small data
- Linear or parametric model
- Willing to answer causal questions

- Goal ?

=> All that matters is inference

Some people crossed the bridge between the two, for e.g. Susan Athey

Inspiring woman (even if she uses R too)

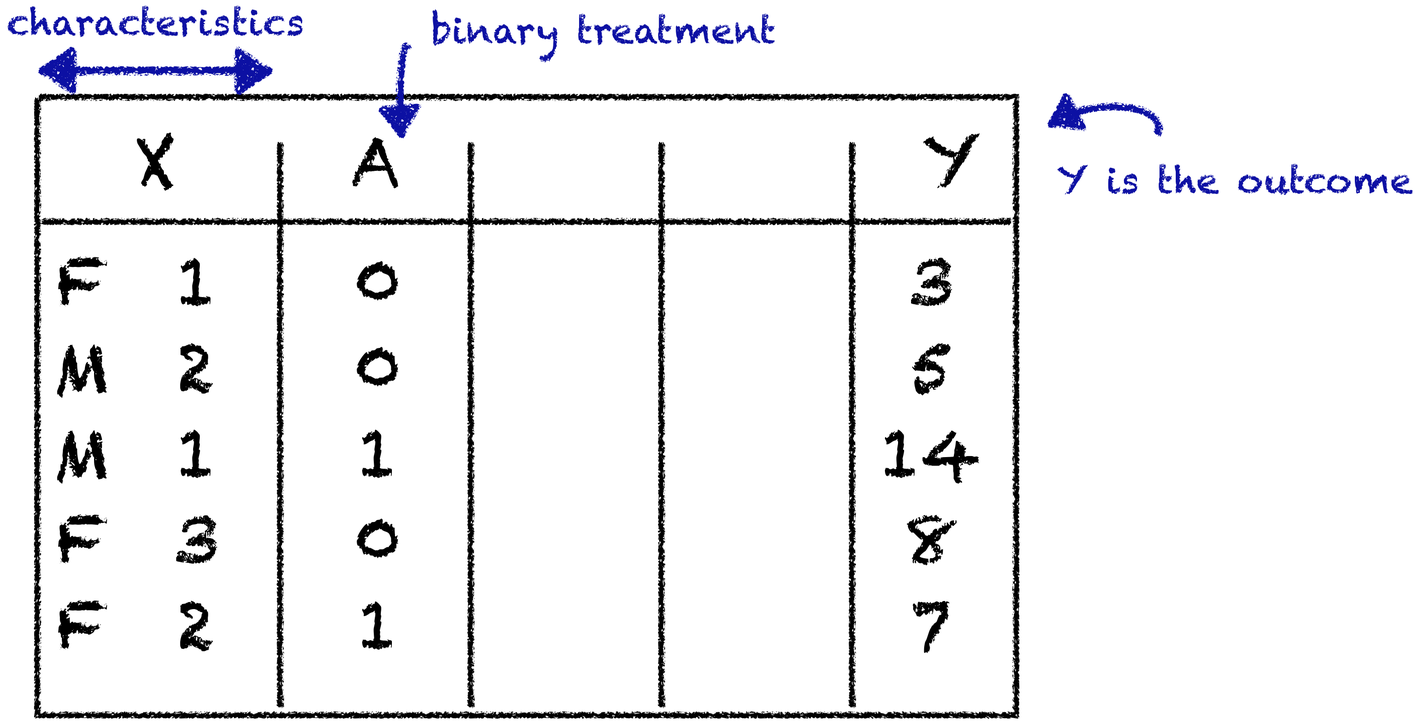
Causal inference

How to frame the problem?

Boileau par Jean-Baptiste Santerre (1678).
— « Ce que l'on conçoit bien s'énonce clairement,
Et les mots pour le dire arrivent aisément.»

Toward formalization — the potential outcomes framework to encode causality

For each individual ***i***, consider each of the possible outcomes for treated $Y^{(1)}$, and control $Y^{(0)}$.



Toward formalization — the potential outcomes framework to encode causality

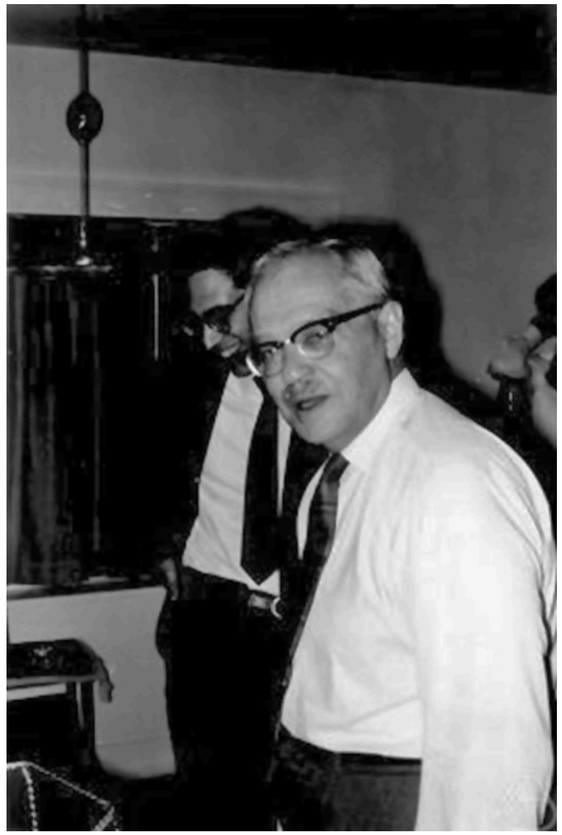
For each individual ***i***, consider each of the possible outcomes for treated $Y^{(1)}$, and control $Y^{(0)}$.

characteristics

binary treatment

X		A	Y(1)	Y(0)	Y
F	1	0	6	3	3
M	2	0	7	5	5
M	1	1	14	3	14
F	3	0	12	8	8
F	2	1	7	7	7

Y is the observed outcome



Source: Wikipedia
Jerzy Neyman à Berkeley en 1969.

Toward formalization — the potential outcomes framework to encode causality

For each individual ***i***, consider each of the possible outcomes for treated $Y^{(1)}$, and control $Y^{(0)}$.

characteristics

binary treatment

X		A	Y(1)	Y(0)	Y
F	1	0	NA	3	3
M	2	0	NA	5	5
M	1	1	14	NA	14
F	3	0	NA	8	8
F	2	1	7	NA	7

Y is the observed outcome

Toward formalization — the potential outcomes framework to encode causality

For each individual ***i***, consider each of the possible outcomes for treated $Y^{(1)}$, and control $Y^{(0)}$.

characteristics

binary treatment

X		A	Y(1)	Y(0)	Y
F	1	0	NA	3	3
M	2	0	NA	5	5
M	1	1	14	NA	14
F	3	0	NA	8	8
F	2	1	7	NA	7

Y is the observed outcome

In a RCT, $$\frac{1}{n_1}\sum_{i=1}^{n} A_i Y_i \rightarrow \mathbb{E}\left[Y \mid A = 1\right] = \mathbb{E}\left[Y^{(1)}\right]$$

Machine-learning versus Causality through the prism of notations

Prediction

$$\mathbb{E}[Y \mid X = x]$$

$\Longrightarrow$ Usual supervised learning

Causality within the potential outcomes framework

- Estimate what is the expected values of Y if everyone gets treatment $\mathbb{E}\left[Y^{(1)}\right]$,
- Or look for average treatment effect (ATE) $\mathbb{E}\left[Y^{(1)} - Y^{(0)}\right]$,
- Or look for individual treatment effect $\mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right]$

$\Longrightarrow$ Rubin, Guido Imbens, Susan Athey, ...

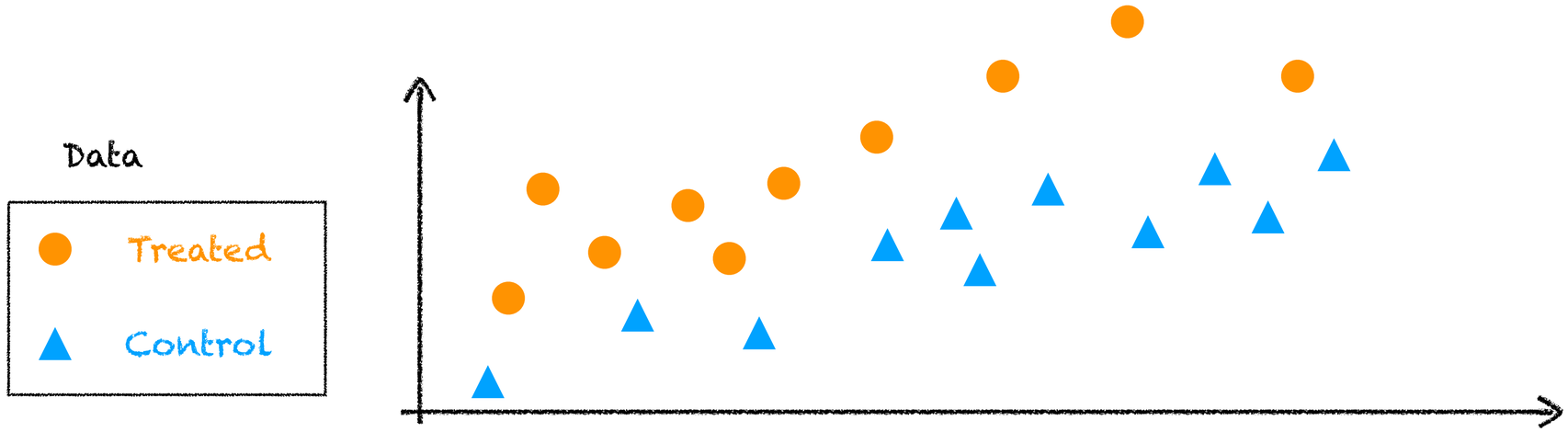
Causality within the SCM framework

- Estimate what is the expected values of Y if everyone gets treatment $\mathbb{E}[Y \mid do(A = 1)]$.

$\Longrightarrow$ Judea Pearl

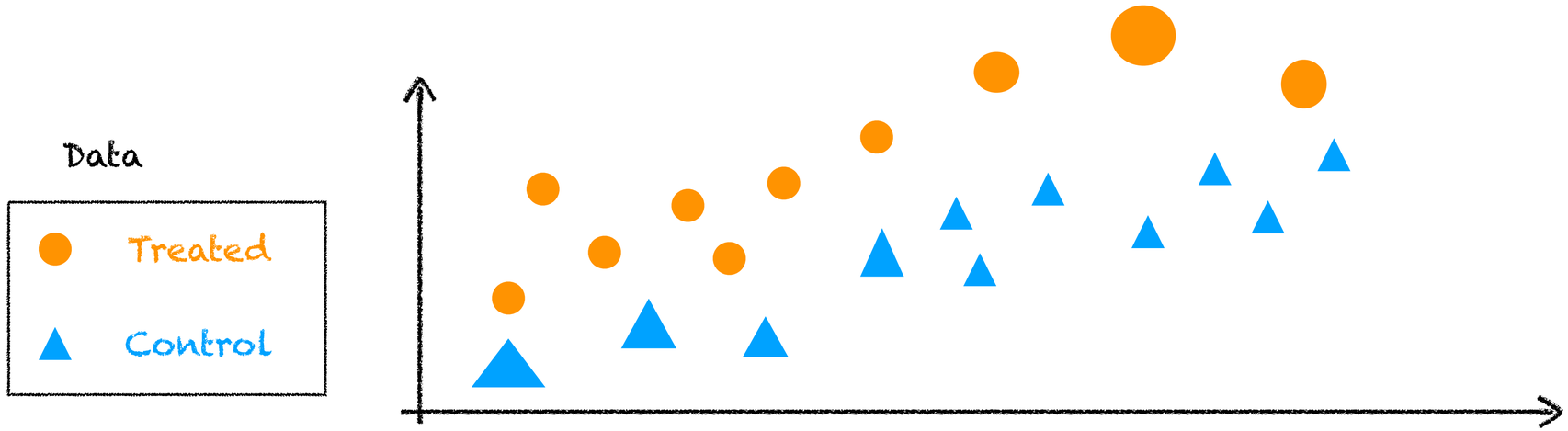
2 main approaches to generalize

1. **Re-weight** the trial individuals — *Inverse Propensity Weighting*



2 main approaches to generalize

1. **Re-weight** the trial individuals — *Inverse Propensity Weighting*



Can you guess the two assumptions I have to use for the approach to be valid?

2 main approaches to generalize

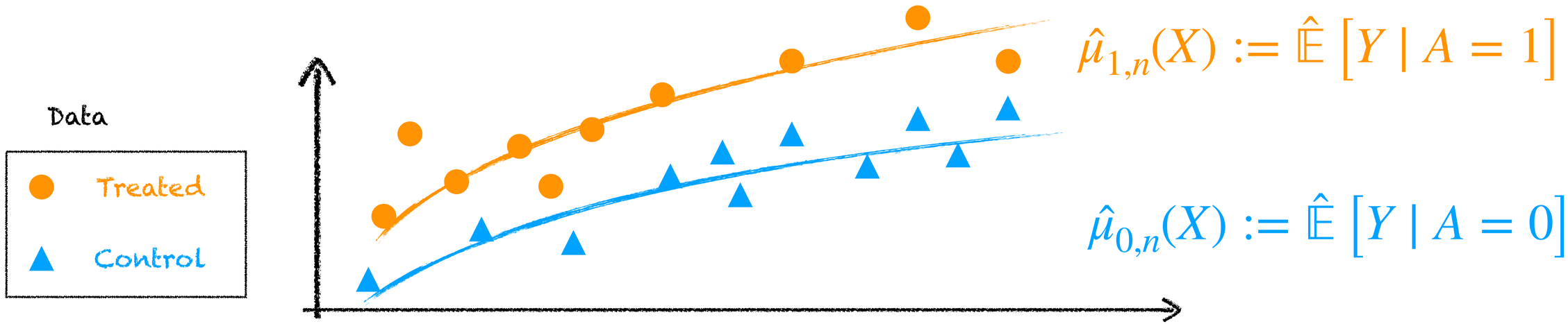
1. **Re-weight** the trial individuals — *Inverse Propensity Weighting*

```
from sklearn.linear_model import LogisticRegression

A = 'intervention'
Y = 'achievement_score'
X = data_with_categ.columns.drop(['schoolid', A, Y])

ps_model = LogisticRegression(C=1e6).fit(data_with_categ[X], data_with_categ[A])

data_ps = data.assign(propensity_score=ps_model.predict_proba(data_with_categ[X])[:, 1])

data_ps[["intervention", "achievement_score", "propensity_score"]].head()
```

Source: Causal Inference for The Brave and True

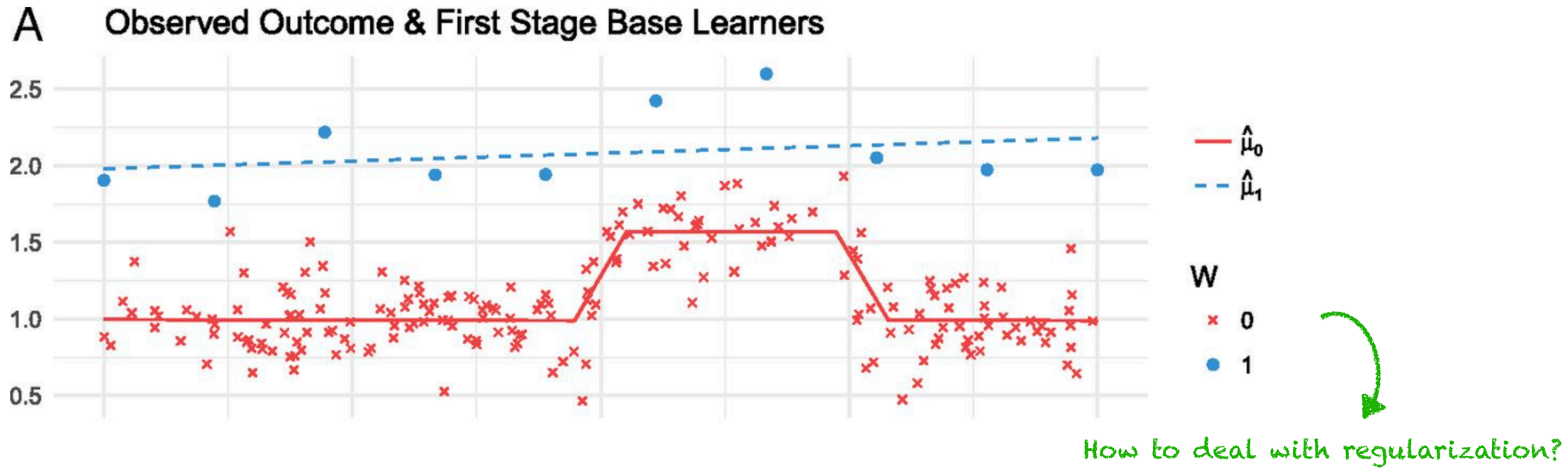
2 main approaches to generalize

1. **Re-weight** the trial individuals — *Inverse Propensity Sampling Weighting*
2. **Model the response** on each group and impute the missing values — *plug-in G-formula*



2 main approaches to generalize

1. **Re-weight** the trial individuals — *Inverse Propensity Weighting*
2. **Model the response** on each group and impute the missing values— *plug-in G-formula*



See Künzel et al. PNAS 2017.

Machine-learning and clinical evidence : how to bind the two?

Clinical evidence is deeply linked to measuring a causal effect.

But also true for humanities, public policy evaluation

Being good at predicting does not imply a causal understanding of phenomenons.

↳ a.k.a ML

Machine-learning and clinical evidence : how to bind the two?

Clinical evidence is deeply linked to measuring a causal effect.

But also true for humanities, public policy evaluation

Being good at predicting does not imply a causal understanding of phenomenons.

↳ a.k.a ML

As of today, the Python language is incredibly good for machine-learning, but is not the most used neither in the causal community, nor in the clinical field.

One has to be cautious when willing to take off-the-shelves algorithm's outputs to interpret it as a new clinical evidence (which directly impacts people's health through new clinical recommendations).