Established in 2013 by a donation from Judea Pearl and continued by support from Microsoft Research and Google, the award provides for a $5,000 cash prize each year.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf are awarded the 2018 Causality in Statistics Education prize for their open-access textbook, *Elements of Causal Inference: Foundations and Learning Algorithms*, that provides accessible coverage of the field, unique original content linking causal inference to machine learning and big data applications, and excellent conceptual and computational exercises for students.

**Recent Award Recipients & Materials**

- **2017:** Ilya Shpitser
- **2016:** Onyebuchi A. Arah and Arvid Sjölander (Arah's Materials, Sjölander's Materials)
- **2015:** Tyler VanderWeele
- **2014:** Maya Petersen and Laura B. Balzer
- **2013:** Felix Elwert

Position compared to other books

First, the present book represents a bias toward a subproblem of causality that may be considered both the most fundamental and the least realistic. This is the cause-effect problem, where the system under analysis contains only two observables.

[...]

And second, our treatment is motivated and influenced by the fields of machine learning and computational statistics. We are interested in how methods thereof can help with the inference of causal structures, and even more so whether causal reasoning can inform the way we should be doing machine learning.
Chapter 1: Statistical and Causal Models

[...] this will help us appreciate how much harder the problems of causal inference are, where the underlying model is no longer a fixed joint distribution of random variables, but a structure that implies multiple such distributions.
“No correlation without causation”

**Principle 1.1 (Reichenbach’s common cause principle)** If two random variables $X$ and $Y$ are statistically dependent ($X \not\perp Y$), then there exists a third variable $Z$ that causally influences both. (As a special case, $Z$ may coincide with either $X$ or $Y$.) Furthermore, this variable $Z$ screens $X$ and $Y$ from each other in the sense that given $Z$, they become independent, $X \perp Y \mid Z$.

\[ \begin{array}{c}
\circlearrowleft Z \\
\downarrow \ X \\
\downarrow \ Y
\end{array} \quad \begin{array}{c}
X \\
\rightarrow \ Y
\end{array} \quad \begin{array}{c}
X \\
\leftarrow \ Y
\end{array} \]

\[ \text{---} \]

\[ ^3 \text{For clarity, we formulate some important assumptions as principles. We do not take them for granted throughout the book; in this sense, they are not axioms.} \]
The Common Cause Principle

Explanation via screening off

Leszek Wroński
September 10, 2010

2 The Principle of the Common Cause: its shapes and content 7

2.1 Probability: the basics ........................................... 7

2.1.1 Screening off ................................................... 12

2.1.2 Observing probabilities and correlations ................. 13

2.2 The plurality of the Principles .................................. 15

2.3 What Reichenbach wrote ......................................... 21

2.3.1 Reichenbach’s argument for the Principle .............. 27

2.3.2 Other problems with Reichenbach’s approach .......... 31

2.4 The PCC after Reichenbach ................................. 41

2.4.1 The argument from conservation principles .......... 43

2.4.2 The “sea levels vs. bread prices” argument .......... 44

2.4.3 Which correlations demand explanation? .............. 47

2.5 An epistemic position .......................................... 51
Venetian sea levels and British bread prices


Two separate causal explanations of the two phenomena?
Melting of sub polar glaciers; high taxes/deteriorating crop levels.

![Internet Explorer vs Murder Rate](image-url)
Peters et al: “In practice…”

1. The random variables we observe are conditioned on others.

Example 6.30, Berkson’s paradox: “Why are handsome men such jerks?”

\[ H := N_H, \]
\[ F := N_F, \]
\[ R := \min(H, F) \oplus N_R, \]

where \( N_H, N_F \overset{iid}{\sim} \text{Ber}(0.5) \) and \( N_R \sim \text{Ber}(0.1). \)
Consider the single people

\[ H := N_H, \]
\[ F := N_F, \]
\[ R := \min(H, F) \oplus N_R, \]

Condition on \( R = 0 \).

If someone is handsome, he is more likely to be unfriendly.

\[ F \not\! \not \! \not \! \not \not \| H \mid R = 0 \]

See also Felix Elwert’s slides. 
https://www.ssc.wisc.edu/~felwert/causality/?page_id=66
Peters et al: “In practice…”

2. The random variables only appear to be dependent. Inference of a dependence as a result of a search procedure over a large number of pairs of random variables without a multiple testing correction.

3. Both random variables may inherit a time dependence and follow a simply physical law, such as exponential growth.
Example: pattern recognition (interventions)

Handwritten digits $X$, class labels $Y$.

Effect of intervening on class labels $Y$?

Model (i); $Y, N_X$ independent

Model (ii); $Z, M_X, M_Y$ independent
Example: phenotype after gene deletion (prediction)
Chapter 2: Assumptions for Causal Inference

(physical) independence of mechanisms
Principle 2.1

- intervenability, autonomy, modularity, invariance, transfer
- independence of information contained in mechanisms
- independence of noises, conditional independence of structures
Running example

A: altitude of city
T: annual average temperature
Intervenability / invariance

\[ p(a, t) = p(a \mid t)p(t) \]

\[ p(a, t) = p(t \mid a)p(a) \]
Intervenability / invariance

\[ p(a, t) = p(a \mid t)p(t) \]

\[ p(a, t) = p(t \mid a)p(a) \]

Choose by thought experiment: effect of interventions.

“Why do we find this description of the effect of interventions plausible, even though the hypothetical intervention is hard or impossible to carry out in practice?”

Do we think of \( p(a \mid t) \) or \( p(t \mid a) \) as invariant?
Arriving at causal structures

Our intuition can be rephrased and postulated in two ways: If $A \rightarrow T$ is the correct causal structure, then

(i) it is in principle **possible to perform a localized intervention** on $A$, in other words, to change $p(a)$ without changing $p(t|a)$, and

(ii) $p(a)$ and $p(t|a)$ are **autonomous**, **modular**, or **invariant** mechanisms or objects in the world.

Identify causal structures through:
- interventions
- checking which decompositions lead to invariant terms.

$$p_{\text{Austria}}(a, t) = p(t \mid a)p_{\text{Austria}}(a)$$
$$p_{\text{Switzerland}}(a, t) = p(t \mid a)p_{\text{Switzerland}}(a)$$
We can change one mechanism (intervene) without affecting the other mechanisms.
Independence of information

\[ p(a, t) = p(t | a)p(a) \]

We expect that \( p(t | a) \) provides no information about \( p(a) \). The mechanism is not influenced by the ensemble of cities to which we apply it.
The mechanism that generates the effect from its cause contains no information about the mechanism generating the cause.

(physical) independence of mechanisms
Principle 2.1

- intervenability, autonomy, modularity, invariance, transfer
- independence of information contained in mechanisms
- independence of noises, conditional independence of structures
Independence of noises

\[ p(a, t) = p(t \mid a)p(a) \]

Write

\[
A := N_A \\
T := f_T(A, N_T)
\]

where \( N_T \perp N_A \).
“[…] a weak dependence of noises may be possible without invalidating the principle of independent mechanisms.”
2.3 Physical Structure Underlying Causal Models

We conclude this chapter with some notes on connections to physics. Readers whose interests are limited to mathematical and statistical structures may prefer to skip this part.
Chapter 3: Cause-Effect Models

SCM = Structural Causal Model

Definition 3.1 (Structural causal models) An SCM $\mathcal{C}$ with graph $C \rightarrow E$ consists of two assignments

\[ C := N_C, \quad E := f_E(C, N_E), \]

where $N_E \perp N_C$, that is, $N_E$ is independent of $N_C$.
The SCM entails a joint distribution $P_{C,E}$ over $C$ and $E$, the observational distribution. The same observational distribution can be generated by different SCMs.

Interventions lead to intervention distributions.

SCMs entail (at least) two types of causal statements:
- the behavior of the system under potential interventions;
- counterfactual statements.
Interventions

Idea: after an intervention, only parts of the data-generating process change.

\[
C := N_C \\
E := f_E(C, N_E)
\]

This is called a \textit{(hard) intervention} and is denoted by \( \text{do}(E := 4) \).

The modified SCM entails a distribution over \( C \) that is denoted by \( P^{\text{do}(E := 4)}_C \) or \( P^{C; \text{do}(E := 4)}_C \).
Soft interventions

Keeps a functional dependence on C.

\[
\begin{align*}
C &:= N_C \\
E &:= f_E(C, N_E)
\end{align*}
\]

\[
\begin{align*}
C &:= N_C \\
E &:= g_E(C) + \tilde{N}_E
\end{align*}
\]

\[
do\left(E := g_E(C) + \tilde{N}_E \right)
\]
Example 3.2 (Cause-effect interventions) Suppose that the distribution $P_{C,E}$ is entailed by an SCM $\mathcal{C}$

$$C := N_C$$
$$E := 4 \cdot C + N_E,$$

with $N_C, N_E \overset{iid}{\sim} \mathcal{N}(0, 1)$, and graph $C \rightarrow E$. Then,

$$P_E^\mathcal{C} = \mathcal{N}(0, 17) \neq \mathcal{N}(8, 1) = P_{E,do(C:=2)}^\mathcal{C} = P_{E|C=2}^\mathcal{C}$$

$$\neq \mathcal{N}(12, 1) = P_{E,do(C:=3)}^\mathcal{C} = P_{E|C=3}^\mathcal{C}. $$
Cause-effect interventions

\[ C := N_C \]
\[ E := 4 \cdot C + N_E \]

Intervening on \( C \) changes the distribution of \( E \). But on the other hand,

\[ P_C^{\mathcal{E}; \text{do}(E:=2)} = \mathcal{N}(0, 1) = P_C^{\mathcal{E}} = P_C^{\mathcal{E}; \text{do}(E:=314159265)} \neq P_C^{\mathcal{E}; \text{do}(E:=2)} \]

\[ P_{C,E}^{\mathcal{E}; \text{do}(C:=\tilde{N}_C)} \quad \text{vs} \quad P_{C,E}^{\mathcal{E}; \text{do}(E:=\tilde{N}_E)} \]
Intervening vs conditioning

\[ P_C^{E; \text{do}(E=2)} = \mathcal{N}(0, 1) = P_C^{E} = P_C^{E; \text{do}(E=314159265)} \neq P_C^{E|E=2}. \quad (3.4) \]
Counterfactuals - eye disease example

B: 1 if a patient turns blind, 0 otherwise.

\( N_B \): latent binary variable, \( N_B \sim \text{Ber}(0.01) \).

T: 1 if treatment is administered, 0 otherwise.

\( N_T \): binary variable indicating whether treatment is administered.

Assume the underlying SCM

\[
\begin{align*}
\mathcal{C}: & \quad T := N_T \\
& \quad B := T \cdot N_B + (1 - T) \cdot (1 - N_B)
\end{align*}
\]
Counterfactual

Suppose a patient is treated and goes blind.

What would have happened if the patient had not been treated? What is the effect of \( do(T := 0) \)?

\[
\begin{align*}
\mathcal{C} : & \quad T := N_T \\
& \quad B := T \cdot N_B + (1 - T) \cdot (1 - N_B)
\end{align*}
\]

\[
\begin{align*}
\mathcal{C}|B = 1, T = 1 : & \quad T := 1 \\
& \quad B := T \cdot 1 + (1 - T) \cdot (1 - 1) = T
\end{align*}
\]

Assignment structure is unchanged, we only update noise distributions.

We have information on the noise variables for the given patient: \( B = T = 1 \) implies that we had \( N_B = 1 \).
do(T := 0)

\( \mathcal{C}|B = 1, T = 1 : \)  
\( T := 1 \)  
\( B := T \cdot 1 + (1 - T) \cdot (1 - 1) = T \)

\( \mathcal{C}|B = 1, T = 1; do(T := 0) : \)  
\( T := 0 \)  
\( B := T \)

\( P^{\mathcal{C}|B=1,T=1;do(T:=0)}(B = 0) = 1. \)

In this case, the causal model is falsifiable if \( N_B \) can be revealed.
For each fixed value $n_E$ of the noise $N_E$, $E$ is a deterministic function of $C$. $E = f_E(C, n_E)$. 

Suppose $C$ takes values in $\mathcal{C}$ and $E$ takes values in $\mathcal{E}$. 

We now view $N_E$ as taking values in the set of functions from $\mathcal{C}$ to $\mathcal{E}$. We write $E = n_E(C)$ and call this the canonical representation of the structural equation relating $C$ and $E$. 
Canonical representation - example 1

\[
C \equiv N_C \\
E \equiv N_E
\]

\[
N_C, N_E \text{ i.i.d. Ber}(0.5)
\]

\[
E = f_{N_E=0}(C), \quad f_{N_E=0}(c) = \begin{cases} 
0 & c = 0 \\
0 & c = 1
\end{cases}
\]

\[
E = f_{N_E=1}(C), \quad f_{N_E=1}(c) = \begin{cases} 
1 & c = 0 \\
1 & c = 1
\end{cases}
\]

We view \( N_E \) as taking values in \( \{ f_{N_E=0}, f_{N_E=1} \} \).
Canonical representation - example 2

\[
C := N_C \\
E := C \cdot (1 - N_E) + (1 - C) \cdot N_E \\
N_C, N_E \text{ i.i.d. Ber}(0.5)
\]

\[
E = g_{N_E=0}(C), \quad g_{N_E=0}(c) = \begin{cases} 
0 & c = 0 \\
1 & c = 1 
\end{cases}
\]

\[
E = g_{N_E=1}(C), \quad g_{N_E=1}(c) = \begin{cases} 
1 & c = 0 \\
0 & c = 1 
\end{cases}
\]

We view \( N_E \) as taking values in \( \{g_{N_E=0}, g_{N_E=1}\} \).
Intervention vs counterfactual

Two SCMs with different canonical representations may induce the same interventional probabilities while differing in their counterfactual statements.

\[
C := N_C \\
E := N_E \\
N_C, N_E \text{ i.i.d. } \text{Ber}(0.5)
\]

\[
C := N_C \\
E := C \cdot (1 - N_E) + (1 - C) \cdot N_E \\
N_C, N_E \text{ i.i.d. } \text{Ber}(0.5)
\]