

Elements of Causal Inference

Chapters 1 & 2

Andreas Kaufmann

MSc Student Computer Science

05. May 2021, Zürich

Introduction

Book info

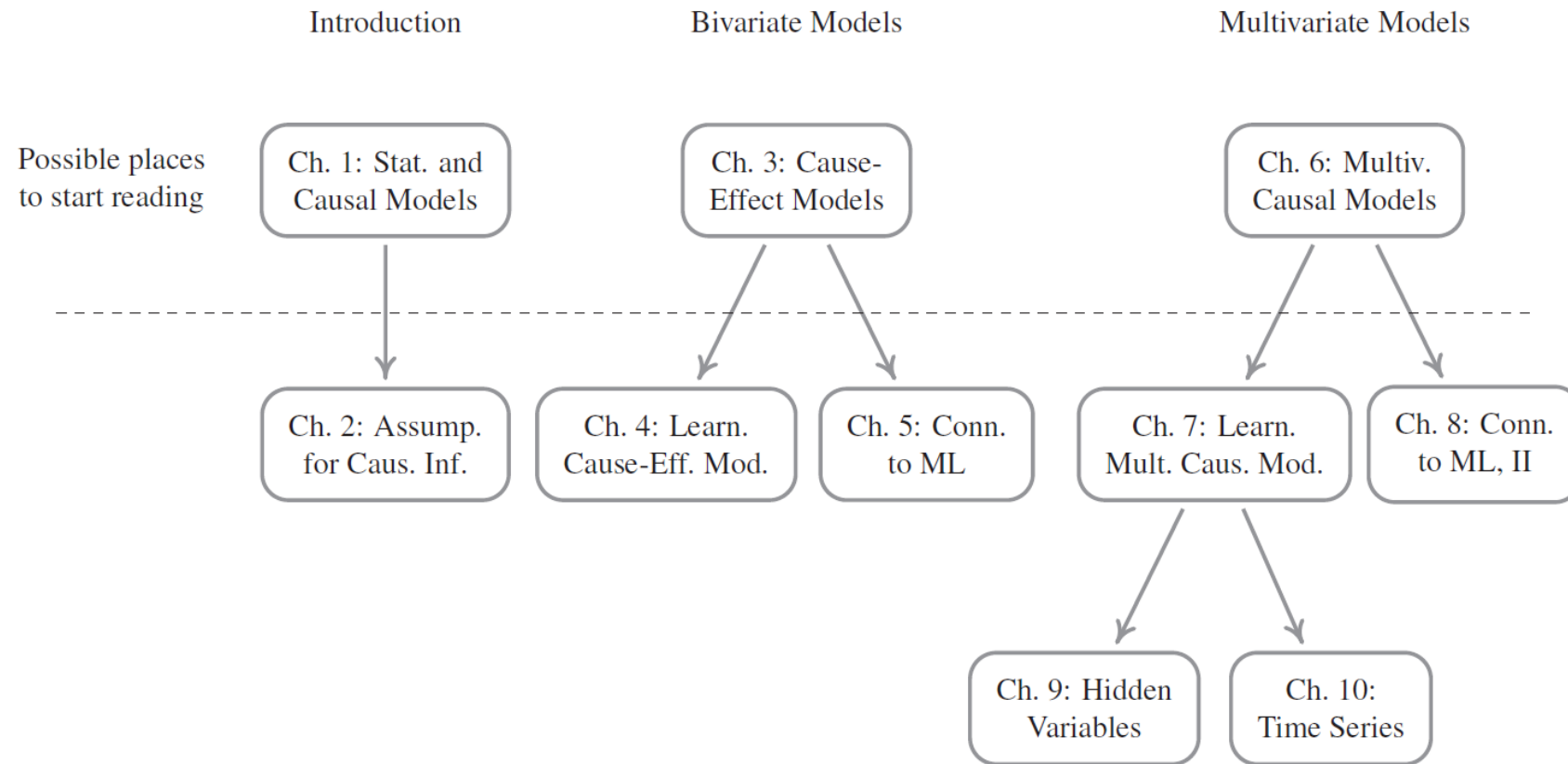
- Released: November 9th 2017
- Written by: Jonas Peters, Dominik Janzin and Bernhard Schölkopf

Topics discussed in the book

- Modeling causal structures
- Interventions and intervention distributions
- Given observed data, how can we infer such models

Introduction

Structure of the book



Content

1. Introduction
2. Probability Theory and Statistics
3. Causal Modeling and Learning
4. The principle of independent mechanisms
5. Connection to physics: Independence of Cause and Mechanism
6. Outlook Chapter 3
7. My take on the book
8. Discussion

Content

1. Introduction

2. Probability Theory and Statistics

3. Causal Modeling and Learning

4. The principle of independent mechanisms

5. Connection to physics: Independence of Cause and Mechanism

6. Outlook Chapter 3

7. My take on the book

8. Discussion

Chapter 1 – Probability Theory and Statistics

Our machine learning models often look as follows. Suppose we have observed data

$$(x_1, y_1), \dots, (x_n, y_n)$$

where $x_i \in \mathcal{X}$ are inputs and $y_i \in \mathcal{Y}$ are outputs. We then most of the time assume that (x_i, y_i) was generated by some unknown random experiment, more precisely that (x_i, y_i) are some realization of random variables (x_i, y_i) that are i.i.d (independent and identically distributed) with a joint distribution $P_{X,Y}$

Questions we might ask:

- Expectation of the output, given the input
- Binary classification
- Density

We use finite datasets to answer these questions.

Chapter 1 – Learning Theory

Example – Binary pattern recognition

We have $\mathcal{Y} = \{\pm 1\}$ and want to learn $f: \mathcal{X} \rightarrow \mathcal{Y}$. We seek to minimize the expected risk

$$R[f] = \int \frac{1}{2} |f(x) - y| dP_{X,Y}(x, y)$$

As $P_{X,Y}$ is unknown, we cannot compute the expected risk, let alone minimize it. Thus we try to do empirical risk minimization. We return the function minimizing the training error

$$R_{\text{emp}}^n[f] = \frac{1}{n} \sum_{i=1}^n \frac{1}{2} |f(x_i) - y_i|$$

over $f \in \mathcal{F}$. When n goes to infinity, risk is minimized (if function class is small)

Content

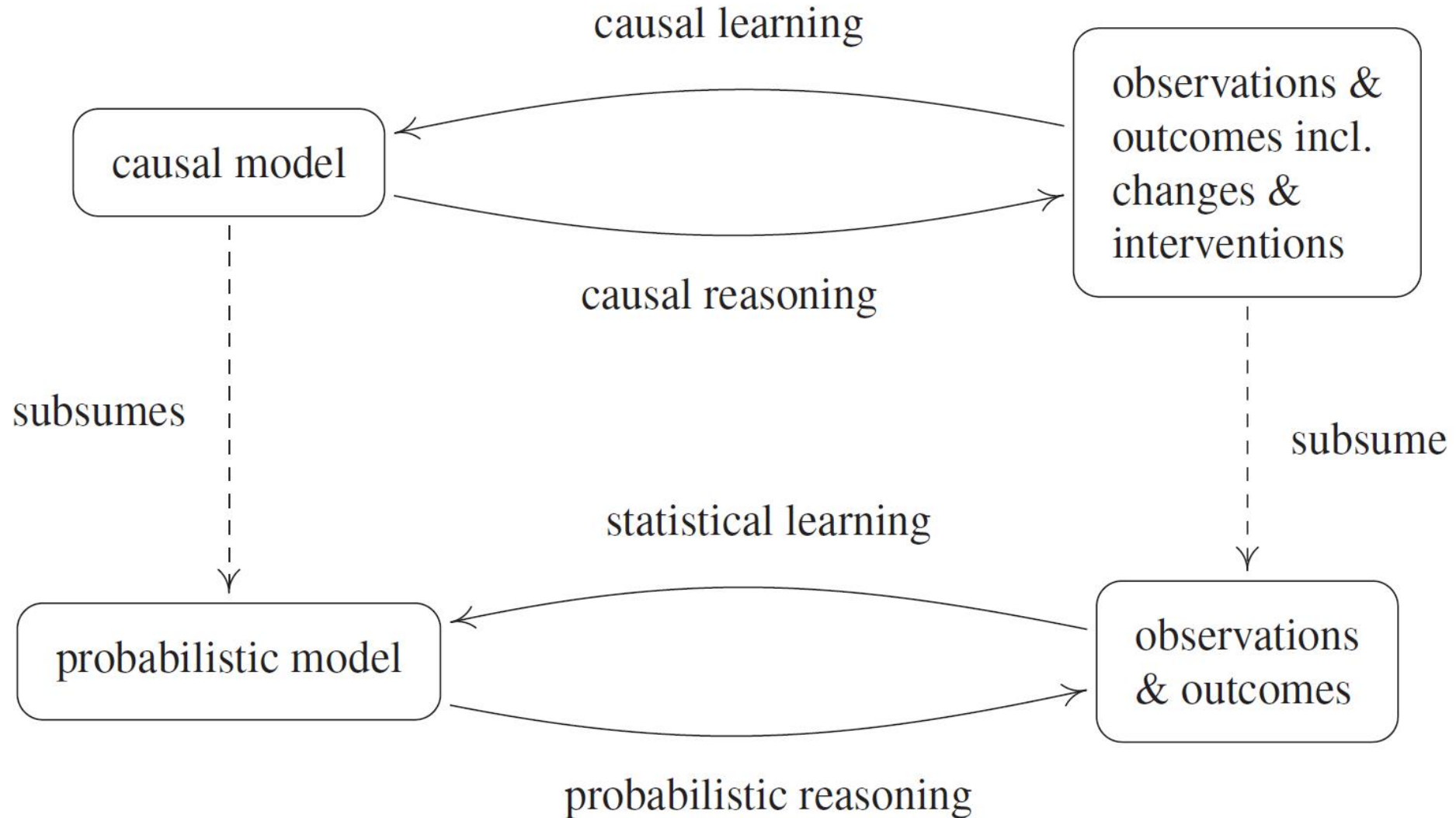
1. Introduction
2. Probability Theory and Statistics
- 3. Causal Modeling and Learning**
4. The principle of independent mechanisms
5. Connection to physics: Independence of Cause and Mechanism
6. Outlook Chapter 3
7. My take on the book
8. Discussion

Chapter 1 – Causal Modeling and Learning

*Definition 1.0 (**Causality**):*

Causality is the influence by which an event contributes to the production of another event.

Chapter 1 - Causal Modeling and Learning



Chapter 1 - Causal Modeling and Learning

Correlation does not imply causation

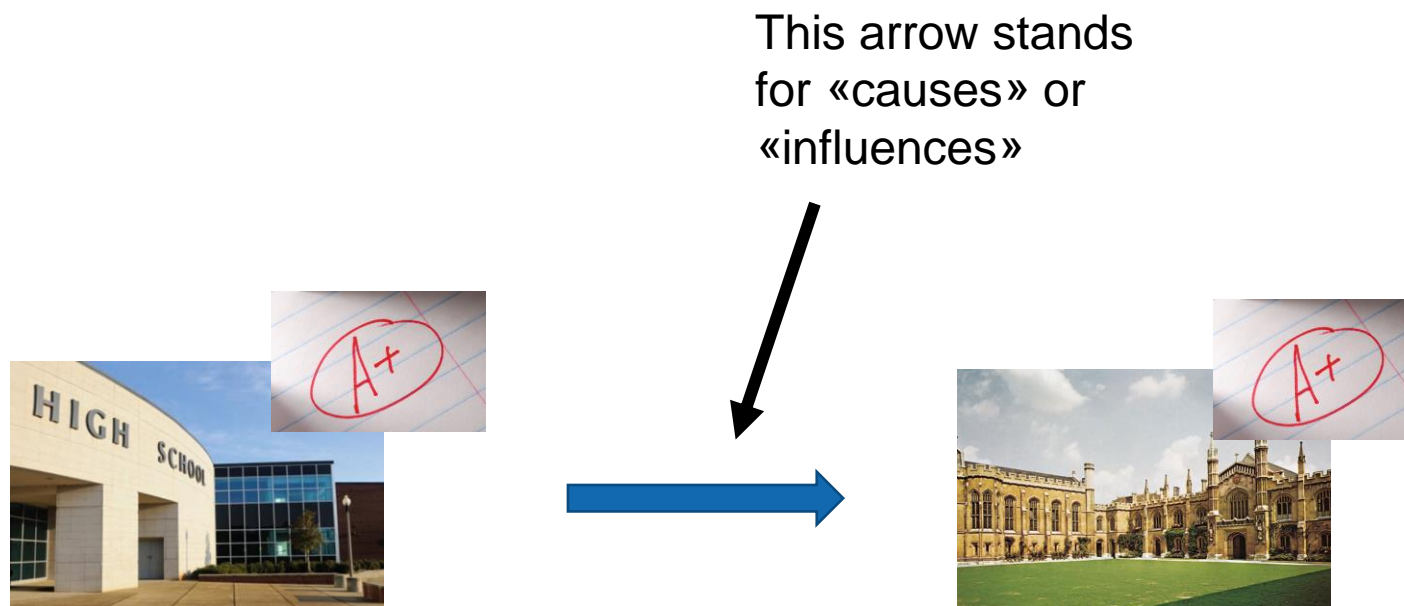
- Statistical properties alone do not determine causal structures



Chapter 1 - Causal Modeling and Learning

Correlation does not imply causation

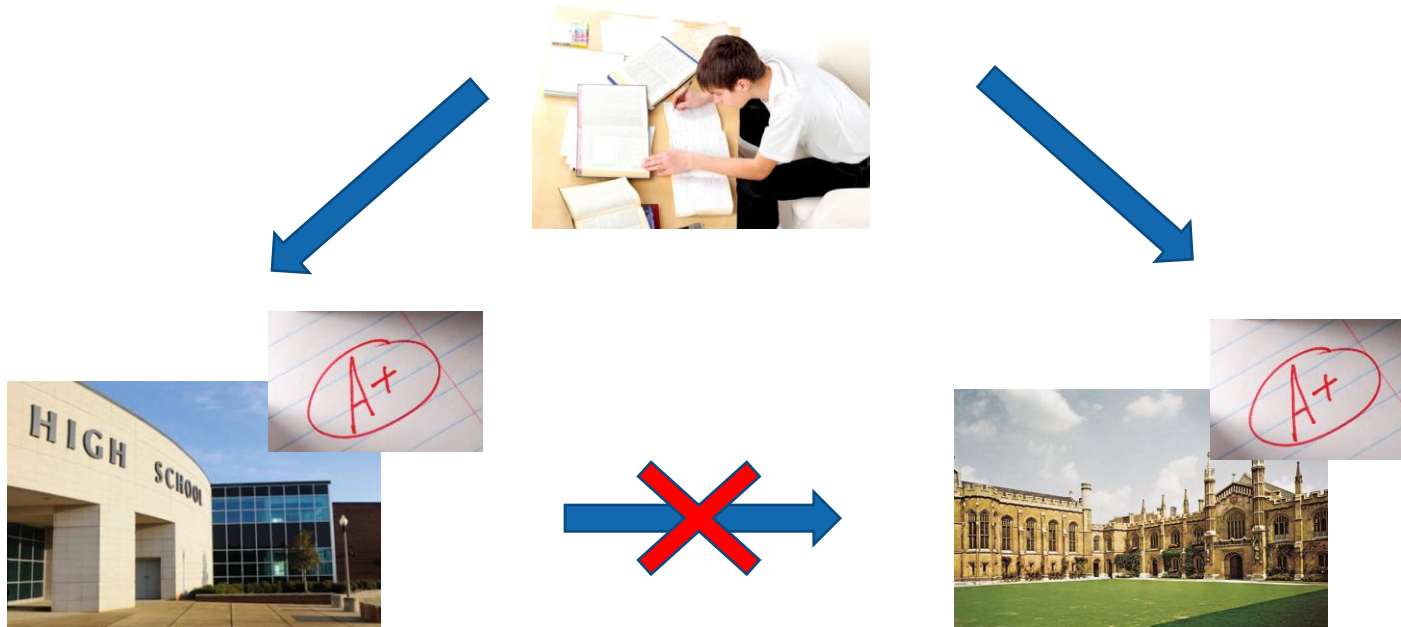
- Statistical properties alone do not determine causal structures



Chapter 1 - Causal Modeling and Learning

Correlation does not imply causation

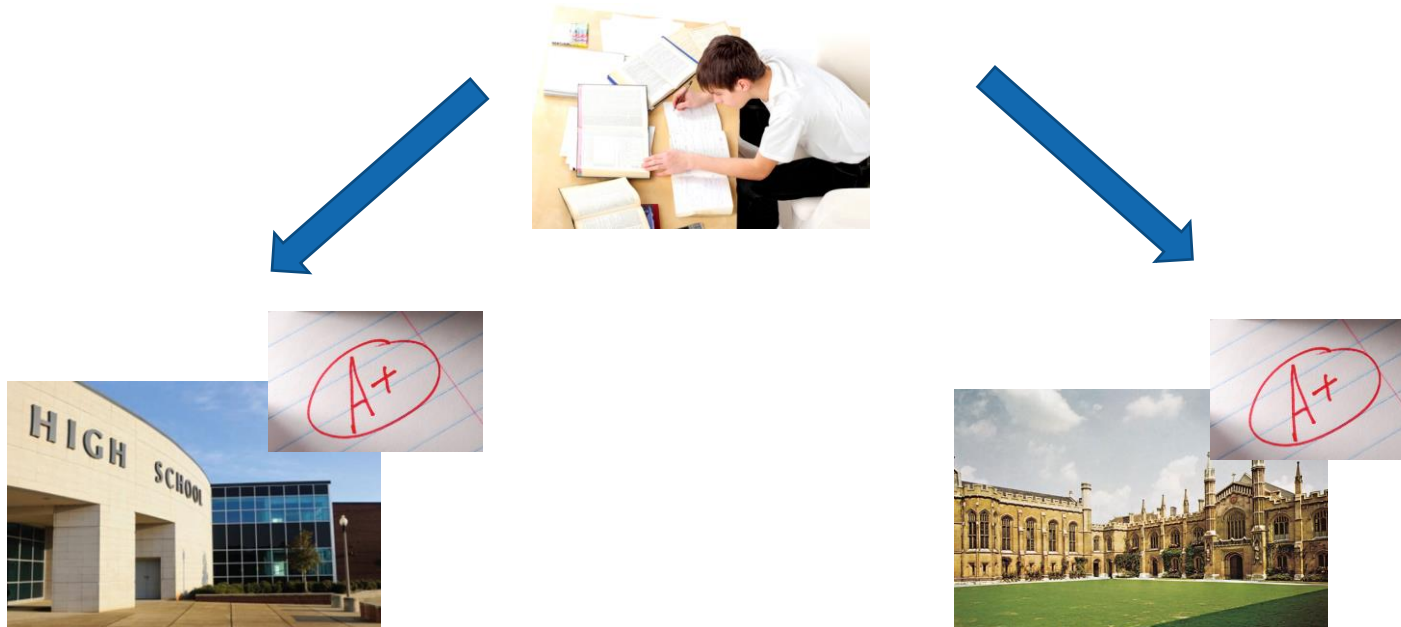
- Statistical properties alone do not determine causal structures



Chapter 1 - Causal Modeling and Learning

Correlation does not imply causation

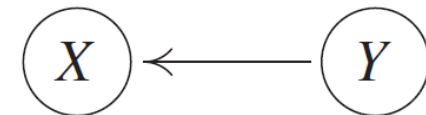
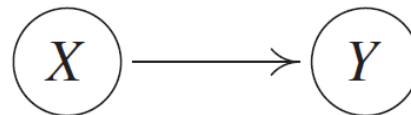
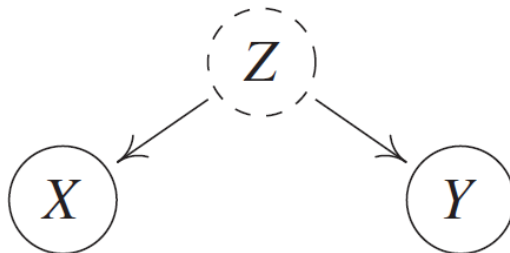
- Statistical properties alone do not determine causal structures



Chapter 1 - Causal Modeling and Learning

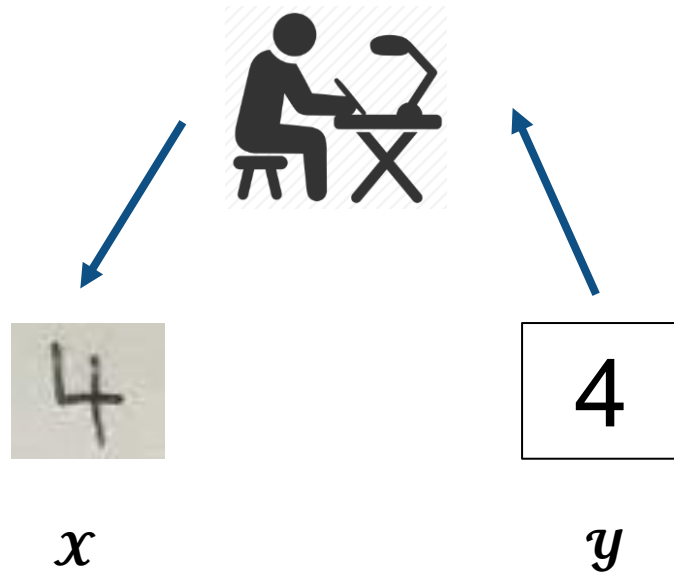
This principle was first understood by Reichenbach [1956]

Principle 1.1 (Reichenbach's common cause principle): If two random variables X and Y are statistically dependent, 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.

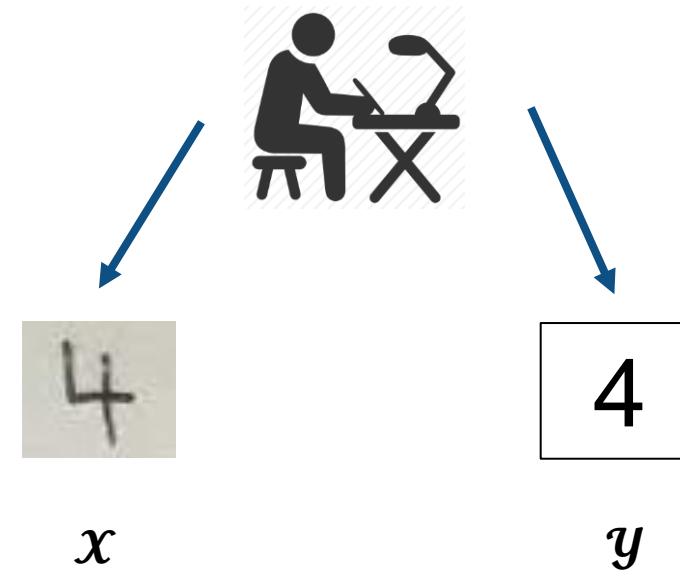


Chapter 1 – Examples for Causal Modeling

Experiment Model 1



Experiment Model 2

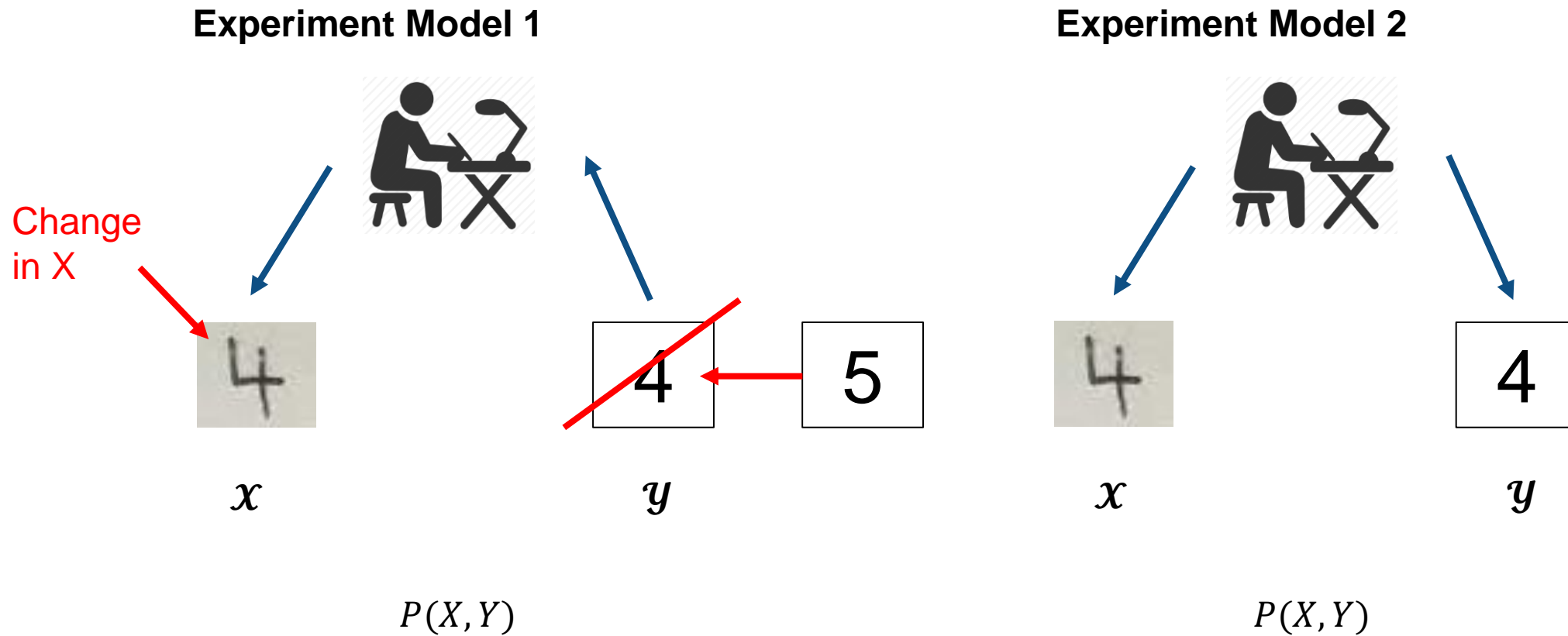


$P(X, Y)$

$P(X, Y)$

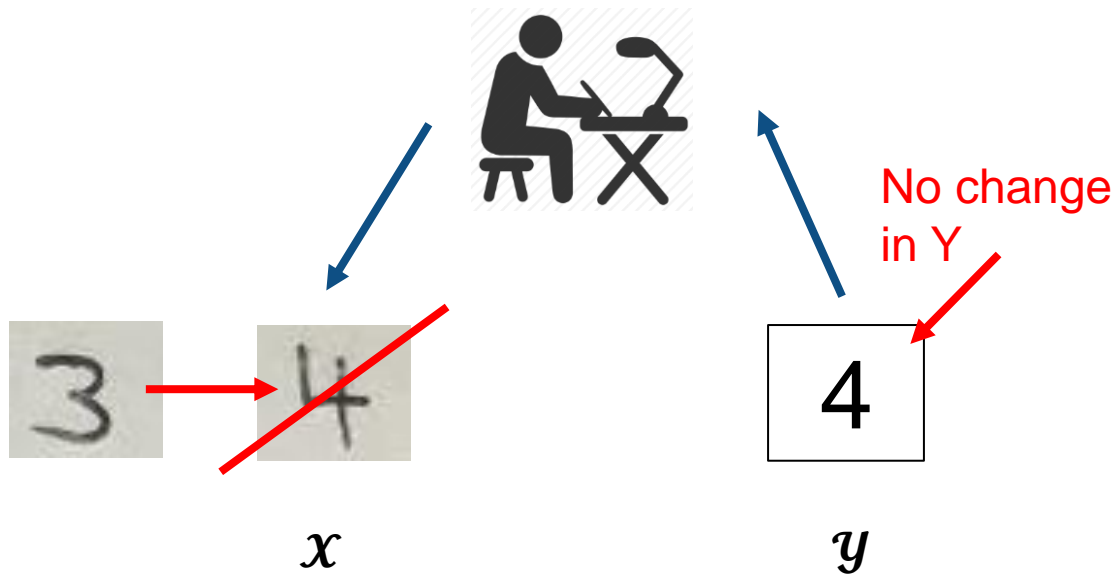
can be equal

Chapter 1 – Examples for Causal Modeling

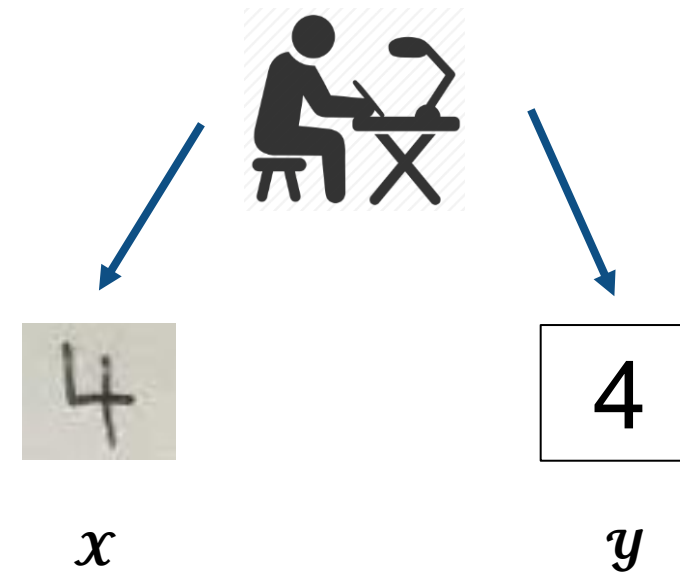


Chapter 1 – Examples for Causal Modeling

Experiment Model 1

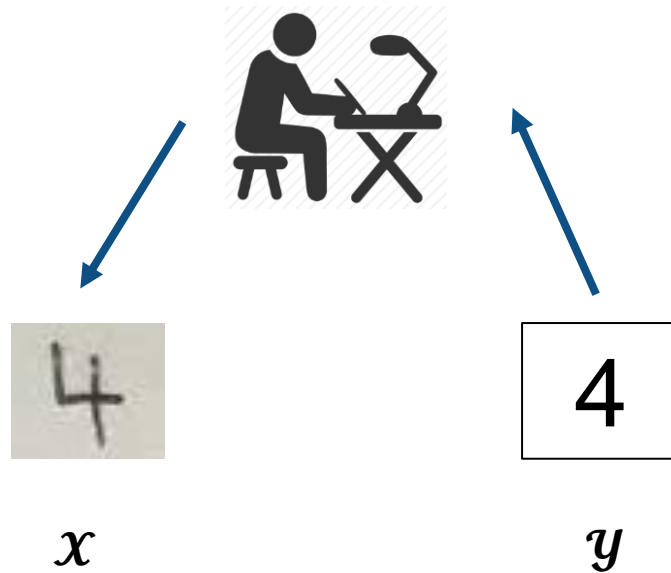


Experiment Model 2



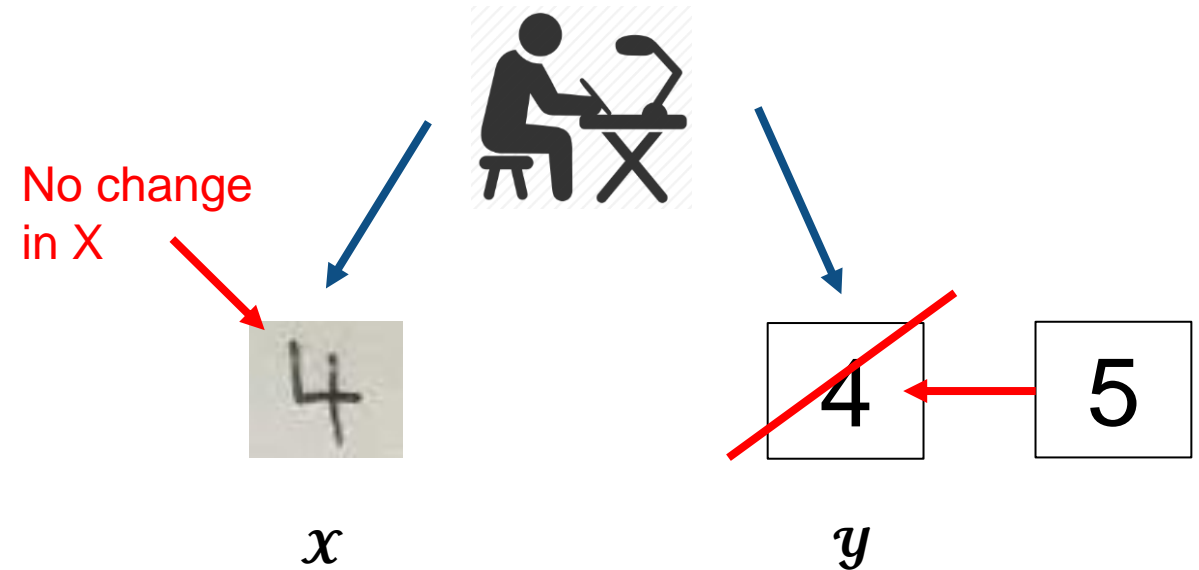
Chapter 1 – Examples for Causal Modeling

Experiment Model 1



$P(X, Y)$

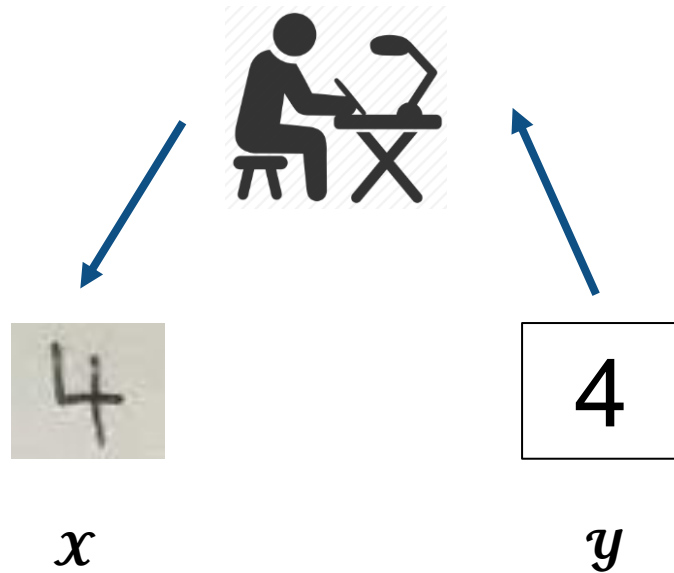
Experiment Model 2



$P(X, Y)$

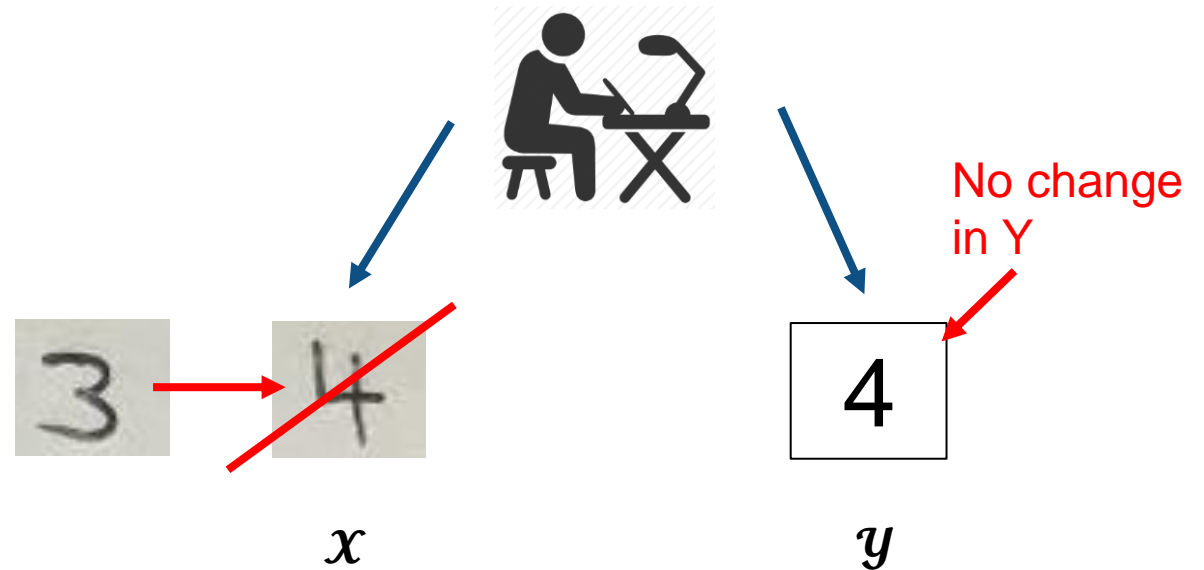
Chapter 1 – Examples for Causal Modeling

Experiment Model 1



$P(X, Y)$

Experiment Model 2

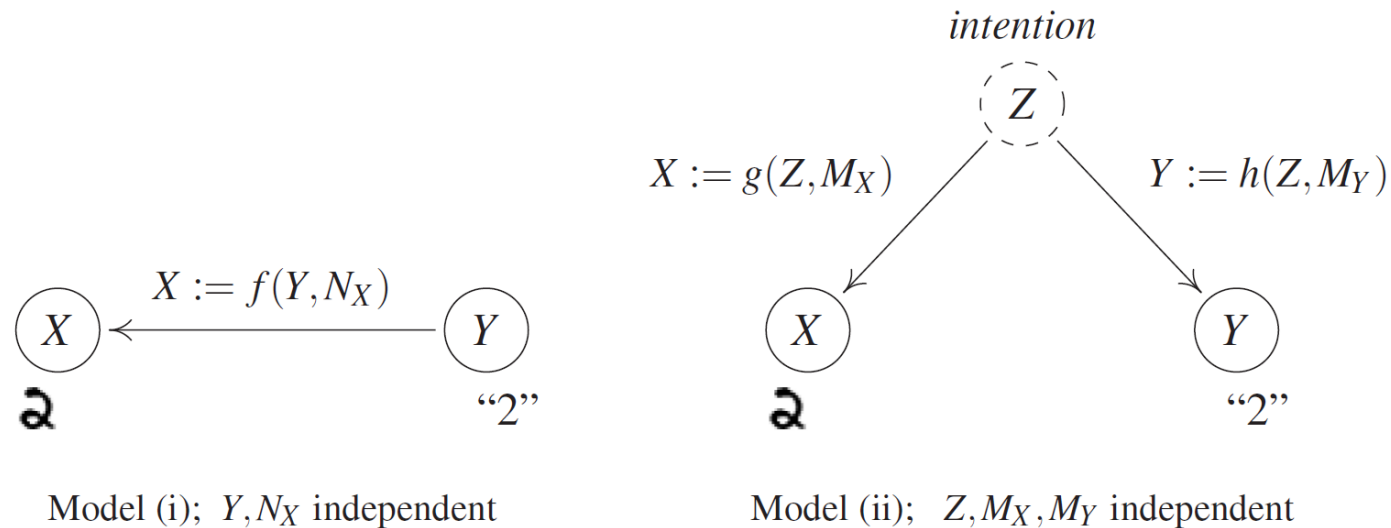


$P(X, Y)$

Chapter 1 – Two Examples for Causal Modeling

1. Pattern recognition

This two models are examples of **Structural Causal Models (SCMs)**



N_X , M_X and M_Y is the noise.

Content

1. Introduction
2. Probability Theory and Statistics
3. Causal Modeling and Learning
- 4. The principle of independent mechanisms**
5. Connection to physics: Independence of Cause and Mechanism
6. Outlook Chapter 3
7. My take on the book
8. Discussion

The two variable problem



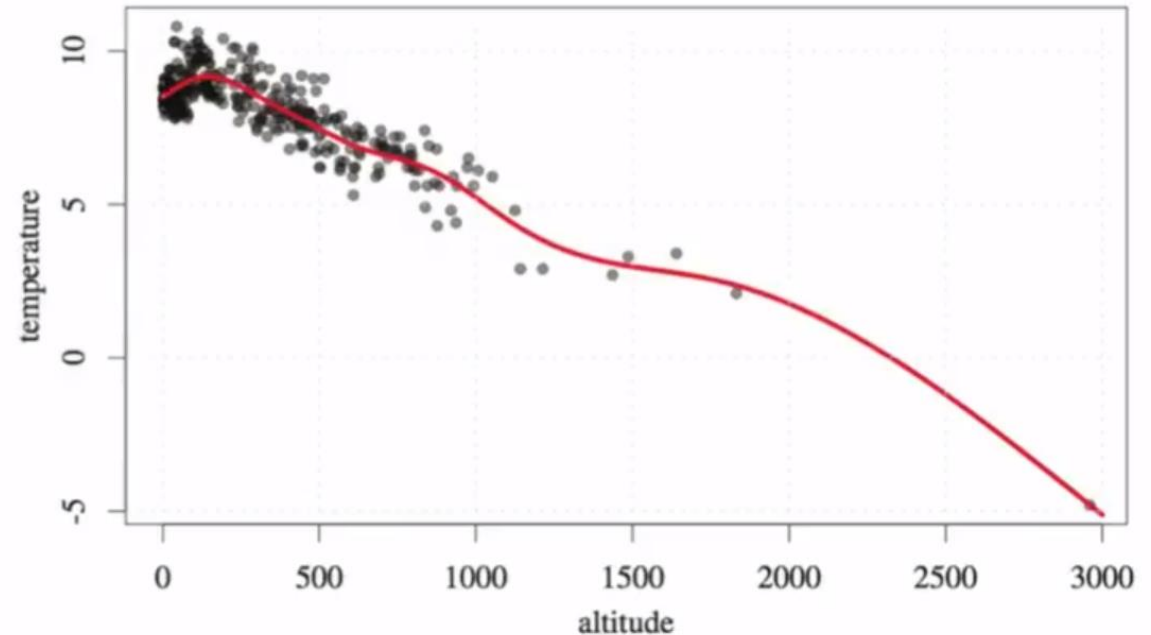
The error of mistaking cause for consequence

Chapter 2 – The Principle of Independent Mechanisms

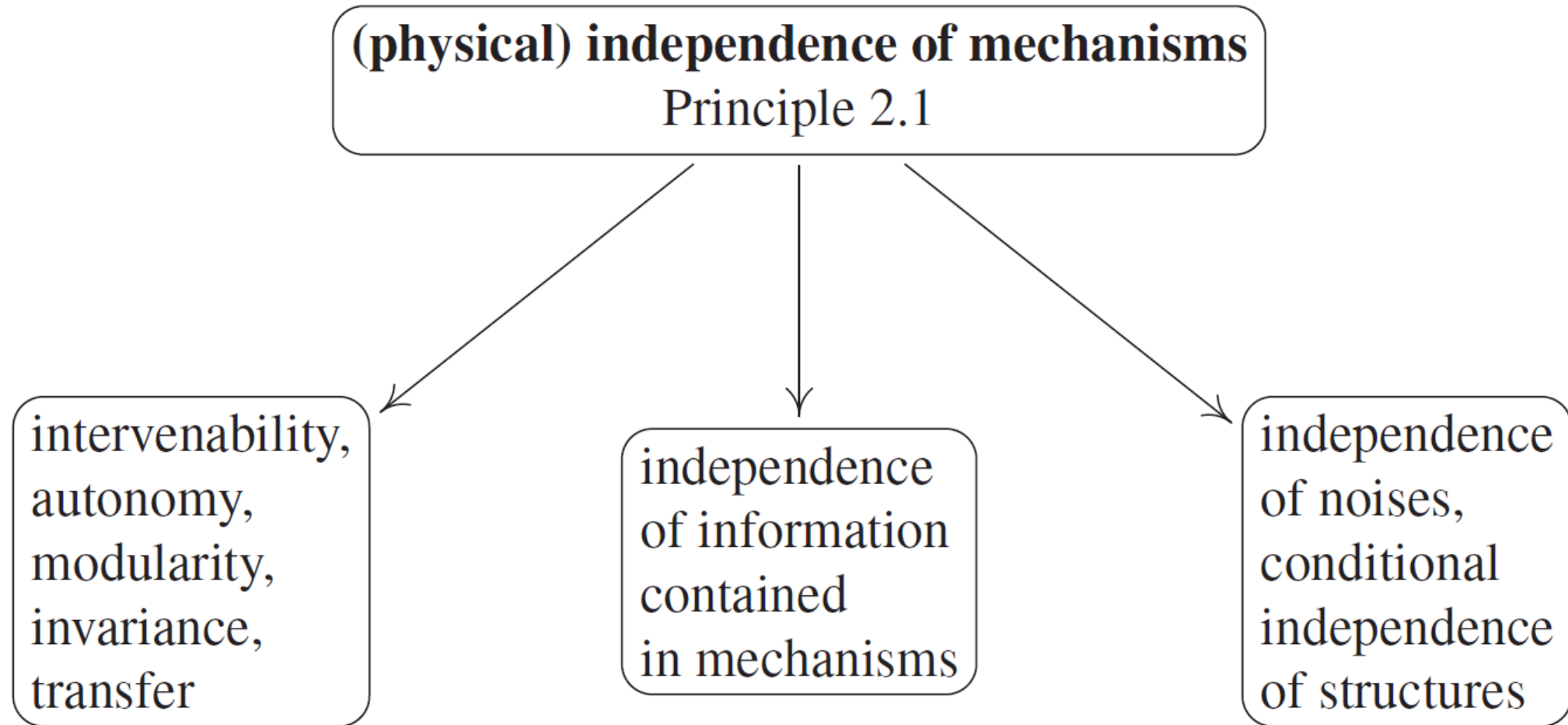
Example: We estimated density $p(a, t)$ of altitude A and average annual temperature T of a set of cities in a country. We can express $p(a, t)$ in two ways:

- $p(a|t) \times p(t)$ (factorization into $T \rightarrow A$)
- $p(t|a) \times p(a)$ (factorization into $A \rightarrow T$)

*Can we decide which factorization is the causal one?
I.e. we ask the question: what is the correct causal structure?*



Chapter 2 – The Principle of Independent Mechanisms

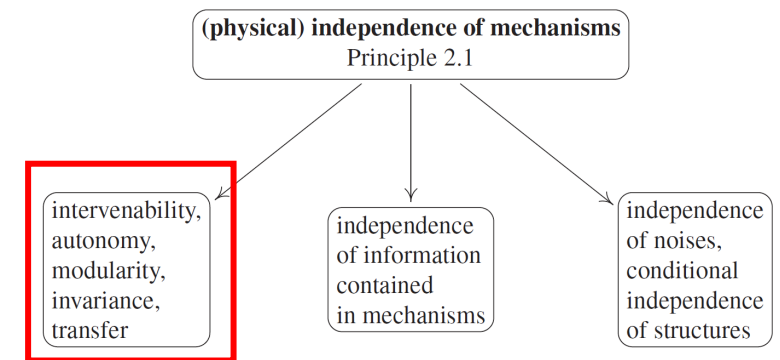


Chapter 2 – The Principle of Independent Mechanisms

1. Effect of interventions: (i.e. changing altitude or changing temperature)

- Change the altitude of a city and find out if the temperature changes over time
- By building a massive heating system, change the Temperature in the City
- We can do this interventions in a hypothetical thought experiment
- It's an argument that we have an independence between the physical mechanism $p(t|a)$ and the distribution of the cities $p(a)$

Our reasoning ends up suggesting that actual interventions may not be the only way to arrive at causal structures. Another way to derive the causal structure is to identify the datasources of $p(a, t)$.



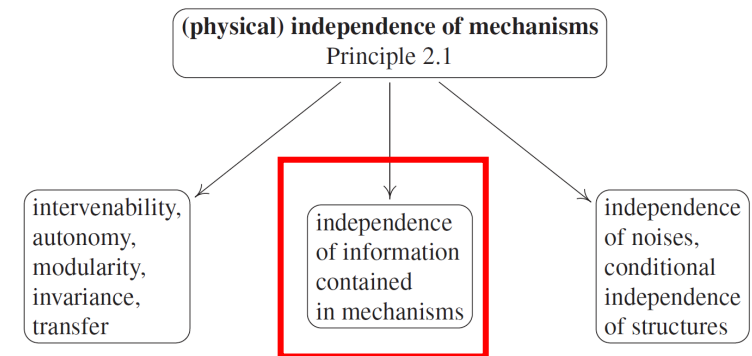
Chapter 2 – The Principle of Independent Mechanisms

2. Identifying the data source

Let $p^O(a, t)$ and $p^S(a, t)$ be the joint distributions of altitude and temperature in Austria and Switzerland, It can still hold that

$$\begin{aligned} p^O(a, t) &= \mathbf{p}(t|\mathbf{a})p^O(a) \\ p^S(a, t) &= \mathbf{p}(t|\mathbf{a})p^S(a) \end{aligned}$$

We could find out that we can fit the two datasets using the invariant conditional $p(t|a)$.



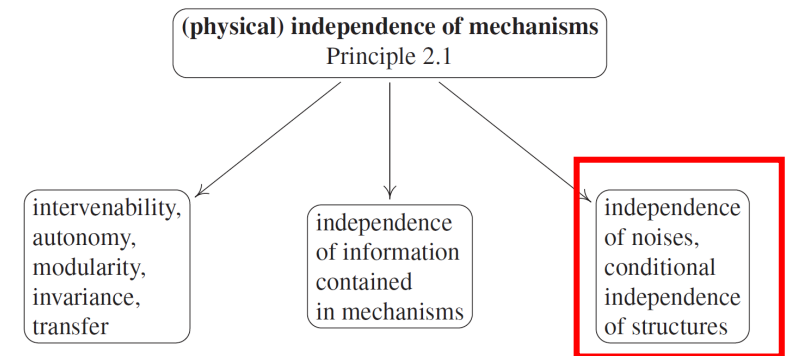
Chapter 2 – The Principle of Independent Mechanisms

3. Independence of noise terms

We can view the distribution with graph $A \rightarrow T$ as a noisy function of the cause A

$$A := N_A$$
$$T := f_T(A, N_T)$$

By making suitable restrictions on the functional form of f_T (Chapter 4.1 and 7.1) allows us to identify which causal structure has entailed the observed $p(a, t)$



Chapter 2 - The Principle of Independent Mechanisms



Notion of Independence:

- Right picture: Object and mechanism by which the information contained in the light arrive at our brain are independent
- Left picture: Independence can be violated by taking a perspective in the right angle. I.e. we receive (perceive) an information which is not there (the 3d structure of a chair)

In this example, the result (our perception) is dependent on object, lightning and viewpoint. Object and lightning are not affected by viewpoint.

Chapter 2 – The Principle of Independent Mechanisms

We like to view all these observations as closely connected instantiations of a general principle of (physically) independent mechanisms. The aspects may help for the problem of causal learning, i.e. they may provide information about causal structures.

Principle 2.1 (Independent mechanisms): The conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions.

The special case of two variables has been referred to as *independence of cause and mechanism (ICM)*. It is obtained by a mechanism that is independent of the mechanism that turns the input into the output.

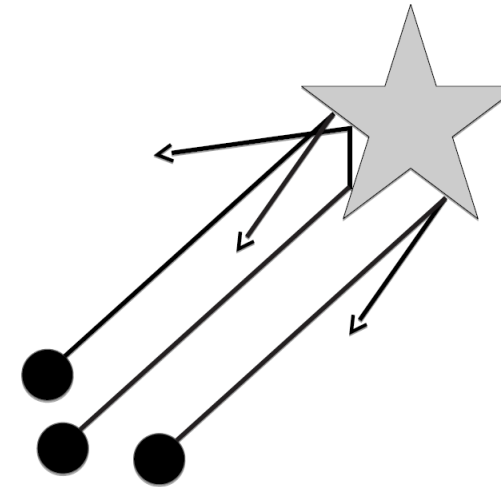
Content

1. Introduction
2. Probability Theory and Statistics
3. Causal Modeling and Learning
4. The principle of independent mechanisms
- 5. Connection to physics: Independence of Cause and Mechanism**
6. Outlook Chapter 3
7. My take on the book
8. Discussion

Physical Structure underlying causal models

Independence of Cause and Mechanism and the Thermodynamic Arrow of Time

- **Cause:** Incoming beam
- **Mechanism:** Scattering at object
- **Effect:** Outgoing beam



Principle 2.2 (Initial state and dynamical law): If s is the initial state of a physical system and M a map describing the effect of applying the system dynamics for some fixed time, then s and M are independent. Here, we assume that the initial state, by definition, is a state that has not interacted with the dynamics before.

Here s is cause and $M(s)$ is the effect. M is the mechanism.

Content

1. Introduction
2. Probability Theory and Statistics
3. Causal Modeling and Learning
4. The principle of independent mechanisms
5. Connection to physics: Independence of Cause and Mechanism

6. Outlook Chapter 3

7. My take on the book
8. Discussion

Outlook Chapter 3

Cause Effect Models

1. Structural Causal Models

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

$$\begin{aligned} C &:= N_C \\ E &:= f_E(C, N_E) \end{aligned}$$

Where N_E independent to N_C .

2. Interventions

1. **Hard intervention:** We set $E := 4$, denoted by $do(E := 4)$, $P_C^{\mathcal{G}, do(E:=4)}$
2. **Soft intervention:** $do(E := g_E(C) + \tilde{N}_E)$

3. Counterfactuals

One of the potential outcomes becomes the actual outcome, the other outcome is the counterfactual outcome (outcome if something different had happened)

Content

1. Introduction
2. Probability Theory and Statistics
3. Causal Modeling and Learning
4. The principle of independent mechanisms
5. Connection to physics: Independence of Cause and Mechanism
6. Outlook Chapter 3
- 7. My take on the book**
8. Discussion

Content

1. Introduction
2. Probability Theory and Statistics
3. Causal Modeling and Learning
4. The principle of independent mechanisms
5. Connection to physics: Independence of Cause and Mechanism
6. Outlook Chapter 3
7. My take on the book
- 8. Discussion**

Appendix

Second Example CH1

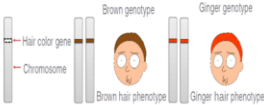
Chapter 1 – Two Examples for Causal Modeling

2. Gene perturbation

In previous example we have seen that different causal structures lead to different intervention distributions. Sometimes we want to predict an outcome under such an intervention.

Definition 1.3 (Gene): A gene is part of an individual's DNA sequence. The DNA is a sequence of genes and all together define the individual.

Definition 1.4 (Phenotype): A phenotype is a physical characteristic of an individual such as blue eyes, brown hair, etc.



ETH zürich Seminar in Advanced Topics in Machine Learning and Data Science by Prof. Fernando Perez-Cruz Andreas Kaufmann | 05.05.2021 | 38/38

Overview different models

Chapter 1 – Two Examples for Causal Modeling

Model	Predict in i.i.d. setting	Predict under changing distr. or intervention	Answer counterfactual questions	Obtain physical insight	Learn from data
Mechanistic/ physical, e.g., Sec. 2.3	yes	yes	yes	yes	?
Structural causal model, e.g., Sec. 6.2	yes	yes	yes	?	?
Causal graphical model, e.g., Sec. 6.5.2	yes	yes	no	?	?
Statistical model, e.g., Sec. 1.2	yes	no	no	no	yes

ETH zürich Seminar in Advanced Topics in Machine Learning and Data Science by Prof. Fernando Perez-Cruz Andreas Kaufmann | 05.05.2021 | 42/38

Historical notes

Chapter 2 – Historical notes

Intellectual antecedent to SEMs:

Sem are nowadays strongly associated with econometrics (Econometrics is the application of statistical methods to economic data in order to give empirical content to economic relationships.) Trygve Haavelmo 1944 laid the conceptual foundations of probabilistic econometrics. I.e. unlike correlation, regression has a natural direction. What direction to perform the regression? Problem of observational equivalence

Connection between: what does make a set of equations or relations structural and properties of invariance and autonomy (Aldrich 1989, Frisch et al 1984):

Here structural relation was aiming to try to capture an underlying structure connecting the variables of the model.

ETH zürich Seminar in Advanced Topics in Machine Learning and Data Science by Prof. Fernando Perez-Cruz Andreas Kaufmann | 05.05.2021 | 43/38

Other Connections to physics

Physical Structure underlying causal models

The Role of Time

This aspect is missing in section 2.1 of the book. Already Simon (1953) recognized that while time ordering can provide a useful asymmetry, it is asymmetry that is important, not the temporal sequence. It is even believed that both classical systems and quantum mechanical systems are invertible. There is work investigating into this (Bennett 1982, Zurek 1989).

ETH zürich Seminar in Advanced Topics in Machine Learning and Data Science by Prof. Fernando Perez-Cruz Andreas Kaufmann | 05.05.2021 | 46/38

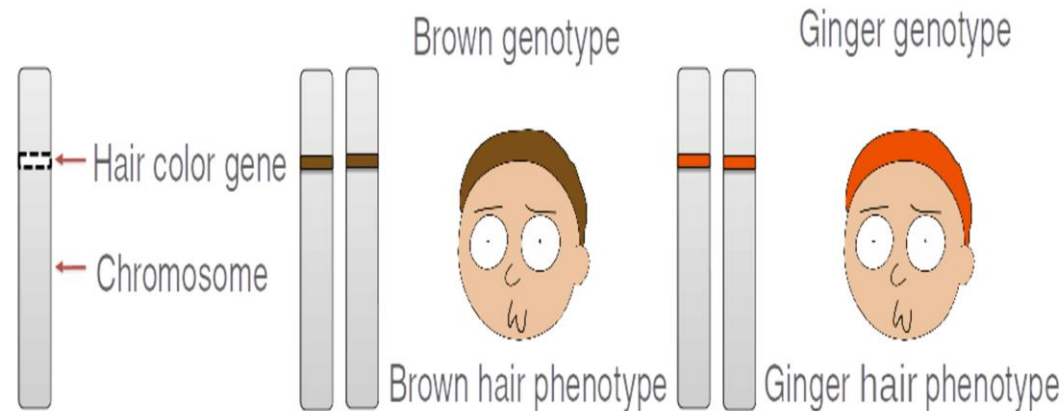
Chapter 1 – Two Examples for Causal Modeling

2. Gene perturbation

In previous example we have seen that different causal structures lead to different intervention distributions. Sometimes we want to predict an outcome under such an intervention.

*Definition 1.3 (**Gene**): A gene is part of an individual's DNA Sequence. The DNA is a sequence of genes and all together define the individual*

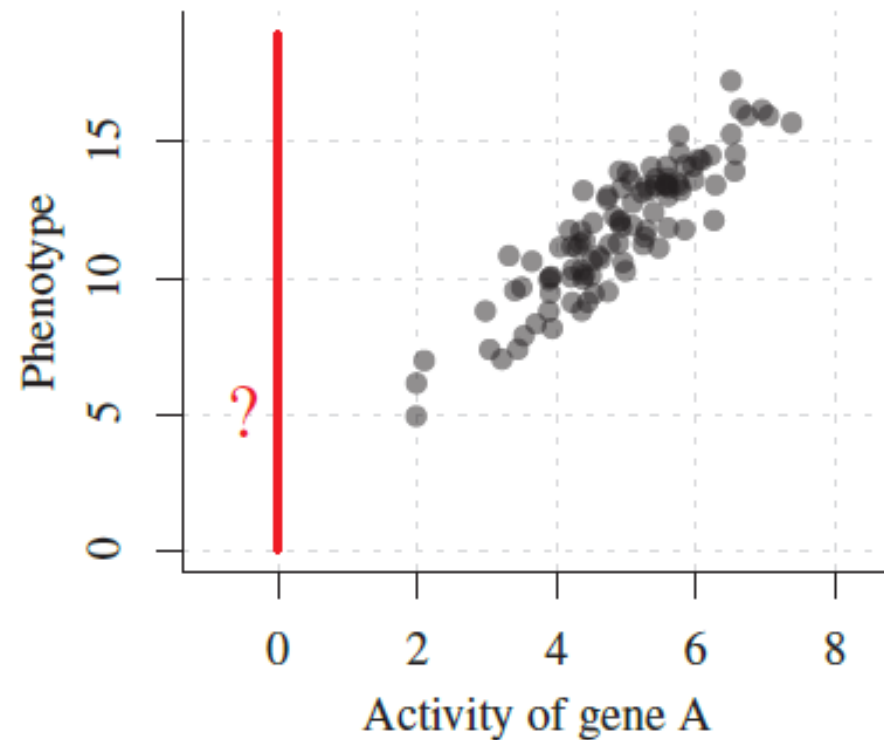
*Definition 1.4 (**Phenotype**): A phenotype is a physical characteristic of an individual such as blue eyes, brown hair, etc.*



Chapter 1 – Two Examples for Causal Modeling

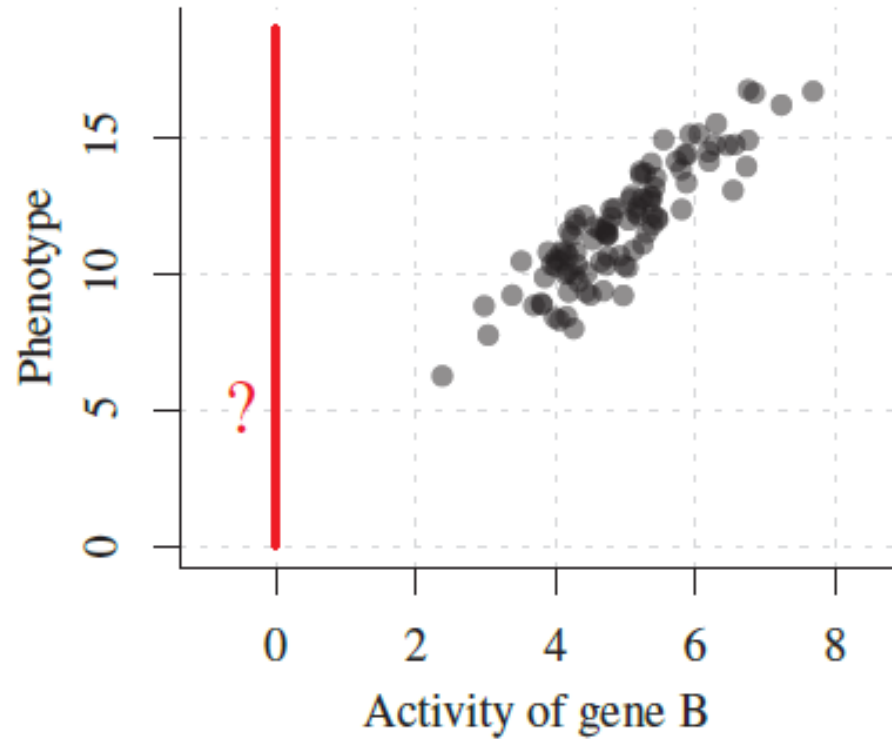
Given: activity data from gene A and measurements of a phenotype.

The diagram shows that clearly there is a correlation between gene A and phenotype.

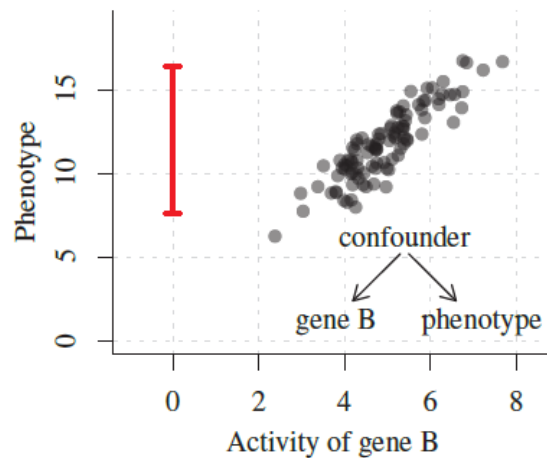
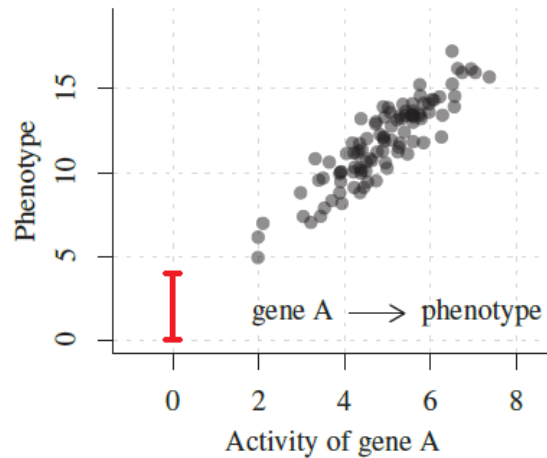


Chapter 1 – Two Examples for Causal Modeling

Similarly we have a correlation between Gene B and the Phenotype



Chapter 1 – Two Examples for Causal Modeling



Possible goal:

Predict phenotype when setting gene A to zero.

We could now want to predict the phenotype when setting gene A to 0.

If we are not willing to employ concepts from causality, we have to answer "I do not know" to the question of predicting a phenotype after deletion of a gene.

Chapter 1 – Two Examples for Causal Modeling

Model	Predict in i.i.d. setting	Predict under changing distr. or intervention	Answer counterfactual questions	Obtain physical insight	Learn from data
Mechanistic/ physical, e.g., Sec. 2.3	yes	yes	yes	yes	?
Structural causal model, e.g., Sec. 6.2	yes	yes	yes	?	?
Causal graphi- cal model, e.g., Sec. 6.5.2	yes	yes	no	?	?
Statistical model, e.g., Sec. 1.2	yes	no	no	no	yes

Chapter 2 – Historical notes

Intellectual antecedent to SEMs:

Sems are nowadays strongly associated with econometrics (Econometrics is the application of statistical methods to economic data in order to give empirical content to economic relationships.) Trygve Haavelmo 1944 laid the conceptual foundations of probabilistic econometrics. I.e. unlike correlation, regression has a natural direction. What direction to perform the regression? Problem of observational equivalence

Connection between: what does make a set of equations or relations structural and properties of invariance and autonomy (Aldrich 1989, Frisch et al 1984):

Here structural relation was aiming to try to capture an underlying structure connecting the variables of the model.

Chapter 2 – Historical notes

Cowels work (endogenous and exogenous variables):

Cowels work distinguished between endogenous and exogenous variables

- endogenous: Can be understood by modeler
- exogenous: Influenced from outside the model (are taken as given)

Koopmans (1950) did research in what should be considered as exogeneous (departmental principle) Example: Weather is exogeneous to economics For causality this means: variables which influence others but are not influenced thereby are called exogenous

Chapter 2 – Historical notes

Causality in economics and econometrics (Hoover 2008):

They discussed a system of the form

$$\begin{aligned} X^i &:= N_X^i \\ Y^i &:= \theta X^i + N_Y^i \end{aligned}$$

Errors N_X^i and N_Y^i are i.i.d. and θ is a parameter. Simon (1953) states that X^i causes Y^i because Y^i knows all about X^i but not vice versa. But Hoover extended this and argued that we cannot infer the correct causal direction on the basis of a single set of data (observational equivalence), but experiments could help us decide. Hoover refers to Simons invariance criterion: The true causal order is the one that is invariant under the right sort of intervention

Chapter 2 – Historical notes

Hurwicz (1962):

Structure is necessary for causality, it is not for prediction

Aldrich 1989):

Argues that autonomous relations are likely to be more stable than others.

Developments in computer science happened separately (Pearl 2009):

They started connecting Bayesian networks and structural equation modeling. For a long time there was a tension between economists and statisticians. The confusion was that statisticians read structural equations as statements about $\mathbb{E}[Y|x]$ while economists read them as $\mathbb{E}[Y|do(x)]$. This might be the reason, why statisticians claim that structural equations have no meaning and economists retort that statistics has no substance. Thus Pearl states that "each parent-child relationship in the network represents an autonomous physical mechanism."

Chapter 2 – Historical notes

Motivation for writing this book:

Most of the work using causal Bayesian networks only exploits the independence of noise terms. This leads to a rich structure of conditional dependences (Coming from Reichenbach's principle 1.1). There is work aiming at that in the caus-effect problem conditional independence is useless since we have only two variables. Janzing and Schölkopf 2010 then formalize independence of mechanism in terms of algorithmic information theory.

Schölkopf et al 2012:

They discuss the question of robustness with respect to changes in the distribution of the cause and connect it to problems in machine learning. They employ a notion of independence of mechanism and input that subsumes both independence under changes and information-theoretic independence.

Physical Structure underlying causal models

The Role of Time

This aspect is missing in section 2.1 of the book. Already Simon (1953) recognized that while time ordering can provide a useful asymmetry, it is asymmetry that is important, not the temporal sequence. It is even believed that both classical systems and quantum mechanical systems are invertible. There is work investigating into this (Bennett 1982, Zurek 1989).

Physical Structure underlying causal models

Physical Laws

As an example we take the gas law:

$$p \cdot V = n \cdot R \cdot T$$

- p = pressure
- V = Volume
- n = Amount of substance
- T = temperature
- R = gas constant

If we change V , p or T will change. If we change T , V or p will change. We can ask the question what causes what?

The gas law refers to an **equilibrium state** of an underlying dynamical system. The equation alone doesn't provide enough information about what interventions in principle are possible and what is their effect. To have this information we need a corresponding acyclic graph representing the causal structure.

Physical Structure underlying causal models

Cyclic assignments

This paragraph looks at cyclic causal models and its justification under assumption of processes taking place in time. Even though time-dependent processes do not have cycles, it is possible that an SCM derived from such a process has cycles. To define interventions becomes a little harder in such systems. But certain types of interventions are still possible. E.g. we can set the value of one variable to a fixed value. This cuts the cycle. It might be impossible to derive an entailed observational distribution:

$$\begin{aligned}X &:= f_X(Y, N_X) \\ Y &:= f_Y(X, N_Y)\end{aligned}$$

We can start with some Y and then substitute the functions into each other back and forth until we converge to some point. This point is then the joint distribution of X and Y . However, such an equilibrium for X, Y need not always exist. (Further details Remark 6.5) If we want to understand general cyclic systems, it may be unavoidable to study systems of differential equations rather than SCMs.

Physical Structure underlying causal models

Feasibility of Interventions

In real systems, there could also be interventions which affect more than one mechanism at a time. If multiple interventions influenced each other, this could be viewed as a violating form of the independence Principle 2.1. We could argue that combined interventions that are "natural" will not violate independence. However, therefore we would have to know the underlying causal structure, which we do not have.

Andreas Kaufmann
Msc Student in Computer Science
ankaufmann@student.ethz.ch

ETH Zurich
Rämistrasse 101
CH-8092 Zürich

www.ethz.ch