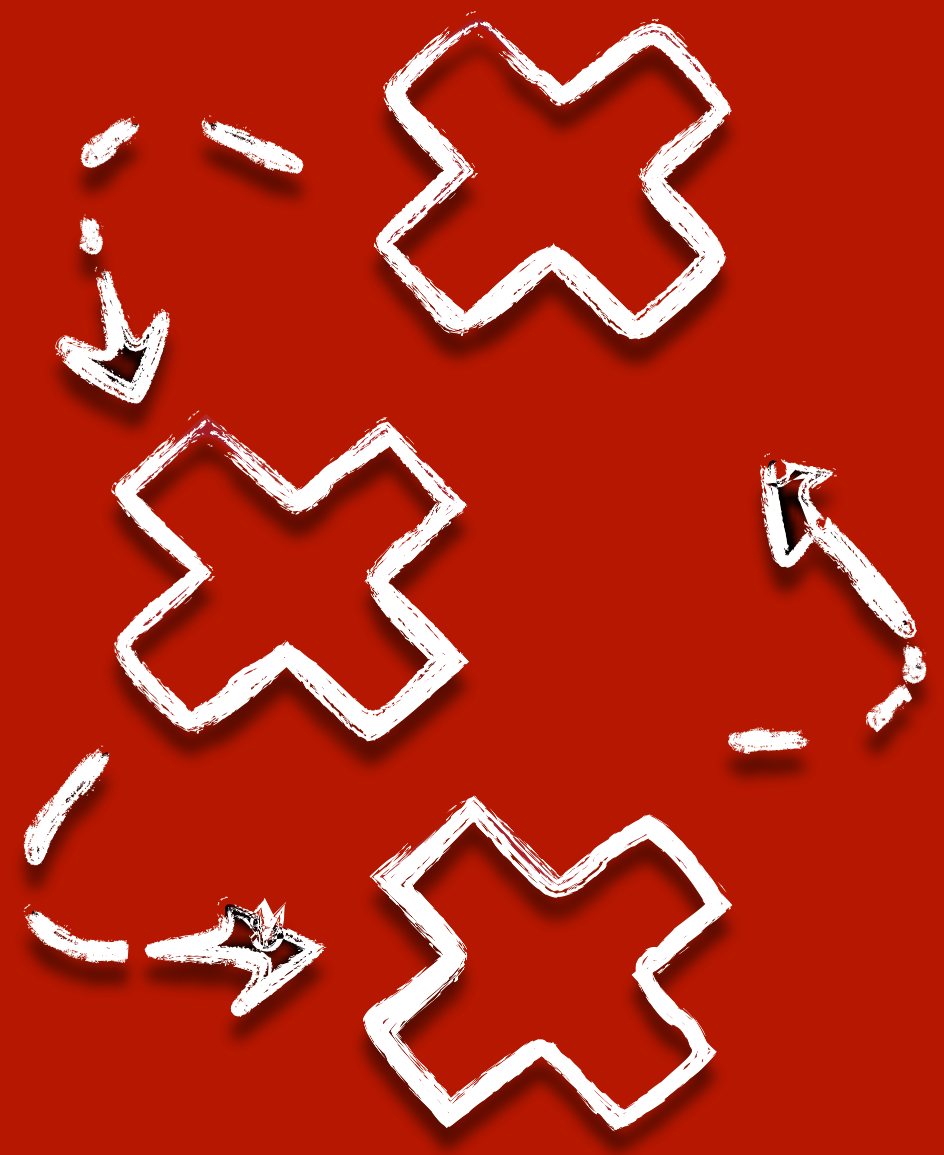


Deep learning 2: Causality & DL

1.3: Causal graphs

Lecturer: Sara Magliacane

UvA - Spring 2022

BNs vs causal BNs - example

- Fire (F) and Alarm (A) with $p(F, A)$ and $A \not\perp\!\!\!\perp F$ can be factorized as:

$$p(F,A) = p(F)\,p(A|F)$$

$$F \rightarrow A$$

CAUSAL

(lighting a fire triggers alarm)

$$p(F,A) = p(A)\,p(F|A)$$

$$F \leftarrow A$$

NOT-CAUSAL

(triggering alarm does not light a fire)

The $do(X = x)$ operator [Pearl 2009]

- We introduce a new operator that can represent a **hypothetical intervention** on the whole population, i.e. a perturbation of the system:

$$\mathrm{do}(X = x)$$

The $do(X = x)$ operator [Pearl 2009]

- We introduce a new operator that can represent a **hypothetical intervention** on the whole population, i.e. a perturbation of the system:

$$\mathrm{do}(X = x)$$

- For a BN $((\mathbf{V}, \mathbf{E}), p)$ and intervention on $i \in \mathbf{V}$, we define the **interventional distribution:**

$$p(X_{\mathbf{V}} \,|\, \mathrm{do}(X_i = x_i)), \text{ which in general } \neq p(X_{\mathbf{V}})$$

The $do(X = x)$ operator [Pearl 2009]

- We introduce a new operator that can represent a **hypothetical intervention** on the whole population, i.e. a perturbation of the system:

$$\mathrm{do}(X = x)$$

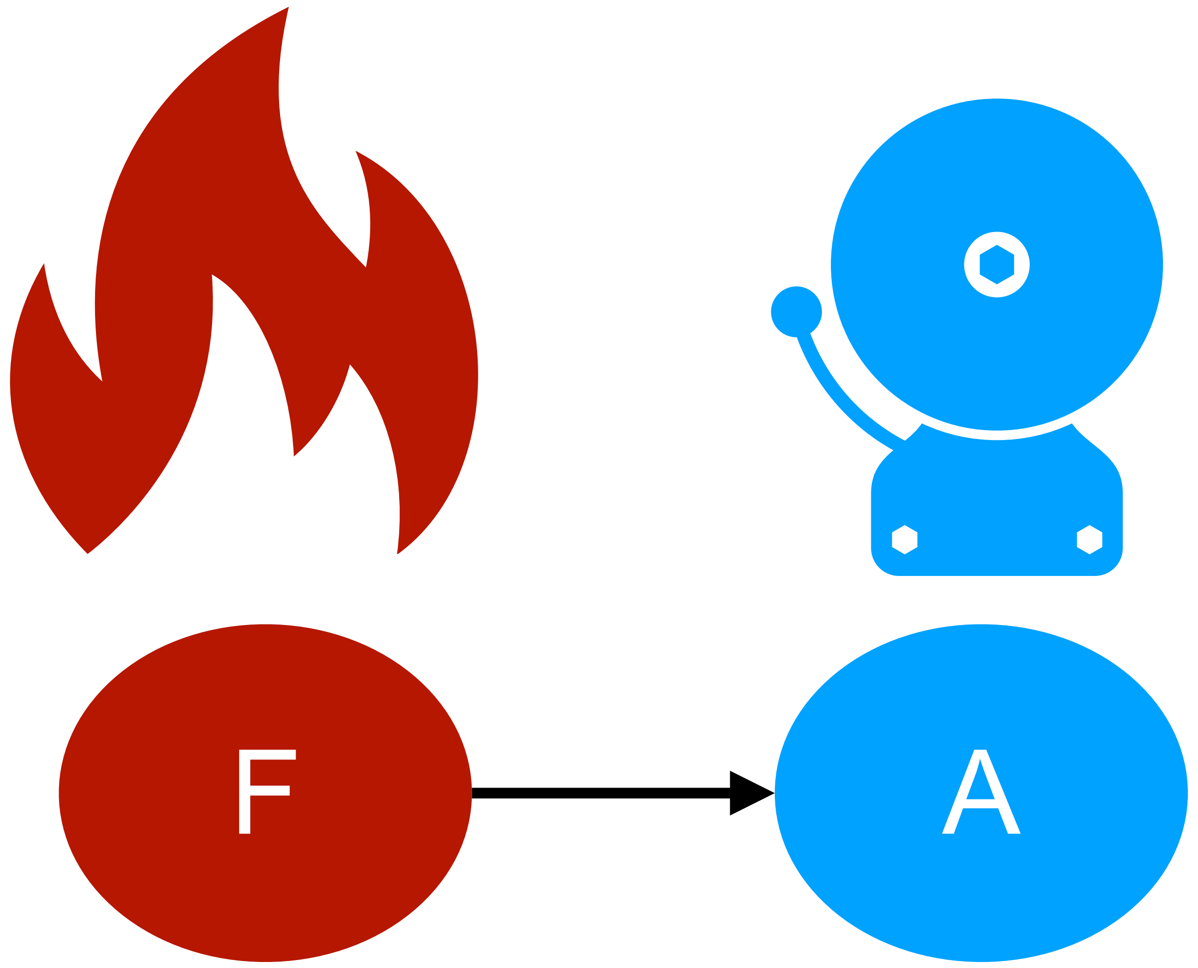
- For a BN $((\mathbf{V}, \mathbf{E}), p)$ and intervention on $i \in \mathbf{V}$, we define the **interventional distribution:**

$$p(X_{\mathbf{V}} \mid \mathrm{do}(X_i = x_i)), \text{ which in general } \neq p(X_{\mathbf{V}})$$

- We can also define conditional and marginal versions:

$$p(X_j \mid \mathrm{do}(X_i = x_i)), \text{ which in general } \neq p(X_j)$$

Seeing is not doing



Causal Bayesian networks

- Given DAG $G = (\mathbf{V}, \mathbf{E})$ and distribution p, (G, p) is a Bayesian network if

$$p(X_1, \ldots, X_p) = \prod_{i \in \mathbf{V}} p(X_i \,|\, \mathbf{X}_{\mathrm{Pa}_G(i)})$$

- If for any $\mathbf{W} \subset \mathbf{V}$:

$$p(X_{\mathbf{V}} \,|\, \mathrm{do}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})) = \begin{cases} 0 \quad \text{if} \quad X_{\mathbf{W}} \neq \tilde{x}_{\mathbf{W}} \\ \underbrace{\prod_{i \in \mathbf{V} \setminus \mathbf{W}} p(X_i \,|\, X_{\mathrm{Pa}_G(i)})}_{\text{doesn't change}} \quad \underbrace{\text{if} \quad X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}}}_{\text{consistent with intervention}} \end{cases}$$

then (G, p) is a **causal Bayesian network**

Causal Bayesian networks

- Given DAG $G = (\mathbf{V}, \mathbf{E})$ and distribution p, (G, p) is a Bayesian network if

$$p(X_1, \ldots, X_p) = \prod_{i \in \mathbf{V}} p(X_i \,|\, \mathbf{X}_{\mathrm{Pa}_G(i)})$$

- If for any $\mathbf{W} \subset \mathbf{V}$:

$$p(X_\mathbf{V} \,|\, \mathrm{do}(X_\mathbf{W} = \tilde{x}_\mathbf{W})) = \prod_{i \in \mathbf{V} \setminus \mathbf{W}} p(X_i \,|\, X_{\mathrm{Pa}_G(i)}) \cdot \underbrace{\mathbf{1}(X_\mathbf{W} = \tilde{x}_\mathbf{W})}_{\text{indicator function}}$$

(G, p) is a **causal Bayesian network**

Parents in G are now direct causes

Truncated factorisation formula [Pearl 2009]

- If for any $\mathbf{W} \subset \mathbf{V}$:

$$p(X_{\mathbf{V}} \mid \mathrm{do}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})) = \underbrace{\prod_{i \in \mathbf{V} \setminus \mathbf{W}} p(X_i \mid X_{\mathrm{Pa}_G(i)})}_{\text{doesn't change}} \cdot \underbrace{\mathbf{1}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})}_{\text{consistent with intervention}}$$

- Includes also observational data $\mathbf{W} = \emptyset$,
- Includes also multiple **intervention targets** $|\mathbf{W}| \geq 1$

BNs vs causal BNs - example

- Fire (F) and Alarm (A) with $p(F, A)$ and $A \not\perp\!\!\!\perp F$ can be factorized as:

$$p(F, A) = p(F)\, p(A \mid F)$$

$$F \rightarrow A$$

do(A=1):

$$p(F, A \mid do(A=1)) = p(F) \cdot \mathbb{1}(A=1)$$

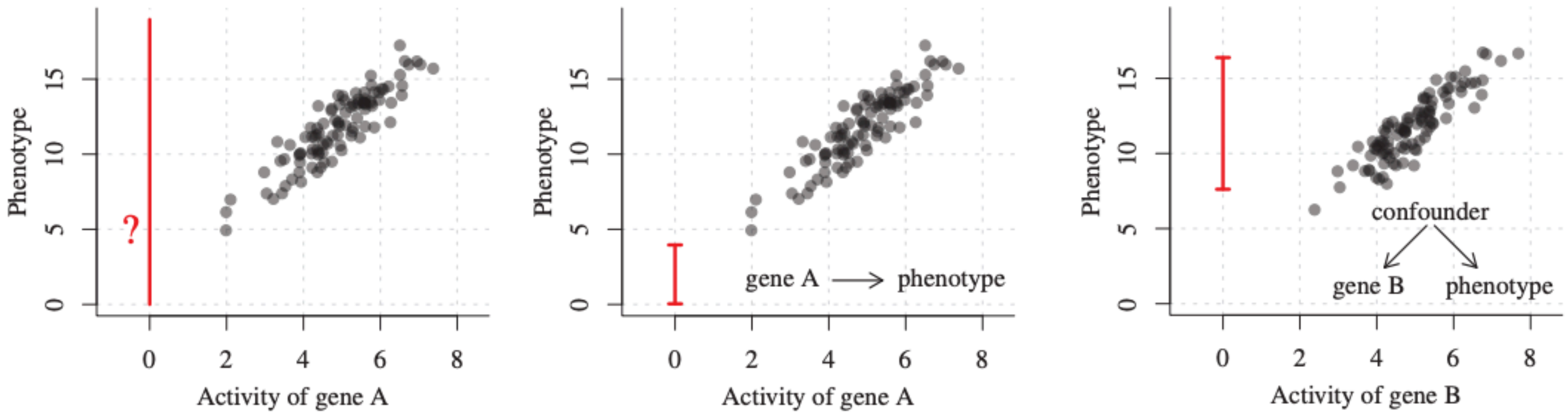
$$p(F, A) = p(A)\, p(F \mid A)$$

$$F \leftarrow A$$

do(A=1):

$$p(F, A \mid do(A=1)) = \mathbb{1}(A=1) \cdot p(F \mid A)$$

Another example of causal effect vs no effect for $\mathrm{do}(X = 0)$



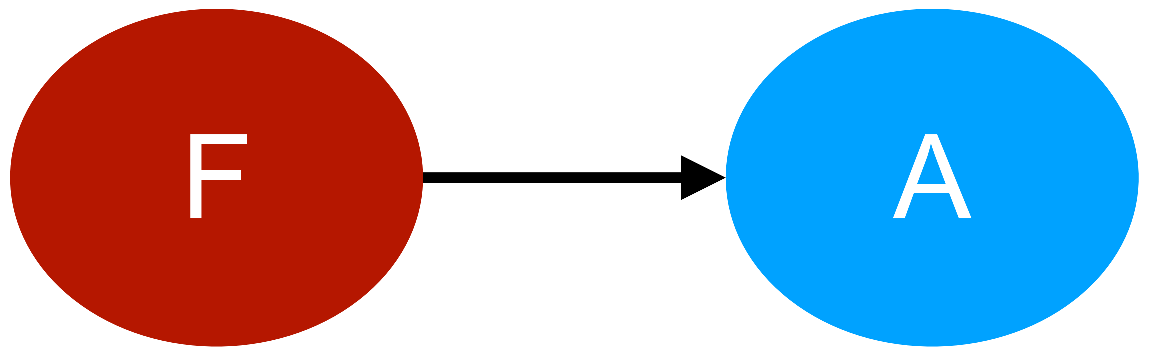
$\mathrm{do}(A = 0)$

$\mathrm{do}(B = 0)$

Fig 1.4 in Elements of Causal Inference (http://web.math.ku.dk/~peters/jonas_files/ElementsOfCausalInference.pdf)

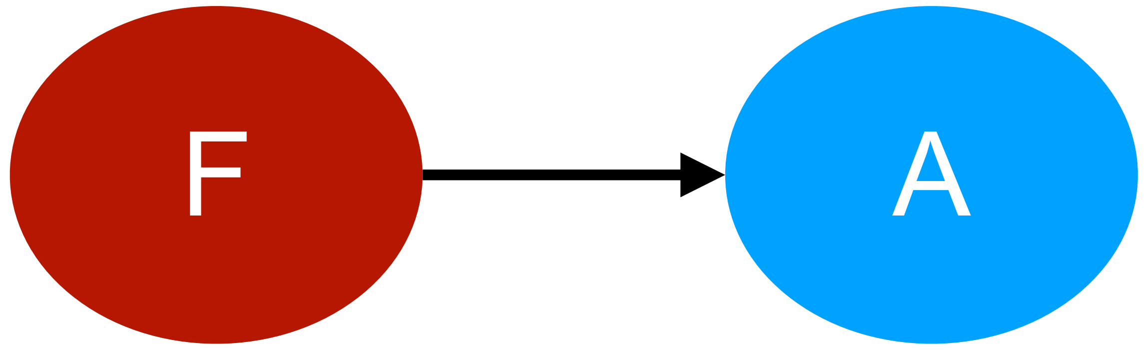
Mutilated/manipulated graphs

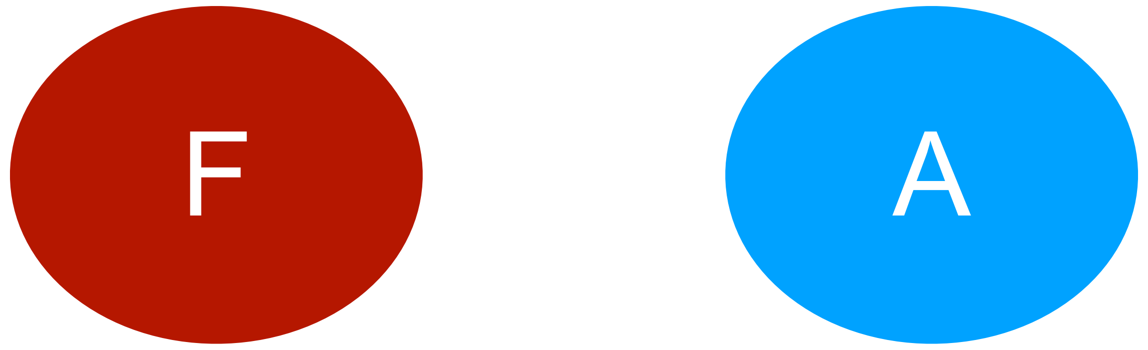
Graphically $p(X_{\mathbf{V}} \mid \mathrm{do}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})) = \prod_{i \in \mathbf{V} \setminus \mathbf{W}} p(X_i \mid X_{\mathrm{Pa}_G(i)}) \cdot \mathbf{1}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})$ can be represented as cutting the incoming edges to $X_{\mathbf{W}}$



Mutilated/manipulated graphs

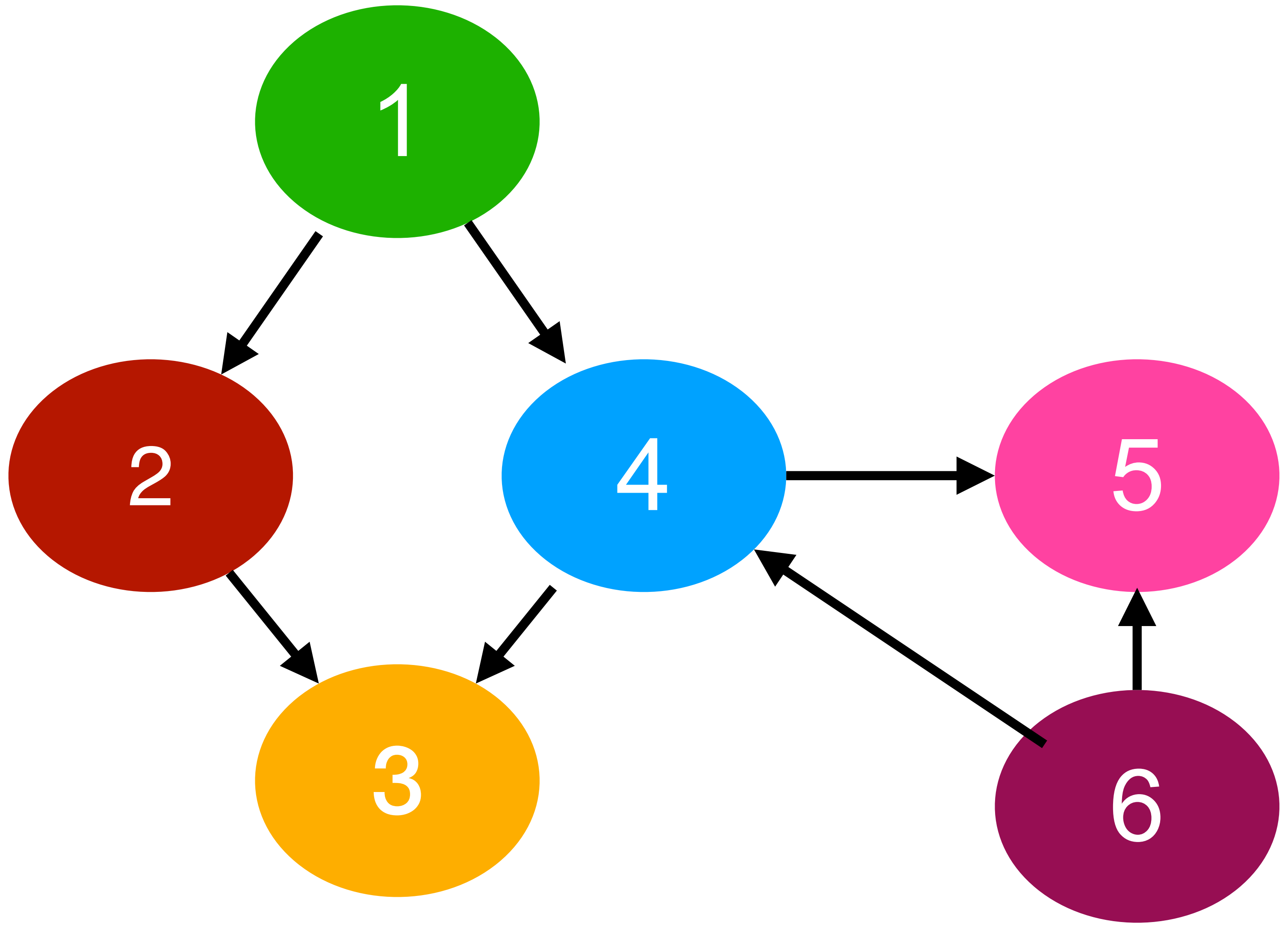
Graphically $p(X_{\mathbf{V}} \mid \mathrm{do}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})) = \prod_{i \in \mathbf{V} \setminus \mathbf{W}} p(X_i \mid X_{\mathrm{Pa}_{\mathrm{G}}(i)}) \cdot \mathbf{1}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})$ can be represented as cutting the incoming edges to $X_{\mathbf{W}}$



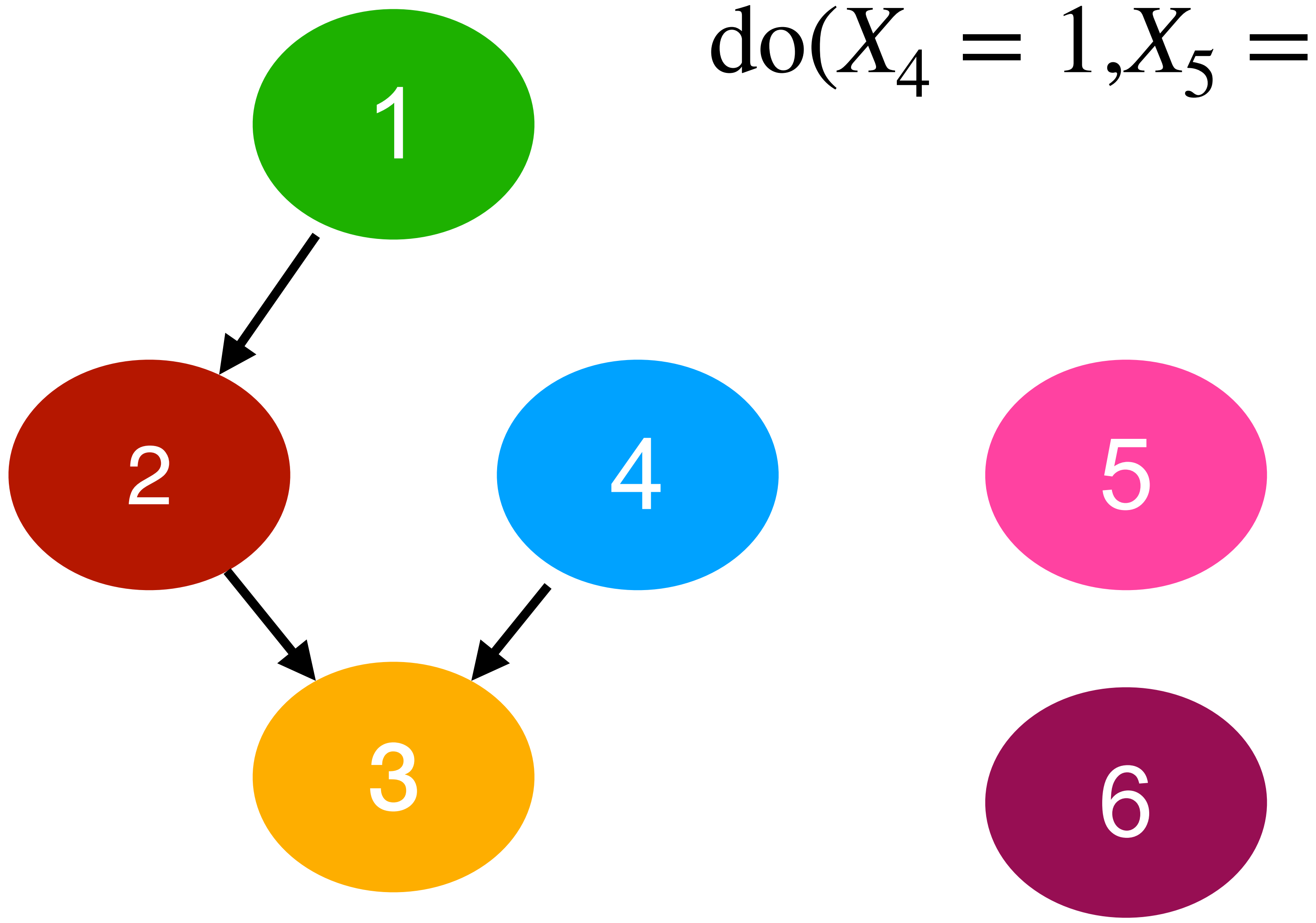


Mutilated/manipulated graphs

Graphically $p(X_{\mathbf{V}} \mid \mathrm{do}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})) = \prod_{i \in \mathbf{V} \setminus \mathbf{W}} p(X_i \mid X_{\mathrm{Pa}_G(i)}) \cdot \mathbf{1}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})$ can be represented as cutting the incoming edges to $X_{\mathbf{W}}$



$\mathrm{do}(X_4 = 1, X_5 = 0)$



Perfect vs soft interventions

- We introduce a new operator that can represent a **hypothetical intervention** on the whole population, i.e. a perturbation of the system:

$$\mathrm{do}(X_i = x_i) \text{ which changes } p(X_i \mid X_{\mathrm{Pa}(i)}) \rightarrow \mathbf{1}(X_i = x_i)$$

- This is called a **perfect** (or surgical) **intervention**

- There are also other types of intervention, e.g. **soft interventions which change** $p(X_i \mid X_{\mathrm{Pa}(i)}) \rightarrow \tilde{p}(X_i \mid X_{\mathrm{Pa}(i)})$

Truncated factorisation formula [Pearl 2009]

- If for any $\mathbf{W} \subset \mathbf{V}$:

$$p(X_{\mathbf{V}} \,|\, \mathrm{do}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})) = \underbrace{\prod_{i \in \mathbf{V} \setminus \mathbf{W}} p(X_i \,|\, X_{\mathrm{Pa}_G(i)})}_{\text{doesn't change from the observational distr.}} \cdot \mathbf{1}(X_{\mathbf{W}} = \tilde{x}_{\mathbf{W}})$$

MODULARITY ASSUMPTION

Causal mechanisms and Modularity

- In a causal BN (G, p), each $p(X_i \mid \mathbf{X}_{\mathrm{Pa}(i)})$ is the **causal mechanism of X_i**
- **Modularity assumption:** intervening on X_j will not change any causal mechanism $p(X_i \mid \mathbf{X}_{\mathrm{Pa}(i)})$ for any $i \neq j$

Causal mechanisms and Modularity

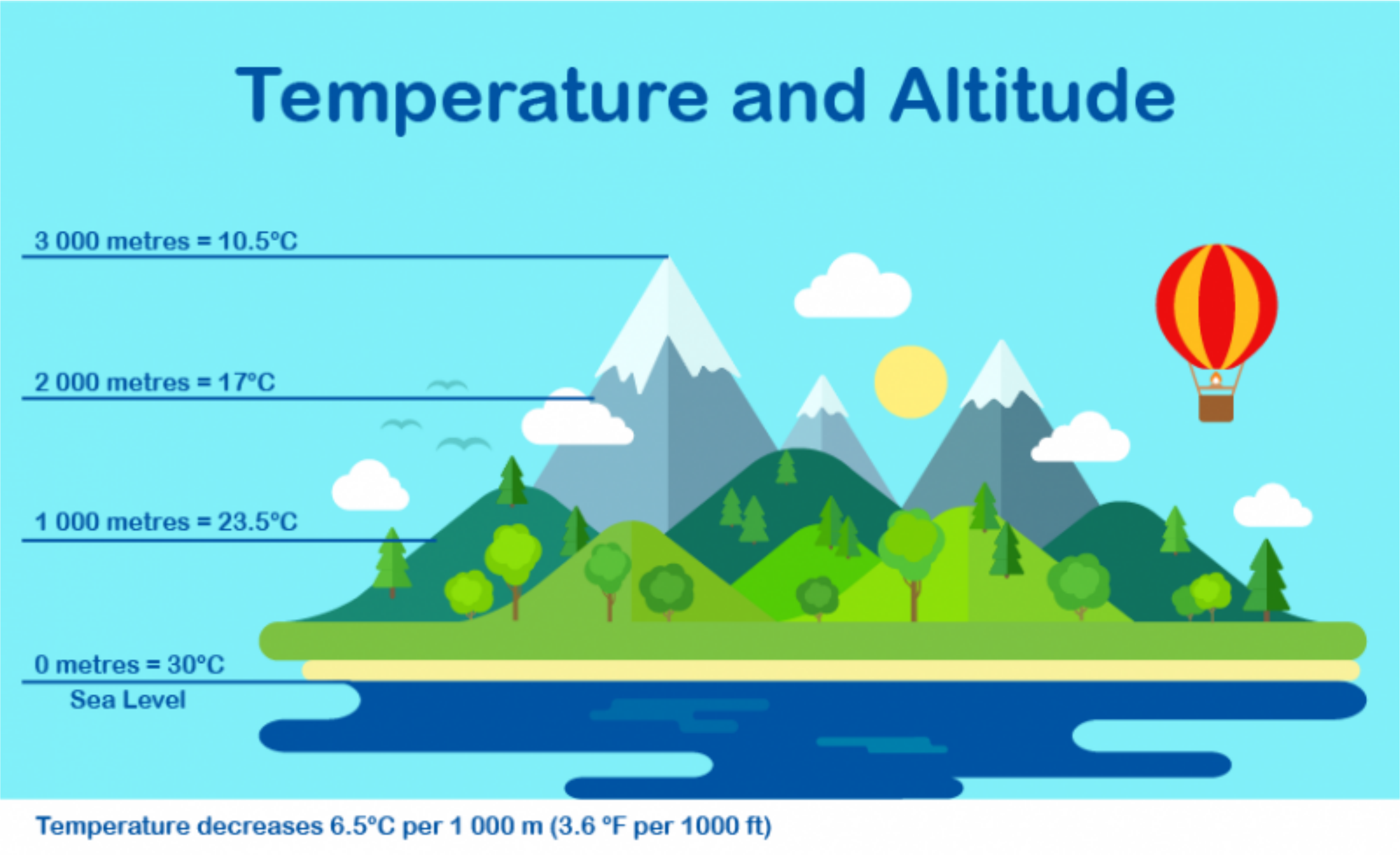
- In a causal BN (G, p), each $p(X_i \mid \mathbf{X}_{\mathrm{Pa}(i)})$ is the **causal mechanism of X_i**
- **Modularity assumption:** intervening on X_j will not change any causal mechanism $p(X_i \mid \mathbf{X}_{\mathrm{Pa}(i)})$ for any $i \neq j$
- **Independent Causal Mechanism Principle:** the generative process is composed of **autonomous models** that do **not inform or influence** each other

Knowing $p(X_j \mid \mathbf{X}_{\mathrm{Pa}(j)})$ Does not give info $p(X_i \mid \mathbf{X}_{\mathrm{Pa}(i)})$ $i \neq j$

Changing **Does not change**

Section 2.1 in Elements of Causal Inference (http://web.math.ku.dk/~peters/jonas_files/ElementsOfCausalInference.pdf)

Modularity example

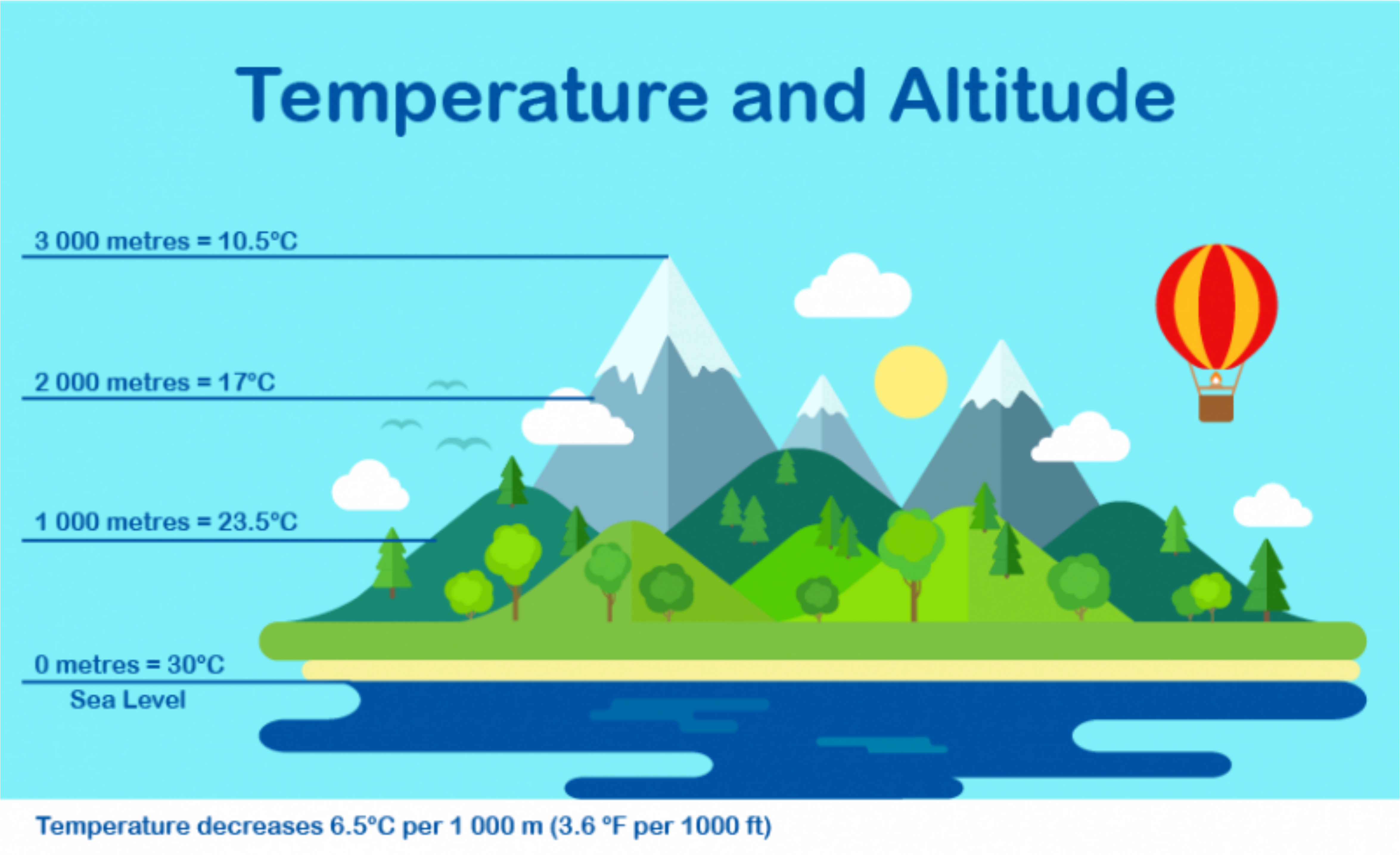


$$P(A, T) = P(T \mid A)P(A)$$

$$P(A, T) = P(A \mid T)P(T)$$

Example from Section 2.1 in Elements of Causal Inference (http://web.math.ku.dk/~peters/jonas_files/ElementsOfCausalInference.pdf)Image from https://lisbdnet.com/how-altitude-affects-climate/

Modularity example



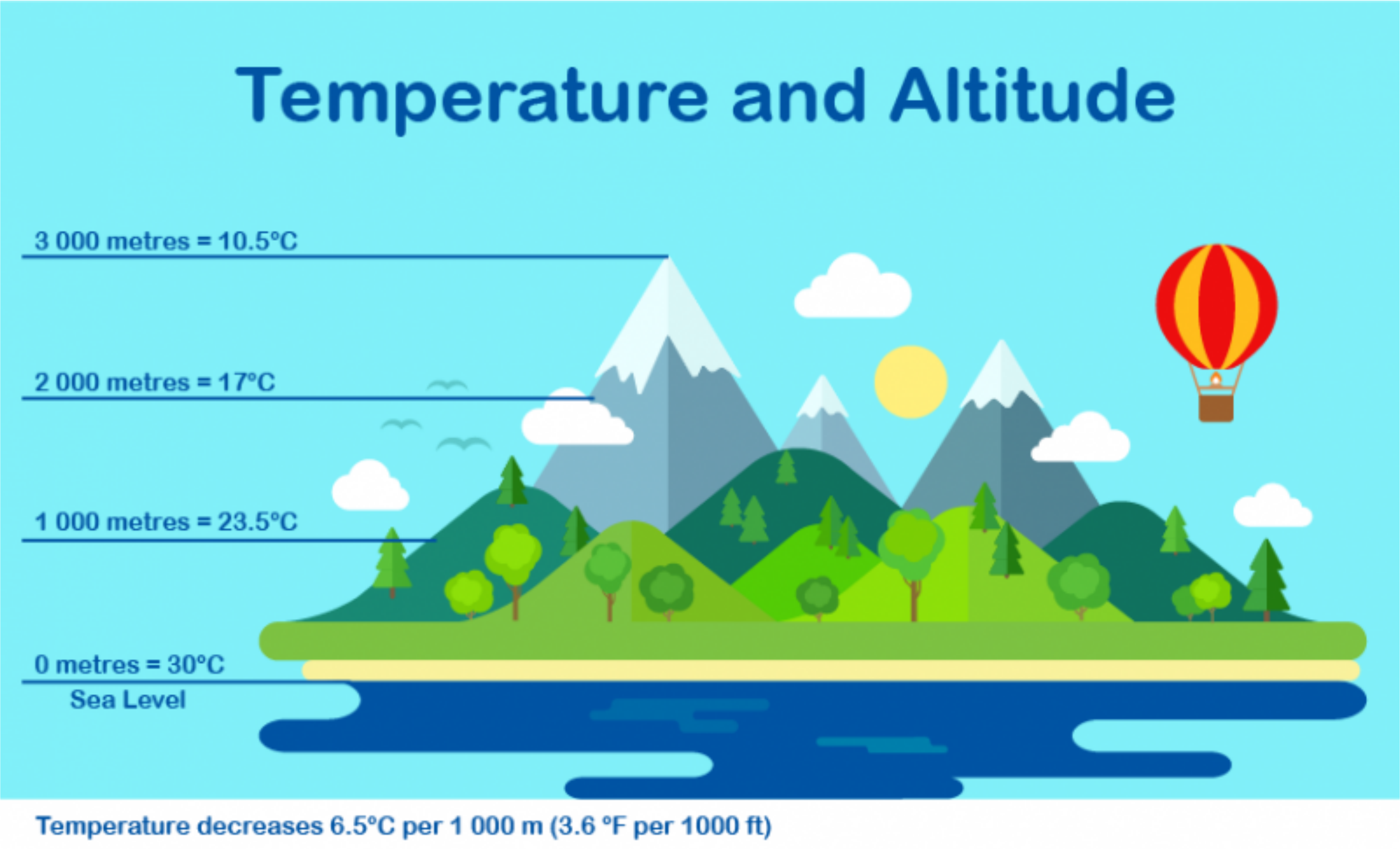
$$P(A, T) = P(T|A)P(A)$$

changing P(A) does not change P(T|A)

$$P(A, T) = P(A|T)P(T)$$

Changing P(T) might change P(A|T)

Modularity example



P(T|A) is an invariant physical mechanism

$$P(A, T) = P(T|A)P(A)$$

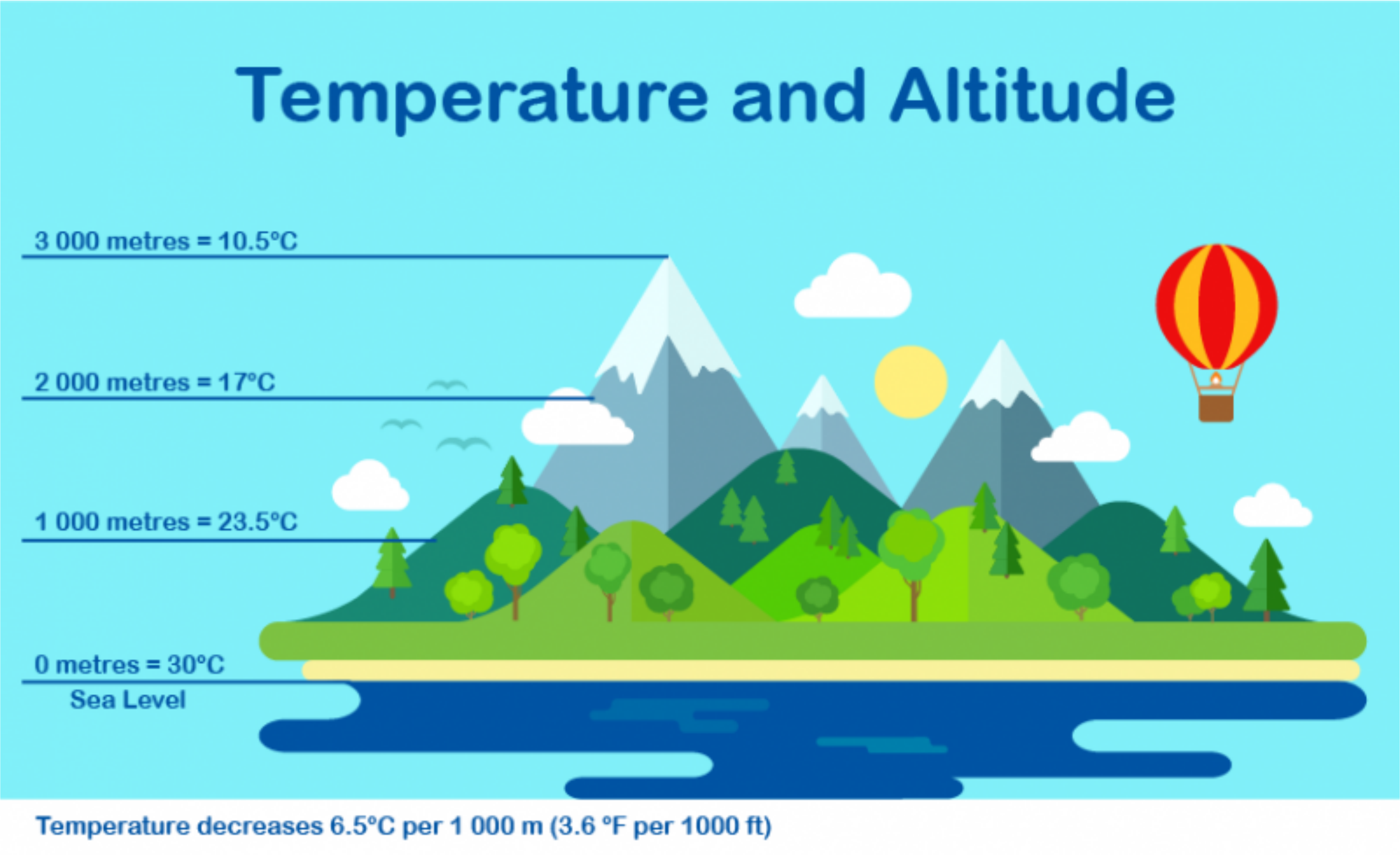
changing P(A) does not change P(T|A)

$$P(A, T) = P(A|T)P(T)$$

Changing P(T) might change P(A|T)

Example from Section 2.1 in Elements of Causal Inference (http://web.math.ku.dk/~peters/jonas_files/ElementsOfCausalInference.pdf)Image from https://lisbdnet.com/how-altitude-affects-climate/

Modularity example



$$P(A, T) = P(T|A)P(A)$$

changing P(A) does not change P(T|A)

The **causal factorisation** allows for **localised/sparse** interventions

Example from Section 2.1 in Elements of Causal Inference (http://web.math.ku.dk/~peters/jonas_files/ElementsOfCausalInference.pdf)Image from https://lisbdnet.com/how-altitude-affects-climate/

Structural **causal** models (SCMs)

- Let (G, p) be a **causal** Bayesian network

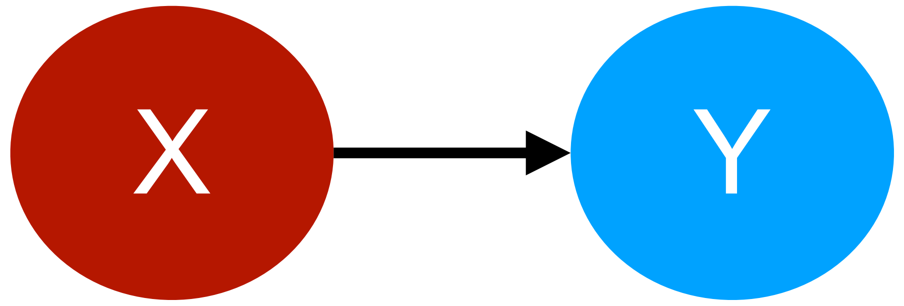
- We can write each variable X_i for $i \in \mathbf{V}$ as a **function of its parents** in G and an additional **noise term** ϵ_i in a **structural equation:**

$$X_i \leftarrow h_i(X_{\mathrm{Pa}(i)}, \epsilon_i)$$

often linear often Gaussian

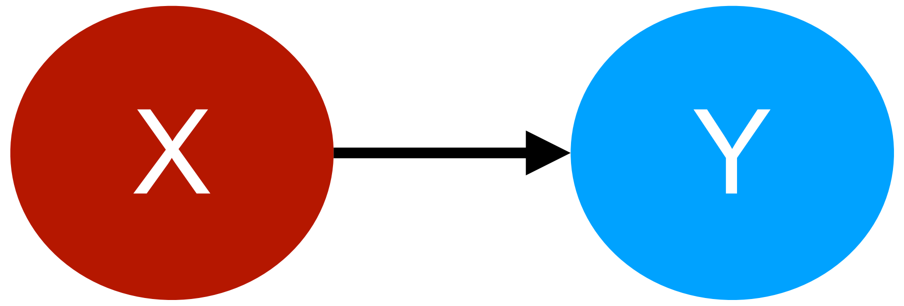
- We often assume noises are **independent of each other** $\forall i \neq j : \epsilon_i \perp\!\!\!\perp \epsilon_j$

SCM: Example 3.2 in Elements of Causal Inference



$$P(X) = \mathcal{N}(0,1)$$
$$P(Y \mid X = x) = \mathcal{N}(4 \cdot x, 1)$$

SCM: Example 3.2 in Elements of Causal Inference

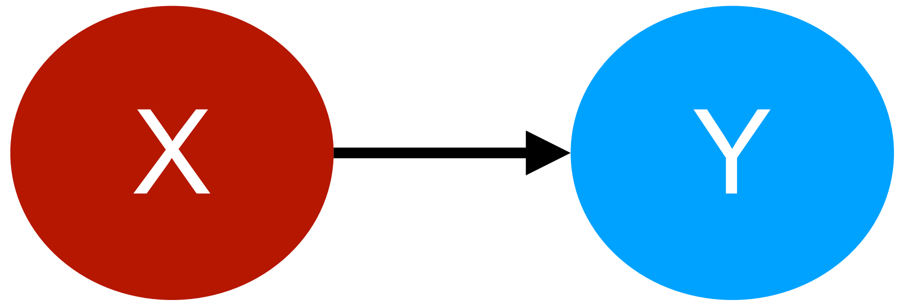


$$P(X) = \mathcal{N}(0,1)$$
$$P(Y|X = x) = \mathcal{N}(4 \cdot x, 1)$$

$$\begin{cases} X \leftarrow \epsilon_X \\ Y \leftarrow 4 \cdot X + \epsilon_Y \end{cases}$$

$$\epsilon_X, \epsilon_Y \sim \mathcal{N}(0,1)$$

SCM: Example 3.2 in Elements of Causal Inference



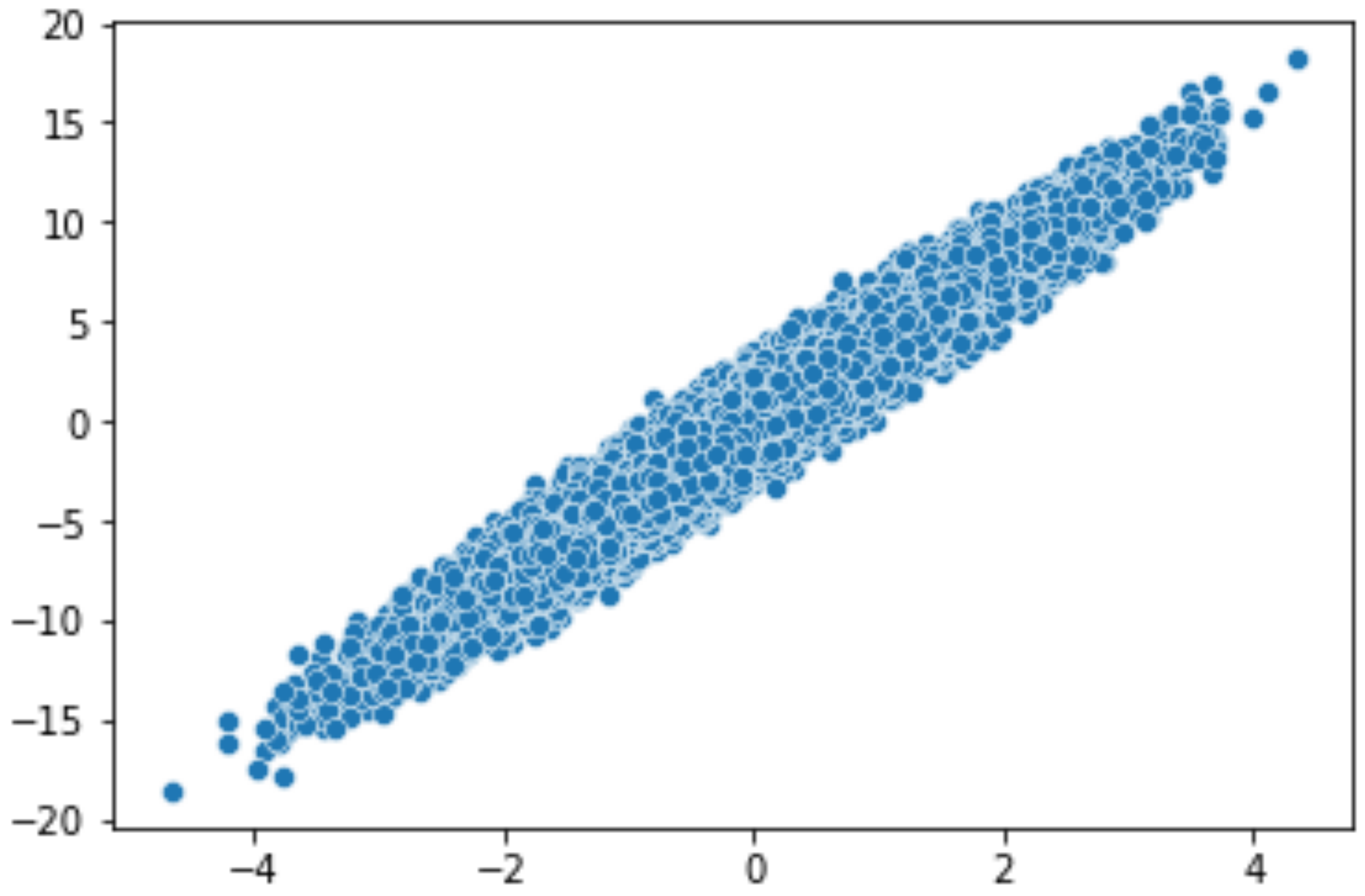
$$P(X) = \mathcal{N}(0,1)$$
$$P(Y \mid X = x) = \mathcal{N}(4 \cdot x, 1)$$

$$\begin{cases} X \leftarrow \epsilon_X \\ Y \leftarrow 4 \cdot X + \epsilon_Y \end{cases}$$

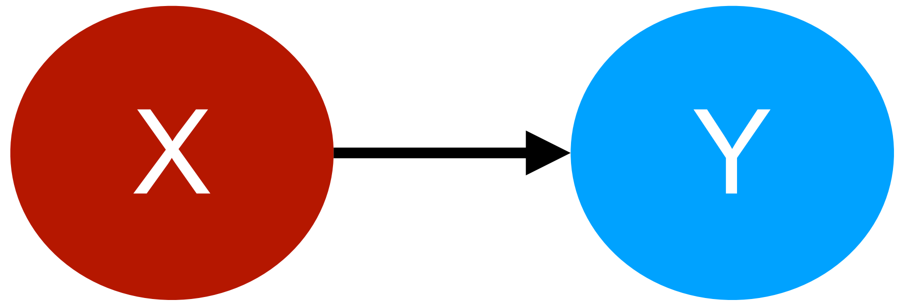
$$\epsilon_X, \epsilon_Y \sim \mathcal{N}(0,1)$$

```
x = randn(n_samples)
y = 4 * x + randn(n_samples)
# plot P(X,Y)
sns.scatterplot(x=x,y=y)
```

```
<AxesSubplot:>
```



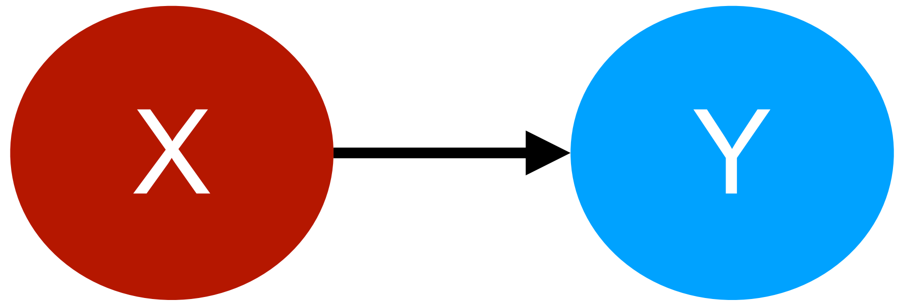
SCM: Example 3.2 in Elements of Causal Inference



$$P(X) = \mathcal{N}(0,1)$$
$$P(Y \mid X = x) = \mathcal{N}(4 \cdot x, 1)$$

$$\begin{cases} X \leftarrow \epsilon_X \\ Y \leftarrow 4 \cdot X + \epsilon_Y \end{cases}$$

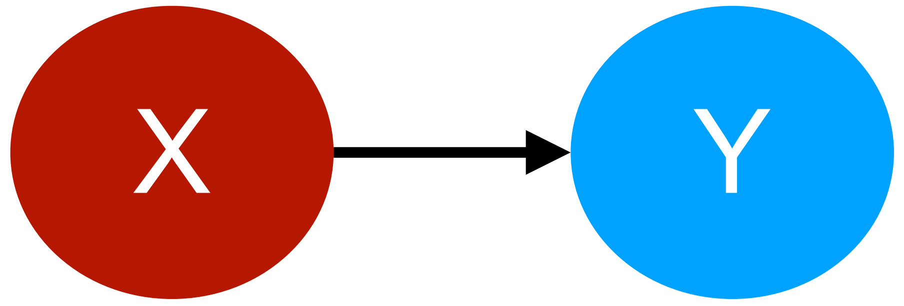
$$\epsilon_X, \epsilon_Y \sim \mathcal{N}(0,1)$$



$$\mathrm{do}(X = 2):$$
$$\begin{cases} X \leftarrow 2 \\ Y \leftarrow 4 \cdot X + \epsilon_Y \end{cases}$$

$$\epsilon_X, \epsilon_Y \sim \mathcal{N}(0,1)$$

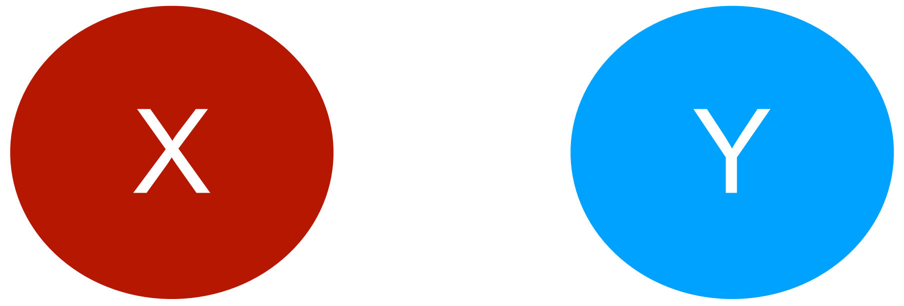
SCM: Example 3.2 in Elements of Causal Inference



$$P(X) = \mathcal{N}(0,1)$$
$$P(Y \mid X = x) = \mathcal{N}(4 \cdot x, 1)$$

$$\begin{cases} X \leftarrow \epsilon_X \\ Y \leftarrow 4 \cdot X + \epsilon_Y \end{cases}$$

$$\epsilon_X, \epsilon_Y \sim \mathcal{N}(0,1)$$

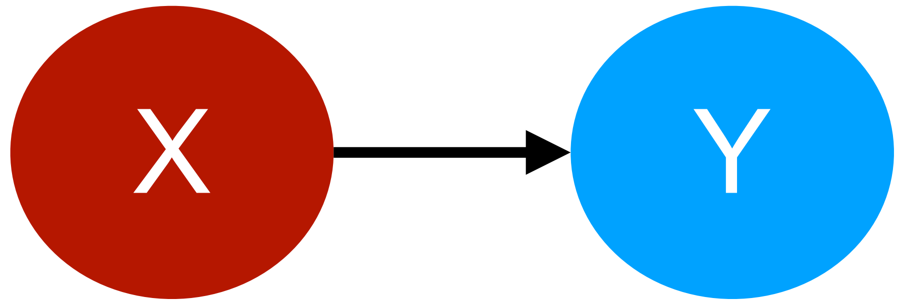


$$\mathrm{do}(Y = 4):$$

$$\begin{cases} X \leftarrow \epsilon_x \\ Y \leftarrow 4 \end{cases}$$

$$\epsilon_X, \epsilon_Y \sim \mathcal{N}(0,1)$$

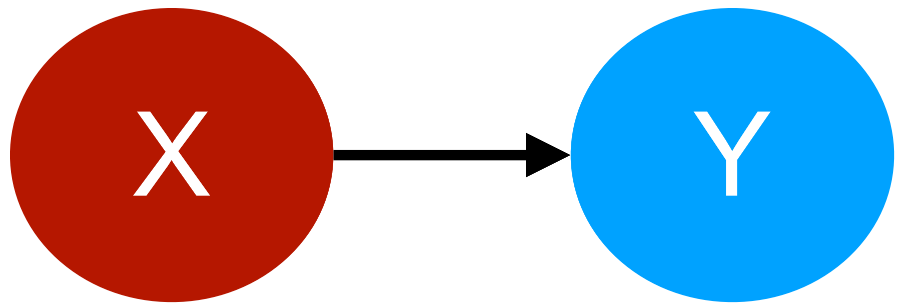
Soft interventions example



$$P(X) = \mathcal{N}(0,1)$$
$$P(Y|X=x) = \mathcal{N}(4 \cdot x, 1)$$

$$\begin{cases} X \leftarrow \epsilon_X \\ Y \leftarrow 4 \cdot X + \epsilon_Y \end{cases}$$

$$\epsilon_X, \epsilon_Y \sim \mathcal{N}(0,1)$$



$$\text{soft intervention on Y :}$$
$$\begin{cases} X \leftarrow \epsilon_X \\ Y \leftarrow 3 \cdot X + 5 \cdot \epsilon_Y \end{cases}$$

$$\epsilon_X, \epsilon_Y \sim \mathcal{N}(0,1)$$

Not in this module: Identification of causal effects

- Given a causal graph G, an **identification strategy** is a formula to estimate an interventional distribution from a combination of observational ones

- **Backdoor criterion, Adjustment criterion**

$$p(x_j \mid \mathrm{do}(x_i)) = \int_{x_{\mathbf{Z}}} p(x_j \mid x_i, x_{\mathbf{Z}}) p(x_{\mathbf{Z}}) dx_{\mathbf{Z}}$$

  - $\mathbf{Z}$ does not contain any descendant of **nodes $r \neq i$ on a directed path from i to j, AND** vs Backdoor: $Z \cap Des(i) = \emptyset$
  - $\mathbf{Z}$ blocks all paths from i to j that are **not directed paths from i to j** vs Backdoor: all backdoor paths

- **Frontdoor criterion**

$$p(x_j \mid \mathrm{do}(x_i')) = \int_{x_{\mathbf{M}}} p(x_{\mathbf{M}} \mid x_i') \int_{x_i} p(x_j \mid x_{\mathbf{M}}, x_i') p(x_i) dx_i$$

  - $\mathbf{M}$ blocks all directed paths from i to j, AND
  - There are **no unblocked backdoor paths** from $i \leftarrow \ldots$ to $\mathbf{M}$, AND
  - $\mathbf{Z} = \{i\}$ **blocks all backdoor paths** from $\mathbf{M} \leftarrow \ldots$ to j

- **Instrumental variables**

  We can exploit the **instrumental variable (IV)** I $\beta = \frac{Cov(I,Y)}{Cov(I,X)}$
  - $I \rightarrow X$, but $I \nrightarrow Y$ directly, $I \perp\!\!\!\perp W$

- **Do-calculus (complete)**