Relationship between causality & machine learning

Speaker: Ziyuan Ye

2022/5/11

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612-634.
Xia, K., Lee, K. Z., Bengio, Y., & Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. Advances in Neural Information Processing Systems, 34.

Structural causal models (SCMs)

- Why we need causal models?
- How to build causal models?
- Causal representation learning

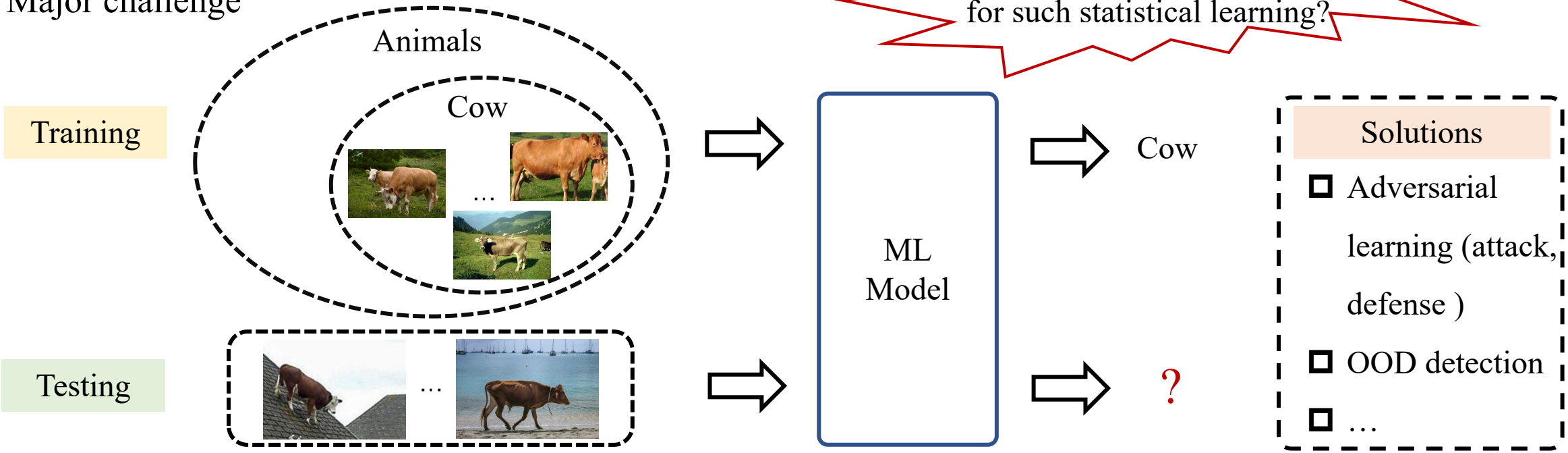
Structural causal models (SCMs)

- Why we need causal models?
- How to build causal models?
- Causal representation learning

Characteristics of current ML models

- Multiple trends at current machine learning:
  - We have amounts of data, often from simulations or large scale human labeling
  - We use high capacity machine learning systems (complex function classes with many adjustable parameters)
  - We employ high performance computer systems
  - The problems are independent and identically distributed (i.i.d)
- Major challenge

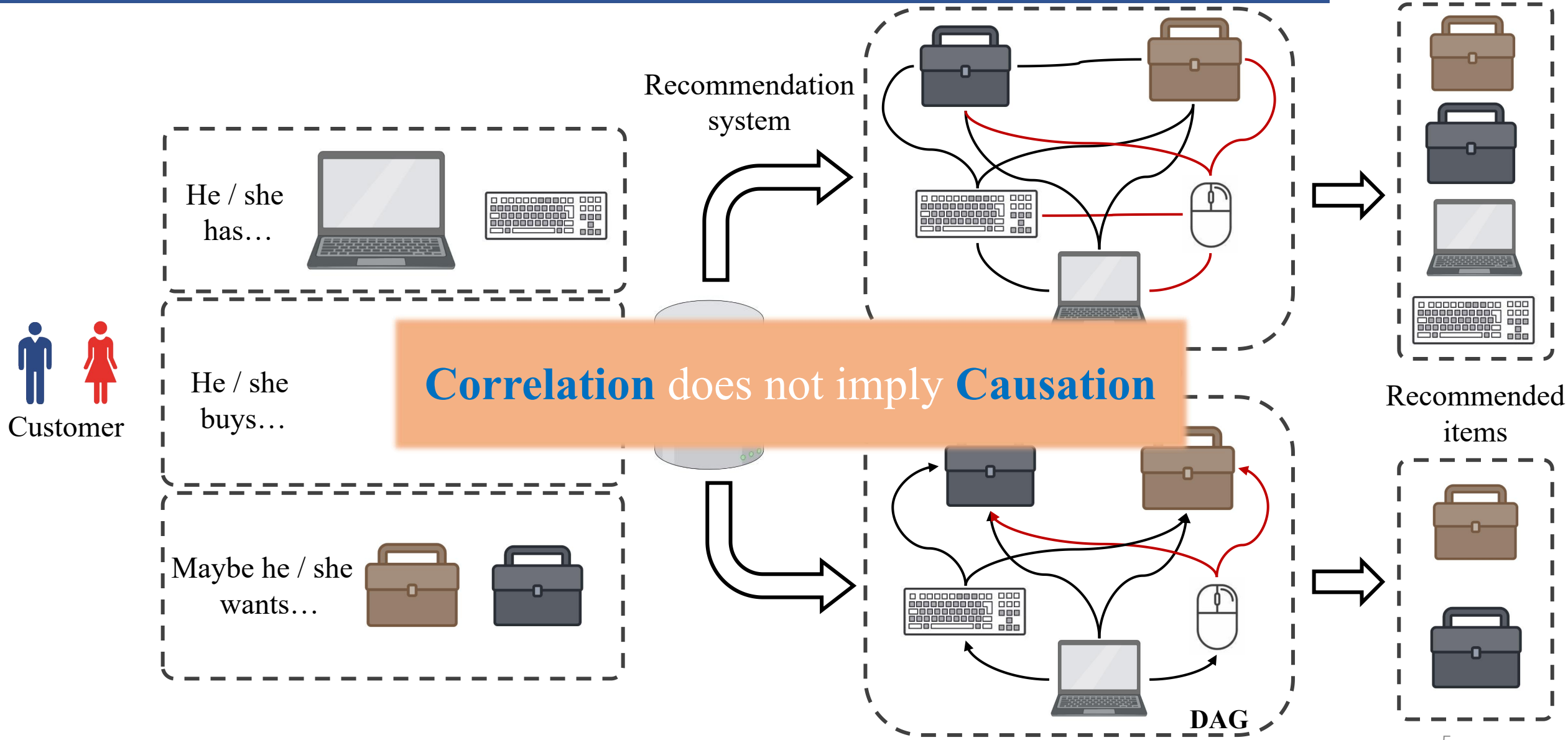
Is the i.i.d assumption reasonable for such statistical learning?



Much of the practice and most theoretical results fail to tackle the hard open problem of generalization across problem.

ML: machine learning; **OOD**: Out of distribution.

From statistical to causal models: A case study

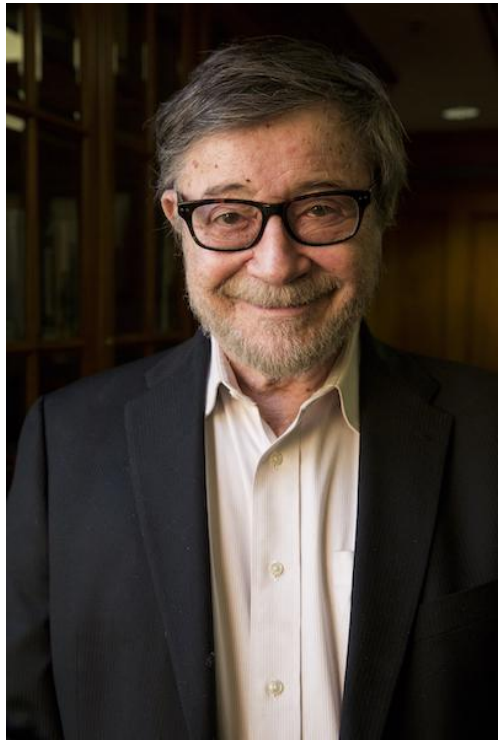


DAG: directed acyclic graphs

Structural causal models (SCMs)

- Why we need causal models?
- How to build causal models?
- Causal representation learning

How to build an intelligent machines?

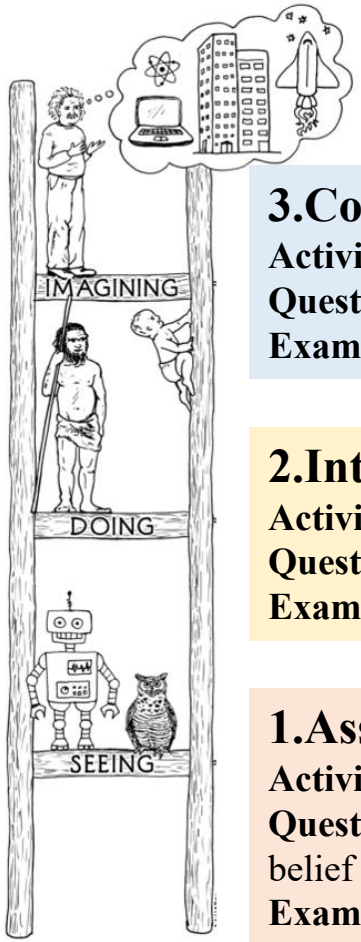


Judea Pearl

- Cognitive Systems Lab
- CS Department, UCLA
- Turing Award Winner

To build truly intelligent machines, teach them cause and effect.

——Judea Pearl



3-level hierarchy of causality

3.Counterfactuals

Activity: Imaging, Retrospection, Understanding
Questions: What if I had done...? Why? (Was it *X* that caused *Y*? What if *X* had not occurred?)
Examples: Was it the aspirin that stopped my headache? What if I had not smoked last year?

2.Intervention

Activity: Doing Intervening
Questions: What if I do…? How? (What would *Y* be if I do *X* ? How can I make Y happen?)
Examples: Was it the aspirin that stopped my headache? What if I had not smoked last year?

1.Association

Activity: Seeing, Observing
Questions: What if I see…? (How are the variables related? How would seeing X change my belief in Y?)
Examples: What does a survey tell us about the election results?

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Hachette UK.

Structural causal models (SCMs)

How can we build a SCM model *M*?

An SCM model M is a 4-tuple $< \boldsymbol{U}, \boldsymbol{V}, F, P(\boldsymbol{U}) >$:

- $\boldsymbol{U}$ is a set of exogenous variables that are determined by factors outside the model;
- $\boldsymbol{V}$ is a set $\{V_1, V_2, ..., V_n\}$ of endogenous variables of interest that are determined by other variables in the models ($\boldsymbol{U} \cup \boldsymbol{V}$) ;
- F is a set of mapping functions $\{f_{V_1}, f_{V_2}, ..., f_{V_n}\}$ such that each f_i is a mapping from $\boldsymbol{U}_{V_i} \cup \boldsymbol{Pa}_{V_i}$ to V_i, where $\boldsymbol{U}_{V_i} \subseteq \mathbf{U}$, $\mathbf{Pa}_{V_i} \subseteq \mathbf{V} \setminus \mathrm{V_i}$ ($\boldsymbol{Pa}_{V_i}$ denotes V_i's parents in the graph);
- $P(\boldsymbol{U})$ is a probability function defined over the domain of $\boldsymbol{U}$.

Observed value $\boldsymbol{X} = \{\boldsymbol{X_1}, ..., \boldsymbol{X_n}\}$, is associated with a directed acyclic graph (DAG).

$$X_i := f_i(Pa_i, U_i),\ (i = 1, ..., n).$$

$$p(X_1, \ldots, X_n)$$

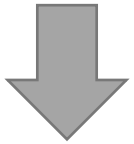
Independent causal mechanisms (ICM)

Causal (or disentangled) factorization

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid \mathbf{PA}_i).$$

Independent Causal Mechanisms (ICM) Principle

- The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.



- Changing (or intervening upon) one mechanism $p(X_i|Pa_i)$ does not change the other mechanisms $p(X_j|Pa_j)$.
- Knowing some other mechanisms $p(X_i|Pa_i)$ does not give us information about a mechanism $p(X_j|Pa_j)$.

Independent causal mechanisms (ICM)

Shanghai 1

Altitude A	Average annual temperature T
0	25.3
50	25.0
150	24.3

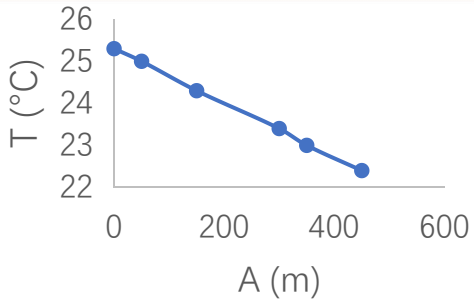
Shanghai 2

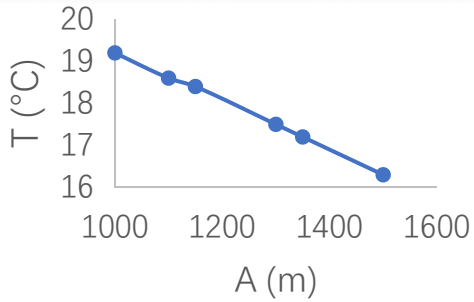
Altitude A	Average annual temperature T
1000	19.2
1100	18.6
1150	18.4

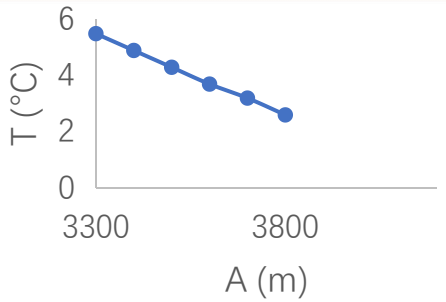
Shanghai 3

Altitude A	Average annual temperature T
3300	5.5
3400	4.9
3500	4.3

For a model to correctly predict the effect of interventions, it needs to be robust with respect to generalizing from an observational distribution to certain interventional distributions.







$p(A,T) = p(A)p(T|A)$ ⇒ $p(T|A)$ can generalize across all places.

$p(A,T) = p(\mathrm{T})p(A|T)$ ⇒ $p(A|T)$ **can't** generalize across all places.

Levels of causal modeling

Table 1 Simple Taxonomy of Models. The Most Detailed Model (Top) Is a Mechanistic or Physical One, Usually in Terms of Differential Equations. At the Other End of the Spectrum (Bottom), We Have a Purely Statistical Model; This Can Be Learned From Data, but It Often Provides Little Insight Beyond Modeling Associations Between Epiphenomena. Causal Models Can Be Seen as Descriptions That Lie in Between, Abstracting Away From Physical Realism While Retaining the Power to Answer Certain Interventional or Counterfactual Questions

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

Structural causal models (SCMs)

- Why we need causal models?
- How to build causal models?
- Causal representation learning

Causal representation learning

What properties should future AI models have?

Robust, transferable, interpretable, explainable, fair

- Learning transferable mechanisms
  - Modularization
    - ✓ Single components can be re-used across a range of environments and tasks (robustness)

Classification | Regression | Generation

- Learning disentangled representations
  - Suppose $X = (X_1, ...X_d)$ is the observation, we want to construct causal variables $S_1, ..., S_n$ (n $\ll d$), Pa_i denotes S_i's parents in the graph. Disentangled representation is

$$\mathrm{p}(S_1, ...S_n) = \prod_{i=1}^{n} p(S_i|Pa_i)$$

    I. An *encoder* q: $\mathbb{R}^d \rightarrow \mathbb{R}^n$ encode the input to latent representation comprising noise variables $U = (U_1, ..., U_n)$;
    II. A *mapping function* $f = (f_1, ..., f_n)$ map U to S, where $S_i := f_i(Pa_i, U_i), (i = 1, ..., n)$;
    III. A *decoder* p: $\mathbb{R}^n \rightarrow \mathbb{R}^d$ decode the disentangled representations.
- Learning interventional world models and reasoning
  - Current representation learning do not take into account causal properties of the variables
  - Future representation learning will move to next level and support intervention, planning, and reasoning (Realizing Konrad Lorenz' notion of thinking as acting in an imagined space)

Current progress

Xia, K., Lee, K. Z., Bengio, Y., & Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. Advances in Neural Information Processing Systems, 34.

Contributions:

1. Their work disentangles the notions of expressivity and learnability, and then verifies that universal approximability is not suitable of learning any SCM by training on data generated by that SCM.
2. They introduce a special type of SCM called a neural causal model (NCM), and formalize a new type of inductive bias to encode structural constraints necessary for performing causal inferences.
3. They develop an algorithm to determine whether a causal effect can be learning from data (i.e., causal identifiability) and estimates the effect whenever identifiability holds (causal estimation).

Current progress

Xia, K., Lee, K. Z., Bengio, Y., & Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. Advances in Neural Information Processing Systems, 34.

Contributions:

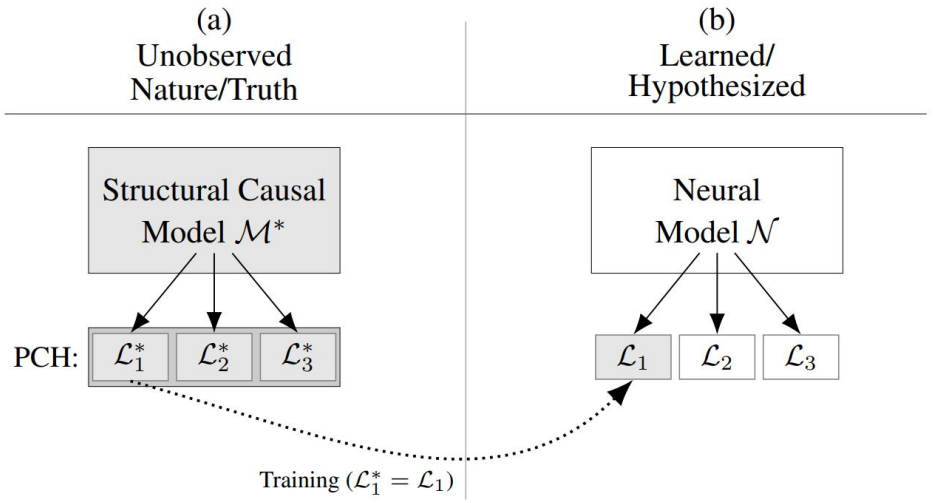
1. Their work disentangles the notions of expressivity and learnability, and then verifies that universal approximability is not suitable of learning any SCM by training on data generated by that SCM.

3.Counterfactuals
Activity: Imaging, Retrospection, Understanding
Questions: What if I had done...? Why? (Was it X that caused Y? What if X had not occurred?)
Examples: Was it the aspirin that stopped my headache? What if I had not smoked last year?

2.Intervention
Activity: Doing Intervening
Questions: What if I do…? How? (What would Y be if I do X ? How can I make Y happen?)
Examples: Was it the aspirin that stopped my headache? What if I had not smoked last year?

1.Association
Activity: Seeing, Observing
Questions: What if I see…? (How are the variables related? How would seeing X change my belief in Y?)
Examples: What does a survey tell us about the election results?

PCH: Pearl Causal Hierarchy (1. seeing 2. doing 3. imaging)



Current progress

Xia, K., Lee, K. Z., Bengio, Y., & Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. Advances in Neural Information Processing Systems, 34.

Contributions:

1. Their work disentangles the notions of expressivity and learnability, and then verifies that universal approximability is not suitable of learning any SCM by training on data generated by that SCM.

$L_1(M)$

The specific distribution $P(\mathbf{V})$, where $\mathbf{X}$ is empty, is defined as layer $L_1(\mathcal{M})$

$L_2(M)$

Definition 2 (Layers 1, 2 Valuations). An SCM $\mathcal{M}$ induces layer $L_2(\mathcal{M})$, a set of distributions over $\mathbf{V}$, one for each intervention $\mathbf{x}$. For each $\mathbf{Y} \subseteq \mathbf{V}$,

$$P^{\mathcal{M}}(\mathbf{y_x}) = \sum_{\{\mathbf{u}|\mathbf{Y_x}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u}), \quad (1)$$

where $\mathbf{Y_x}(\mathbf{u})$ is the solution for $\mathbf{Y}$ after evaluating $\mathcal{F}_{\mathbf{x}} := \{f_{V_i} : V_i \in \mathbf{V} \setminus \mathbf{X}\} \cup \{f_X \leftarrow x : X \in \mathbf{X}\}$.

$L_3(M)$

Definition 9 (Layer 3 Valuation). An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ induces a family of joint distributions over counterfactual events $\mathbf{Y_x}, \ldots, \mathbf{Z_w}$, for any $\mathbf{Y}, \mathbf{Z}, \ldots, \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y_x}, \ldots, \mathbf{z_w}) = \sum_{\substack{\{\mathbf{u} \mid \mathbf{Y_x}(\mathbf{u})=\mathbf{y}, \\ \ldots, \mathbf{Z_w}(\mathbf{u})=\mathbf{z}\}}} P(\mathbf{u}). \quad (7)$$

$P^{(L_i)}$ *consistency*

Definition 4 ($\mathrm{P}^{(L_i)}$-Consistency). Consider two SCMs, $\mathcal{M}_1$ and $\mathcal{M}_2$. $\mathcal{M}_2$ is said to be $\mathrm{P}^{(L_i)}$-consistent (for short, L_i-consistent) w.r.t. $\mathcal{M}_1$ if $L_i(\mathcal{M}_1) = L_i(\mathcal{M}_2)$. ■

Current progress

Xia, K., Lee, K. Z., Bengio, Y., & Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. Advances in Neural Information Processing Systems, 34.

Contributions:

1. Their work disentangles the notions of expressivity and learnability, and then verifies that universal approximability is not suitable of learning any SCM by training on data generated by that SCM.
2. They introduce a special type of SCM called a neural causal model (NCM), and formalize a new type of inductive bias to encode structural constraints necessary for performing causal inferences.

An SCM model M is a 4-tuple $< \boldsymbol{U}, \boldsymbol{V}, F, P(\boldsymbol{U}) >$:

- $\boldsymbol{U}$ is a set of exogenous variables that are determined by factors outside the model;
- $\boldsymbol{V}$ is a set $\{V_1, V_2, ..., V_n\}$ of endogenous variables of interest that are determined by other variables in the models $(\boldsymbol{U} \cup \boldsymbol{V})$;
- F is a set of mapping functions $\{f_{V_1}, f_{V_2}, ..., f_{V_n}\}$ such that each f_i is a mapping from $\boldsymbol{U}_{V_i} \cup \boldsymbol{Pa}_{V_i}$ to V_i, where $\boldsymbol{U}_{V_i} \subseteq \mathbf{U}$, $\mathbf{Pa}_{V_i} \subseteq \mathbf{V} \setminus \mathrm{V_i}$.
- $P(\boldsymbol{u})$ is a probability function defined over the domain of $\boldsymbol{U}$

An NCM model $\widehat{M}(\theta)$ is a 4-tuple $< \widehat{\boldsymbol{U}}, \boldsymbol{V}, \widehat{F}, P(\widehat{\boldsymbol{U}}) >$ with parameters $\boldsymbol{\theta} = \{\theta_{V_i} : V_i \in \boldsymbol{V}\}$:

- $\widehat{\boldsymbol{U}} \subseteq \{\widehat{\boldsymbol{U}}_{\mathbf{C}} : \mathbf{C} \subseteq \mathbf{V}\}$, where each $\widehat{\boldsymbol{U}}$ is associated with some subset of variables $\boldsymbol{C} \subseteq \mathbf{V}$.
- $\widehat{F} = \{\hat{f}_{V_i} : V_i \in \boldsymbol{V}\}$ is a set of mapping functions $\{f_{V_1}, f_{V_2}, ..., f_{V_n}\}$ such that each f_i is a feedforward neural network parameterized by $\theta_{V_i} \in \boldsymbol{\theta}$ mapping $\boldsymbol{U}_{V_i} \cup \boldsymbol{Pa}_{V_i}$ to V_i for some $\mathbf{Pa}_{V_i} \subseteq \mathbf{V}$ and $\boldsymbol{U}_{V_i} = \{\widehat{\boldsymbol{U}}_{\mathbf{C}} : \widehat{\boldsymbol{U}}_{\mathbf{C}} \in \widehat{\boldsymbol{U}}, V_i \in \boldsymbol{C}\}$.
- $P(\widehat{\boldsymbol{U}})$ is a standard uniform distribution $\widehat{\boldsymbol{U}} \sim Unif(0, 1)$.

Current progress

Xia, K., Lee, K. Z., Bengio, Y., & Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. Advances in Neural Information Processing Systems, 34.

Contributions:

1. Their work disentangles the notions of expressivity and learnability, and then verifies that universal approximability is not suitable of learning any SCM by training on data generated by that SCM.
2. They introduce a special type of SCM called a neural causal model (NCM), and formalize a new type of inductive bias to encode structural constraints necessary for performing causal inferences.

Definition 4 ($\mathrm{P}^{(L_i)}$-Consistency). Consider two SCMs, $\mathcal{M}_1$ and $\mathcal{M}_2$. $\mathcal{M}_2$ is said to be $\mathrm{P}^{(L_i)}$-consistent (for short, L_i-consistent) w.r.t. $\mathcal{M}_1$ if $L_i(\mathcal{M}_1) = L_i(\mathcal{M}_2)$. ■

Theorem 1 (NCM Expressiveness). *For any SCM* $\mathcal{M}^* = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, *there exists an NCM* $\widehat{M}(\boldsymbol{\theta}) = \langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, P(\widehat{\mathbf{U}}) \rangle$ *s.t.* $\widehat{M}$ *is* L_3*-consistent w.r.t.* $\mathcal{M}^*$. ■

Intuitive assumption: An NCM can be trained on the observed data and act as a proxy for the true SCM M^* , and inferences about other quantities of M^* can be done through computation directly in $\widehat{M}(\theta)$.

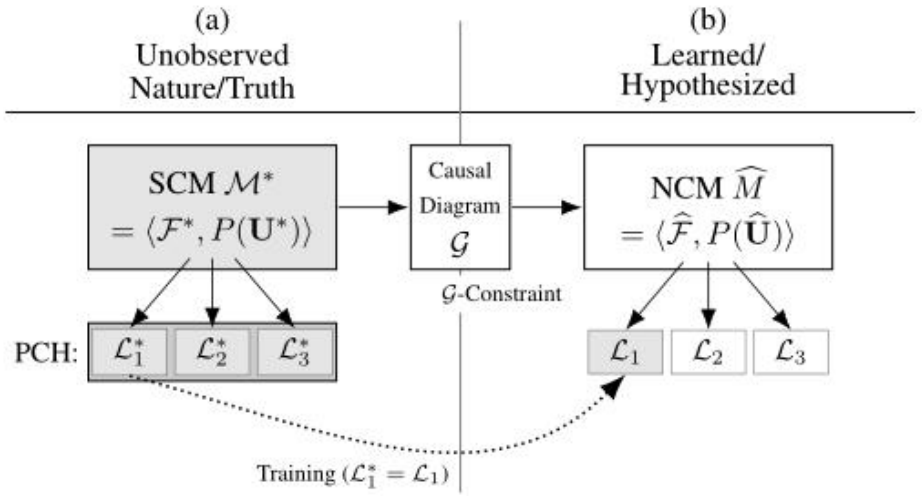
Unfortunately this assumption fails in almost all case.

Current progress

Xia, K., Lee, K. Z., Bengio, Y., & Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. Advances in Neural Information Processing Systems, 34.

Contributions:

1. Their work disentangles the notions of expressivity and learnability, and then verifies that universal approximability is not suitable of learning any SCM by training on data generated by that SCM.
2. They introduce a special type of SCM called a neural causal model (NCM), and formalize a new type of inductive bias to encode structural constraints necessary for performing causal inferences.
3. They develop an algorithm to determine whether a causal effect can be learning from data (i.e., causal identifiability) and estimates the effect whenever identifiability holds (causal estimation).

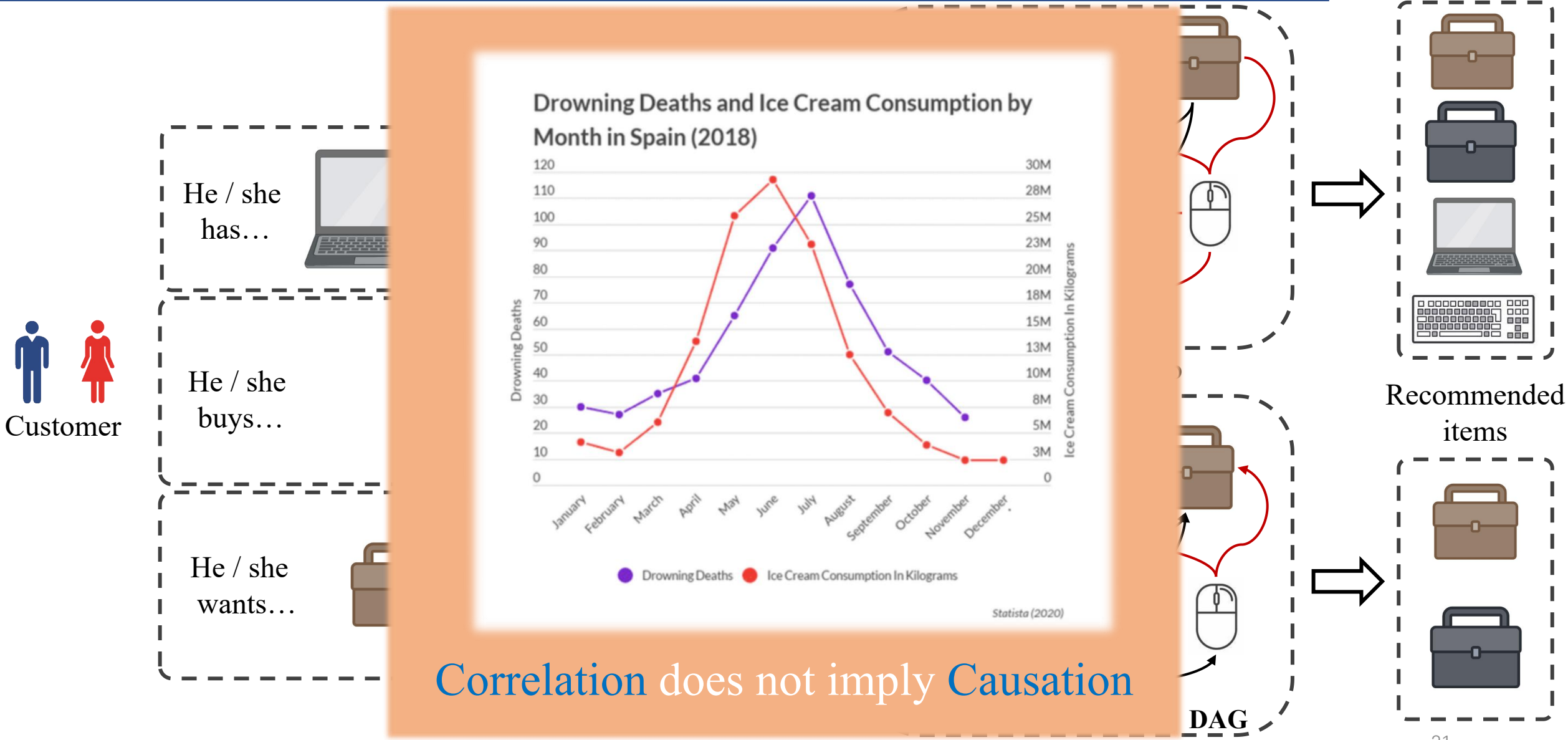


Definition 5 ($\mathcal{G}$-Consistency). Let $\mathcal{G}$ be the causal diagram induced by SCM $\mathcal{M}^*$. For any SCM $\mathcal{M}$, we say that $\mathcal{M}$ is $\mathcal{G}$-consistent (w.r.t. $\mathcal{M}^*$) if $\mathcal{G}$ is a CBN for $L_2(\mathcal{M})$. ■

CBN: causal bayesian network

Thanks for your attention!

From statistical to causal models: A case study



DAG: directed acyclic graphs