Causal AI: the Way of Change in the age of AI

Professor Javen Qinfeng Shi

November 2024

All forms are non-stable and forever changing

The way of change is causality disguised as cause-and-effect

Do NOT trust a decision derived from correlation alone, even if the data are distortion-free and noise-free

Simpson's Reversal & Paradox

Kidney Stone Treatment

	Treatment A	Treatment B
Small stones	Group 1 **93% (81/87)**	Group 2 87% (234/270)
Large stones	Group 3 **73% (192/263)**	Group 4 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Quiz: Which treatment is better?

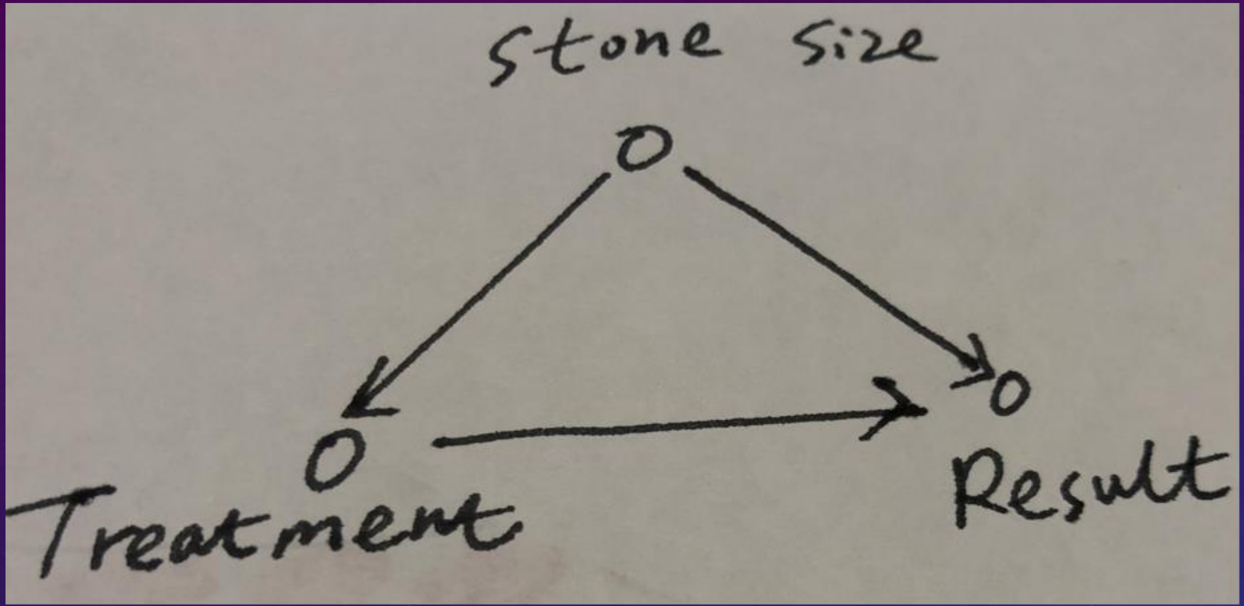
https://en.wikipedia.org/wiki/Simpson%27s_paradox

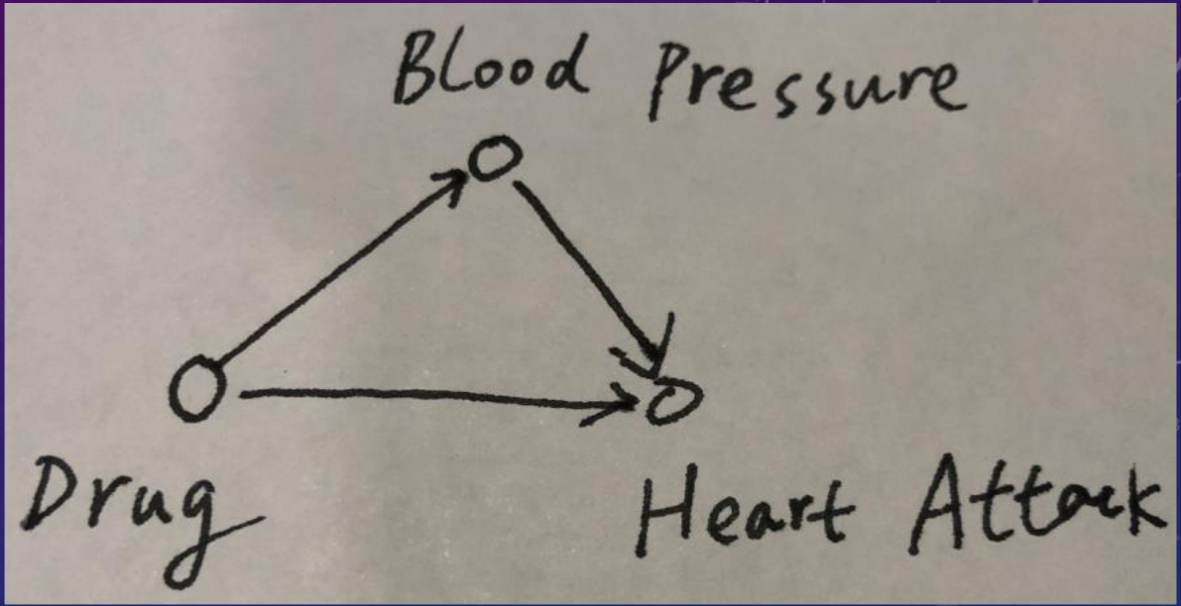
Simpson's Reversal & Paradox

Heart disease drugs (fictitious data to make a point)

	Treatment with Drug A	Treatment with Drug B
Low blood pressure	Group 1 **93% (81/87)**	Group 2 87% (234/270)
High blood pressure	Group 3 **73% (192/263)**	Group 4 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Quiz: Which drug is better?





When you know the causal diagram, you know which one is the right decision.

Paradoxes are treasures

- Not all things can be seen and known
  - Use causal AI to discover, and/or acknowledge the existence of unseen and unknown

- Situation may reduce your ability (vision/IQ/EQ) temporarily
  - Basketball & gorilla example

- Mind is a choice maker
  - Reaction often leads to a regrettable choice
  - Instinct is NOT Intuition

- No learning comes without desire and persistence
  - Desire ~= Reward
  - 1M-10M episodes for RL to converge

Correlation whispers the presence of causation

Reichenbach's Common Cause Principle

If A correlates with B, then

- A-> B
- Or A <- B
- Or A<- C -> B, where C is the common cause

Causality Brings

- **A Deeper Truth**: Beyond Mere Correlation
  - Find (root) causes (explainability & accountability)
  - Reveal the Right choice (amidst spurious correlation & Simpson's paradox)
  - Uncover hidden variables & mechanism
- **Proactive Change**: Shaping, Not Just Adapting
  - It offers how to change/intervene to achieve specific outcomes (rather than merely adapting to survive)
- **A Holistic View:** Oneness and Evolution
  - See diverse data, tasks and domains as one
  - Naturally evolute to long term optimal with ease (free of fear, and anxiety)

Causal AI As Structural Causal Models (SCM)

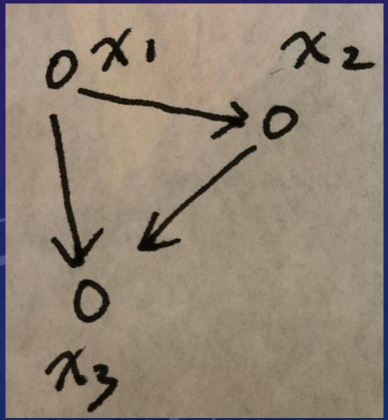
A SCM example

$$x_1 = n_1$$

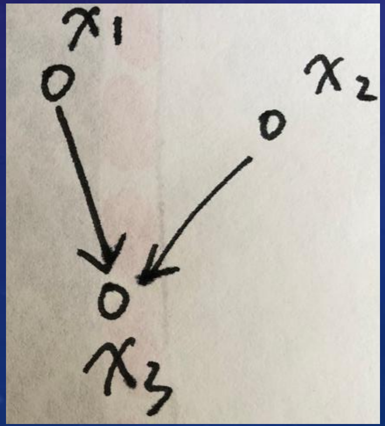
$$x_2 = 2x_1^2 + n_2$$

$$x_3 = 3x_2 + 2x_1 + n_3$$

Intervention



$do(x_2 = 2)$



Counterfactual

Observe $(x_1, x_2, x_3) = (-1, 3, 8)$

What x_3 would have been had x_1 been 2 ?

Recover $(n_1, n_2, n_3) = (-1, 1, 1)$

$$x_1 = 2$$

$$x_2 = 2 \times 4 + 1 = 9$$

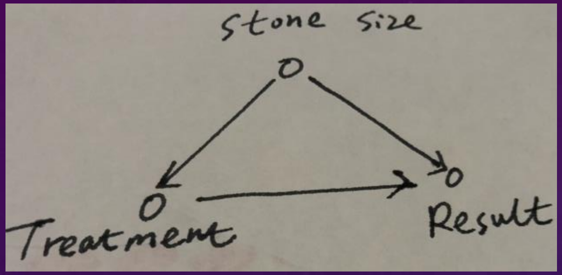
$$x_3 = 3 \times 9 + 2 \times 2 + 1 = 32$$

Causal AI's Key Research Questions

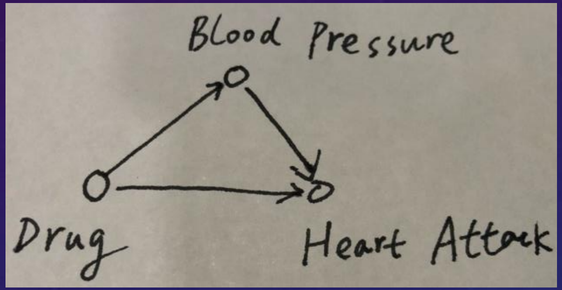
Forward question: Uncover the underlying causal mechanism (which then estimates the future consequence of interventions, or counterfactual effects in the past)

Reverse question: How to suggest a sequence (among infinite many) of interventions in a changing environment, given resources or budgets, to bring a desirable future outcome?

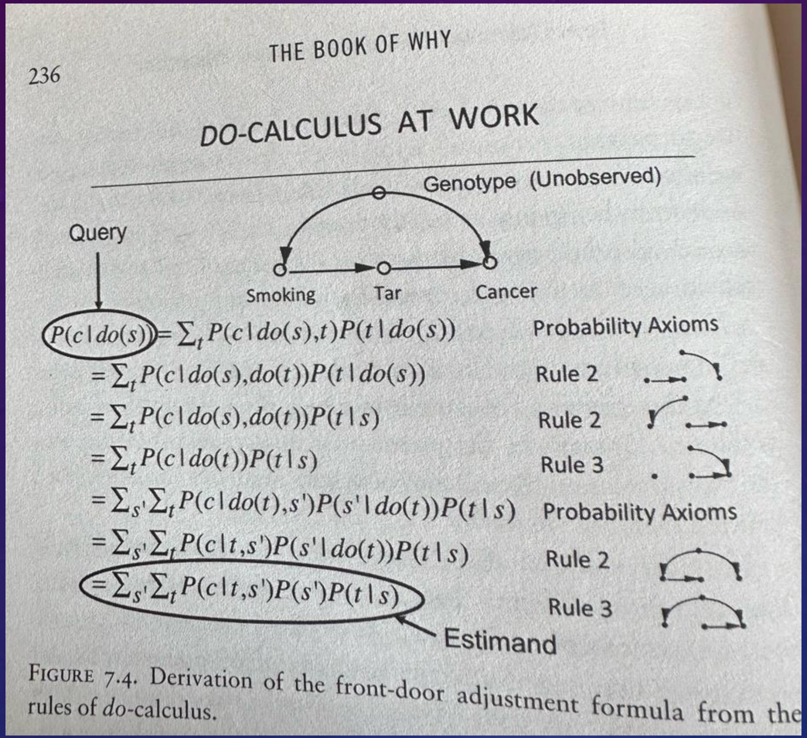
Estimate Causal Effects: Backdoor, Front Door And IV



Back-door path and adjustable



No back-door path



236 THE BOOK OF WHY

DO-CALCULUS AT WORK

Genotype (Unobserved)

Query

Smoking Tar Cancer

$P(c \mid do(s)) = \sum_t P(c \mid do(s), t) P(t \mid do(s))$ Probability Axioms

$= \sum_t P(c \mid do(s), do(t)) P(t \mid do(s))$ Rule 2

$= \sum_t P(c \mid do(s), do(t)) P(t \mid s)$ Rule 2

$= \sum_t P(c \mid do(t)) P(t \mid s)$ Rule 3

$= \sum_{s'} \sum_t P(c \mid do(t), s') P(s' \mid do(t)) P(t \mid s)$ Probability Axioms

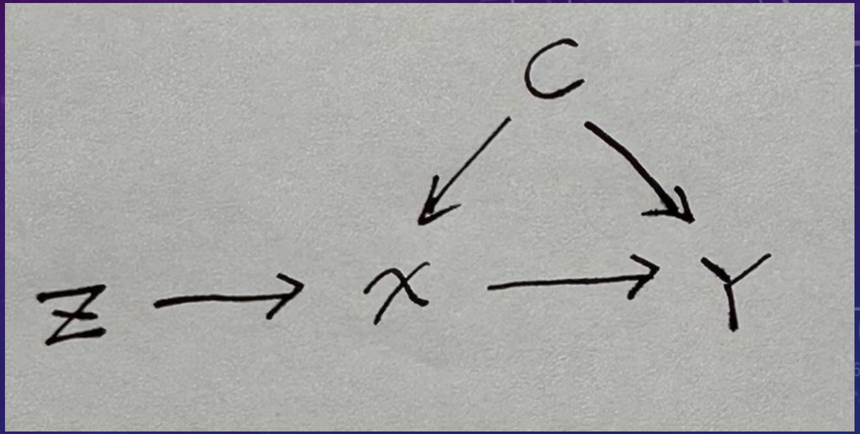
$= \sum_{s'} \sum_t P(c \mid t, s') P(s' \mid do(t)) P(t \mid s)$ Rule 2

$= \sum_{s'} \sum_t P(c \mid t, s') P(s') P(t \mid s)$ Rule 3

Estimand

FIGURE 7.4. Derivation of the front-door adjustment formula from the rules of *do*-calculus.

Back-door exists but unadjustable, but front-door path and adjustable



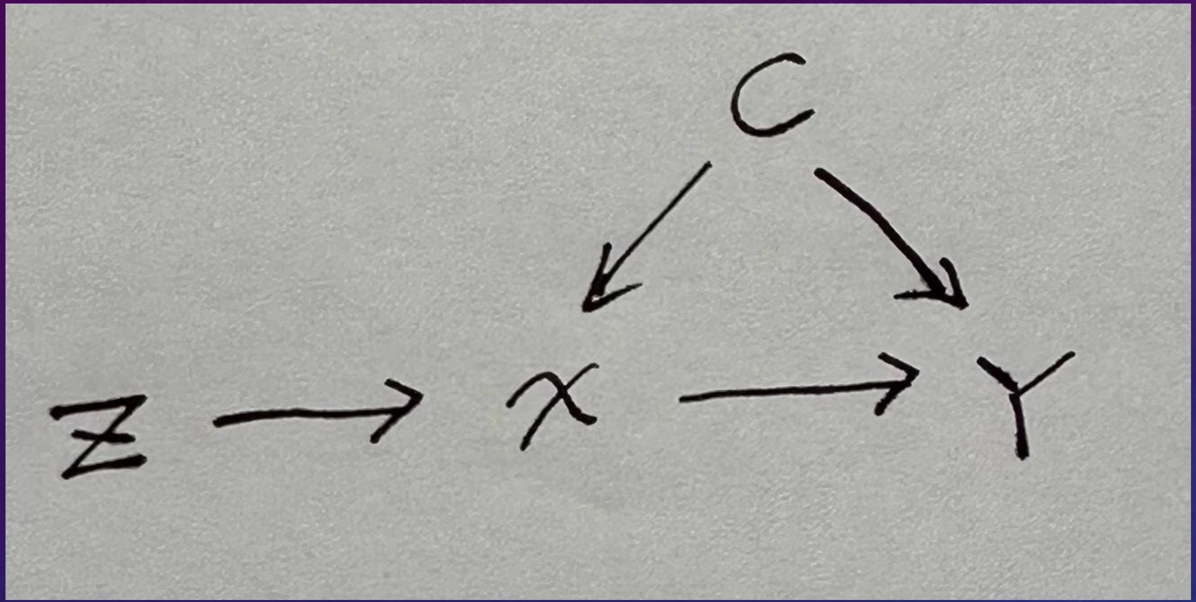
Back-door exists but unadjustable and front-door does not exist

Instrumental Variable (IV) exists

- Strong IV available
- Strong IV unavailable, but multiple weak IVs available

Middle image from The Book of Why

Estimate Causal Effects: IV & 2SLS



To estimate the effect Y of the treatment X.

C unobservable (i.e. can not adjust or control C)

Strong IV Z observable

Stage 1: Regress each column of **X** on **Z**, ($X = Z\delta + \text{errors}$):

$$\hat{\delta} = (Z^{\mathrm{T}} Z)^{-1} Z^{\mathrm{T}} X,$$

and save the predicted values:

$$\widehat{X} = Z\hat{\delta} = Z(Z^{\mathrm{T}} Z)^{-1} Z^{\mathrm{T}} X = P_Z X.$$

In the second stage, the regression of interest is estimated as usual, values from the first stage:

Stage 2: Regress **Y** on the predicted values from the first stage:

$$Y = \widehat{X}\beta + \text{noise},$$

which gives

$$\beta_{2\mathrm{SLS}} = \left(X^{\mathrm{T}} P_Z X\right)^{-1} X^{\mathrm{T}} P_Z Y.$$

Two-stage least squares (2SLS)

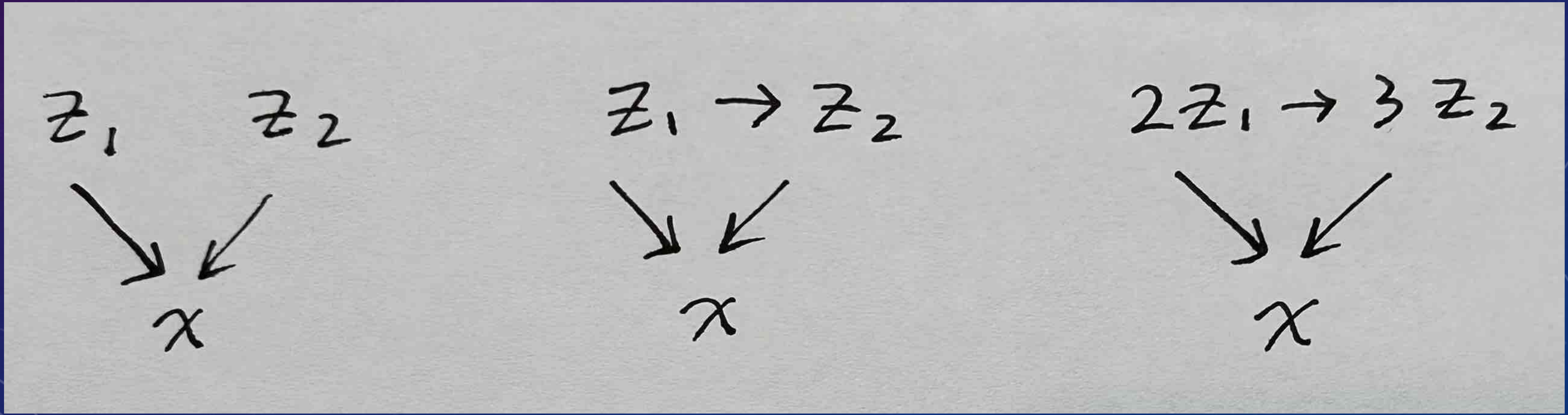
https://en.wikipedia.org/wiki/Instrumental_variables_estimation

What if no IV (nor its candidates) available, and the structure is unknown?

Is it possible to uncover the unknown causal mechanism? How?

Uncover Unknown Causal Mechanism

Is it even possible to uniquely recover what is unobservable and unknown (i.e. Z), purely from some observation (i.e. X)?



Changes Make It Possible

$$z_2 = 3u z_1$$

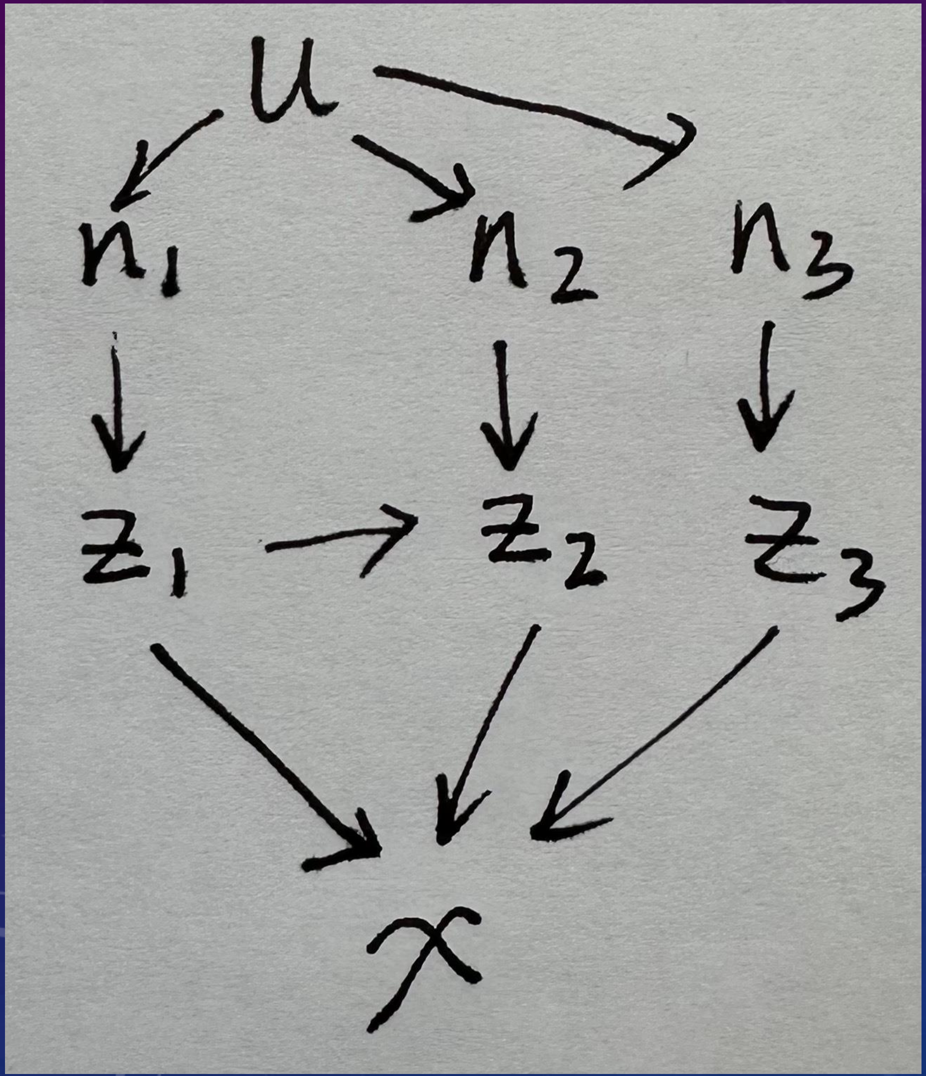
$$z_2 = (3u+2) z_1 + 4$$

$$z_2 = (3u+2) z_1$$

$$u \to z \to x$$

$$u \to n \to z \to x$$

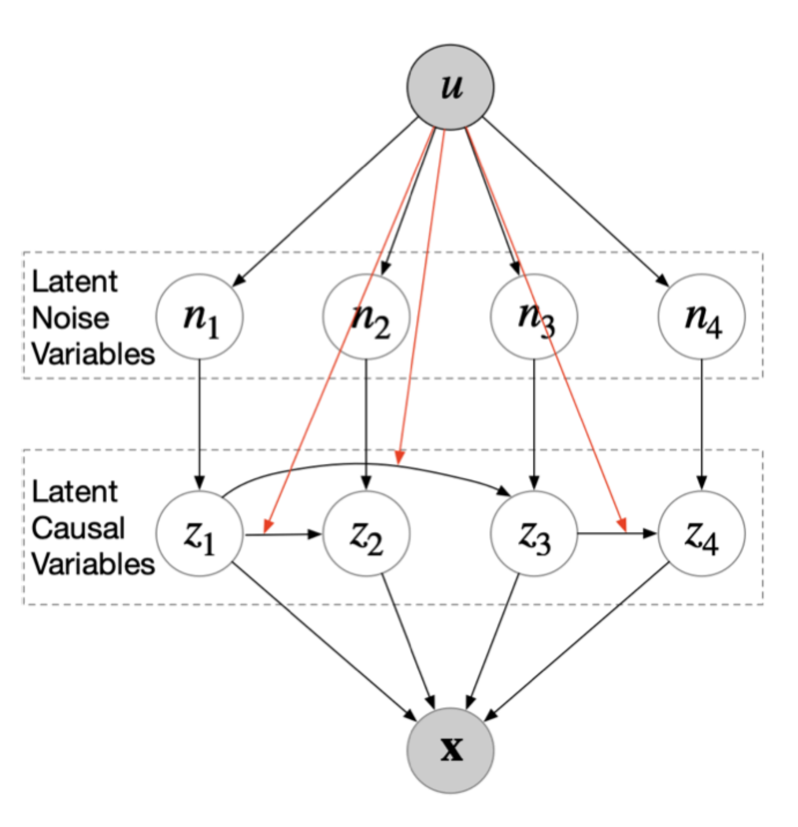
Sufficient changes in n and z (via u), enable to recover unknown n and z uniquely (a.k.a. identifiability in causal theory)



$$z_2 = (3u + \cancel{2})\, z_1 + \cancel{4} + n_2$$

Identifying Weight-Variant Latent Causal Models

Yuhang Liu[1] Zhen Zhang[1] Dong Gong[2] Mingming Gong[3] Biwei Huang[4] Anton van den Hengel[1]
Kun Zhang[5] Javen Qinfeng Shi[1]



$$n_i :\sim \mathcal{N}(\beta_{i,1}(\mathbf{u}), \beta_{i,2}(\mathbf{u})), \tag{1}$$

$$z_i := \boldsymbol{\lambda}_i^T(\mathbf{u})(\mathbf{z}) + n_i, \tag{2}$$

$$\mathbf{x} := \mathbf{f}(\mathbf{z}) + \boldsymbol{\varepsilon} \tag{3}$$

where

- each noise term n_i is Gaussian distributed with mean $\beta_{i,1}(\mathbf{u})$ and variance $\beta_{i,2}(\mathbf{u})$; both $\beta_{i,1}$ and $\beta_{i,2}$ can be nonlinear mappings. Moreover, the distribution of n_i is modulated by the observed variable $\mathbf{u}$.
- In Eq. (2), $\boldsymbol{\lambda_i}(\mathbf{u})$ denote the vector corresponding to the causal weights from $\mathbf{z}$ to z_i, *e.g.* $\boldsymbol{\lambda_i}(\mathbf{u}) = [\lambda_{1,i}(\mathbf{u}), \lambda_{2,i}(\mathbf{u}), ..., \lambda_{n,i}(\mathbf{u})]$, and each $\lambda_{j,i}$ could be a nonlinear mapping.
- In Eq. (3), $\mathbf{f}$ denote a nonlinear mapping, and $\boldsymbol{\varepsilon}$ is independent noise with probability density function $p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$.

Theorem 4.1. *Suppose latent causal variables* $\mathbf{z}$ *and the observed variable* $\mathbf{x}$ *follow the generative models defined in Eq.* (1)*- Eq.* (3)*, with parameters* $(\mathbf{f}, \boldsymbol{\lambda}, \boldsymbol{\beta})$*. Assume the following holds:*

(i) The set $\{\mathbf{x} \in \mathcal{X} | \varphi_{\varepsilon}(\mathbf{x}) = 0\}$ *has measure zero (i.e., has at the most countable number of elements), where* $\varphi_{\boldsymbol{\varepsilon}}$ *is the characteristic function of the density* $p_{\boldsymbol{\varepsilon}}$*.*

(ii) The function $\mathbf{f}$ *in Eq.* (3) *is bijective.*

(iii) There exist $2n+1$ *distinct points* $\mathbf{u}_{\mathbf{n},0}, \mathbf{u}_{\mathbf{n},1}, ..., \mathbf{u}_{\mathbf{n},2n}$ *such that the matrix*

$$\mathbf{L_n} = (\boldsymbol{\eta}_\mathbf{n}(\mathbf{u}_{\mathbf{n},1}) - \boldsymbol{\eta}_\mathbf{n}(\mathbf{u}_{\mathbf{n},0}), ..., \boldsymbol{\eta}_\mathbf{n}(\mathbf{u}_{\mathbf{n},2n}) - \boldsymbol{\eta}_\mathbf{n}(\mathbf{u}_{\mathbf{n},0})) \tag{8}$$

of size $2n \times 2n$ *is invertible.*

(iv) There exist $k+1$ *distinct points* $\mathbf{u}_{\mathbf{z},0}, \mathbf{u}_{\mathbf{z},1}, ..., \mathbf{u}_{\mathbf{z},k}$ *such that the matrix*

$$\mathbf{L_z} = (\boldsymbol{\eta}_\mathbf{z}(\mathbf{u}_{\mathbf{z},1}) - \boldsymbol{\eta}_\mathbf{z}(\mathbf{u}_{\mathbf{z},0}), ..., \boldsymbol{\eta}_\mathbf{z}(\mathbf{u}_{\mathbf{z},k}) - \boldsymbol{\eta}_\mathbf{z}(\mathbf{u}_{\mathbf{z},0})) \tag{9}$$

of size $k \times k$ *is invertible.*

(v) The function class of $\lambda_{i,j}$ *can be expressed by a Taylor series: for each* $\lambda_{i,j}$*,* $\lambda_{i,j}(\mathbf{0}) = 0$*,*

then the true latent causal variables $\mathbf{z}$ *are related to the estimated latent causal variables* $\hat{\mathbf{z}}$ *by the following relationship:* $\mathbf{z} = \mathbf{P}\hat{\mathbf{z}} + \mathbf{c}$*, where* $\mathbf{P}$ *denotes the permutation matrix with scaling,* $\mathbf{c}$ *denotes a constant vector.*

IDENTIFIABLE LATENT POLYNOMIAL CAUSAL MODELS THROUGH THE LENS OF CHANGE

Yuhang Liu[1], Zhen Zhang[1], Dong Gong[2], Mingming Gong[3], Biwei Huang[4]
Anton van den Hengel[1], Kun Zhang[5], Javen Qinfeng Shi[1]
[1] Australian Institute for Machine Learning, The University of Adelaide, Australia
[2] School of Computer Science and Engineering, The University of New South Wales, Australia
[3] School of Mathematics and Statistics, The University of Melbourne, Australia
[4] Halicioğlu Data Science Institute (HDSI), University of California San Diego, USA
[5] Department of Philosophy, Carnegie Mellon University, USA
`yuhang.liu01@adelaide.edu.au`

$\mathbf{n}$ follows an exponential family given $\mathbf{u}$, and assuming $\mathbf{z}$ and $\mathbf{x}$ are generated as follows:

$$p_{(\mathbf{T},\boldsymbol{\eta})}(\mathbf{n}|\mathbf{u}) := \prod_i \frac{1}{Z_i(\mathbf{u})} \exp[\sum_j (T_{i,j}(n_i)\eta_{i,j}(\mathbf{u}))], \tag{1}$$

$$z_i := \mathrm{g}_i(\mathrm{pa}_i, \mathbf{u}) + n_i, \tag{2}$$

$$\mathbf{x} := \mathbf{f}(\mathbf{z}) + \boldsymbol{\varepsilon}, \tag{3}$$

with

$$\mathrm{g}_i(\mathbf{z},\mathbf{u}) = \boldsymbol{\lambda}_i^T(\mathbf{u})[\mathbf{z}, \mathbf{z}\bar{\otimes}\mathbf{z}, ..., \underbrace{\mathbf{z}\bar{\otimes}...\bar{\otimes}\mathbf{z}}], \tag{4}$$

where

- in Eq. 1, $Z_i(\mathbf{u})$ denotes the normalizing constant, and $T_{i,j}(n_i)$ denotes the sufficient statistic for n_i, whose the natural parameter $\eta_{i,j}(\mathbf{u})$ depends on $\mathbf{u}$. Here we focus on two-parameter (e.g., $j \in \{1,2\}$) exponential family members, which include not only Gaussian, but also inverse Gaussian, Gamma, inverse Gamma, and beta distributions.
- In Eq. 2, pa_i denotes the set of parents of z_i.
- In Eq. 3, $\mathbf{f}$ denotes a nonlinear mapping, and $\boldsymbol{\varepsilon}$ is independent noise with probability density function $p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$.
- In Eq. 4, where $\boldsymbol{\lambda_i}(\mathbf{u}) = [\lambda_{1,i}(\mathbf{u}), \lambda_{2,i}(\mathbf{u}), ...)]$, $\bar{\otimes}$ represents the Kronecker product with all distinct entries, e.g., for 2-dimension case, $z_1\bar{\otimes}z_2 = [z_1^2, z_2^2, z_1z_2]$.

Theorem 3.1 *Suppose latent causal variables $\mathbf{z}$ and the observed variable $\mathbf{x}$ follow the causal generative models defined in Eq. 1 - Eq. 4. Assume the following holds:*

(i) The set $\{\mathbf{x} \in \mathcal{X} | \varphi_{\varepsilon}(\mathbf{x}) = 0\}$ has measure zero (i.e., has at the most countable number of elements), where $\varphi_{\boldsymbol{\varepsilon}}$ is the characteristic function of the density $p_{\boldsymbol{\varepsilon}}$,

(ii) The function $\mathbf{f}$ in Eq. 3 is bijective,

(iii) The random vector $\mathbf{u}$ [1] takes up at least $2\ell + 1$ distinct vectors denoted as $\mathbf{u}_0, \mathbf{u}_1, ..., \mathbf{u}_{2\ell}$ such that the matrix

$$\mathbf{L} = (\boldsymbol{\eta}(\mathbf{u} = \mathbf{u}_1) - \boldsymbol{\eta}(\mathbf{u} = \mathbf{u}_0), ..., \boldsymbol{\eta}(\mathbf{u} = \mathbf{u}_{2\ell+1}) - \boldsymbol{\eta}(\mathbf{u} = \mathbf{u}_0)) \tag{5}$$

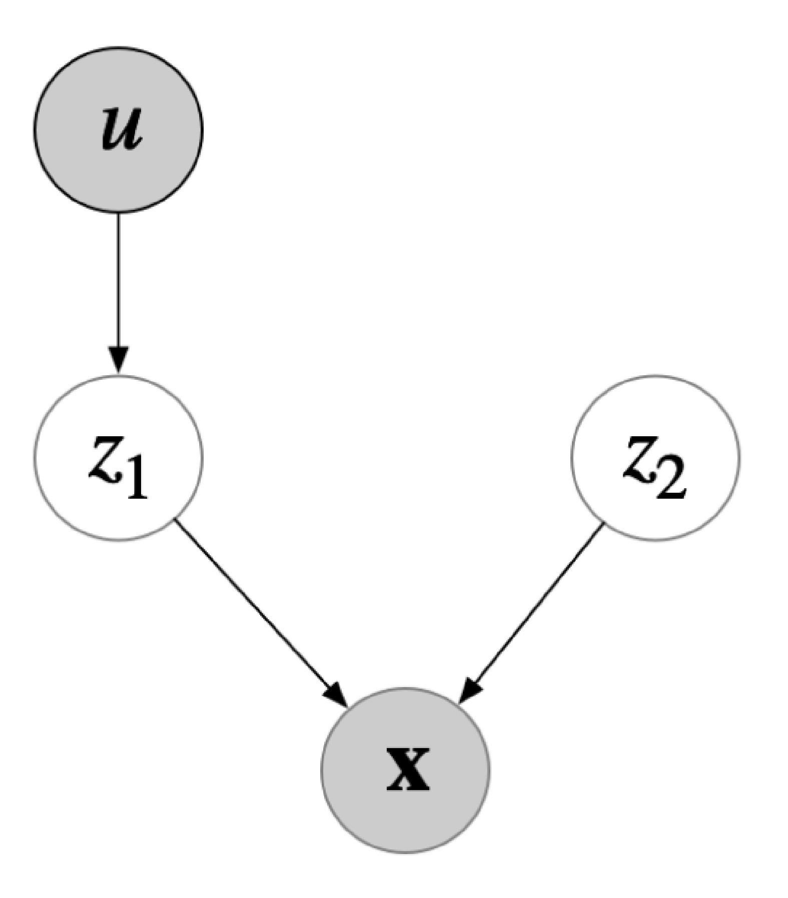
of size $2\ell \times 2\ell$ is invertible. Here $\boldsymbol{\eta}(\mathbf{u}) = [\eta_{i,j}(\mathbf{u})]_{i,j}$,

(iv) The function class of $\lambda_{i,j}$ can be expressed by a Taylor series: for each $\lambda_{i,j}$, $\lambda_{i,j}(\mathbf{u} = \mathbf{0}) = 0$,

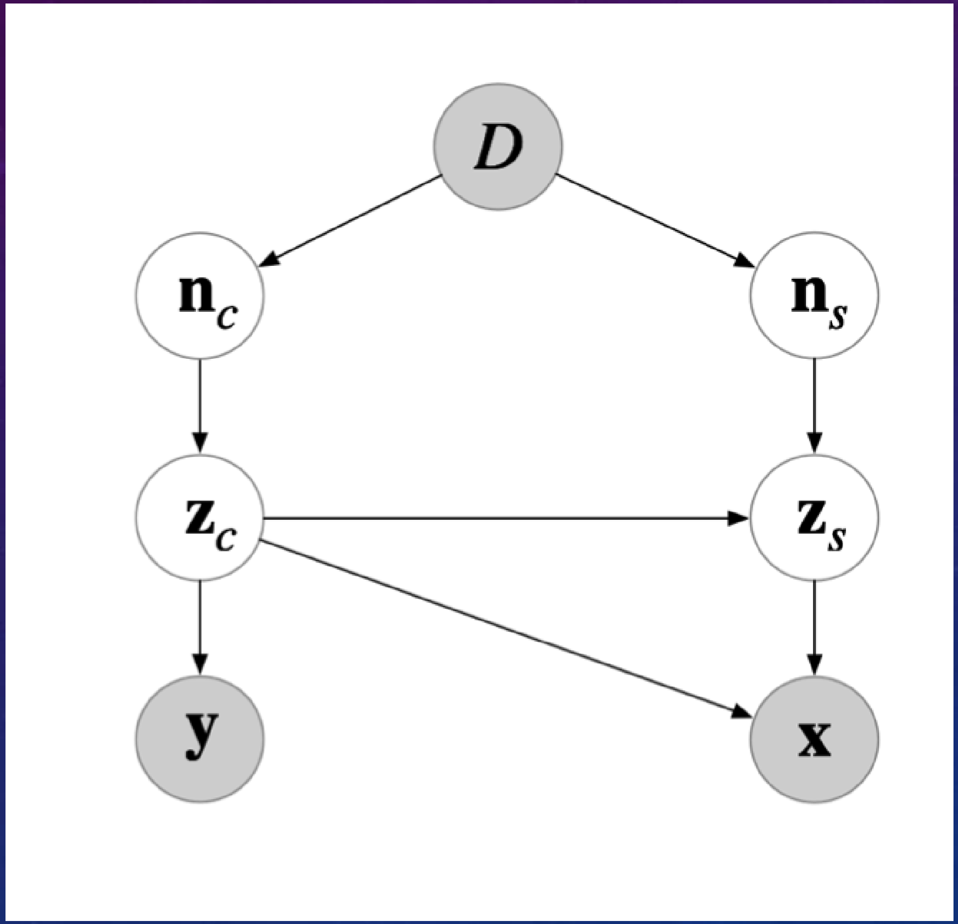
then the true latent causal variables $\mathbf{z}$ are related to the estimated latent causal variables $\hat{\mathbf{z}}$, which are learned by matching the true marginal data distribution $p(\mathbf{x}|\mathbf{u})$, by the following relationship: $\mathbf{z} = \mathbf{P}\hat{\mathbf{z}} + \mathbf{c}$, where $\mathbf{P}$ denotes the permutation matrix with scaling, $\mathbf{c}$ denotes a constant vector.

Quiz

By now you may be familiar that under certain assumptions facilitated by $\mathbf{u}$, the latent variable z_1 is identifiable. Given such assumptions and the identifiability of z_1, can you uniquely recover z_2 given $\mathbf{u}$ and $\mathbf{x}$? If yes, please explain why. If no, does it matter for predicting $\mathbf{x}$ across different domains $\mathbf{u}$, and also explain why.

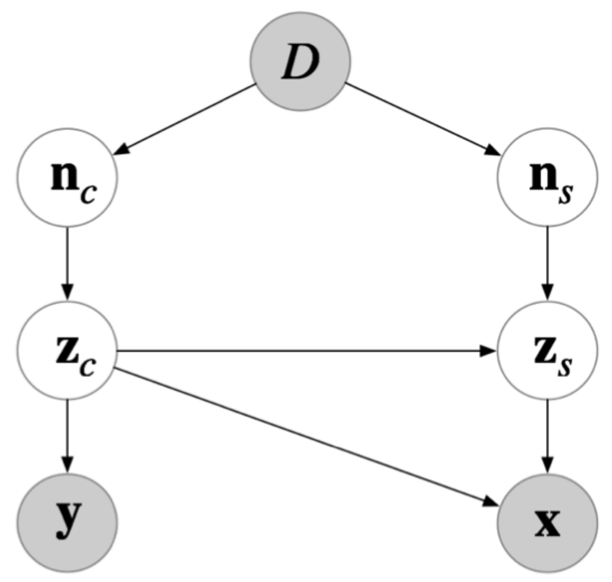


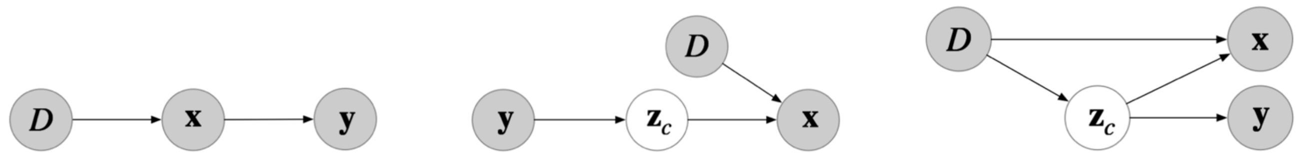
Disentangle contents from styles to improve generalisation



Identifying Latent Causal Content for Multi-Source Domain Adaptation

Yuhang Liu[1], Zhen Zhang[1], Dong Gong[2], Mingming Gong[3], Biwei Huang[4], Kun Zhang[5], Javen Qinfeng Shi[1]





(a) Covariate Shift (b) Conditional Shift (c) Latent Covariate Shift

Figure 1: The illustration of three different assumptions for MSDA. (a) Covariate Shift: $p_D(\mathbf{x})$ changes across domains, while $p_D(\mathbf{y}|\mathbf{x})$ is invariant across domains. (b) Conditional Shift: $p_D(\mathbf{y})$ is invariant, while $p_D(\mathbf{x}|\mathbf{y})$ changes across domains. (c) Latent Covariate Shift: $p_D(\mathbf{z}_c)$ changes across domains while $p_D(\mathbf{y}|\mathbf{z}_c)$ is invariant across domains.

Table 2: Classification results on TerraIncognita.

Methods	Accuracy				
	→L28	→L43	→L46	→L7	Average
ERM	54.1 ± 2.8	62.3 ± 0.7	44.7 ± 0.9	74.5 ± 2.6	58.9
MCDA ((Saito et al., 2018))	54.9 ± 4.1	61.2 ± 1.2	42.7 ± 0.3	64.8 ± 8.1	55.9
M3SDA (Peng et al., 2019)	62.3 ± 1.4	62.7 ± 0.4	41.3 ± 0.3	57.4 ± 0.9	55.9
LtC-MSDA (Wang et al., 2020)	51.9 ± 5.7	54.6 ± 1.3	45.7 ± 1.0	69.1 ± 0.3	55.3
T-SVDNet (Li et al., 2021)	58.2 ± 1.7	61.9 ± 0.3	45.6 ± 2.0	68.2 ± 1.1	58.5
IRM (Arjovsky et al., 2019)	57.5 ± 1.7	60.7± 0.3	42.4 ± 0.6	74.1 ± 1.6	58.7
IWCDAN (Tachet des Combes et al., 2020)	58.1 ± 1.8	59.3 ± 1.9	43.8± 1.5	58.9 ± 3.8	55.0
LaCIM (Sun et al., 2021)	58.2 ± 3.3	59.8 ± 1.6	46.3 ± 1.1	70.8 ± 1.0	58.8
iLCC-MSDA(Ours)	**64.3 ± 3.4**	**63.1 ± 1.6**	44.7 ± 0.4	**80.8 ± 0.4**	**63.2**
iLCC-MSDA(Ours) with $\beta = 1$	56.3 ± 4.3	61.5 ± 0.7	45.2 ± 0.3	80.1 ± 0.6	60.8
iLCC-MSDA(Ours) with $\gamma = 0$	54.8 ± 1.4	58.9 ± 1.8	**46.8 ± 1.4**	73.1 ± 0.6	58.4

Table 1: Classification results and ablation study on PACS data.

PACS ($D_{\mathcal{KL}} = 0.3$)					
Methods	Accuracy				
	→Art	→Cartoon	→Photo	→Sketch	Average
ERM	82.3 ± 0.3	81.3 ± 0.9	94.9 ± 0.2	76.2 ± 0.7	83.6
MCDA ((Saito et al., 2018))	76.6 ± 0.6	85.1 ± 0.3	96.6 ± 0.1	70.1 ± 1.3	82.1
M3SDA (Peng et al., 2019)	79.6 ± 1.0	**86.6 ± 0.5**	**97.1 ± 0.3**	83.3 ± 1.0	86.6
LtC-MSDA (Wang et al., 2020)	82.7 ± 1.3	84.9 ± 1.4	96.9 ± 0.2	75.3 ± 3.1	84.9
T-SVDNet (Li et al., 2021)	81.8 ± 0.3	86.5 ± 0.2	95.9 ± 0.2	80.7 ± 0.8	86.3
IRM (Arjovsky et al., 2019)	79.6 ± 0.7	77.0 ± 2.2	94.6 ± 0.2	71.7 ± 2.3	80.7
IWCDAN (Tachet des Combes et al., 2020)	84.0 ± 0.5	78.1 ± 0.7	96.0 ± 0.1	75.5 ± 1.9	83.4
LaCIM (Sun et al., 2021)	63.1 ± 1.5	72.6 ± 1.0	82.7 ± 1.3	71.5 ± 0.9	72.5
iLCC-MSDA(Ours)	**86.4 ± 0.8**	81.1 ± 0.8	95.9 ± 0.1	**86.0 ± 1.0**	**87.4**
PACS ($D_{\mathcal{KL}} = 0.5$)					
ERM	85.4± 0.6	76.4 ± 0.5	94.4 ± 0.4	85.0 ± 0.6	85.3
MCDA ((Saito et al., 2018))	81.6 ± 0.1	76.8 ± 0.1	93.6 ± 0.1	84.1 ± .6	84.0
M3SDA (Peng et al., 2019)	81.2 ± 1.2	77.5 ± 1.3	94.5 ± 0.5	84.3 ± 0.5	84.4
LtC-MSDA (Wang et al., 2020)	85.2 ± 1.5	75.2 ± 2.6	94.9 ± 0.6	85.1 ± 2.7	85.1
T-SVDNet (Li et al., 2021)	84.8 ± 0.3	77.6 ± 1.7	94.2 ± 0.2	86.4 ± 0.2	85.6
IRM (Arjovsky et al., 2019)	81.5 ± 0.3	71.1 ± 1.3	94.2 ± 0.1	78.7 ± 0.7	81.4
IWCDAN (Tachet des Combes et al., 2020)	79.2 ± 1.6	72.6 ± 0.7	**95.6 ± 0.1**	82.1 ± 2.2	82.4
LaCIM (Sun et al., 2021)	67.4 ± 1.6	66.6 ± 0.6	81.0 ± 1.2	82.3 ± 0.6	74.3
iLCC-MSDA(Ours)	**89.0 ± 0.7**	**77.6 ± 0.5**	95.0 ± 0.3	**87.4 ± 1.6**	**87.3**
PACS ($D_{\mathcal{KL}} = 0.7$)					
ERM	86.1 ± 0.6	**76.8 ± 0.3**	94.6 ± 0.4	81.3 ± 2.0	84.7
MCDA ((Saito et al., 2018))	80.8 ± 0.7	74.1 ± 1.2	94.4 ± 0.4	77.9 ± 0.4	81.8
M3SDA (Peng et al., 2019)	82.7 ± 1.3	76.2 ± 1.0	94.5 ± 0.7	80.8 ± 1.2	83.6
LtC-MSDA (Wang et al., 2020)	83.7 ± 1.6	74.6 ± 1.4	95.0 ± 0.7	80.8 ± 0.6	83.5
T-SVDNet (Li et al., 2021)	83.3 ± 0.8	74.7 ± 0.6	95.2 ± 0.3	74.5 ± 3.3	81.9
IRM (Arjovsky et al., 2019)	84.3 ± 0.8	73.3 ± 1.8	94.3 ± 0.1	69.4 ± 4.6	80.3
IWCDAN (Tachet des Combes et al., 2020)	76.3 ± 0.8	73.9 ± 1.6	93.1 ± 0.5	77.6 ± 3.8	80.2
LaCIM (Sun et al., 2021)	63.6 ± 0.9	68.7 ± 1.4	77.5 ± 3.8	77.8 ± 2.2	71.9
iLCC-MSDA(Ours)	**90.7 ± 0.3**	74.2 ± 0.7	**95.8 ± 0.3**	**83.0 ± 2.2**	**86.0**
iLCC-MSDA(Ours) with $\beta = 1$	90.2 ± 0.5	73.4 ± 0.8	95.7 ± 0.4	82.7 ± 0.7	85.5
iLCC-MSDA(Ours) with $\gamma = 0$	81.1 ± 1.5	70.0 ± 1.6	92.0 ± 0.5	59.6 ± 0.7	75.7

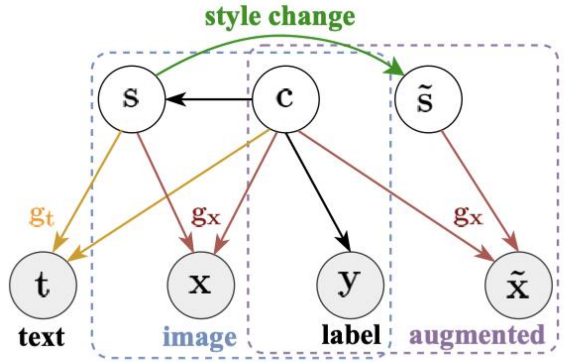
Creating new image-to-image pairings, or text-to-image pairings, can be hard to control the quality

Creating new text-to-text pairings is easy, and the quality is almost self-guaranteed. Why?

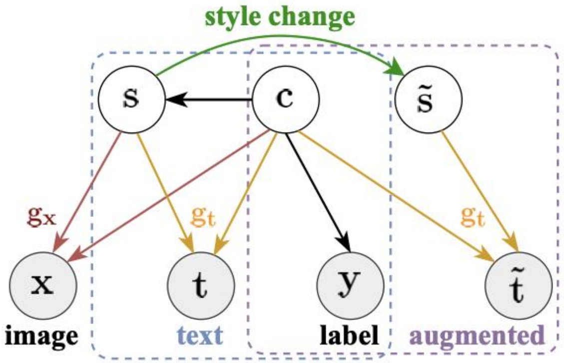
CLAP: Isolating Content from Style through Contrastive Learning with Augmented Prompts

Yichao Cai, Yuhang Liu, Zhen Zhang, and Javen Qinfeng Shi

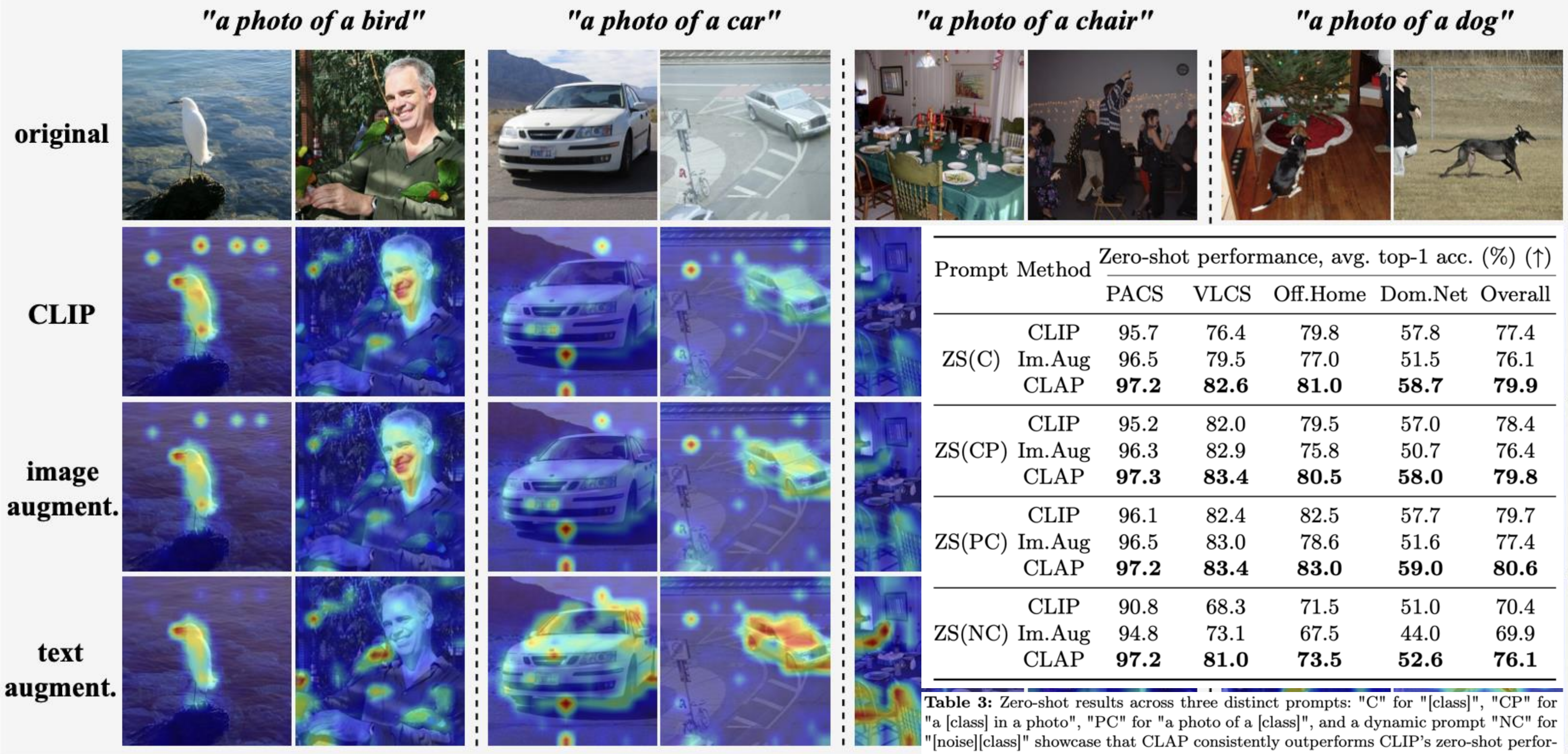
AIML, The University of Adelaide, Adelaide SA 5000, Australia
{yichao.cai,yuhang.liu01,zhen.zhang02,javen.shi}@adelaide.edu.au



(a) Image augmentation



(b) Text augmentation



Prompt	Method	PACS	VLCS	Off.Home	Dom.Net	Overall
		Zero-shot performance, avg. top-1 acc. (%) (↑)				
ZS(C)	CLIP	95.7	76.4	79.8	57.8	77.4
	Im.Aug	96.5	79.5	77.0	51.5	76.1
	CLAP	**97.2**	**82.6**	**81.0**	**58.7**	**79.9**
ZS(CP)	CLIP	95.2	82.0	79.5	57.0	78.4
	Im.Aug	96.3	82.9	75.8	50.7	76.4
	CLAP	**97.3**	**83.4**	**80.5**	**58.0**	**79.8**
ZS(PC)	CLIP	96.1	82.4	82.5	57.7	79.7
	Im.Aug	96.5	83.0	78.6	51.6	77.4
	CLAP	**97.2**	**83.4**	**83.0**	**59.0**	**80.6**
ZS(NC)	CLIP	90.8	68.3	71.5	51.0	70.4
	Im.Aug	94.8	73.1	67.5	44.0	69.9
	CLAP	**97.2**	**81.0**	**73.5**	**52.6**	**76.1**

Table 3: Zero-shot results across three distinct prompts: "C" for "[class]", "CP" for "a [class] in a photo", "PC" for "a photo of a [class]", and a dynamic prompt "NC" for "[noise][class]" showcase that CLAP consistently outperforms CLIP's zero-shot performance across all datasets, whereas image augmentation exhibits mixed outcomes.

(c) CAM visualization examples

Is it always possible to remove all correlations (i.e. getting a complete causal graph/model)?

Would (remaining) correlation(s) in a partial causal model be harmful?

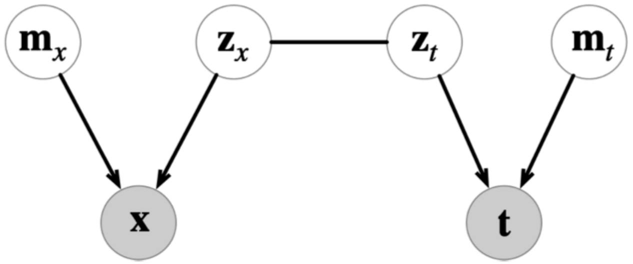


Figure 1. Illustration of the proposed latent partial causal model: The latent space is partitioned into $(\mathbf{m}_x, \mathbf{z}_x)$ and $(\mathbf{m}_t, \mathbf{z}_t)$, where $\mathbf{m}_x$ and $\mathbf{m}_t$ represent modality-specific latent variables. An undirected edge between $\mathbf{z}_x$ and $\mathbf{z}_t$ is employed to model latent shared patterns. The observations $\mathbf{x}$ (*e.g.*, images) and $\mathbf{t}$ (*e.g.*, text) are generated by two distinct generative processes, $\mathbf{g}_x(\mathbf{m}_x, \mathbf{z}_x)$ and $\mathbf{g}_t(\mathbf{m}_t, \mathbf{z}_t)$, respectively.

Beyond DAGs: A Latent Partial Causal Model For Multimodal Learning

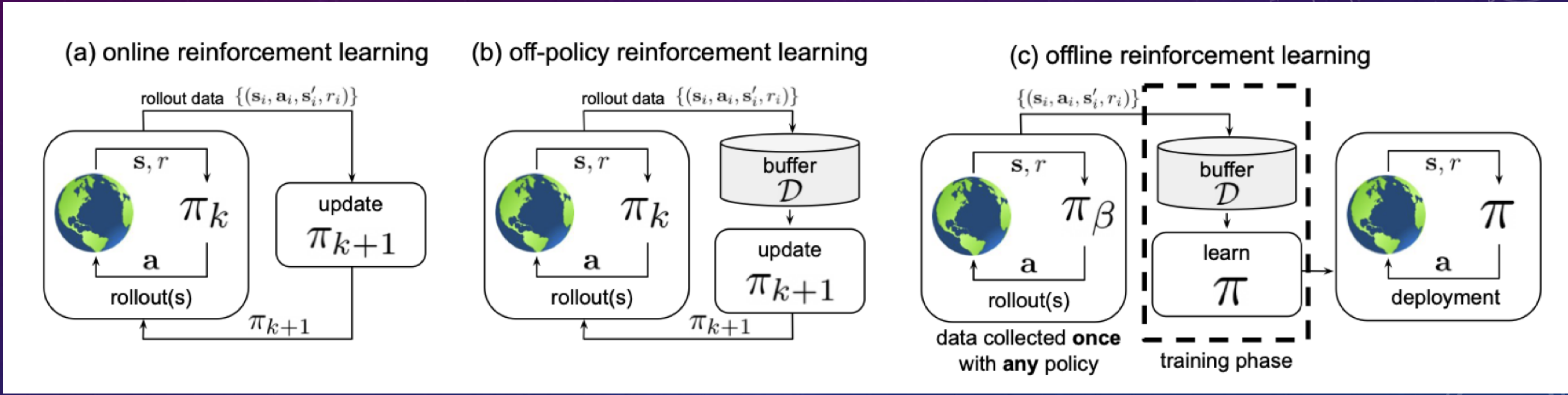
No need to remove all correlation. Partial causation can be ok

The Reverse Question

How to suggest a sequence (among infinite many) of interventions in a changing environment, given resources or budgets, to bring a desirable future outcome?

This differs from RL, even from causal RL. Both can help a bit, but are far from enough.

If just adapt to the environment, reinforcement learning (RL) is sufficient



Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems

Sergey Levine[1,2], Aviral Kumar[1], George Tucker[2], Justin Fu[1]
[1]UC Berkeley, [2]Google Research, Brain Team

MDP And POMDP

Definition 2.1 (Markov decision process)**.** The Markov decision process is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, d_0, r, \gamma)$, where $\mathcal{S}$ is a set of states $\mathbf{s} \in \mathcal{S}$, which may be either discrete or continuous (i.e., multi-dimensional vectors), $\mathcal{A}$ is a set of actions $\mathbf{a} \in \mathcal{A}$, which similarly can be discrete or continuous, T defines a conditional probability distribution of the form $T(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ that describes the dynamics of the system,[1] d_0 defines the initial state distribution $d_0(\mathbf{s}_0)$, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ defines a reward function, and $\gamma \in (0, 1]$ is a scalar discount factor.

We will use the fully-observed formalism in most of this article, though the definition for the partially observed Markov decision process (POMDP) is also provided for completeness. The MDP definition can be extended to the partially observed setting as follows:

Definition 2.2 (Partially observed Markov decision process)**.** The partially observed Markov decision process is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, T, d_0, E, r, \gamma)$, where $\mathcal{S}$, $\mathcal{A}$, T, d_0, r, and γ are defined as before, $\mathcal{O}$ is a set of observations, where each observation is given by $\mathbf{o} \in \mathcal{O}$, and E is an emission function, which defines the distribution $E(\mathbf{o}_t|\mathbf{s}_t)$.

Expected Return

Expected return

$$J(\pi) = \mathbb{E}_{\tau \sim p_\pi(\tau)} \left[\sum_{t=0}^{H} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

Probability for a trajectory

$$p_\pi(\tau) = d_0(\mathbf{s}_0) \prod_{t=0}^{H} \pi(\mathbf{a}_t | \mathbf{s}_t) T(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)$$

Policy Gradient

$$\frac{\partial J(\pi_\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} E_{\tau \sim \pi_\theta(\tau)}[R(\tau)]$$

$$= E_{\tau \sim \pi_\theta(\tau)}\left[\frac{\partial \lg \pi_\theta(\tau)}{\partial \theta} \cdot R(\tau)\right]$$

$$= E_{\tau \sim \pi_\theta(\tau)}\left(\sum_t \underline{\frac{\partial \lg \pi_\theta(a_t \mid s_t)}{\partial \theta}}\right) \cdot \underline{R(\tau)}$$

Algorithm 1 On-policy policy gradient with Monte Carlo estimator

initialize θ_0
for iteration $k \in [0, \ldots, K]$ **do**
  sample trajectories $\{\tau_i\}$ by running $\pi_{\theta_k}(\mathbf{a}_t|\mathbf{s}_t)$ ▷ each τ_i consists of $\mathbf{s}_{i,0}, \mathbf{a}_{i,0}, \ldots, \mathbf{s}_{i,H}, \mathbf{a}_{i,H}$
  compute $\mathcal{R}_{i,t} = \sum_{t'=t}^{H} \gamma^{t'-t} r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$
  fit $b(\mathbf{s}_t)$ to $\{\mathcal{R}i, t\}$ ▷ use constant $b_t = \frac{1}{N} \sum_i \mathcal{R}i, t$, or fit $b(\mathbf{s}_t)$ to $\{\mathcal{R}i, t\}$
  compute $\hat{A}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}) = \mathcal{R}_{i,t} - b(\mathbf{s}_t)$
  estimate $\nabla_{\theta_k} J(\pi_{\theta_k}) \approx \sum_{i,t} \nabla_{\theta_k} \log \pi_{\theta_k}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t}) \hat{A}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$
  update parameters: $\theta_{k+1} \leftarrow \theta_k + \alpha \nabla_{\theta_k} J(\pi_{\theta_k})$
end for

Q Learning (Bellman Equation)

$$V^{\pi}(s) = \sum_{a} \pi(s,a) \sum_{s'} p(s'|s,a) \left(\underline{r_{ss'}^{a}} + \gamma \cdot \underline{V^{\pi}(s')} \right)$$

$$Q^{\pi}(s,a) = \sum_{s'} p(s'|s,a) \left[\underline{r_{ss'}^{a}} + \gamma \cdot \max_{a'} \underline{Q^{\pi}(s',a')} \right]$$

$$\text{or } \gamma \cdot V^{\pi}(s') . \text{ or } \sum_{a'} \pi(s',a') Q^{\pi}(s',a')$$

$$V^{*}(s) = \max_{a} \sum_{s'} p(s'|s,a) \left[r_{ss'}^{a} + \gamma \cdot V^{*}(s') \right]$$

$$Q^{*}(s,a) = \sum_{s'} p(s'|s,a) \left[r_{ss'}^{a} + \gamma \cdot \max_{a'} Q^{*}(s',a') \right]$$

Algorithm 2 Generic Q-learning (includes FQI and DQN as special cases)

initialize ϕ_0
initialize $\pi_0(\mathbf{a}|\mathbf{s}) = \epsilon\mathcal{U}(\mathbf{a}) + (1-\epsilon)\delta(\mathbf{a} = \arg\max_{\mathbf{a}} Q_{\phi_0}(\mathbf{s},\mathbf{a}))$ ▷ Use ϵ-greedy exploration
initialize replay buffer $\mathcal{D} = \emptyset$ as a ring buffer of fixed size
initialize $\mathbf{s} \sim d_0(\mathbf{s})$
for iteration $k \in [0, \ldots, K]$ **do**
  for step $s \in [0, \ldots, S-1]$ **do**
    $\mathbf{a} \sim \pi_k(\mathbf{a}|\mathbf{s})$ ▷ sample action from exploration policy
    $\mathbf{s}' \sim p(\mathbf{s}'|\mathbf{s},\mathbf{a})$ ▷ sample next state from MDP
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s},\mathbf{a},\mathbf{s}',r(\mathbf{s},\mathbf{a}))\}$ ▷ append to buffer, purging old data if buffer too big
  end for
  $\phi_{k,0} \leftarrow \phi_k$
  for gradient step $g \in [0, \ldots, G-1]$ **do**
    sample batch $B \subset \mathcal{D}$ ▷ $B = \{(\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}'_i, r_t)\}$
    estimate error $\mathcal{E}(B, \phi_{k,g}) = \sum_i \left(Q_{\phi_{k,g}} - (r_i + \gamma \max_{\mathbf{a}'} Q_{\phi_k}(\mathbf{s}',\mathbf{a}'))\right)^2$
    update parameters: $\phi_{k,g+1} \leftarrow \phi_{k,g} - \alpha \nabla_{\phi_{k,g}} \mathcal{E}(B, \phi_{k,g})$
  end for
  $\phi_{k+1} \leftarrow \phi_{k,G}$ ▷ update parameters
end for

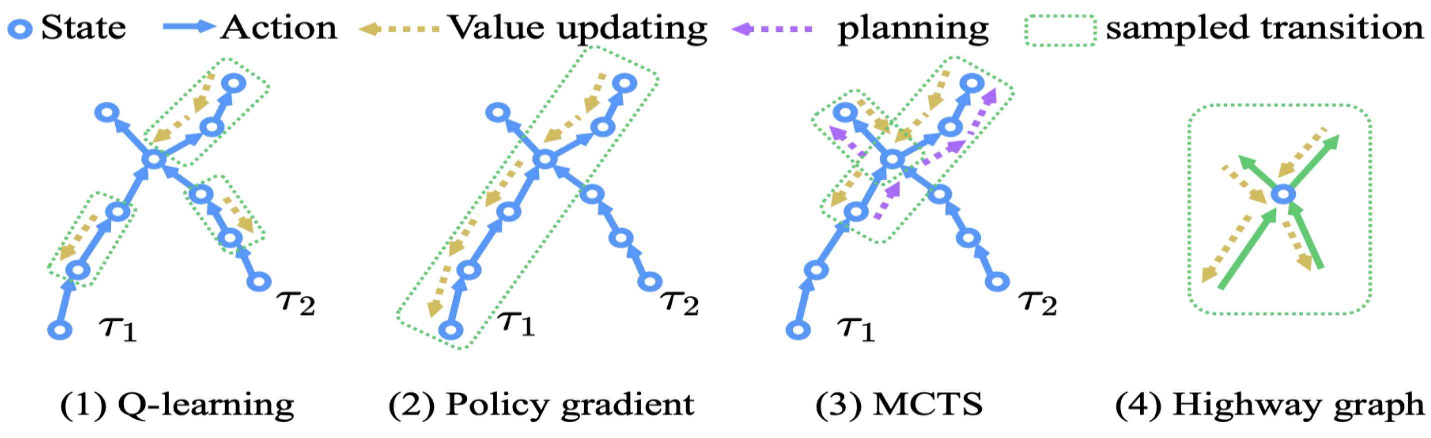
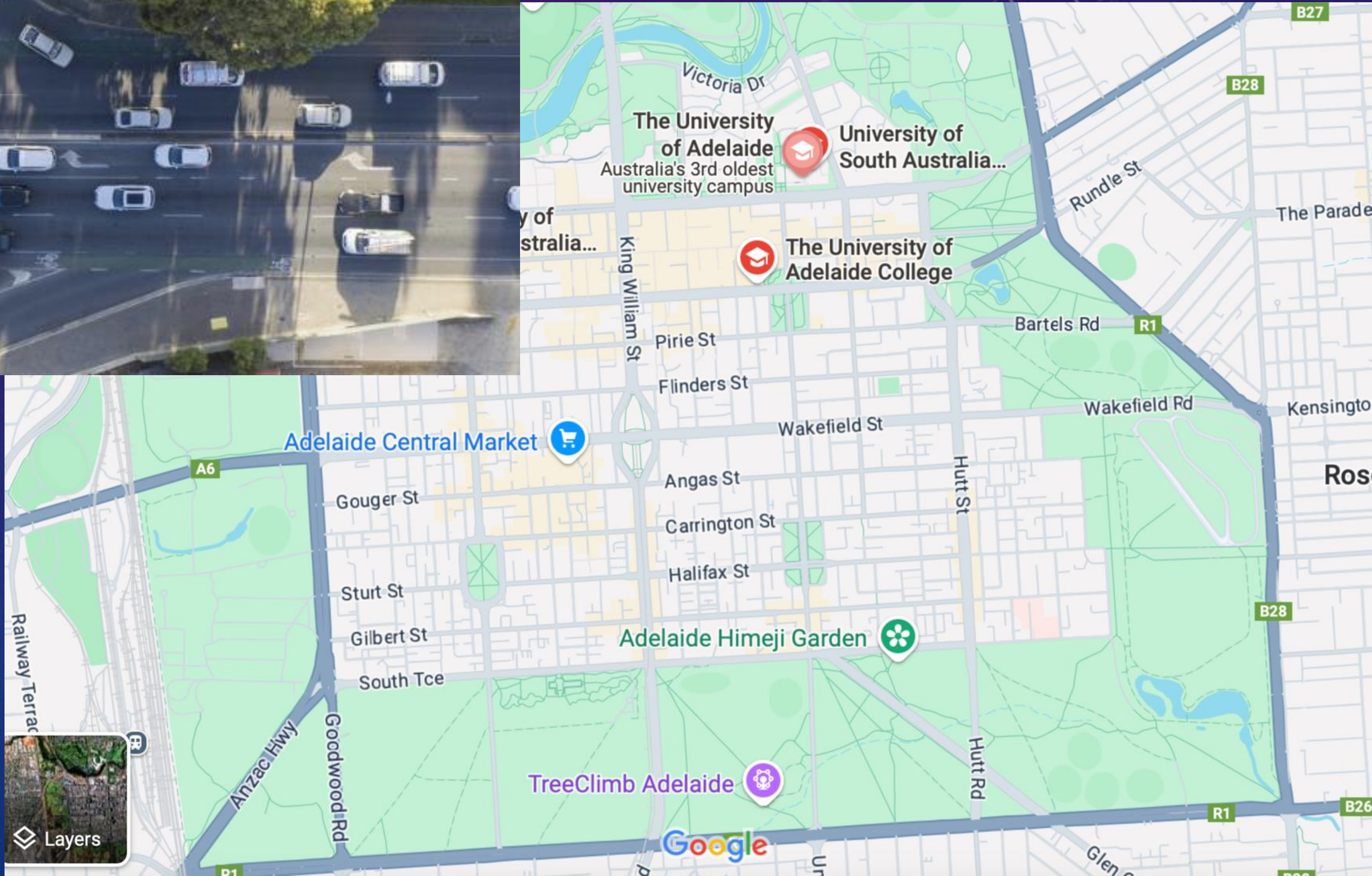


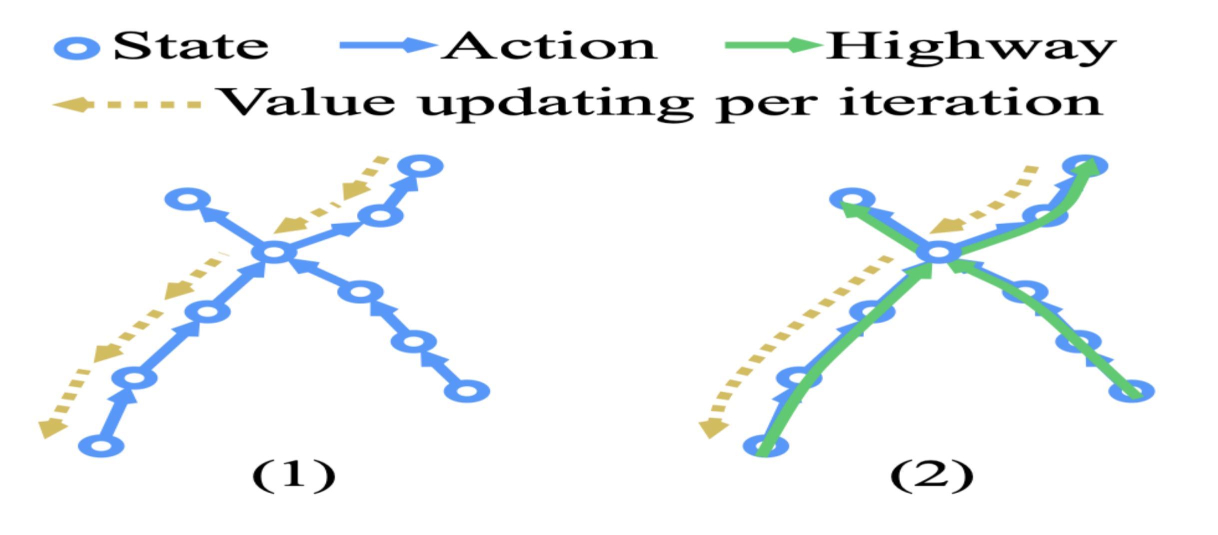
Figure 3: Types of value updating.

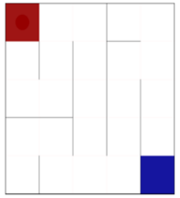


Victoria Dr
The University of Adelaide
Australia's 3rd oldest university campus
University of South Australia...
Rundle St
The Parade
The University of Adelaide College
King William St
Bartels Rd
R1
B27
B28
Pirie St
Flinders St
Wakefield St
Wakefield Rd
Kensingto
Adelaide Central Market
A6
Gouger St
Angas St
Carrington St
Hutt St
Halifax St
Sturt St
Gilbert St
Adelaide Himeji Garden
South Tce
Railway Terrac
Anzac Hwy
Goodwood Rd
Hutt Rd
TreeClimb Adelaide
Layers
Google
B26

Highway Graph to Accelerate Reinforcement Learning

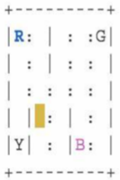
Zidu Yin, Zhen Zhang, Dong Gong, Stefano V. Albrecht, Javen Q. Shi



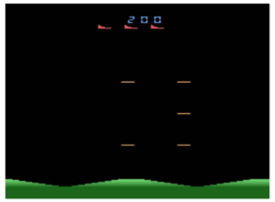


Simple maze (5 x 5)

Google research football

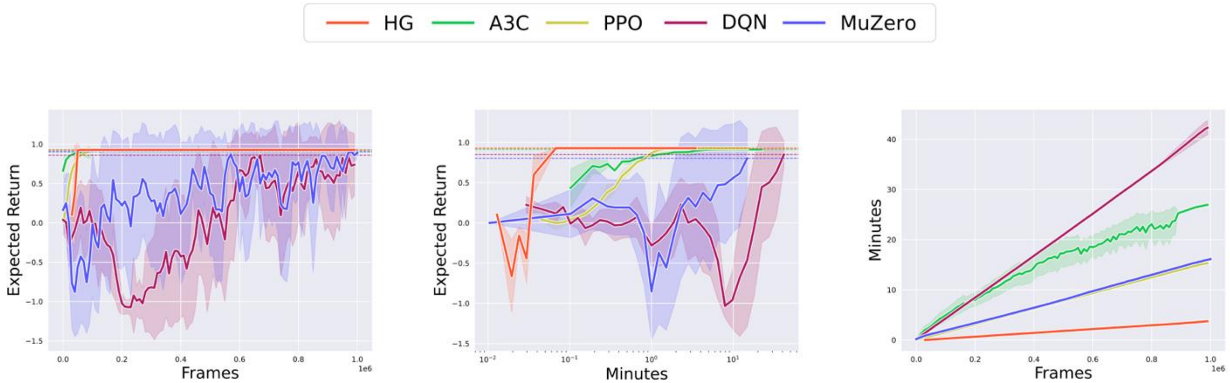


Toy text

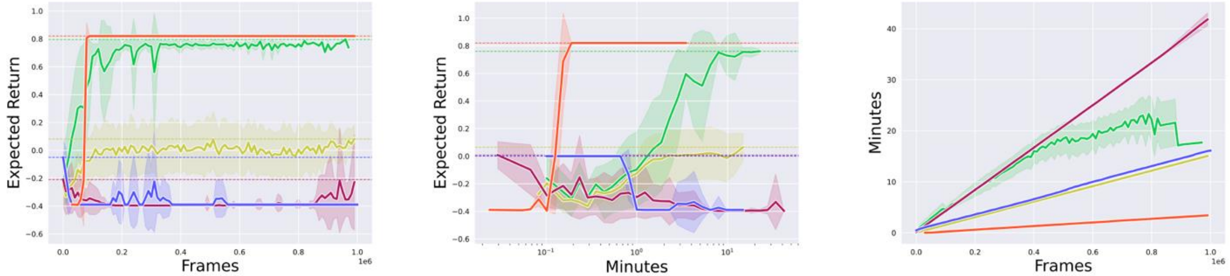


Atari games

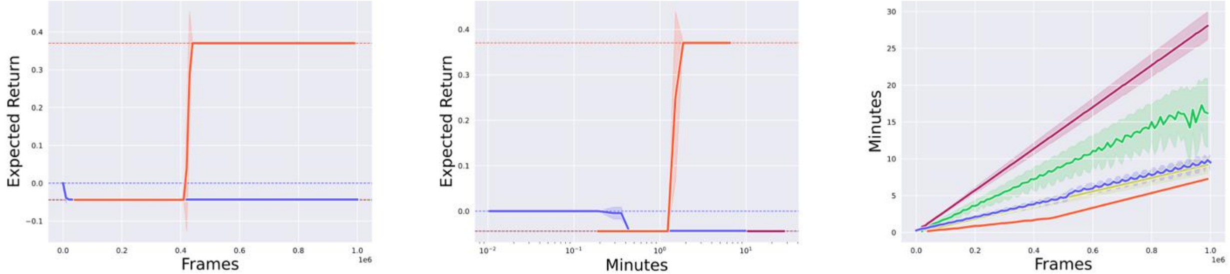
		Simple maze (5 x 5)	Google research football	Toy text	Atari games
Baselines	Converged time (min)	14.86 ± 2.90 (A3C)	79.81 ± 0.42 (IMPALA)	41.67 ± 0.18 (DQN)	251.26 ± 17.64 (MFEC)
	Eval. return	0.76 ± 0.03	1.01 ± 0.09	-88.18 ± 0.73	764.46 ± 239.06
Highway graph	Converged time (min)	0.15 ± 0.02 (99.0 speedup)	7.33 ± 0.10 (10.9 speedup)	3.82 ± 0.23 (10.9 speedup)	1.51 ± 0.10 (166.4 speedup)
	Eval. return	0.83 ± 0.00	1.03 ± 0.62	5.99 ± 0.74	2201.62 ± 918.52



(a) Simple maze 3x3



(b) Simple maze 5x5



(c) Simple maze 15x15

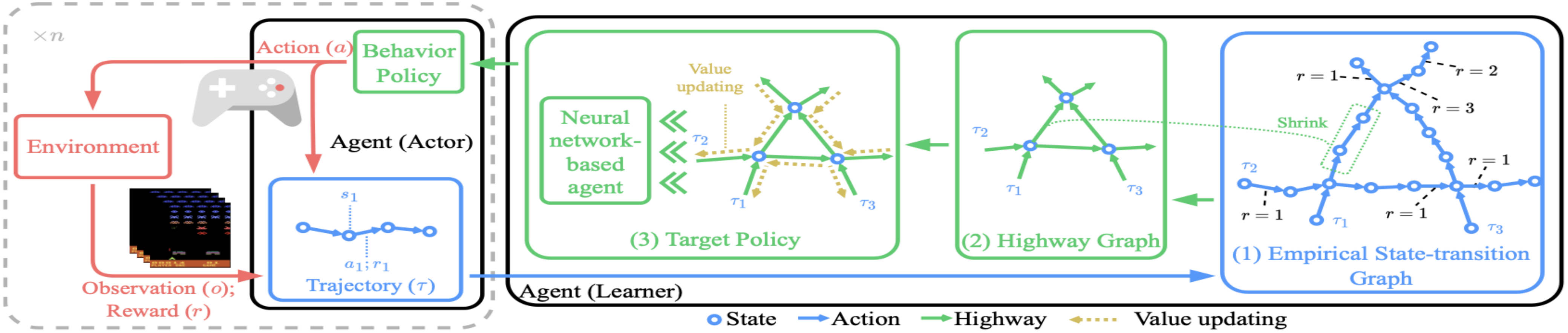
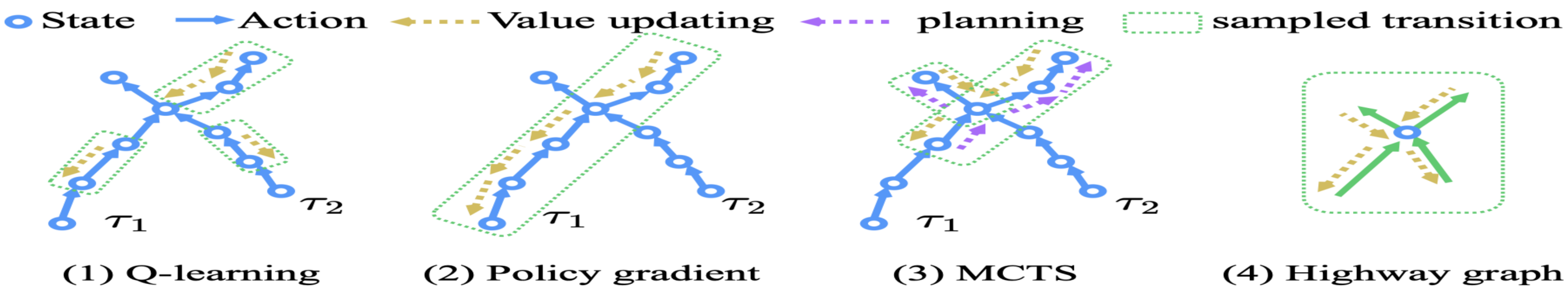
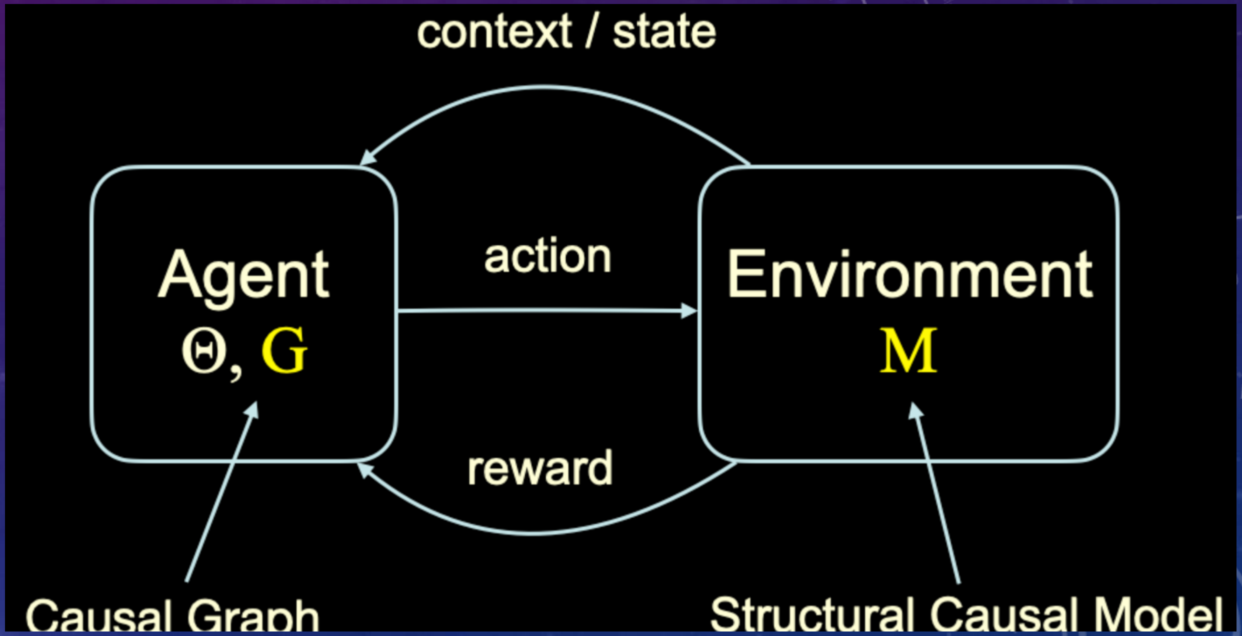


Figure 4: Overall data flow of our highway graph RL method. The actor (on the left) sends the sampled transitions by the behavior policy to the learner (on the right) which (1) constructs the empirical state-transition graph with rewards (in Section 3.1); (2) converts the empirical state-transition graph to the corresponding highway graph (Section 3.2); (3) updates the value of state-actions in the highway graph by an improved value iteration algorithm and re-parameterize the highway graph to a neural network-based agent as the new behavior policy (Section 3.3).



If wish to proactively change, beyond merely adapting to the environment

- Casual RL https://crl.causalai.net
- Dynamic Treatment Regime

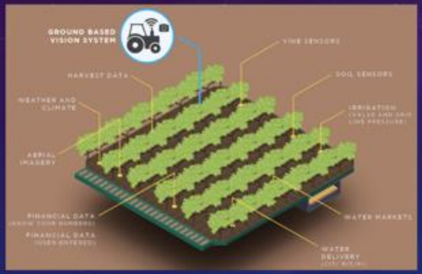


Both can help, but are not sufficient to answer the reverse question, which we are working.

AI's Impact on Society: Inspirational AI Successes

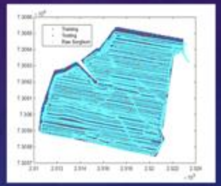
- Agricultural productivity
- Disaster response
- Energy material discovery& Decarbonization
- Wave and wind energy harvesting
- Mining exploration
- Health
- ...

Viticulture&Wine (VitiVisor)

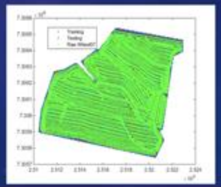


Crop Yield & Soil Variability

Sorghum Yield



Wheat Yield



Predict Adjusted Yield via Genotype

Australian Grains Technologies (AGT)

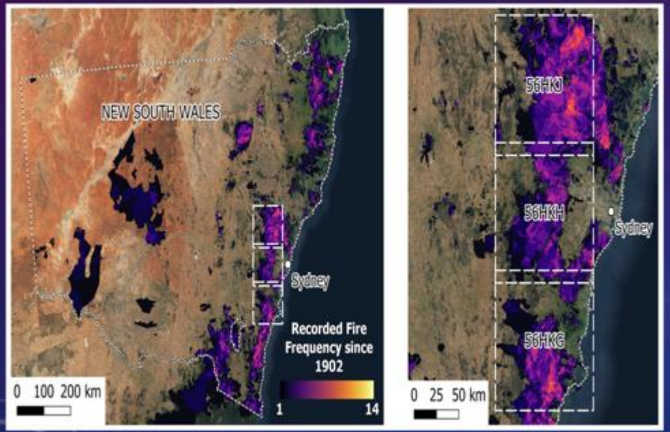
300,000 gene varieties (called lines)

- > 6 billion marker data points

We achieved

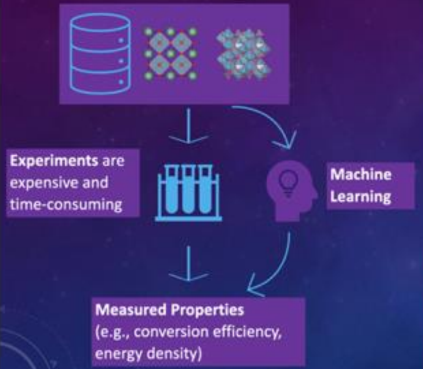
- Using only 302 genetic markers (others use 17k markers) beating the current STOA
- ~ 7% testing error on the predicted adjusted yield

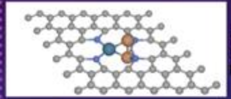
Predict bushfire burn areas/fire scars



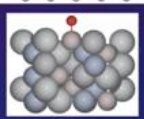
Fuel load & risk assessment mobile app

ENERGY-MATERIAL DISCOVERY & DECARBONIZATION

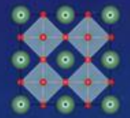




Few-atom catalysts for CO_2 recycling[1]



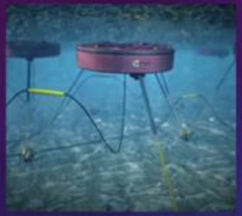
High-entropy alloys for lithium batteries



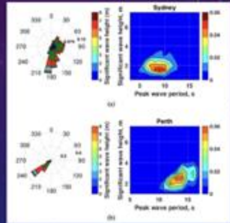
Perovskite solar cells

1. Li et al. Nano Energy 115 (2023) 108695

WAVE AND WIND ENERGY HARVESTING

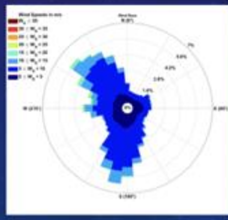


Wave energy harvesting via continuous action RL and optimization (Carnegie Clean Energy)



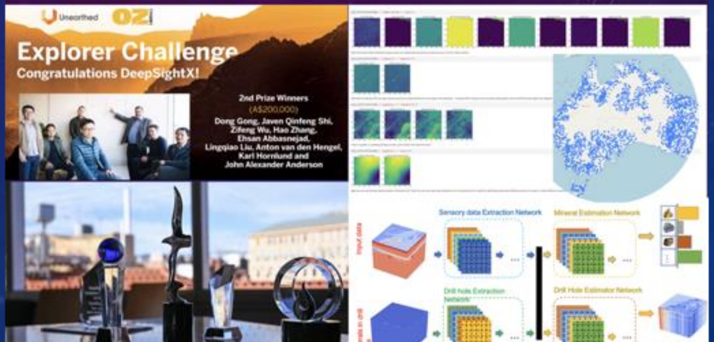
Short-term wind speed forecasting

Wind Turbine Power output Prediction



https://scienceexchange.caltech.edu/topics/sustainability/wind-energy-advantages-disadvantages

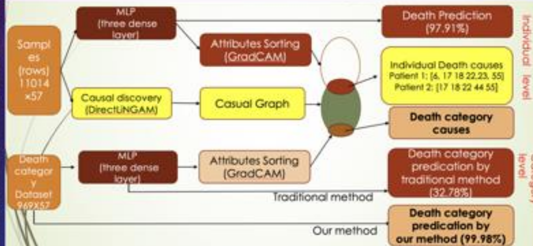
Predict Where to Drill



Predict newborn survival and leading causes

Women's and Children's Hospital

https://www.wch.sa.gov.au
https://www.abc.net.au/news/2022-09-27/new-womens-and-childrens-hospital-to-be-at-thebarton-barracks/101476264



AI’s Impact On Society: The Future Of Work?

- Will AI and robotics replace jobs?
- Humans are afraid of losing income (not jobs)
- Why income is linked with jobs, in the past?
- The Future of Life