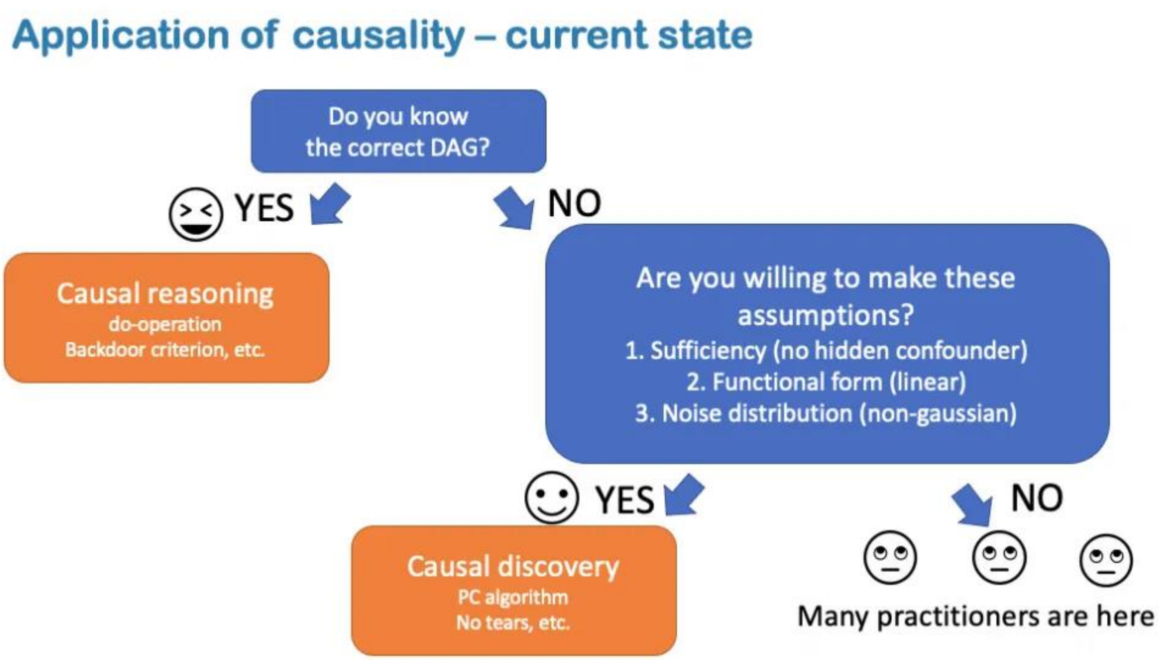
3.2 Causal Deep Learning
ZhaoZhi Qian (@QianZhaoZhi) & Jeroen Berrevoets (@J_Berrevoets)
Research Pillar: Causal deep learning
https://www.vanderschaar-lab.com/causal-deep-learning/

Causality: the study of cause and effect

- Philosophical schools for causality
  - Structural equations framework (Pearl)
    - https://ftp.cs.ucla.edu/pub/stat_ser/r370.pdf
  - Potential outcomes framework (Rubin)
    - https://www.causalconversations.com/post/po-introduction/
  - Probabilistic causation (Reichenbach)
    - https://plato.stanford.edu/entries/causation-probabilistic/#:~:text=Reichenbach%20says%20that%20the%20common,result%20from%20a%20causal%20relationship.
  - Counterfactual theory (Lewis)
    - https://plato.stanford.edu/entries/causation-counterfactual/#:~:text=1.-,Lewis's%201973%20Counterfactual%20Analysis,would%20have%20happened%20without%20it.
  - …
- **Application of causality: current state**



  - Bayesian network:
    - A connecting B: Causality
  - Estimating conditional probability
    - We do not know most of the correct DAG
      - No knowledge

            - Too many variables
                - Invisible
    - Automated causal discovery
        - Stringent assumptions
            - Sufficiency (no hidden confounder)
            - Functional form (linear)
            - Noise distribution (non-gaussian)
        - Thus, for most practitioners, casual theory cannot be applied
            - Causal deep learning

What is casual deep learning?

- Ladder causation
    - Association:
        - Deep learning methods
        - Powerful and expressive
        - Unable to generalization
    - Interventions
        - Lot of rooms between intervention and association
            - Casual Deep Learning may hello
                - E.g. improving **association** robustness -> domain generationlisation -> to motivate **correct** objectives -> to **intervention**
        - Mostly causality
        - Principled, but many assumptions
        - e.g. Clinical trials
    - Counterfactual

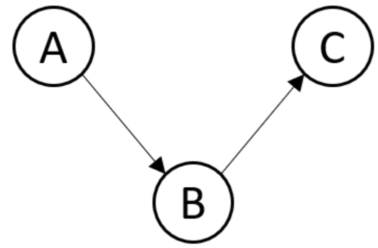
Causal deep learning

- Use tools and concepts from causality to inform deep learning
    - Loss function, regularization
    - Inductive bias, architecture design
    - Auxiliary task, self-supervision
    - Theoretical foundation and formalism
- Focus on empirically verifiable tasks and metrics
    - Real data validation
    - Quantitative model comparison
    - Iterative model development
    - Beyond one-dataset-one-task setting
        - Different environments (demographics, geographical locations)
        - Different actions (variation in practice)
        - Different tasks (multiple endpoints, labels)

How can we CDL?

- Moving up the ladder
    - Improving association (CASTLE)
    - Domain generalization (CAS)
    - Correct goal definition (MCM)

        - Interventions (DECAF)
- Preliminaries on graphs
    - Considering bayesian network (Pearl)
        - Causal structure,
        - A can impact B and then C



    - Some properties:
        - Resembles p(A,B,C) but much more compact
        - A does not give us information on C once we know BL
            - $A \perp C \mid B$
    - When the network is causal, this happens:
        - Knowing this graph, can mutilate it by intervening on a variable
        - Observe a new distribution $p^{do(B:=b)}(A,B,C)$
    - BNs gives structure, causal graphs gives knowledge – both can be used
- Association using structures (BNs)
    - Using structuring to improve supervised learning?
        - Regularization
            - L_2 makes sure the weights of a model are small
            - L_1 encourages sparse weights
        - Only one-size-fits all solutions – smarter solution:
            - Just like L_1 and L_2, adda differentiable regularizer to a loss function:
                - $tr(e^{A \circ A}) - d$
            - Which is 0 when A is DAG
                - https://papers.nips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html
                - http://proceedings.mlr.press/v108/zheng20a/zheng20a.pdf

Association using structure (BNs)

- CASTLE
    - **CASTLE: Regularization via Auxiliary Causal Graph Discovery**
    - https://arxiv.org/pdf/2009.13180.pdf
- Set A to be based on the input layers of a neural network
- With DAG,incorporate the regulariser as a reconstruction based regulariser which only reconstructs adjacent features

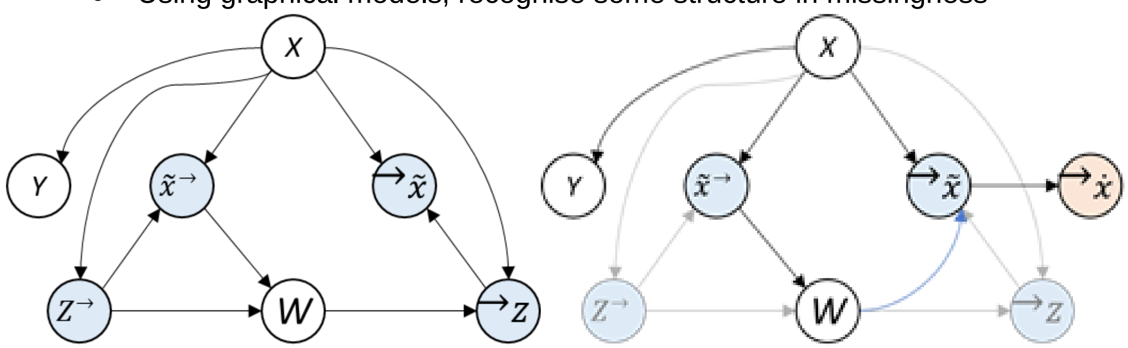
Domain generalization using causal graphs

- CAS
- Causal structure should remain constant over domains
    - Exploiting:
        - Given a DAG (learned or assumed), learn a model in the training domain, respecting that DAG
        - Model then at an advantage in the target domain – same DAG governing
- **Improving Model Robustness Using Causal Knowledge**
    - https://arxiv.org/abs/1911.12441

- **Exploiting Causal Structure for Robust Model Selection in Unsupervised Domain Adaptation**
  - https://ieeexplore.ieee.org/document/9503312

Defining correct goals using causal graphs

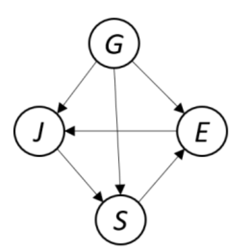
- MCM
- Imputing missing data when estimating treatment effects:
  - Missing data may cause treatment selection – $Z^{\rightarrow}$
    - Transfusing blood if blood type is not measured
    - Decide to transfuse Type O
  - treatment may cause missingness ${}^{\rightarrow}Z$
    - Many treatment decisions require blood test to be performed
    - Much less likely when not getting that treatment
  - Consider a binary vector $Z \in \{0,1\}^d$ indicating when a variable in X is missing, then above realization leads to two factors in Z
  - Using graphical models, recognise some structure in missingness



    - Transformation of structure
      - X is moving directly to X out into W
      - $X^{\rightarrow}$: should not impute
      - ${}^{\rightarrow}X$: should impute
      - Collateral structure, no influence of ${}^{\rightarrow}X$
        - Only conditioning will influence
    - Because of the structure, conditioning on entire covariate set, either imputed or not imputed, introduces bias in the model
      - **To Impute or not to Impute? Missing Data in Treatment Effect Estimation**
        - https://arxiv.org/abs/2202.02096

Using interventions to generate fair data ~DECAF

- Interventions to change a distribution;
  - Distribution:
  - *G* = gender
  - *E* = education
  - *J* = job
  - *S* = salary



- Why is gender influencing salary?
  - Dataset including the variables in the graphical model

- DECAF: propose method mitigate the issue
    - **DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks**
        - https://arxiv.org/abs/2110.12884

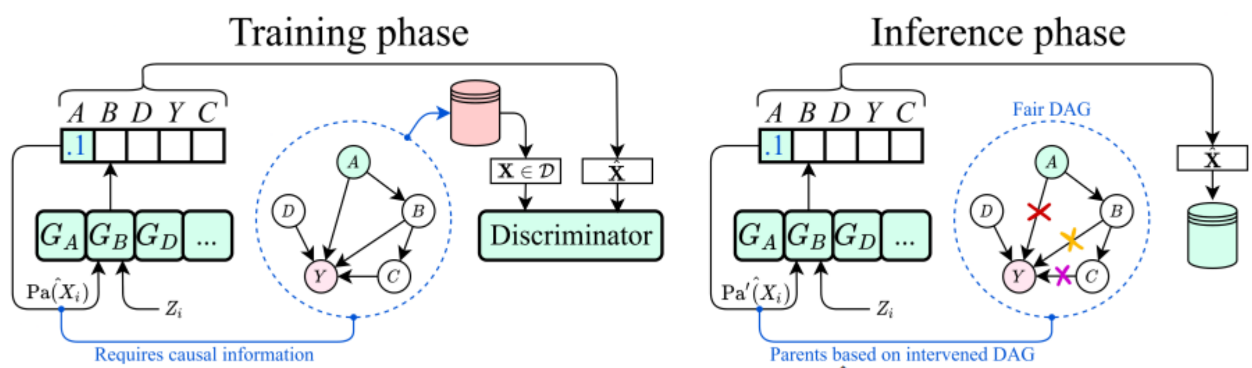


Figure 2: **Architecture.** *Training phase*— Each component in $\hat{\mathbf{X}}$ is generated sequentially as a function (where the function is that component's generator G_i) of the component's parents. Parental knowledge is provided by the DAG governing the data. *Inference phase*— As the component-wise generation of the generator network is independent of the DAG governing the data, we can easily replace (or intervene on) the DAG governing parental information. The resulting synthetic data (right) will be governed by the intervened DAG. *FTU is achieved by removing: ✗; DP: ✗✗✗; e.g. CF when* $R = C$*: ✗✗*.

- DEAF trains a generative model for each feature
    - Respecting the topological order of a causal graph
        - Input and output
    - At inference phase, can simply intervene on the graph and sample new data that respects the new graph

Summary

Using causality we can generate synthetic fair data for downstream models	**DECAF generates fair data** from intervened causal graphs	▲ Using **interventions**
Causality can yield new insights into architecture and learning objectives	**Causal structure dictates** what to impute and what not to impute	▲ Motivating **correct** objectives
DAGs remain constant across domains	**CAS** and other methods exploit this to **enhance generalization**	▲ Domain generalization
Discussed how we can use causality to regularize	**CASTLE regularizes** the hypothesis space with auxiliary DAG discovery	▲ Improving **association**

Q&A

- Temporal data to the model

    - In short terms, recurrent neural structure – be cautious
    - Needs to be dealt separately in a longer term
        - Causality in a time series is submitting
- Data size matters?
    - Larger sample sizes with assumptions favoring
    - Proper causal analysis with larger size will allow higher confidence
    - Causal DL analysis should be fine with small sample size
- GAN:
    - Allowing new regularization function in computation itself
    - Enforcing the theorem
    - Using some sort of invitation that enforces independence between treatment and imputed values