Causal Inference & Reasoning with Causal Bayesian Networks
Neyman-Rubin Framework

- *Potential Outcome Framework*: for each unit $k$ and each treatment $i$, there is a potential outcome on an attribute $U, U_{ik}$, representing the outcome unit $k$ would display if it received treatment $i$.

- Treatment effect can be defined in terms of potential outcome.
  
  e.g., a drug (indexed by 1) has an effect on a patient $k$ vs. a placebo (indexed by 0) on recovery if the potential recovery $U_{1k} \neq U_{0k}$.

- A counterfactual framework – for each unit $k$, only one $U_{ik}$ can be observed.
Example: Fisher’s Tea-Tasting Lady

- Treatment variable $Z$ – 70 possible values, corresponding to 70 possible arrangements of 8 cups of tea.

- Potential responses: $U_{Z=i}$, $1 \leq i \leq 70$.

- Fisher’s null hypothesis: no effect of $Z$ on $U$, i.e., (the probability distribution of) $U_{Z=i}$ does not vary with $i$.

- Upon complete randomization, the null hypothesis implies that $P(U = Z) = 1/70$. 
Potential response $Y_{X=x}$ can be interpreted as the value $Y$ would take given that $X$ is intervened to take value $x$. Let us write $P_{X=x}(Y)$ to represent the probability distribution of $Y$ given that $X$ is intervened to take value $x$.

Multivariate generalization: $P_{X=x}(Y)$.

Causal knowledge is useful for estimating such post-intervention quantities.

A well-developed machinery for this purpose is *Causal Bayesian Networks*. 
Directed Acyclic Graphs (DAGs)

- A set of nodes and a set of directed edges between nodes.
  - standard terminology: parent/child, ancestor/descendant, adjacency, path, directed path, etc.

- No directed cycles, i.e., no directed path from one variable back to itself.
Probabilistic Interpretation of DAGs

- Take vertices as random variables.
- *(Local) Markov condition*: every variable is probabilistically independent of its non-descendants given its parents.

![Diagram of a directed acyclic graph (DAG) with variables X1 to X5 and corresponding independence statements.

\[ I(X_1, X_2) \]
\[ I(X_2, \{X_1, X_3\}) \]
\[ I(X_3, \{X_2, X_4, X_5\} \mid X_1) \]
\[ I(X_4, X_3 \mid X_1, X_2) \]
\[ I(X_5, \{X_1, X_2, X_3\} \mid X_4) \]
Given a DAG $G$, choose any ancestral order of the variables: $X_1, \ldots, X_n$, s.t. if $i<j$, then $X_j$ is not an ancestor of $X_i$.

For a joint distribution $P$ that is Markov to (i.e., satisfies the Markov condition with) $G$, we can factorize it as follows:

$$P(X_1, \ldots, X_n) = P(X_1) P(X_2|X_1) \ldots P(X_n|X_1, \ldots, X_{n-1})$$

Chain Rule

$$= \prod_i P(X_i | \text{Pa}(X_i))$$

Exploit Local Markov Property

Conversely, factorization implies the Markov condition.
Example

\[ P(X_1, X_2, X_3, X_4, X_5) \]
\[ = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)P(X_5|X_1, X_2, X_3, X_4) \]
\[ = P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_1, X_2)P(X_5|X_4) \]
Because of the factorization, to specify a joint distribution of the variables in a DAG, it suffices to specify the distribution of each variable conditional on its parents: $P(X_i \mid \text{Pa}(X_i))$.

A DAG together with these local conditional distributions is known as a *Bayesian Network*. 
Example of a Bayesian Network

P(F|C) P(R|C,A)
P(W)
P(A|W)P(C|W)

P(W)
P(C|W)
P(A|W)
P(F|C)
P(R|C,A)
Causal Interpretation of DAGs

- parent/child $\rightarrow$ direct cause/effect (relative to the set of variables in the graph).

![Diagram showing causal relationships between Smoking, Radon, Yellow Teeth, Lung Cancer, and Weight Loss]
Surgical Intervention

Intervention on a cause

- Smoking
- Radon

Intervention on an effect

- Yellow Teeth
- Lung Cancer
- Weight Loss
Causal Bayesian Network

- Recall the factorization of a Bayesian network:
  \[ P(V) = \prod_{Y \in V} P(Y | Pa_G(Y)) \]

- A causal Bayesian network over \( V \) is a DAG \( G \) and a set of distributions \( \{P_{X=x}(V) | X \subseteq V\} \) that satisfies the following intervention postulate:
  \[ P_{X=x}(V) = \prod_{Y \in V \setminus X} P(Y | Pa_G(Y)) \]
  for values of \( V \) consistent with \( X=x \), where
  \[ P(Y | Pa_G(Y)) = P_{\emptyset}(Y | Pa_G(Y)) \]
  denoting distributions in the absence of interventions.

- A causal Bayesian network provides a link between post-intervention distributions and pre-intervention distributions, the latter of which can be estimated from passive observations.
The intervention postulate assumes the following:

- Interventions are effective (arrow-breaking) and local (no fat hand);
- The post-intervention joint distribution factorizes according to the post-intervention causal DAG.
- Modularity: the local conditional distribution for a variable $Y$ is invariant under interventions of other variables (including $Y$'s causes);

The intervention postulate forms the basis of reasoning about consequences of interventions with the help of a causal diagram.
Causal Reasoning

\[ P_{s=yes}(V) = P_{s=yes}(S) P_{s=yes}(YT | S) P_{s=yes}(R) P_{s=yes}(LC | S, R) P_{s=yes}(WL | LC) \]
\[ = P(YT | S) P(R) P(LC | S, R) P(WL | LC), \text{ for values consistent with } S=yes \]

From which we can derive, e.g., \( P_{s=yes}(LC) = P(LC | S = yes) \)

\[ P_{YT=yes}(V) = P_{YT=yes}(YT) P_{YT=yes}(S) P_{YT=yes}(R) P_{YT=yes}(LC | S, R) P_{YT=yes}(WL | LC) \]
\[ = P(S) P(R) P(LC | S, R) P(WL | LC), \text{ for values consistent with } YT=yes \]

From which we can derive, e.g., \( P_{YT=yes}(LC) = P(LC) \)
Recall: Simpson’s Reversal

\[ P_{X=1}(Y=1) \]
\[ = P_{X=1}(Y=1 \mid X=1) \]
\[ = P_{X=1}(Y=1 \mid X=1, Z=1) P_{X=1}(Z=1 \mid X=1) \]
\[ + P_{X=1}(Y=1 \mid X=1, Z=0) P_{X=1}(Z=0 \mid X=1) \]
\[ = P_{X=1}(Y=1 \mid X=1, Z=1) P_{X=1}(Z=1) \]
\[ + P_{X=1}(Y=1 \mid X=1, Z=0) P_{X=1}(Z=0) \]
\[ = P(Y=1 \mid X=1, Z=1) P(Z=1) \]
\[ + P(Y=1 \mid X=1, Z=0) P(Z=0) \]
Interventionist Interpretation of Arrows

- Take the notion of “direct cause” as primitive?

- Assuming there is a causal Bayesian network for $V$, we can derive that there is a uniquely minimal one: $X$ is a parent of $Y$ iff.
  \[
  P_{X=x_1, Z=z}(Y) \neq P_{X=x_2, Z=z}(Y)
  \]
  for some $x_1 \neq x_2$ and $z$, where $Z = V \setminus \{X, Y\}$.

- It is natural to refer to this minimal causal DAG the true causal structure for $V$. This agrees well with the standard interventionist interpretation of direct cause.
A special case of the intervention postulate obtains when there is no intervention:

$$P_{\emptyset}(V) = P(V) = \prod_{Y \in V} P(Y \mid \text{Pa}_G(Y))$$

Equivalently the pre-intervention distribution satisfies the Markov condition with the causal DAG $G$.

This is known as the *Causal Markov Condition* (CMC): every variable is probabilistically independent of its non-effects (non-descendants in the causal graph) given its direct causes (parents in the causal graph).
Intuitions

- No causal relation or common cause, no dependence: e.g., I(S, R).
- Common cause screens off collateral effects: e.g., I(YT, LC | S).
- Intermediate cause screens off remote cause: e.g., I(R, W | LC).
- The CMC is a generalization of these intuitions.
The conditional independence relations explicitly stated in the CMC can entail others (in virtue of the probability calculus).

These entailed conditional independence relations are in principle testable with observational data.

A graphical criterion called *d-separation* captures precisely the (conditional) independence relations entailed by a causal structure.
Given an path \( u \), \( X \) is a collider on \( u \) if arrows collide at \( X \) on \( u \). Otherwise it is a non-collider on \( u \).

The notion of collider/non-collider is relative to a path! A node can be a collider on one path and a non-collider on another.
“No gas” and “Battery dead” are independent.

However, after we learn that “Car dead” is true, they become dependent.

This structure, which we call unshielded collider, is very important in causal discovery from observational data.
In a DAG $G$, a path between $X$ and $Y$ is active given $Z$ iff.

(i) every non-collider on the path is not in $Z$;
(ii) every collider on the path has a descendant in $Z$.

$X$ and $Y$ are $d$-separated by $Z$ if no path between a variable in $X$ and a variable in $Y$ is active given $Z$. Otherwise they are $d$-connected by $Z$.

Theorem: $X$ and $Y$ are $d$-separated by $Z$ in a DAG $G$ if and only if $G$ entails (via the Markov condition) that $X$ and $Y$ are independent conditional on $Z$. 
X and Y are d-separated (by empty set)
X and Y are d-separated by \( \{Z\} \)
X and Y are d-separated by \( \{Z, V\} \)
X and Y are d-connected by \( \{U, S\} \)
The CMC reformulated: \( d \)-separation \( \Rightarrow \) conditional independence. But it does not say anything about \( d \)-connection.

- *Causal Faithfulness Condition (CFC):* All conditional independence relations that hold (in the pre-intervention population) are *entailed* by (the Markov condition of) the causal graph.

- In other words, \( d \)-connection implies dependence.

- CMC and CFC: conditional independence \( \Leftrightarrow \) \( d \)-separation.
Proposition 1: in a DAG $G$, if $X$ is not a descendant of $Y$ and there is no directed path from $X$ to $Y$, then $X$ and $Y$ are $d$-separated by empty set or by some non-descendants of $X$.

Proposition 2: in a DAG $G$, if there is a directed path from $X$ to $Y$, then $X$ and $Y$ are $d$-connected given any set of non-descendants of $X$ (including the empty set).

Hints: proposition 1 may be used to argue that assuming CMC, Suppes-type definition of genuine causal relevance gives sufficient conditions; proposition 2 may be used to argue that assuming CFC, the definition gives necessary conditions.
Causal Inference under CMC & CFC

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex 1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Ex 2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>Ex 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Ex 4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>

Data

A is independent of B
A is NOT independent of B conditional on C

Independence Facts

A is d-separated from B
A is NOT d-separated from B by C

D-separation features
Bad News: Markov Equivalence

- Different DAGs may share the exact same d-separation structures. They are called *Markov Equivalent*.

- In this sense, CMC and CFC do not reduce causality to probabilities.

An Equivalence Class of DAGs

A is d-separated from B
A is NOT d-separated from B by C
...

D-separation features
Example
Markov equivalent networks may share common features.

Even though we cannot distinguish between Markov equivalent causal DAGs based on patterns of conditional independence, their common substructure can be inferred.
A Pattern (or PDAG) can contain both directed edges (that represent the common directions shared by the equivalent DAGs) and undirected edges (that indicate ambiguity about the direction).
An Equivalence Class of DAGs represented by a Pattern

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex 1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Ex 2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>Ex 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Ex 4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Independence Facts

A is independent of B
A is NOT independent of B conditional on C

D-separation features

A is d-separated from B
A is NOT d-separated from B by C

...
So far we have assumed that all variables in $V$ are observed. Given this assumption, algorithms are available that are sound and complete in extracting causal information from an oracle of conditional independence.

More realistically, we only observe a subset of $O \subseteq V$. We are only entitled to assume that there is some set of latent variables $L$, such that there is a causal Bayesian network over $O \cup L$, and the CMC and CFC hold of $O \cup L$.

Causal inference gets much more difficult and complicated in the presence of latent variables, but not hopeless.
A PAG is a graph that can contain four kinds of edges: o⎯o, o→, →, ↔, representing an (infinite) set of DAGs that entail the exact same observed conditional independence relations.

- \( A \circ \circ B \): Either \( A \) is a cause of \( B \), or \( B \) a cause of \( A \), or there is a latent common cause.
- \( A \circ \rightarrow B \): \( B \) is NOT a cause of \( A \)
- \( A \rightarrow B \): \( A \) is a cause of \( B \)
- \( A \leftrightarrow B \): There is a latent common cause of \( A \) and \( B \)
Since a PAG gives partial information about the underlying causal Bayesian network, not all post-intervention probabilities are identifiable based on a PAG.

But the partial information is often enough to identify some post-intervention probabilities.

Rules for when we can identify a post-intervention probability given a PAG are available.
Scheines’ Example: College Plan

- Real data set created by Sewell and Shah (1968);
- 10,138 instances from Wisconsin high schools;

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEX</td>
<td>male, female</td>
</tr>
<tr>
<td>IQ (intelligent quotient)</td>
<td>low, lower middle, upper middle, high</td>
</tr>
<tr>
<td>CP (college plan)</td>
<td>yes, no</td>
</tr>
<tr>
<td>PE (parental encouragement)</td>
<td>low, high</td>
</tr>
<tr>
<td>SES (socioeconomic status)</td>
<td>low, lower middle, upper middle, high</td>
</tr>
</tbody>
</table>
The PAG Learned from Data

- SES
- IQ
- SEX
- PE
- CP
Estimating Causal Effect

\[ P_{PE}(CP) = \sum_{SES} P(CP \mid PE, SES)P(SES) \]

\[ P_{PE=high}(CP = yes) = 0.517 \]
\[ P_{PE=low}(CP = yes) = 0.081 \]
A bunch of causal discovery algorithms are implemented in the Tetrad package: www.phil.cmu.edu/projects/tetrad

A fine online course if you want to know more but do not want to consult me any more: www.phil.cmu.edu/projects/csr

A single best reference in this field is
Pearl, J. (2000) *Causality: Models, Reasoning and Inference*

For computational aspects,