Causal analysis
THE BASIC PRINCIPLES

- **Causation**
  = encoding of behavior under interventions

- **Interventions**
  = surgeries on mechanisms

- **Mechanisms**
  = stable functional relationships
  = equations + graphs
WHAT'S IN A CAUSAL MODEL?

Oracle that assigns truth value to causal sentences:

Action sentences: \( B \) if we do \( A \).

Counterfactuals: \( \neg B \Rightarrow B \) if it were \( A \).

Explanation: \( B \) occurred because of \( A \).

Optional: with what probability?
There are 4 interesting points

1. It qualifies as a causal model because it contains the information to confirm or refute all action, counterfactual and explanatory sentences concerned with the operation of the circuit.
   For example, we can figure out what the output would be like if we set Y to zero, or if we change this OR gate to a NOR gate or if we perform any of the billions combinations of such actions.

2. Logical functions (Boolean input-output relation) is insufficient for answering such queries

3. These actions were not specified in advance, they do not have special names and they do not show up in the diagram. In fact, the great majority of the action queries that this circuit can answer have never been considered by the designer of this circuit.

4. So how does the circuit encode this extra information? Through two encoding tricks:
   4.1 The symbolic units correspond to stable physical mechanisms (i.e., the logical gates)
   4.2 Each variable has precisely one mechanism that determines its value.
CAUSAL MODELS AT WORK
(The impatient firing-squad)

\[ U \text{ (Court order)} \]

\[ C \text{ (Captain)} \]

\[ A \quad B \text{ (Riflemen)} \]

\[ D \text{ (Death)} \]
functional equality means that the left hand side is determined by the right hand side and not the other way around.
S1. (prediction): If rifleman-A shot, the prisoner is dead,
   \[A \Rightarrow D\]

S2. (abduction): If the prisoner is alive, then the Captain did not signal,
   \[\neg D \Rightarrow \neg C\]

S3. (transduction): If rifleman-A shot, then B shot as well,
   \[A \Rightarrow B\]
Shooting with no signal constitutes a blatant violation of one mechanism in the story: rifleman-A's commitment to follow the Captain's signal.

Violation renders this mechanism inactive, hence we must excise the corresponding equation from the model, using this knife, and replace it by a new mechanism: $A = \text{TRUE}$
S4. (action):

If the captain gave no signal and Mr. A decides to shoot, the prisoner will die:

\[ \neg C \Rightarrow D_A, \]

and \( B \) will not shoot:

\[ \neg C \Rightarrow \neg B_A \]
MUTILATION IN SYMBOLIC CAUSAL MODELS

Model $M_A$ (Modify $A=C$):

$C = U$  \hspace{1cm} (U)

$A = C$  \hspace{1cm} (C)

$B = C$  \hspace{1cm} (A)

$D = A \lor B$  \hspace{1cm} (D)

Facts: $\neg C$

Conclusions: ?

S4. (action): If the captain gave no signal and $A$ decides to shoot, the prisoner will die and $B$ will not shoot, $\neg C \Rightarrow D_A \land \neg B_A$
3-STEPS TO COMPUTING COUNTERFACTUALS

S5. If the prisoner is dead, he would still be dead if A were not to have shot. \( D \Rightarrow D_{\neg A} \)

Abduction

TRUE

Action

TRUE

Prediction

TRUE

\( A \rightarrow B \rightarrow C \rightarrow D \rightarrow U \)

\( A \rightarrow B \rightarrow D \rightarrow U \)

\( A \rightarrow B \rightarrow D \rightarrow U \)

\( A \rightarrow B \rightarrow D \rightarrow U \)

\( A \rightarrow B \rightarrow D \rightarrow U \)

\( A \rightarrow B \rightarrow D \rightarrow U \)
Consider now our counterfactual sentence
S5: If the prisoner is Dead, he would still be dead if \( A \) were not to have shot. \( D \implies D_{\neg A} \)

The antecedent \( \{\neg A\} \) should still be treated as interventional surgery, but only after we fully account for the evidence given: \( D \).
This calls for three steps
1. Abduction: Interpret the past in light of the evidence
2. Action: Bend the course of history (minimally) to account for the hypothetical antecedant \( \neg A \).
3. Prediction: Project the consequences to the future.
**COMPUTING PROBABILITIES OF COUNTERFACTUALS**

\[ P(S5). \text{The prisoner is dead. How likely is it that he would be dead if A were not to have shot. } P(D_{\neg A}|D) = ? \]

**Abduction**

- \[ P(u) \]
- \[ P(u|D) \]
- \[ P(u|D) \]
- \[ P(D_{\neg A}|D) \]

**Action**

- \[ P(u|D) \]
- \[ P(u|D) \]
- \[ P(D_{\neg A}|D) \]

**Prediction**

- \[ P(u|D) \]
- \[ P(u|D) \]
- \[ P(D_{\neg A}|D) \]
Suppose $X_1, X_2, X_3, X_4,$ and $X_5$ are unobserved. Can we find $P(x_j|\bar{x}_i)$?
Figure 3.8: Typical models in which the effect of $X$ on $Y$ is identifiable. Dashed arcs represent confounding paths, and $Z$ represents observed covariates.
Figure 3.9: Typical models in which $P(y|\hat{x})$ is not identifiable.
D-separation

• X is d-sep from Y given Z if
  \[ p(x,y|z) = p(x|z)p(y|z) \]

**Theorem 5.2.3**
(d-Separation in General Linear Model)
For any linear model structured according to a diagram \( D \), which may include cycles and bidirected arcs, the partial correlation \( \rho_{XY,Z} \) vanishes if the nodes corresponding to the set of variables \( Z \) d-separate node \( X \) from node \( Y \) in \( D \).
General structural equations

\[ x_i = f_i(pa_i, \epsilon_i), \quad i = 1, \ldots, n, \]

Linear structural equations

\[ x_i = \sum_{k \neq i} \alpha_{ik} x_k + \epsilon_i, \quad i = 1, \ldots, n, \]

• Looks like a regression model-why not?
  – Independence of \( x_k \) and noise not assumed.
The coefficient of $X$ in the regression of $Y$ on $X, Z_1, \ldots, Z_k$

$$y = ax + b_1 Z_1 + \ldots + b_k z_k$$

is given by

$$a = r_{YX \cdot Z_1 Z_2 \ldots Z_k}$$

\[
B_1 = \{ \rho_{32.1} = 0, \rho_{41.3} = 0, \rho_{42.3} = 0, \rho_{51.43} = 0, \\
\rho_{52.43} = 0 \}\n\]

\[
B_2 = \{ \rho_{32.1} = 0, \rho_{41.3} = 0, \rho_{42.1} = 0, \rho_{51.3} = 0, \\
\rho_{52.1} = 0 \}\n\]
Wright Rule (1923):

\[ r_{XY} = \text{Sum of products of path coefficients along all collider-free paths between } X \text{ and } Y. \]

If there is an edge \( X \xrightarrow{\alpha} Y \) in the model then:

\[ r_{XY} = \alpha + I_{YX} \]

where \( I_{YX} \) is independent of \( \alpha \).

Thus, \( \alpha = r_{YX} \) if \( X \) and \( Y \) are \( d \)-separated in \( G_{\alpha} \)
Theorem 5.3.1
(Single-Door Criterion for Direct Effects)
Let $G$ be any path diagram in which $\alpha$ is the path coefficient associated with link $X \rightarrow Y$, and let $G_\alpha$ denote the diagram that results when $X \rightarrow Y$ is deleted from $G$. The coefficient $\alpha$ is identifiable if there exists a set of variables $Z$ such that (i) $Z$ contains no descendant of $Y$ and (ii) $Z$ $d$-separates $X$ from $Y$ in $G_\alpha$. If $Z$ satisfies these two conditions, then $\alpha$ is equal to the regression coefficient $r_{YX.Z}$. Conversely, if $Z$ does not satisfy these conditions, then $r_{YX.Z}$ is not a consistent estimand of $\alpha$ (except in rare instances of measure zero).

Figure 5.7: The identification of $\alpha$ with $r_{YX.Z}$.
Theorem 5.3.2 (Back-Door Criterion)
For any two variables $X$ and $Y$ in a causal diagram $G$, the total effect of $X$ on $Y$ is identifiable if there exists a set of measurements $Z$ such that
1. no member of $Z$ is a descendant of $X$; and
2. $Z$ $d$-separates $X$ from $Y$ in the subgraph $G_X$ formed by deleting from $G$ all arrows emanating from $X$.

Moreover, if the two conditions are satisfied, then the total effect of $X$ on $Y$ is given by $r_{Y|X,Z}$. 

Figure 5.8: $\alpha + \beta \gamma = r_{Y|X,Z_2}$
Instrumental variables:

Figure 5.9: Graphical identification of $\alpha$ using instrumental variable $Z$, $\alpha = r_{YZ}/r_{XZ}$.

Figure 5.10 Graphical identification of $\alpha$, $\beta$, and $\gamma$.

\[
\begin{align*}
\alpha \beta &= r_{YX \cdot Z} \\
\beta &= r_{YX \cdot Z}/r_{WX} \\
\gamma &= r_{YZ \cdot X}
\end{align*}
\]
BUCKET ELIMINATION PROCEDURE

1. Start by searching for identifiable causal effects among pairs of variables in the graph, using the back-door criterion and Theorem 5.3.1. These can be either direct effects, total effects, or partial effects (i.e., effects mediated by specific sets of variables).

2. For any such identified effect, collect the path coefficients involved and put them in a bucket.

3. Begin labeling the coefficients in the buckets according to the following procedure:
   (a) if a bucket is a singleton, label its coefficient \( I \) (denoting identifiable);
   (b) if a bucket is not a singleton but contains only a single unlabeled element, label that element \( I \).

4. Repeat this process until no new labeling is possible.

5. List all labeled coefficients; these are identifiable.
1. When are two structural equation models observationally indistinguishable?

2. When do regression coefficients represent path coefficients?

3. When would the addition of a regressor introduce bias?

4. How can we tell, prior to taking any data, which path coefficients can be identified?

5. When can we dispose of the linearity-normality assumption and still extract causal information from the data?
Probability of Causation

- Action CONTINGENCY - \( P(x|\text{action}) - P(x|\text{no action}) \)
- Obs CONTINGENCY - \( P(x|\text{observe } y) - P(x|\text{observe } \sim y) \)

**Example**

\( \Delta P = P(\text{lung cancer }|\text{ yellowed fingers}) - P(\text{lung cancer }|\text{ not-yellowed fingers}) \)

is not very informative for the Surgeon General, who is interested in what causes lung cancer, and in the changes in the probability of lung cancer that various interventions would bring about, \( \Delta P \) is irrelevant.

**Better:** **Conditional CONTINGENCY**

- \( \Delta P_{\text{Smoking}} = P(\text{lung cancer }|\text{ yellowed fingers, smoking}) - P(\text{lung cancer }|\text{ not-yellowed fingers, smoking}) \)
- \( \Delta P_{\text{Not-smoking}} = P(\text{lung cancer }|\text{ yellowed fingers, not-smoking}) - P(\text{lung cancer }|\text{ not yellowed fingers, not smoking}) \)

Both of these are zero
ΔP doesn’t capture people’s judgments

• When \( P(e \mid c) = P(e \mid \neg c) = 1 \) OR 0, people do not judge that \( c \) does not produce \( e \); rather, they tend to withhold causal judgment altogether.

• If \( P(e \mid c) - P(e \mid \neg c) = P(e \mid c') - P(e \mid \neg c') > 0 \), and \( P(e \mid \neg c) > P(e \mid \neg c') \), people tend to judge that \( c \) has greater power than \( c' \) to produce \( e \).

• In contrast, if \( P(e \mid c) - P(e \mid \neg c) = P(e \mid c') - P(e \mid \neg c') < 0 \), and \( P(e \mid \neg c) > P(e \mid c') \), then people tend to judge that \( c' \) has more power than \( c \) to prevent \( e \).

• People tend to weigh frequencies of events that estimate \( P(e \mid c) \) more than those that estimate \( P(e \mid \neg c) \).
Box 1. Bayes nets, the Markov assumption and conditional independence

The graph above represents the claim that smoking is a cause of yellowed teeth and lung cancer, but that lung cancer does not cause yellowed teeth and yellowed teeth do not cause lung cancer. It also represents claims about the conditional probability relations among the three variables: for all values of Y, S and L (for example, all combinations of present or absent)

\[ Pr(Y, S, L) = Pr(Y|L, S) \cdot Pr(L|S) \cdot Pr(S) = Pr(Y|S) \cdot Pr(L|S) \cdot Pr(S) \]

where \( Pr(Y = \text{present}|L = \text{absent}, S = \text{present}) \), for example, represents the probability of yellowed teeth among smokers without lung cancer. The first equality is necessarily true, but the second is an assumption, the Markov factorization, which says that the joint distribution of all variables is equal to a product of the conditional distributions of each variable on its parents in the graph. The Markov factorization is equivalent, in this example, to the claim that \( Pr(Y|S, L) = Pr(Y|S) \).
Box 2. Graphical representations of interventions

Starting with the causal system represented by the directed structure:

```
  Smoking (S)
   /     \
  /       \
Yellow teeth (Y)  Lung cancer (L)
```

with the joint probability distribution $\Pr(Y,S,L) = \Pr(Y|S)\cdot\Pr(L|S)\cdot\Pr(S)$, we imagine an intervention that forces everyone to brush their teeth daily with stain removing paste. We reconceive the system above with an expanded structure:
and probability distribution \( \Pr(Y,S,L,I) = \Pr(Y|S,I) \cdot \Pr(L|S) \cdot \Pr(S) \)
with the understanding that \( \Pr(Y,S,L,I = \text{no intervention}) = \Pr(Y|S) \cdot \Pr(L|S) \cdot \Pr(S) \), the original distribution, and \( \Pr(Y,S,L,I = \text{intervention}) = \Pr(Y|I = \text{intervention}) \cdot \Pr(L|S) \cdot \Pr(S) = 0 \) unless \( Y = \text{present} \), and equals \( \Pr(L|S) \cdot \Pr(S) \) otherwise. The act of intervention fixes the value of \( Y \) to present, and thus makes \( Y \) independent of \( S \). The system after the intervention can be more simply represented by the graph:
Box 3. Comparing Bayesian learning and constraint-based learning of Bayes nets

Bayesian learning

(1) Prior probability distribution \( \Pr(G; \theta) \) over all directed acyclic graphs \( G \) and probability distributions \( \theta \) on the variables (vertices in \( G \)), with a Markov factorization for \( G \).

(2) Likelihood function \( L(D; G, \theta) \) giving the probability of the observations \( D \) conditional on the truth of \( G, \theta \).

(3) Compute the probability of any graph \( G \) conditional on the data by using Bayes Theorem and integrating over \( \theta \)

\[
\Pr(G|D) = \frac{\int \Pr(G; \theta)L(D; G, \theta)d\theta}{\Pr(D)}
\]

(4) Find the graphs \( G \) such that for all other graphs \( G^* \), \( \Pr(G|D) \geq \Pr(G^*|D) \)

---

True unknown structure

TRENDS in Cognitive Sciences
Fig. 1. Normative casual model in making judgments (see text for details).
Constraint-based learning

(1) Form the complete undirected graph, U
(2) Estimate from the data which pairs of variables are independent and remove the corresponding edges to form U1.
(3) Estimate from the data which pairs of variables in U1 are independent conditional on one of their adjacent neighbors and remove their edges, forming U2; continue to form U3, U4,... until no more edges are removed.
(4) Orient $X \rightarrow Z \rightarrow Y$ as $X \rightarrow Z \leftarrow Y$ if $Z$ was not conditioned on when removing the $X \rightarrow Y$ edge.
(5) Orient remaining undirected edges so as to avoid creating colliders: $\rightarrow V \leftarrow$
Box 5. From noisy-or gates to Cheng models

In a noisy-or gate an effect \( E \) is assumed to be a Boolean function of its potential causes \( A, U \), and parameters:

\[
E = qa A \oplus qu U
\]  

(1)

where \( qa, qu, E \) and \( U \) all take values \( \{0, 1\} \), and \( \oplus \) is Boolean sum (\( = 1 \) if and only if either argument is 1). So, letting \( A, \) etc., stand for \( A = 1 \), etc., and \( \sim A \), etc., stand for \( A = 0 \), etc., (1) implies:

\[
\Pr(E) = \Pr(qa A \oplus qu U)
\]

\[
= \Pr(qa \cdot A) + \Pr(qu \cdot U) - \Pr(qa \cdot A) \cdot \Pr(qu \cdot U)
\]  

(2)

Therefore \( \Pr(E|A, \sim U) = \Pr(qa) \), Cheng’s causal power of \( A \) to generate \( E \).

If \( A, U, qa \) and \( qu \) are all independent in probability, (2) becomes:

\[
\Pr(E) = \Pr(qa) \cdot \Pr(A) + \Pr(qu) \cdot \Pr(U) - \Pr(qa) \cdot \Pr(A) \cdot \Pr(qu) \cdot \Pr(U)
\]  

(3)

and (3) implies, with some algebra:

\[
\Pr(qa) = [\Pr(E|A, \sim U) - \Pr(E| \sim A, \sim U)]/[1 - \Pr(E| \sim A)]
\]  

(4)

which is Cheng’s formula for estimating the generative causal power of \( A \) when \( U \) is thought to be independent of \( A \).

\[
\Pr(E) = \Pr(E|A,U) \Pr(A) \Pr(U)
\]
The Rescorla–Wagner procedure estimates that the associative strength of potential cause $C_i$ with the effect, $E$, after trial $t + 1$ is $V_i = V_i + \Delta V_i$; where $\Delta V_i$ is given by:

$$\Delta V_i^t = \begin{cases} 
0, & \text{if the cause, } C_i \text{ does not appear in case } t; \\
\alpha_i \beta_1 \left( \lambda - \sum_{\text{Cause } C_i \text{ appears in case } t} V_j \right), & \text{if both } C_i \text{ and } E \text{ appear in case } t; \\
\alpha_i \beta_2 \left( 0 - \sum_{\text{Cause } C_i \text{ appears in case } t} V_j \right), & \text{if } C_i \text{ appears and } E \text{ does not in case } t.
\end{cases}$$
Fig. 2. Graph of an (extended) Cheng model for which the generative causal power of C can be estimated from observations of A, C and E without a focal set.
Bayesian networks

**Nodes:** variables

**Links:** direct dependencies

Each node has a conditional probability distribution

**Data:** observations of $X_1$, ..., $X_4$

**Four random variables:**
- $X_1$: coughing
- $X_2$: high body temperature
- $X_3$: flu
- $X_4$: lung cancer
Causal Bayesian networks

**Nodes:** variables

**Links:** causal mechanisms

Each node has a conditional probability distribution

**Data:** observations of and interventions on $X_1, ..., X_4$

Four random variables:

- $X_1$  coughing
- $X_2$  high body temperature
- $X_3$  flu
- $X_4$  lung cancer

(Pearl; Glymour & Cooper)
Inference in causal graphical models

• Explaining away or “discounting” in social reasoning (Kelley; Morris & Larrick)

• “Screening off” in intuitive causal reasoning (Waldmann, Rehder & Burnett, Blok & Sloman, Gopnik & Sobel)

  – Better in chains than common-cause structures; common-cause better if mechanisms clearly independent

• Understanding and predicting the effects of interventions (Sloman & Lagnado; Gopnik & Schulz)
Learning graphical models

• Structure learning: what causes what?

• Parameter learning: how do causes work?
Bayesian learning of causal structure

Data $d$

Causal hypotheses $h$

$d_1 = \{X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1\}$

$d_2 = \{X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1\}$

$d_3 = \{X_1 = 0, X_2 = 1, X_3 = 0, X_4 = 1\}$

$d_4 = \{X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1\}$

1. What is the most likely network $h$ given observed data $d$?

2. How likely is there to be a link $X_4 \rightarrow X_2$?

$P(h_i \mid d) = \frac{P(d \mid h_i)P(h_i)}{\sum_j P(d \mid h_j)P(h_j)}$

$P(X_4 \rightarrow X_2 \mid d) = \sum_{h_j \in H} P(X_4 \rightarrow X_2 \mid h_j)P(h_j \mid d)$

(Bayesian model averaging)
Bayesian Occam’s Razor

For any model $M$, $\sum_{\text{all } d \in D} p(D = d \mid M) = 1$

*Law of “conservation of belief”: A model that can predict many possible data sets must assign each of them low probability.*
Learning causation from contingencies

<table>
<thead>
<tr>
<th></th>
<th>$C$ present $(c^+)$</th>
<th>$C$ absent $(c^-)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$ present $(e^+)$</td>
<td>a</td>
<td>c</td>
</tr>
<tr>
<td>$E$ absent $(e^-)$</td>
<td>b</td>
<td>d</td>
</tr>
</tbody>
</table>

e.g., “Does injecting this chemical cause mice to express a certain gene?”

Subjects judge the extent $C$ to which causes $E$ (rate on a scale from 0 to 100)
Two models of causal judgment

• Delta-P (Jenkins & Ward, 1965):
  \[ \Delta P \equiv P(e^+ | c^+) - P(e^+ | c^-) \]

• Power PC (Cheng, 1997):
  \[ Power \equiv \frac{\Delta P}{1 - P(e^+ | c^-)} \]
Judging the probability that C → E
(Buehner & Cheng, 1997; 2003)

- Independent effects of both ΔP and causal power.
- At ΔP=0, judgments decrease with base rate. (“frequency illusion”)
Learning causal strength
(parameter learning)

Assume this causal structure:

\[ \Delta P \text{ and causal power are maximum likelihood estimates of the strength parameter } w_1, \text{ under different parameterizations for } P(E|B,C): } \]

- linear \( \rightarrow \Delta P \)
- Noisy-OR \( \rightarrow \) causal power
Learning causal structure
(Griffiths & Tenenbaum, 2005)

• Hypotheses:

  - $h_1$: [Diagram of causal structure]
  - $h_0$: [Diagram of causal structure]

• Bayesian causal support: $\log \frac{P(d \mid h_1)}{P(d \mid h_0)}$

  - $P(d \mid h_0) = \int_0^1 P(d \mid w_0) \ p(w_0 \mid h_0) \ dw_0$
  - $P(d \mid h_1) = \int_0^1 \int_0^1 P(d \mid w_0, w_1) \ p(w_0, w_1 \mid h_1) \ dw_0 \ dw_1$

(noisy-OR)

(assume uniform parameter priors, but see Yuille et al., Danks et al.)
Buehner and Cheng (1997)

People

$\Delta P$ (r = 0.89)

Power (r = 0.88)

N/A

Support (r = 0.97)
Implicit background theory

• Injections may or may not cause gene expression, but gene expression does not cause injections.
  – No hypotheses with $E \rightarrow C$

• Other naturally occurring processes may also cause gene expression.
  – All hypotheses include an always-present background cause $B \rightarrow C$

• Causes are generative, probabilistically sufficient and independent, i.e. each cause independently produces the effect in some proportion of cases.
  – Noisy-OR parameterization
Sensitivity analysis

\[ P(e^+|c^+) = 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \]

\[ P(e^+|c^-) = 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \quad 1.00 \quad 0.75 \quad 0.50 \quad 0.25 \quad 0.00 \]

People

Support (Noisy-OR)

\[ \chi^2 \]

Support (generic parameterization)
Generativity is essential

\[
P(e^+|c^+) = \frac{8}{8} \quad \frac{6}{8} \quad \frac{4}{8} \quad \frac{2}{8} \quad \frac{0}{8}
\]

\[
P(e^+|c^-) = \frac{8}{8} \quad \frac{6}{8} \quad \frac{4}{8} \quad \frac{2}{8} \quad \frac{0}{8}
\]

- Predictions result from “ceiling effect”
  - ceiling effects only matter if you believe a cause increases the probability of an effect
Different parameterizations for different kinds of mechanisms

“Does C cause E?”

“Is there a difference in E with C vs. not-C?”

“Does C prevent E?”
Blicket detector
(Sobel, Gopnik, and colleagues)

See this? It’s a blicket machine.
Blickets make it go.

Let’s put this one on the machine.

Oooh, it’s a blicket!
“Backwards blocking”  
(Sobel, Tenenbaum & Gopnik, 2004)

- Initially: Nothing on detector – detector silent (A=0, B=0, E=0)
- Trial 1: A B on detector – detector active (A=1, B=1, E=1)
- Trial 2: A on detector – detector active (A=1, B=0, E=1)
- 4-year-olds judge if each object is a blicket
  A: a blicket (100% say yes)
  B: probably not a blicket (34% say yes)

(cf. “explaining away in weight space”, Dayan & Kakade)
Possible hypotheses?
Bayesian causal learning

With a uniform prior on hypotheses, generic parameterization:

Probability of being a blicket:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.32</td>
<td>0.32</td>
</tr>
</tbody>
</table>

0.34    0.34
A stronger hypothesis space

- Links can only exist from blocks to detectors.
- Blocks are blickets with prior probability $q$.
- Blickets always activate detectors, detectors never activate on their own (i.e., deterministic OR parameterization, no hidden causes).

$$P(h_{00}) = (1 - q)^2 \quad P(h_{01}) = (1 - q)q \quad P(h_{10}) = q(1 - q) \quad P(h_{11}) = q^2$$
Manipulating prior probability
(Tenenbaum, Sobel, Griffiths, & Gopnik)
Learning more complex structures

- Tenenbaum et al., Griffiths & Sobel: detectors with more than two objects and noisy mechanisms
- Steyvers et al., Sobel & Kushnir: active learning with interventions (c.f. Tong & Koller, Murphy)
- Lagnado & Sloman: learning from interventions on continuous dynamical systems
Inferring hidden causes

Common unobserved cause

Independent unobserved causes

One observed cause

The “stick ball” machine

(Kushnir, Schulz, Gopnik, & Danks, 2003)
Bayesian learning with unknown number of hidden variables

(Griffiths et al 2006)
Common unobserved cause

Independent unobserved causes

One observed cause

\[ \alpha = 0.3 \]

\[ \omega = 0.8 \]

\[ r = 0.94 \]
Inferring latent causes in classical conditioning
(Courville, Daw, Gordon, Touretzky 2003)

e.g.,
A noise
X tone
B click
US shock

Training:
A US
A X
B US

Test:
X
X B

(a) Sigmoid belief network
(b) Marginal likelihood

(a) Few A-X trials
(b) Many A-X trials
(c) Model size over trials
Inferring latent causes in perceptual learning (Orban, Fiser, Aslin, Lengyel 2006)

Learning to recognize objects and segment scenes:

A

True combos

Scene

Mixture combos

B

\[ x_1 \rightarrow w_{x_1} \rightarrow w_{y_1} \]

\[ x_2 \rightarrow w_{x_2} \rightarrow w_{y_2} \]

\[ x_2 \rightarrow w_{x_2} \rightarrow w_{y_3} \]

\[ x_2 \rightarrow w_{x_2} \rightarrow w_{y_4} \]
Inferring latent causes in sensory integration (Kording et al. 2006, NIPS 06)
Summary: causal inference & learning

- Human causal induction can be explained using core principles of graphical models.
  - Bayesian inference (explaining away, screening off)
  - Bayesian structure learning (Occam’s razor, model averaging)
  - Active learning with interventions
  - Identifying latent causes
Summary: causal inference & learning

• Crucial constraints on hypothesis spaces come from abstract prior knowledge, or “intuitive theories”.
  – What are the variables?
  – How can they be connected?
  – How are their effects parameterized?

• Big open questions…
  – How can these theories be described formally?
  – How can these theories be learned?