Models and theories in cognitive science

Tom Griffiths
Department of Psychology
Program in Cognitive Science
University of California, Berkeley

Marr's three levels

Computation

"What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?"

Representation and algorithm

"What is the representation for the input and output, and the algorithm for the transformation?"

Implementation

"How can the representation and algorithm be realized physically?"

Marr on the computational level

…an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism land the hardware) in which it is embodied. In a similar vein, trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds’ wings make sense.

Questions

- How does one go about conducting a computational-level analysis?
- What is the equivalent of aerodynamics for cognition?
- What are the consequences of this kind of approach?

Questions

- How does one go about conducting a computational-level analysis?

- What is the equivalent of aerodynamics for cognition?

- What are the consequences of this kind of approach?

An approach to analyzing cognition

Identify the underlying computational problem

Find a good solution to that problem

Compare human cognition to that solution

Directly relates cognition and computation

(Marr, 1982; Shepard, 1987; Anderson, 1990)

Questions

- How does one go about conducting a computational-level analysis?
- What is the equivalent of aerodynamics for cognition?
- What are the consequences of this kind of approach?

A theory of induction

Posterior probability

Likelihood

Prior probability

$$P(h \mid d) = \frac{P(d \mid h) P(h)}{\sum_{h'} P(d \mid h') P(h')}$$

Sum over space of hypotheses

h: hypothesis
d: data

Statistics tells us what structure we can infer from data

Questions

- How does one go about conducting a computational-level analysis?

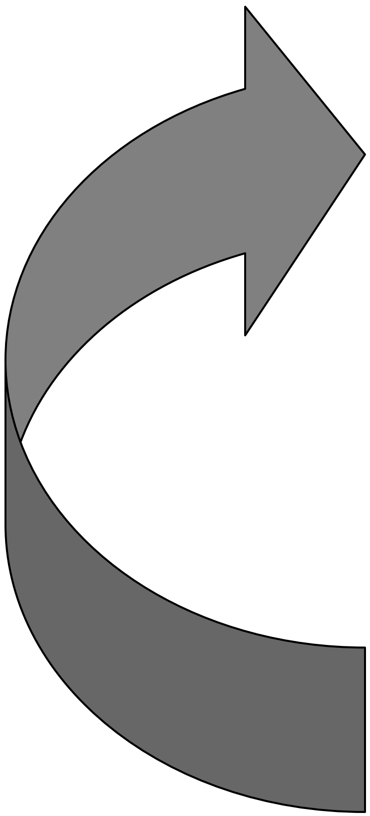
- What is the equivalent of aerodynamics for cognition?

- What are the consequences of this kind of approach?

Results of computational level analysis

1. Connections between problems in cognitive science and problems in statistics.

Human learning	**Machine learning**
Categorization	Density estimation
Causal learning	Graphical models
Function learning	Regression
Representations	Nonparametric Bayes
Language	Probabilistic grammars
Experiment design	Inference algorithms
…	…



Human learning

Machine learning

Results of computational level analysis

1. Connections between problems in cognitive science and problems in statistics.

2. A characterization of the inductive biases of human learners.

The importance of inductive biases

“pecora”

Identifying inductive biases

Posterior probability

Likelihood

Prior probability

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h'} P(d \mid h')P(h')}$$

Sum over space of hypotheses

h: hypothesis
d: data

(more generally… sources of regularization)

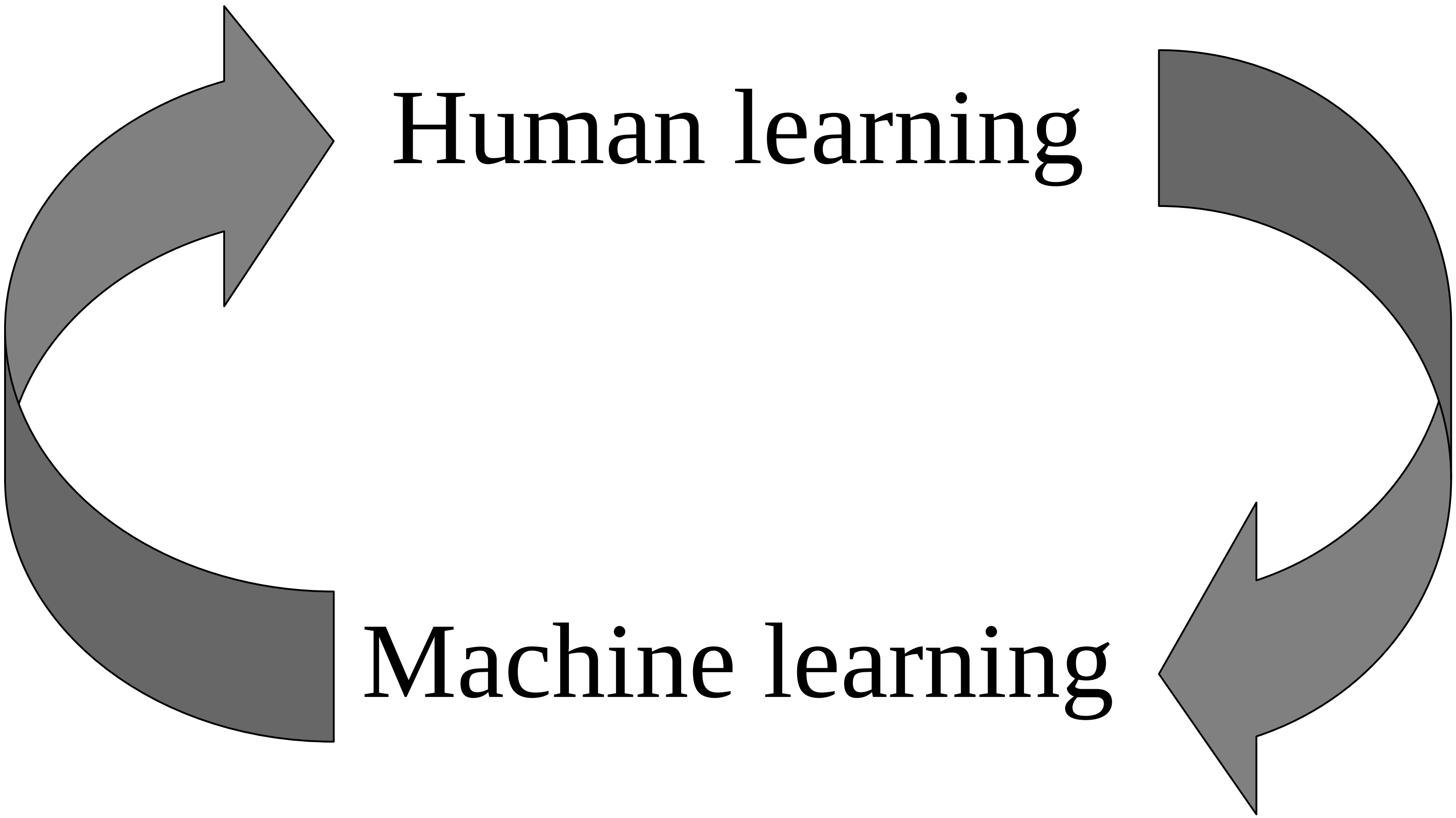
Results of computational level analysis

1. Connections between problems in cognitive science and problems in statistics.

2. A characterization of the inductive biases of human learners.

3. Some understanding of where those inductive biases come from.

Human learning	**Machine learning**
Categorization	Density estimation
Causal learning	Graphical models
Function learning	Regression
Representations	Nonparametric Bayes
Language	Probabilistic grammars
Experiment design	Inference algorithms
…	…


Human learning
Machine learning

Human learning	**Machine learning**
Categorization	Density estimation
Causal learning	Graphical models
Function learning	Regression
Representations	Nonparametric Bayes
Language	Probabilistic grammars
Experiment design	Inference algorithms
…	…

“The contagion spread rapidly and before its progress could be arrested, sixteen persons were affected of which two died. Of these sixteen, eight were under my care. On this occasion I used for the first time the affusion of cold water in the manner described by Dr. Wright. It was first tried in two cases ... [then] employed in five other cases. It was repeated daily, and of these seven patients, the whole recovered.”

Currie (1798)

Medical Reports on, the Effects of Water, Cold and Warm, as a Remedy in Fevers and Febrile Diseases

	Treated	*Untreated*
Recovered	7	7
Died	0	2

"Does the treatment cause recovery?"

	C present (c^+)	C absent (c^-)
E present (e^+)	a	c
E absent (e^-)	b	d

"Does C cause E?"

(rate on a scale from 0 to 100)

Two models of causal judgment

- Delta-P (Jenkins & Ward, 1965):

$$\Delta P \equiv P(e^{+} \mid c^{+}) - P(e^{+} \mid c^{-})$$

- Power PC (Cheng, 1997):

$$Power \equiv \frac{\Delta P}{1 - P(e^{+} \mid c^{-})}$$

	Treated	*Untreated*
Recovered	7	7
Died	0	2

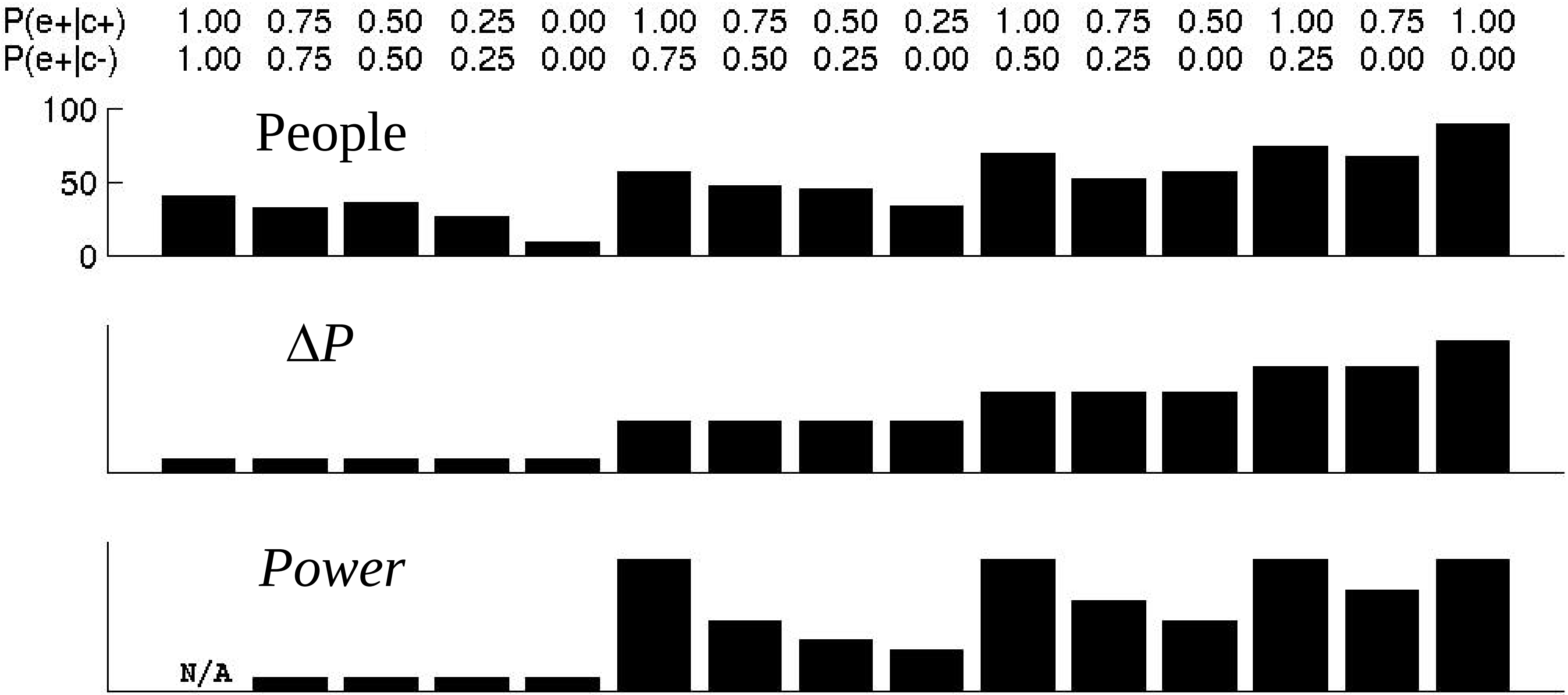
$$P(e^+|c^+) = 7/7 = 1.00$$

$$P(e^+|c^-) = 7/9 = 0.78$$

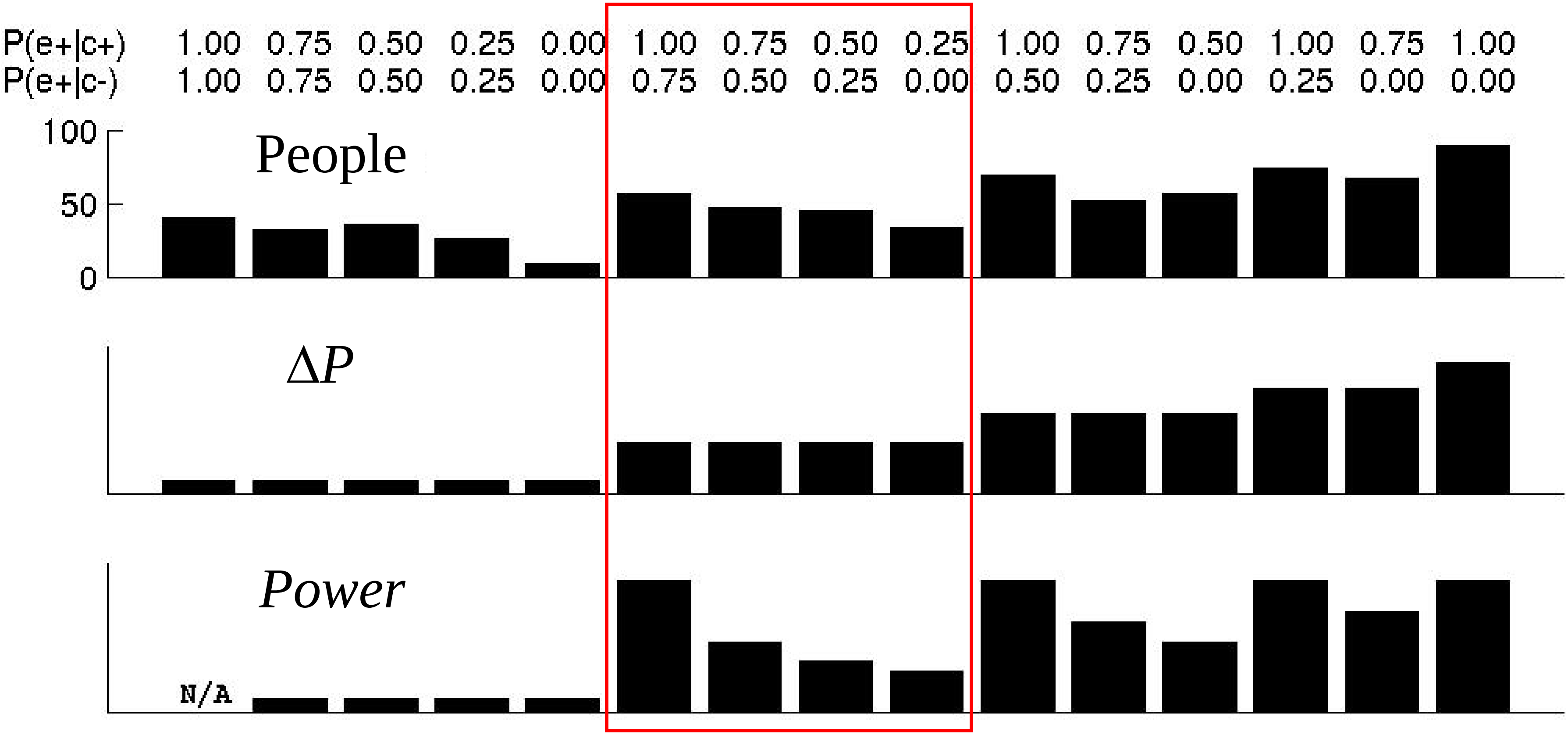
$$\Delta P \equiv P(e^+ \mid c^+) - P(e^+ \mid c^-) = 1.00\text{-}0.78 = 0.22$$

$$Power \equiv \frac{\Delta P}{1 - P(e^+ \mid c^-)} = 0.22/0.22 = 1.00$$

Buehner and Cheng (1997)

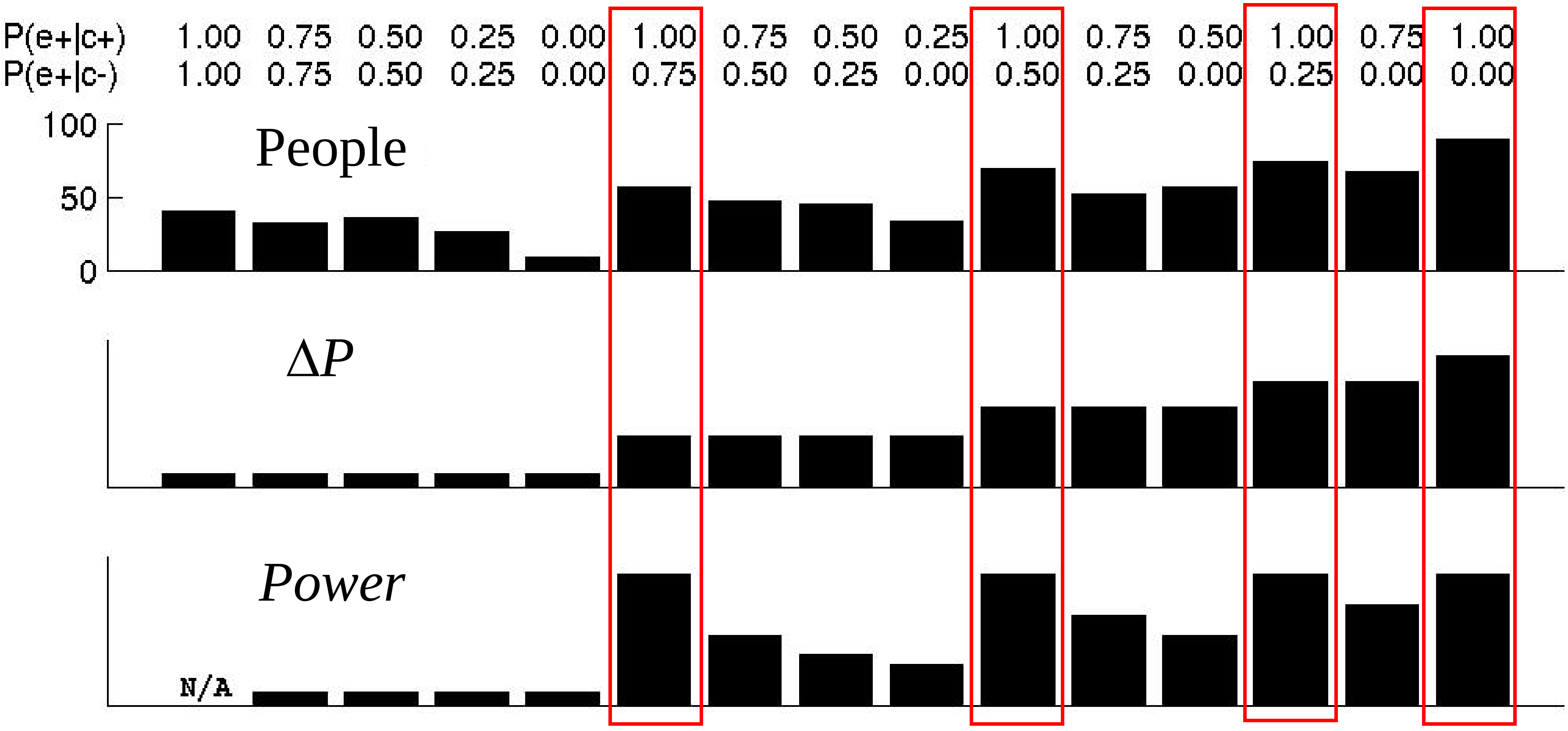


Buehner and Cheng (1997)



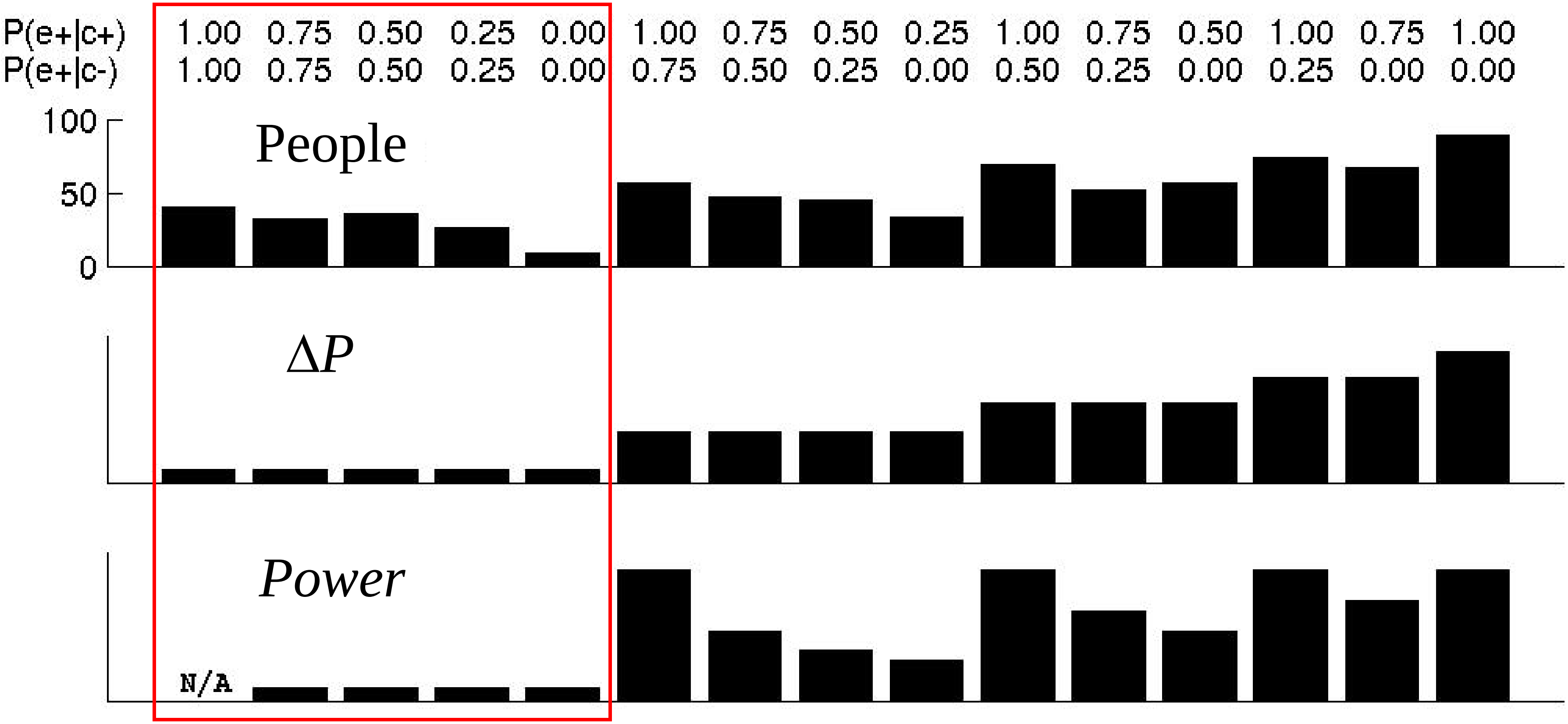
Constant ΔP, changing judgments

Buehner and Cheng (1997)

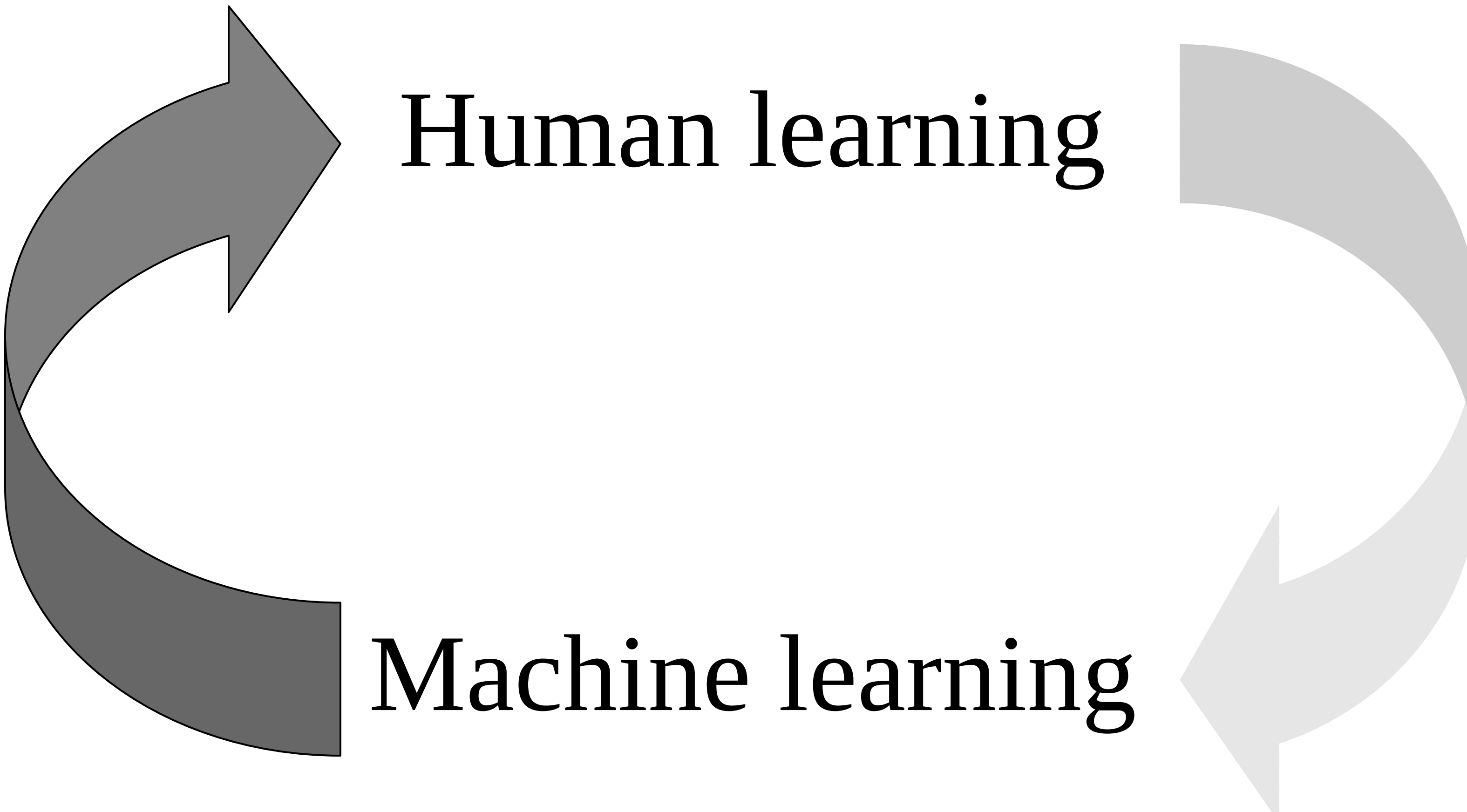


Constant causal power, changing judgments

Buehner and Cheng (1997)



ΔP = 0, changing judgments

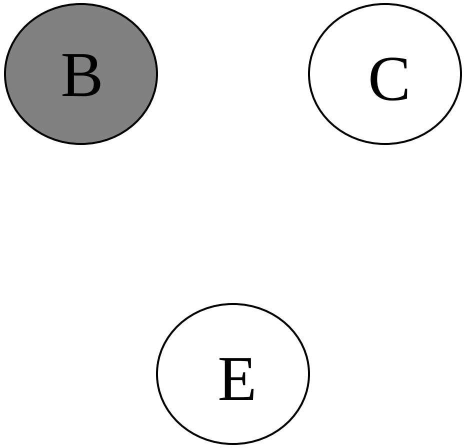

Human learning
Machine learning

Causal graphical models

(Pearl, 2000; Spirtes, Glymour, & Schienes, 1993)

Causal graphical models

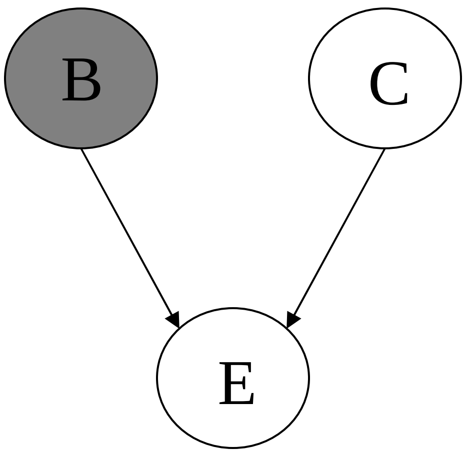
(Pearl, 2000; Spirtes, Glymour, & Schienes, 1993)



- Variables

Causal graphical models

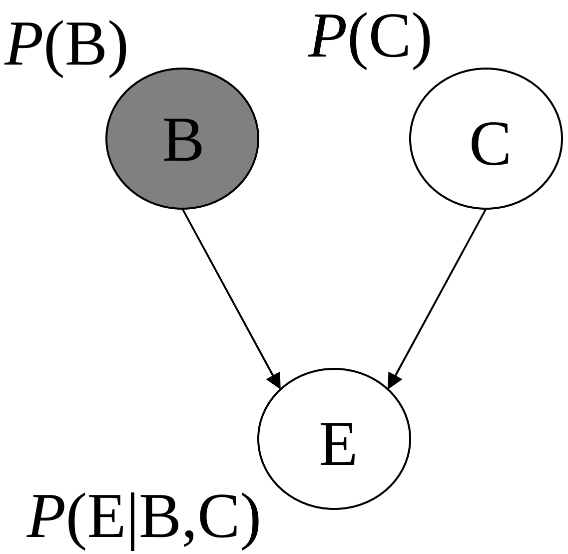
(Pearl, 2000; Spirtes, Glymour, & Schienes, 1993)



- Variables
- Structure

Causal graphical models

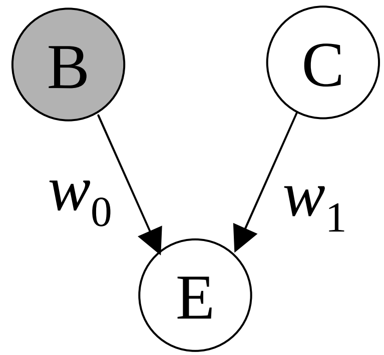
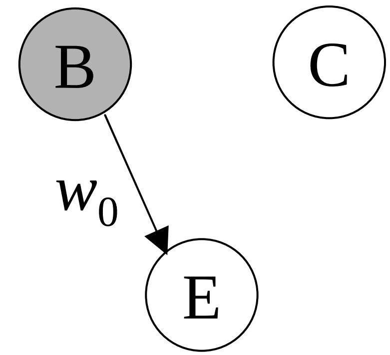
(Pearl, 2000; Spirtes, Glymour, & Schienes, 1993)



- Variables
- Structure
- Conditional probabilities

Defines probability distribution over variables
(for both observation, and intervention)

Conditional probabilities

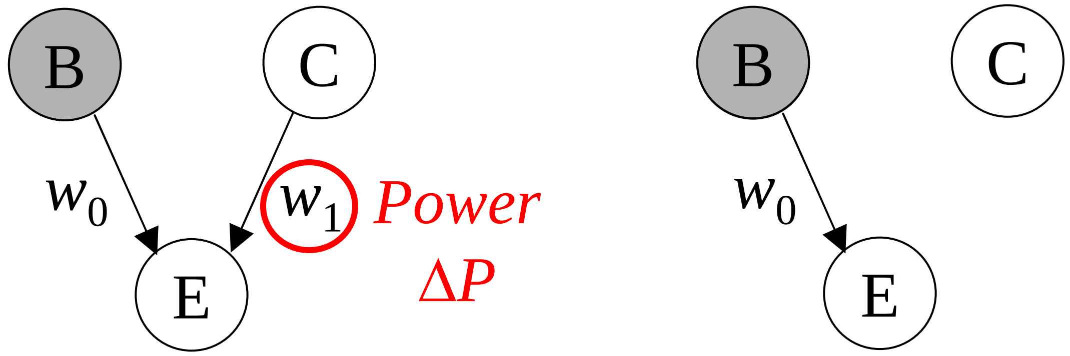
- Structures: $h_1 =$  $h_0 =$ 

w_0, w_1: strength parameters for B, C

- Parameterization: "Noisy-OR" (Pearl, 1988)

C	B	P(E = 1 | C, B)	P(E = 1| C, B)
0	0	0	0
1	0	w_1	0
0	1	w_0	w_0
1	1	$w_1 + w_0 - w_1 w_0$	w_0

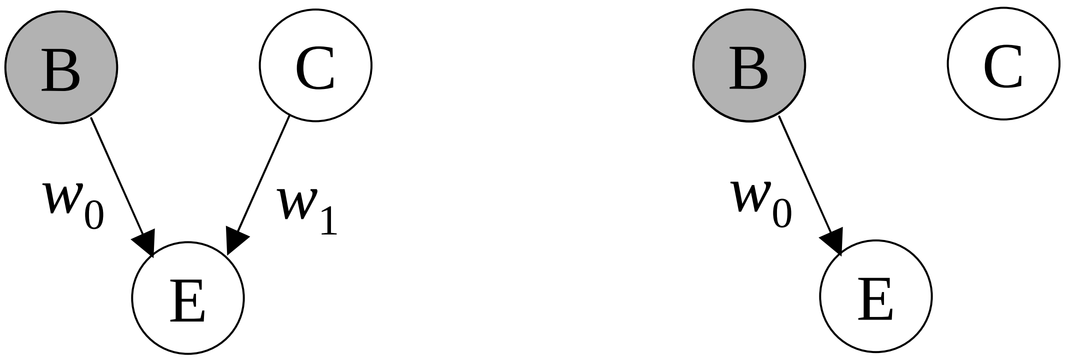
Causal learning



- **Structure:** does a relationship exist?
- **Strength:** how strong is the relationship?

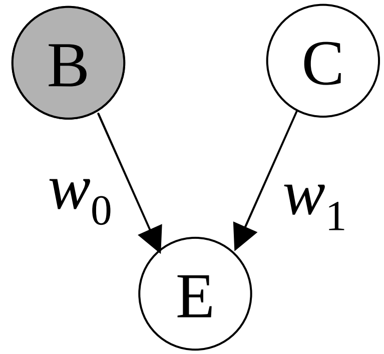
(Griffiths & Tenenbaum, 2005)

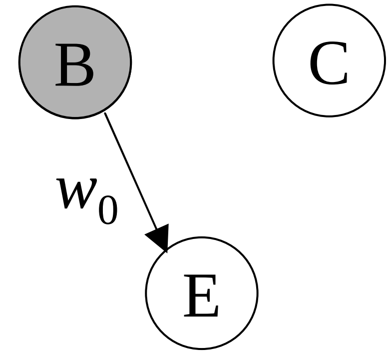
Causal learning



- **Structure:** does a relationship exist?
- **Strength:** how strong is the relationship?

(Griffiths & Tenenbaum, 2005)

- Hypotheses: $h_1 =$ 

$h_0 =$ 

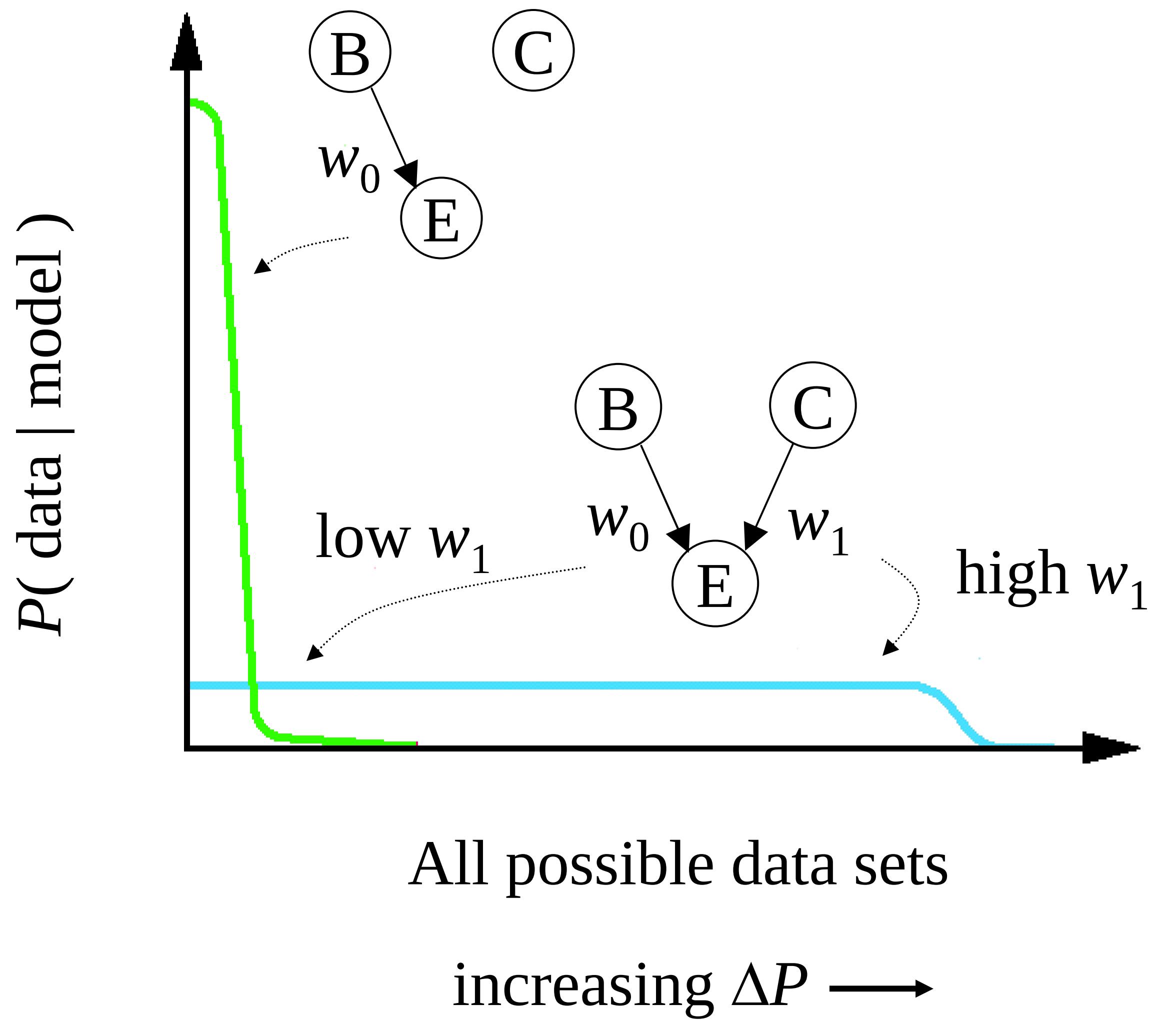
- Bayesian structure learning:

$$\text{support} = \frac{P(\text{data} \mid h_1)}{P(\text{data} \mid h_0)}$$

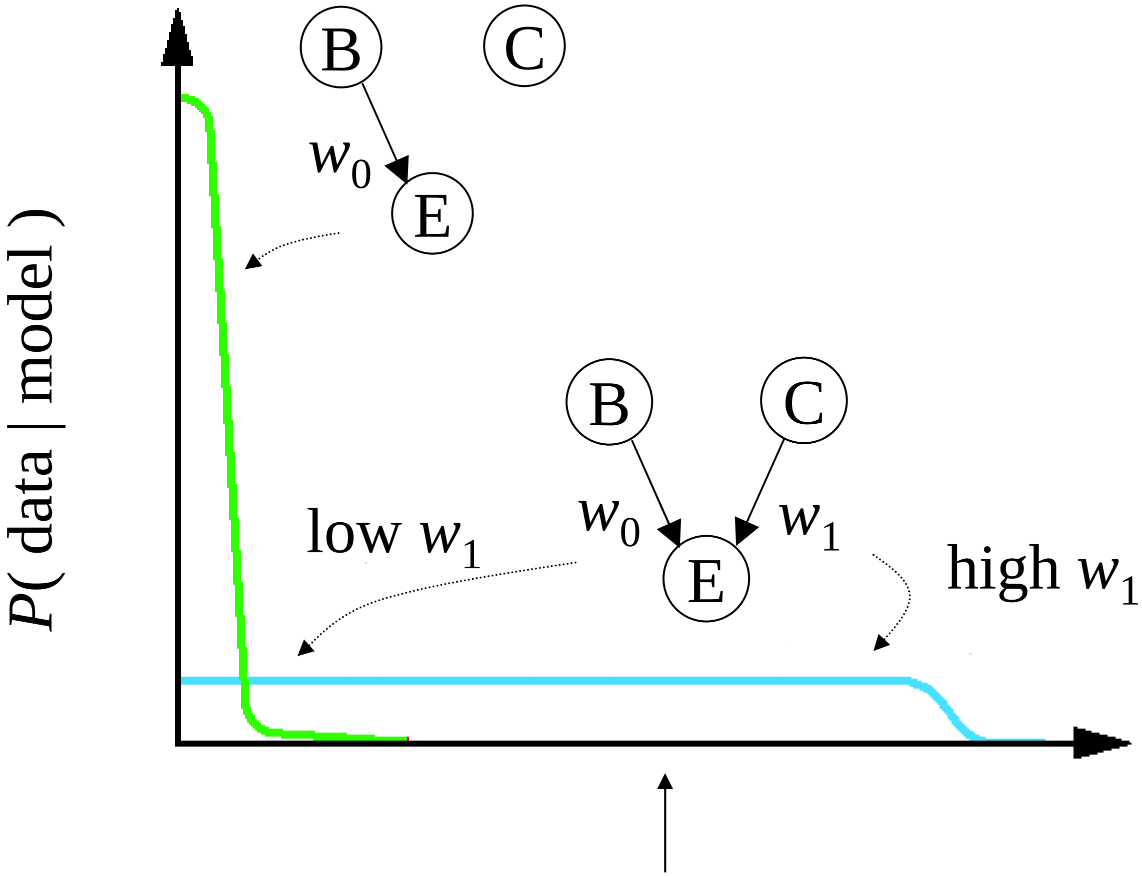
$$P(\text{data} \mid h_0) = \int_0^1 P(\text{data} \mid w_0)\ p(w_0 \mid h_0)\ dw_0$$

$$P(\text{data} \mid h_1) = \int_0^1 \int_0^1 \underbrace{P(\text{data} \mid w_0, w_1)}_{\text{noisy-OR}}\ p(w_0, w_1 \mid h_1)\ dw_0\ dw_1$$

Bayesian Occam's Razor



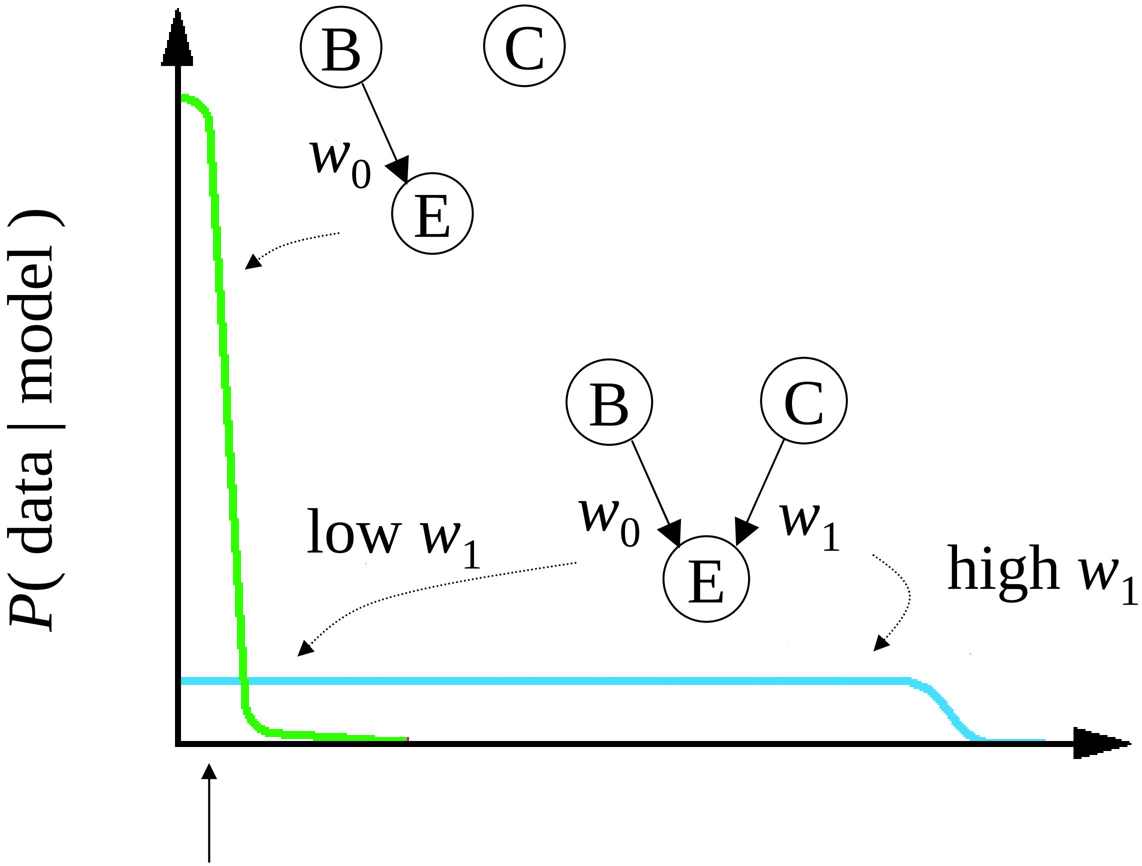
Bayesian Occam's Razor



$$P(e^+ \mid c^+) = 80/100$$

$$P(e^+ \mid c^-) = 20/100$$

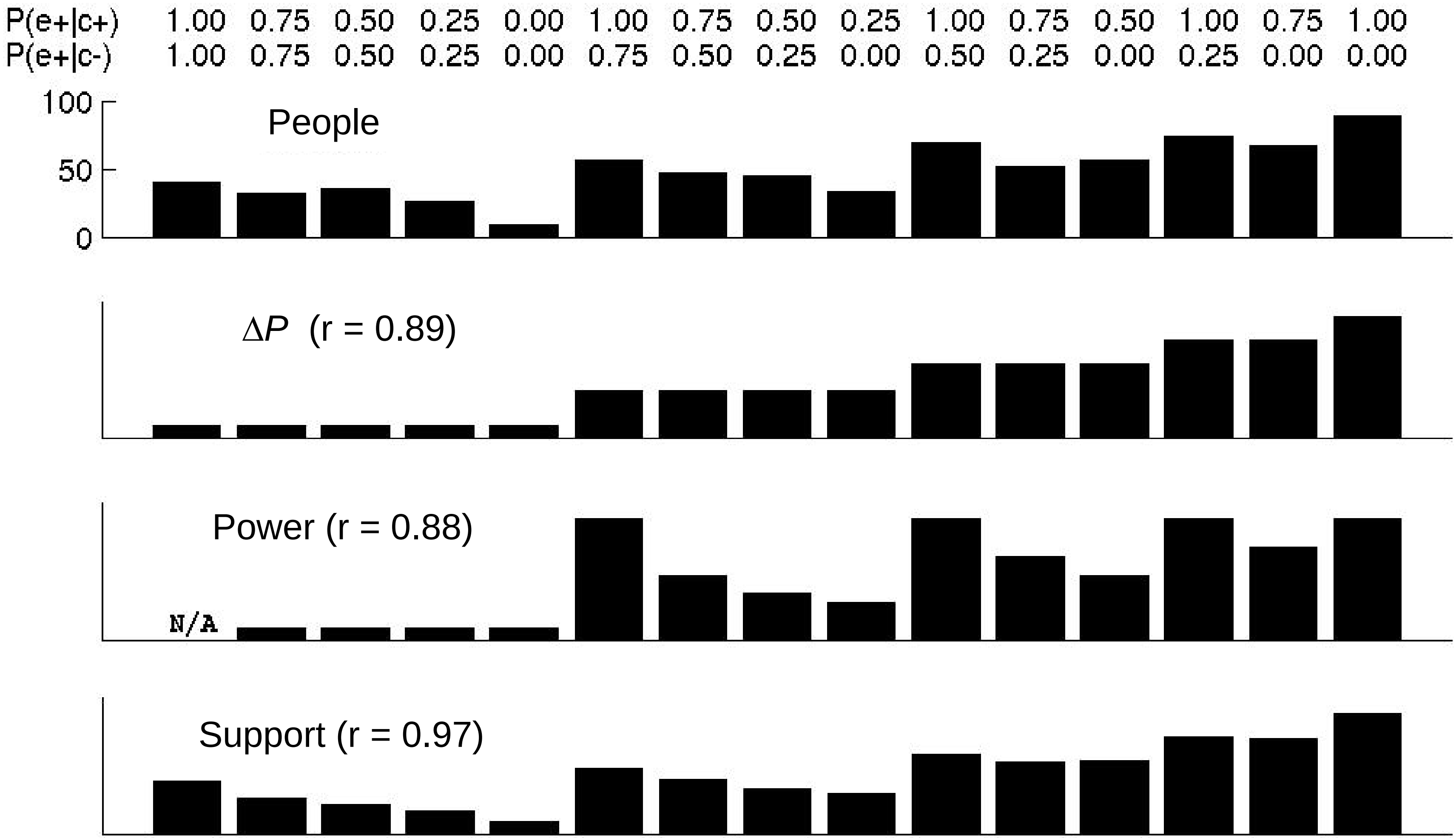
Bayesian Occam's Razor



$$P(e^+ \mid c^+) = 80/100$$

$$P(e^+ \mid c^-) = 77/100$$

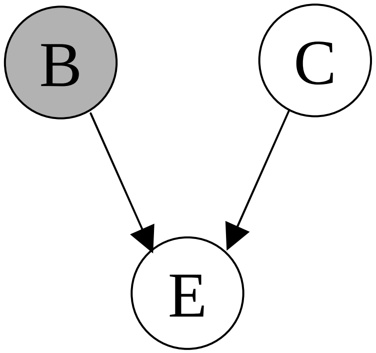
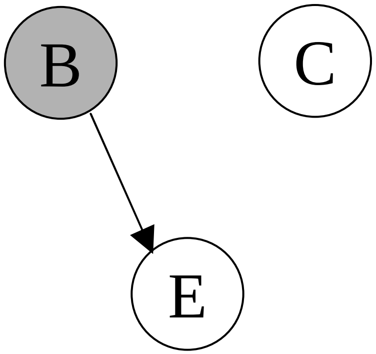
Buehner and Cheng (1997)



Assumptions guiding inference

- What assumptions are responsible for this?
  - alternative model: Bayes with arbitrary $P(E|B,C)$

Conditional probabilities

- Structures: $h_1 =$  $h_0 =$ 

- Parameterization: Generic

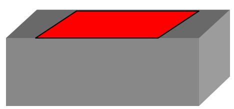
C	B	P(E = 1 | C, B)	P(E = 1| C, B)
0	0	p_{00}	p_0
1	0	p_{10}	p_0
0	1	p_{01}	p_1
1	1	p_{11}	p_1

Assumptions guiding inference

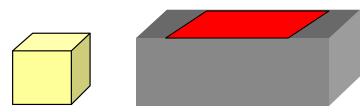
- What assumptions are responsible for this?
  - alternative model: Bayes with arbitrary $P(E|B,C)$

- Critical assumption: causes increase the probability of their effects (as in noisy-OR)
- People have strong intuitions about the nature of causality, beyond statistical dependence

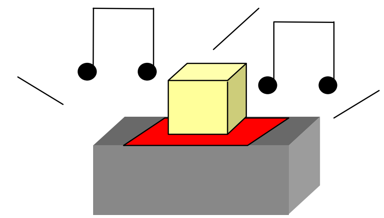
The blicket detector



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



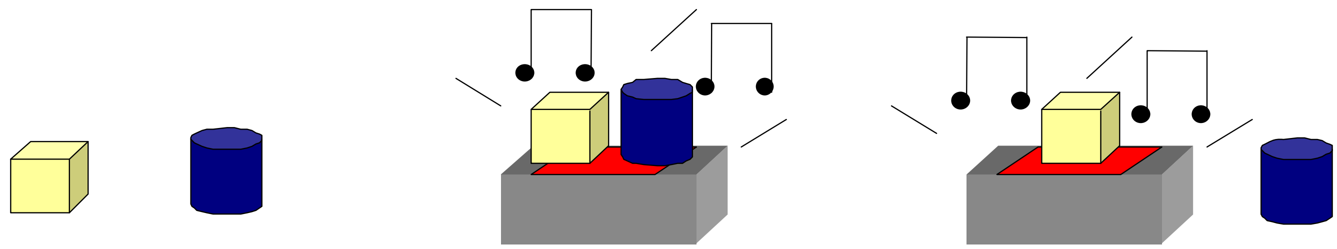
Let’s put this one on the machine.



Oooh, it’s a blicket!

“Backwards blocking”

(Sobel, Tenenbaum & Gopnik, 2004)



A *B* AB Trial A Trial

– Two objects: A and B
– Trial 1: A B on detector – detector active
– Trial 2: A on detector – detector active
– 4-year-olds judge whether each object is a blicket
  - A: a blicket (100% say yes)
  - B: probably not a blicket (34% say yes)

Bayesian inference

- Evaluating causal models in light of data:

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

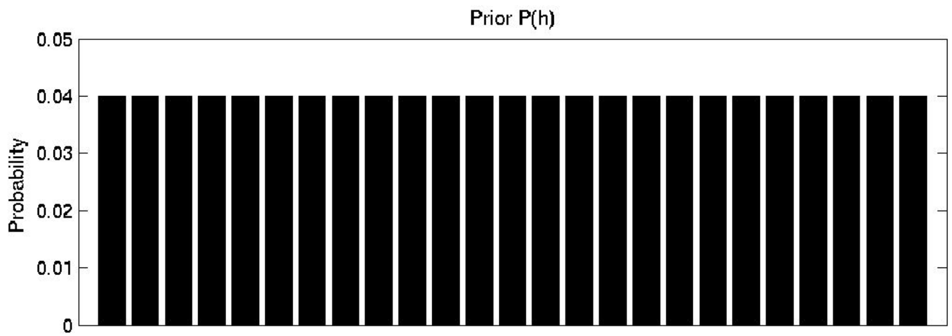
- Inferring a particular causal relation:

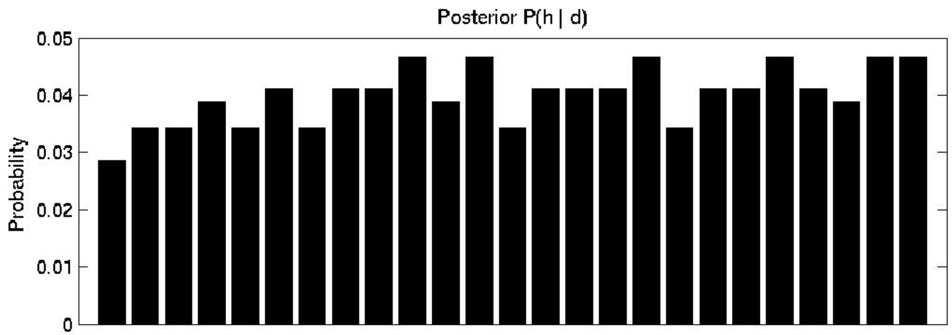
$$P(A \rightarrow E \mid d) = \sum_{h_j \in H} P(A \rightarrow E \mid h_j)P(h_j \mid d)$$

Bayesian inference

With a uniform prior on hypotheses, and the generic parameterization (with uniform prior), integrating over parameters

(Cooper & Herskovits, 1992)





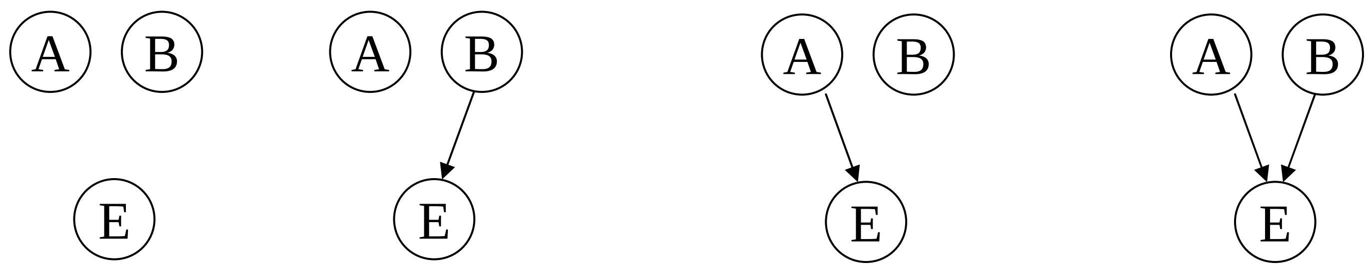
Probability of being a blicket

A	B
0.32	0.32
0.34	0.34

Two key assumptions

- A restricted hypothesis space

$P(h_{00}) = (1 - q)^2$ $P(h_{01}) = (1 - q)\, q$ $P(h_{10}) = q(1 - q)$ $P(h_{11}) = q^2$

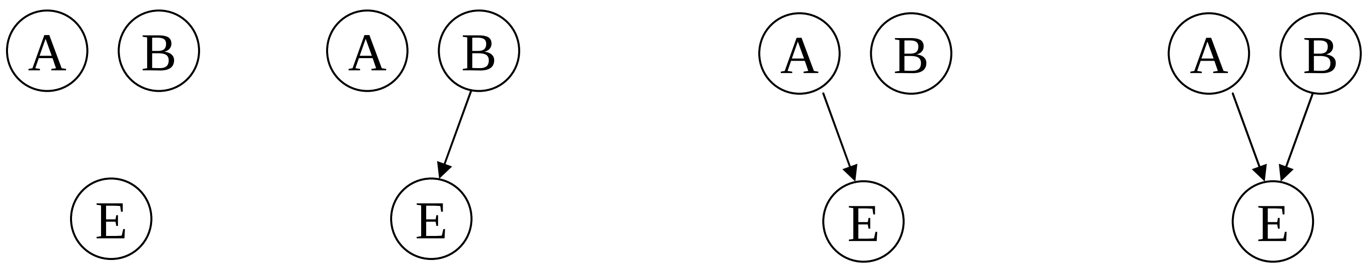


- Detectors follow a deterministic “activation law”
  - always activate if a blicket is on the detector
  - never activate otherwise

(Tenenbaum & Griffiths, 2003; Griffiths, 2005)

Modeling backwards blocking

	$P(h_{00}) = (1-q)^2$	$P(h_{01}) = (1-q)\,q$	$P(h_{10}) = q(1-q)$	$P(h_{11}) = q^2$

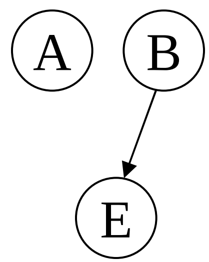


$P(E{=}1 \mid A{=}0, B{=}0)$:	0	0	0	0
$P(E{=}1 \mid A{=}1, B{=}0)$:	0	0	1	1
$P(E{=}1 \mid A{=}0, B{=}1)$:	0	1	0	1
$P(E{=}1 \mid A{=}1, B{=}1)$:	0	1	1	1

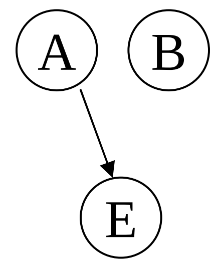
$$P(B \to E \mid d) = P(h_{01}) + P(h_{11}) = q$$

Modeling backwards blocking

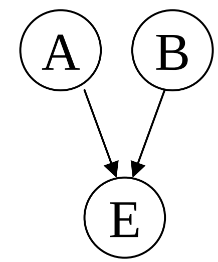
$P(h_{01}) = (1 - q)\, q$



$P(h_{10}) = q(1 - q)$



$P(h_{11}) = q^2$

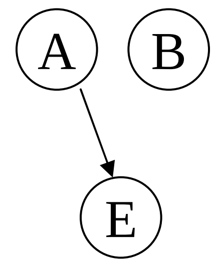


$P(E{=}1 \mid A{=}1, B{=}1)$:	1	1	1

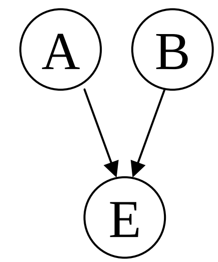
$$P(B \to E \mid d) = \frac{P(h_{01}) + P(h_{11})}{P(h_{01}) + P(h_{10}) + P(h_{11})} = \frac{q}{q + q(1-q)}$$

Modeling backwards blocking

$P(h_{10}) = q(1 - q)$



$P(h_{11}) = q^2$

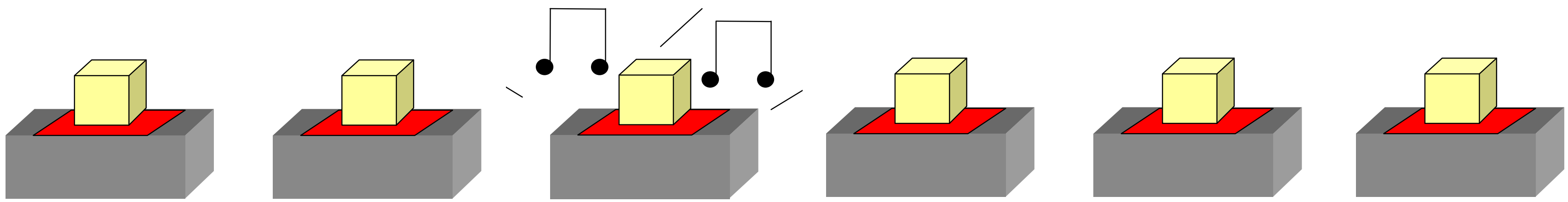


	h_{10}	h_{11}
$P(E{=}1 \mid A{=}1, B{=}0)$:	1	1
$P(E{=}1 \mid A{=}1, B{=}1)$:	1	1

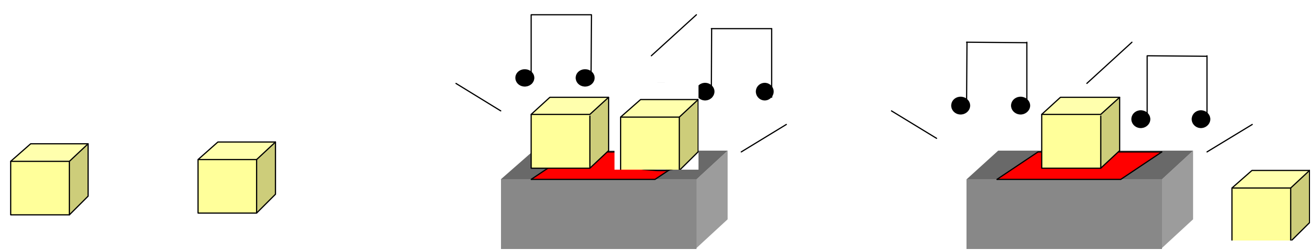
$$P(B \rightarrow E \mid d) = \frac{P(h_{11})}{P(h_{10}) + P(h_{11})} = q$$

Manipulating the prior

I. Pre-training phase: Establish baserate for blickets (q)



II. Backwards blocking phase:



A *B* AB Trial A Trial

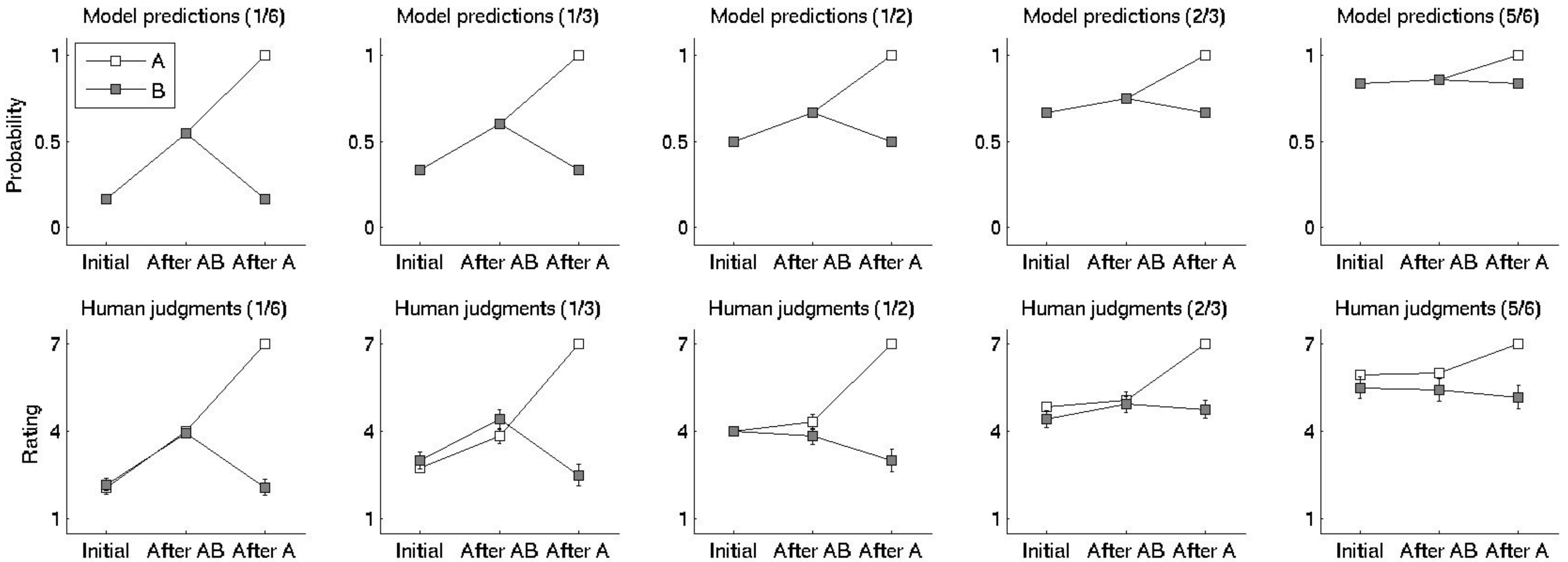
After each trial, adults judge probability that each object is a blicket.

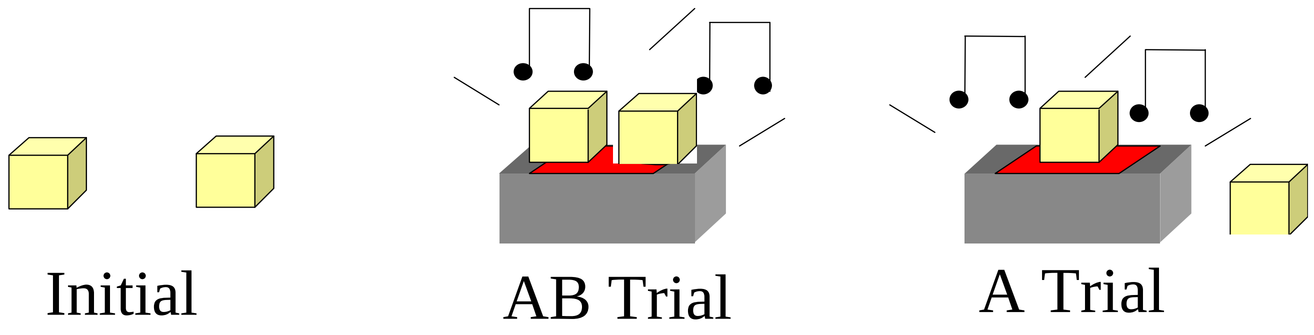
Manipulating the prior

- Expose to different base-rates
  - q = 1/6, 1/3, 1/2, 2/3, 5/6
- Test with backwards blocking
- Model makes two qualitative predictions:
  - evaluation of both A and B as blickets will increase with baserate
  - evaluation of B will increase after AB Trial, then return to baserate after A Trial

(Tenenbaum, Sobel, Griffiths, & Gopnik, submitted)

Manipulating the prior





Theory-based causal induction

Theory

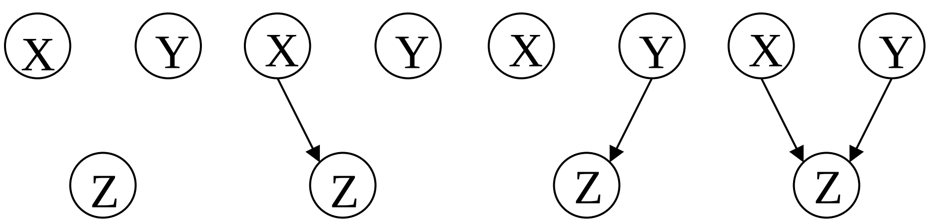
Ontology

Plausible relations

Functional form

generates

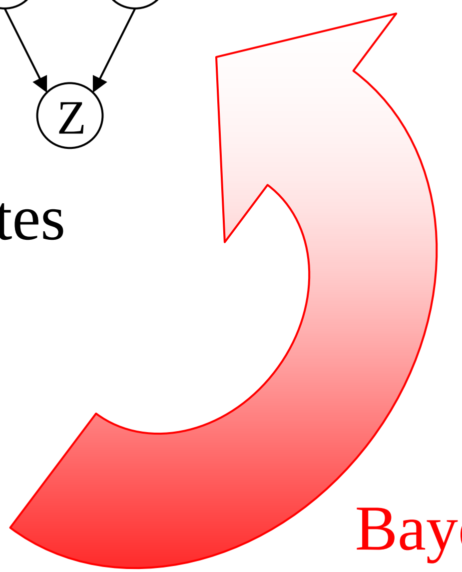
Hypothesis space



generates

Data

Case	X	Y	Z
1	1	0	1
2	0	1	1
3	1	1	1
4	0	0	0
...			



Bayesian inference

Learning causal theories

Theory

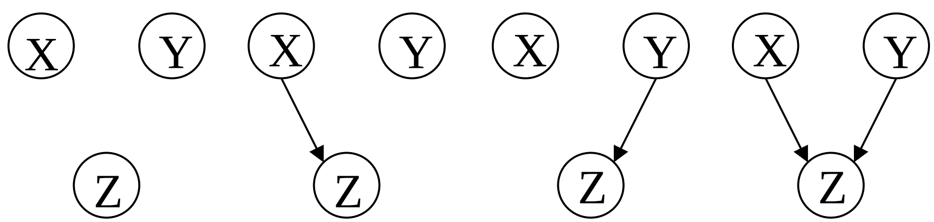
Ontology

Plausible relations

Functional form

generates

Hypothesis space



generates

Data

Case	X	Y	Z
1	1	0	1
2	0	1	1
3	1	1	1
4	0	0	0

...

Learning causal theories

Bayesian inference

Theory

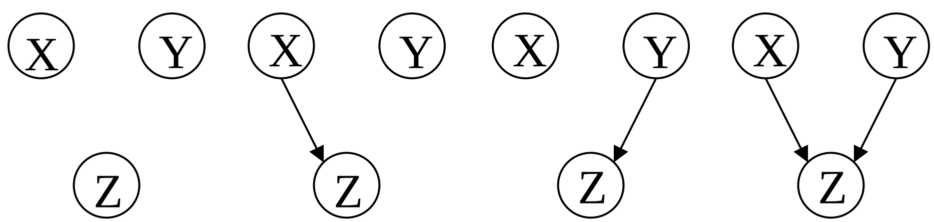
Ontology

Plausible relations

Functional form

generates

Hypothesis space



generates

Data

Case	X	Y	Z
1	1	0	1
2	0	1	1
3	1	1	1
4	0	0	0
	...		

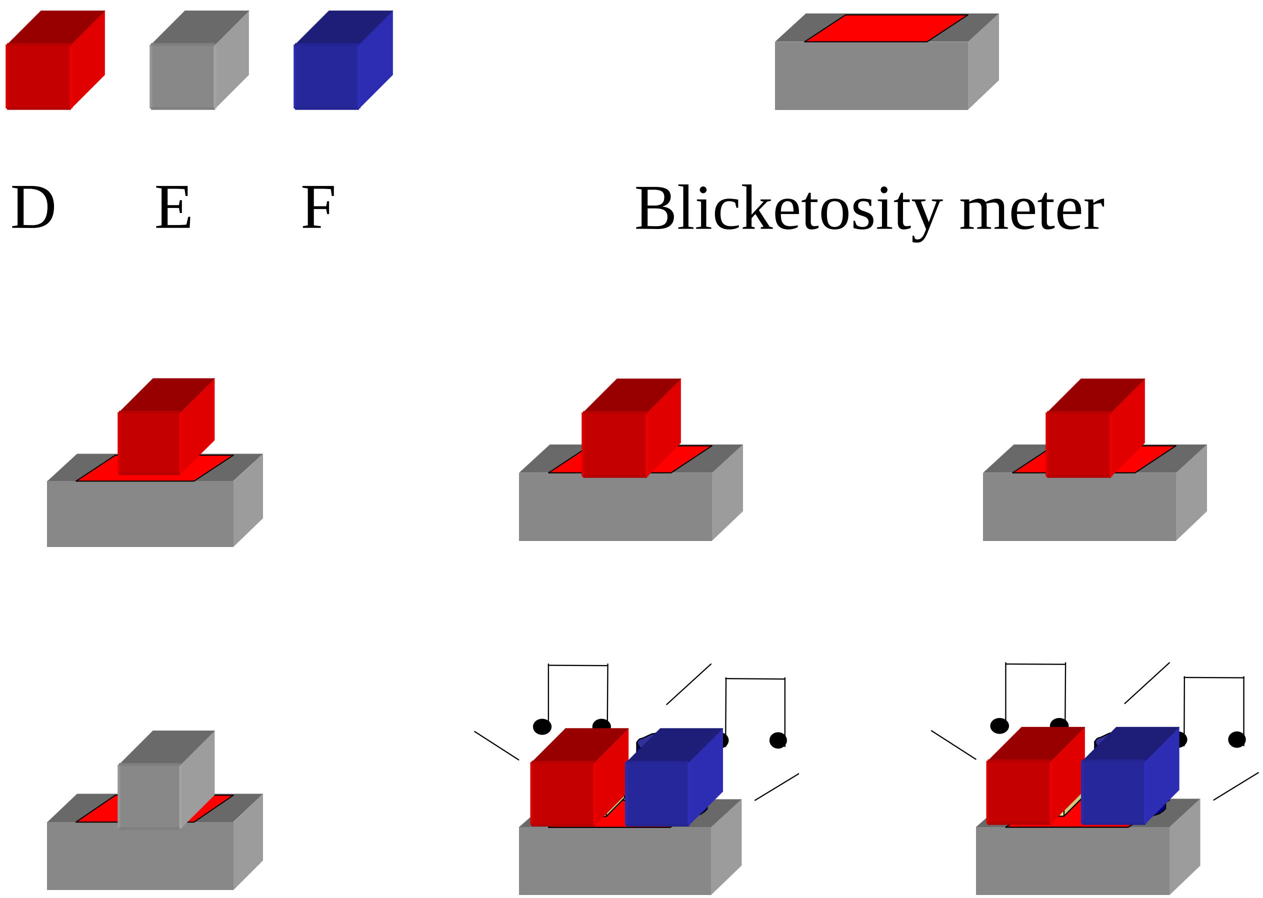
Case	X	Y	Z
1	1	0	1
2	0	1	1
3	1	1	1
4	0	0	0
	...		

Case	X	Y	Z
1	1	0	1
2	0	1	1
3	1	1	1
4	0	0	0
	...		

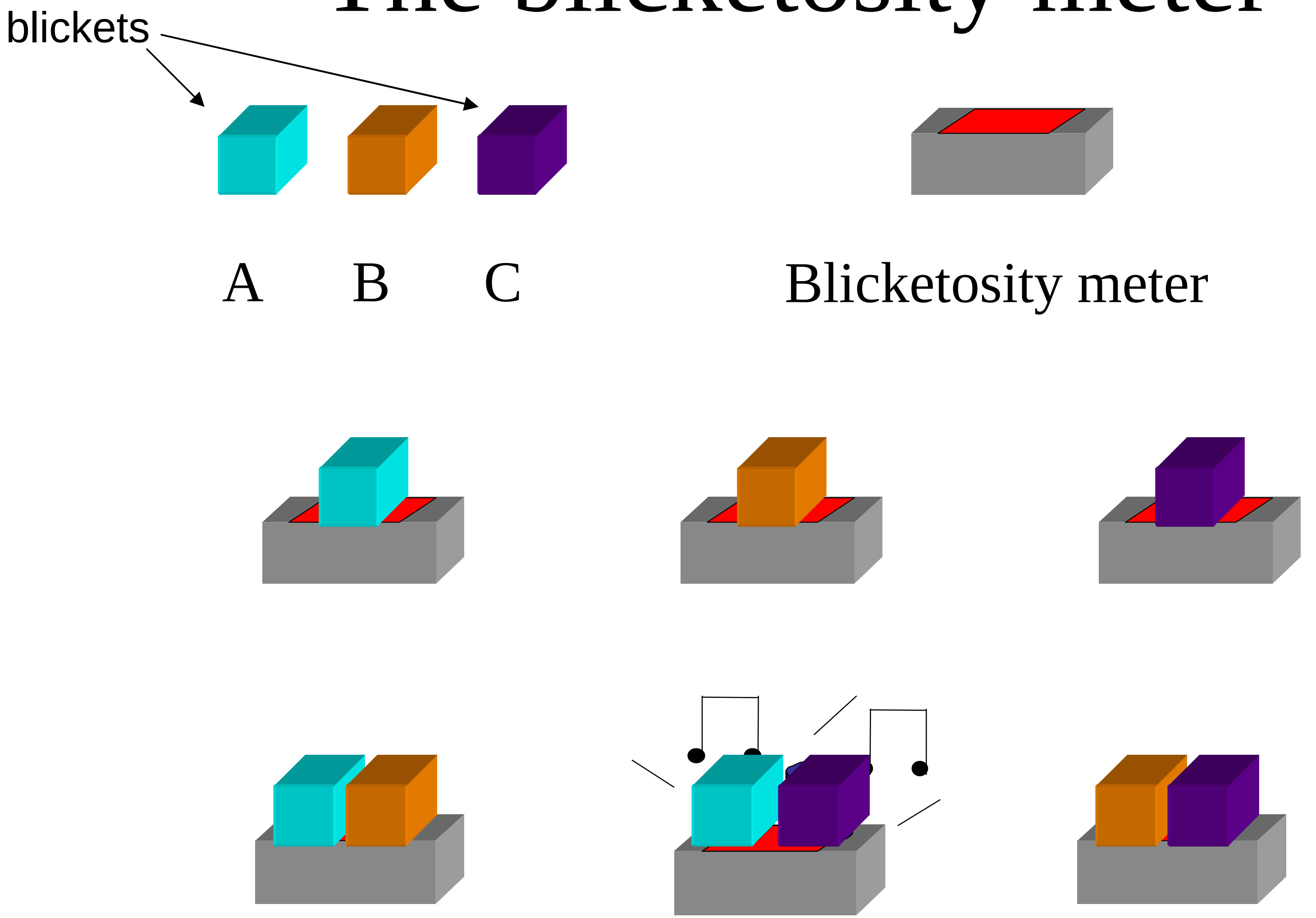
Case	X	Y	Z
1	1	0	1
2	0	1	1
3	1	1	1
4	0	0	0
	...		

Case	X	Y	Z
1	1	0	1
2	0	1	1
3	1	1	1
4	0	0	0
	...		

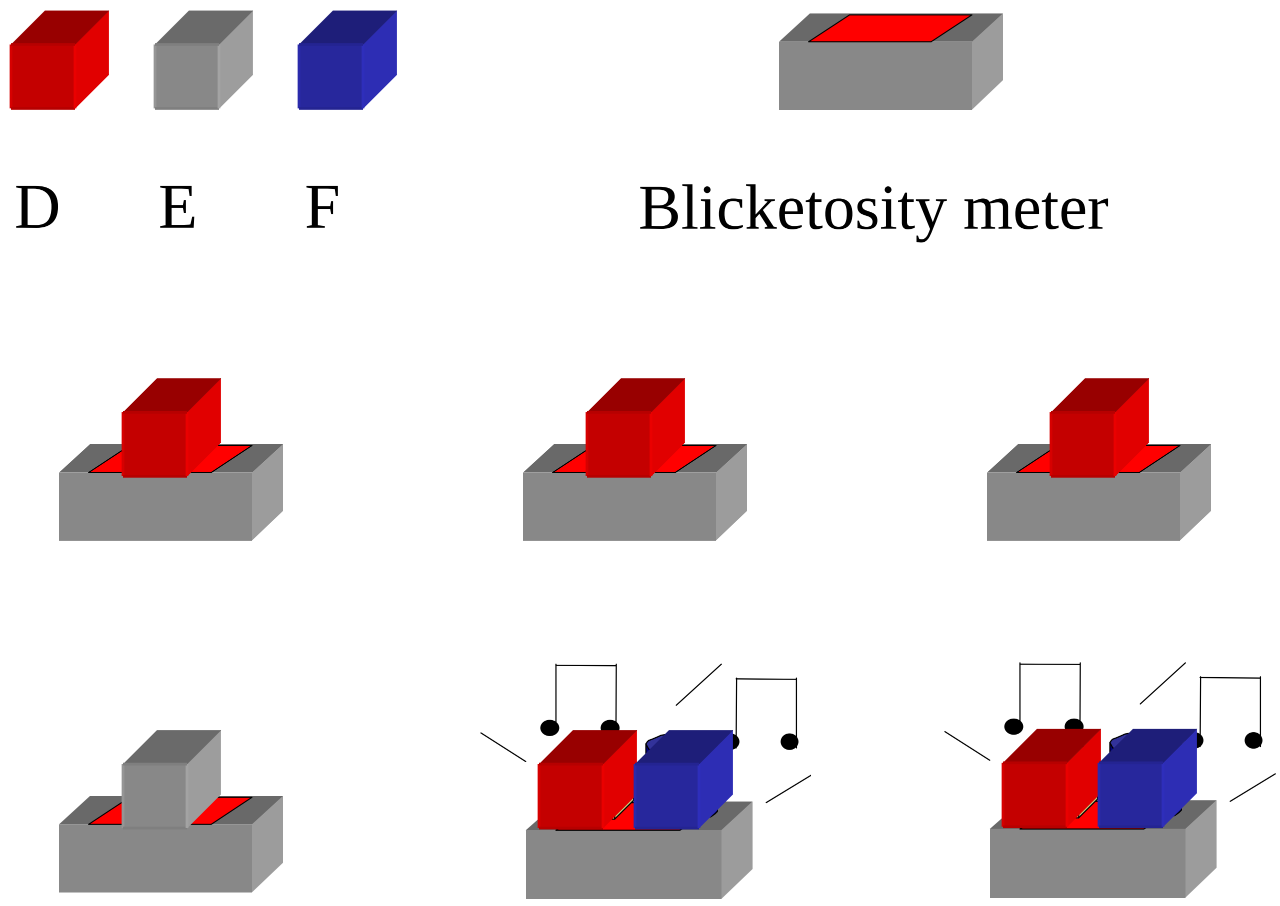
The blicketosity meter



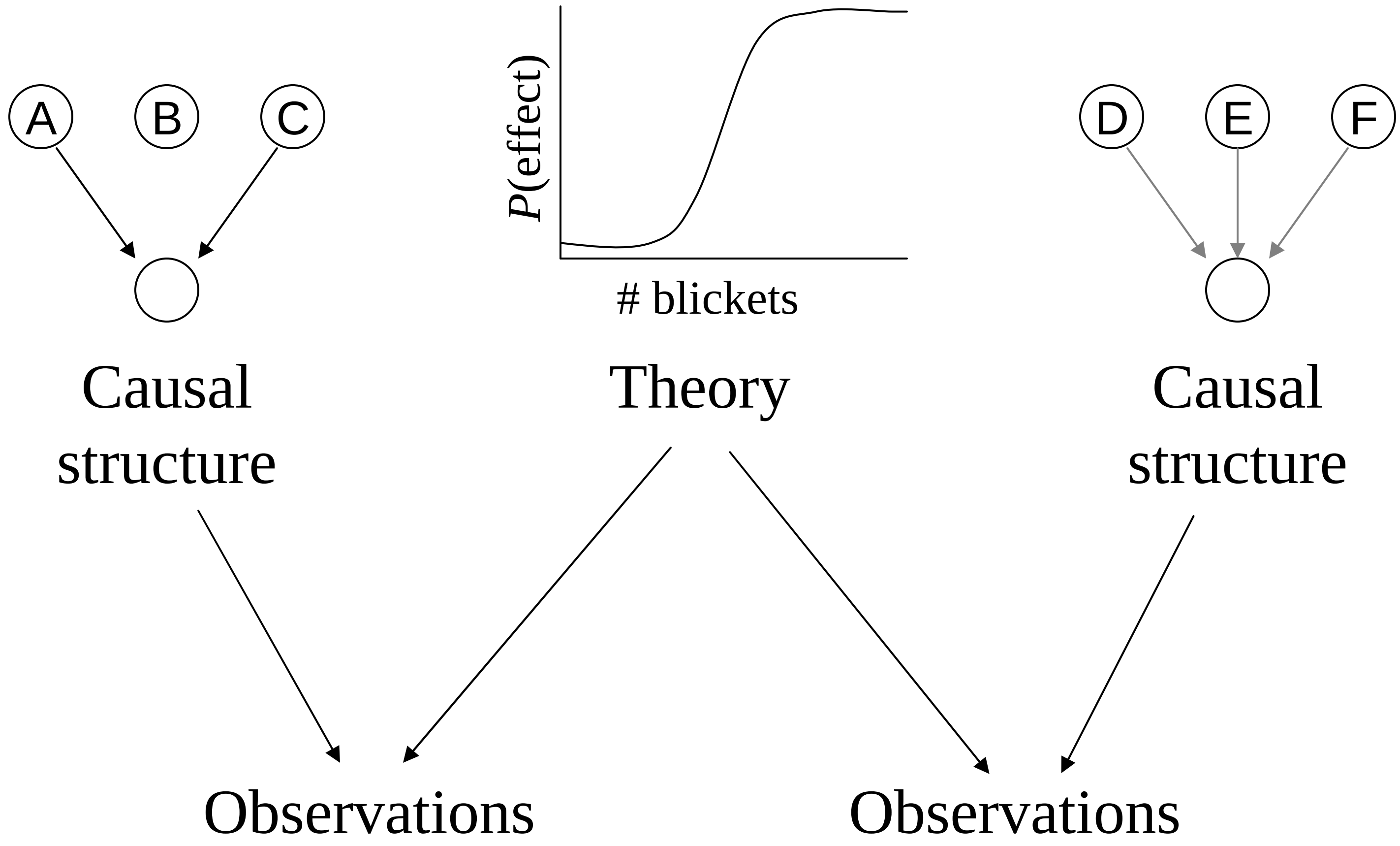
The blicketosity meter



The blicketosity meter

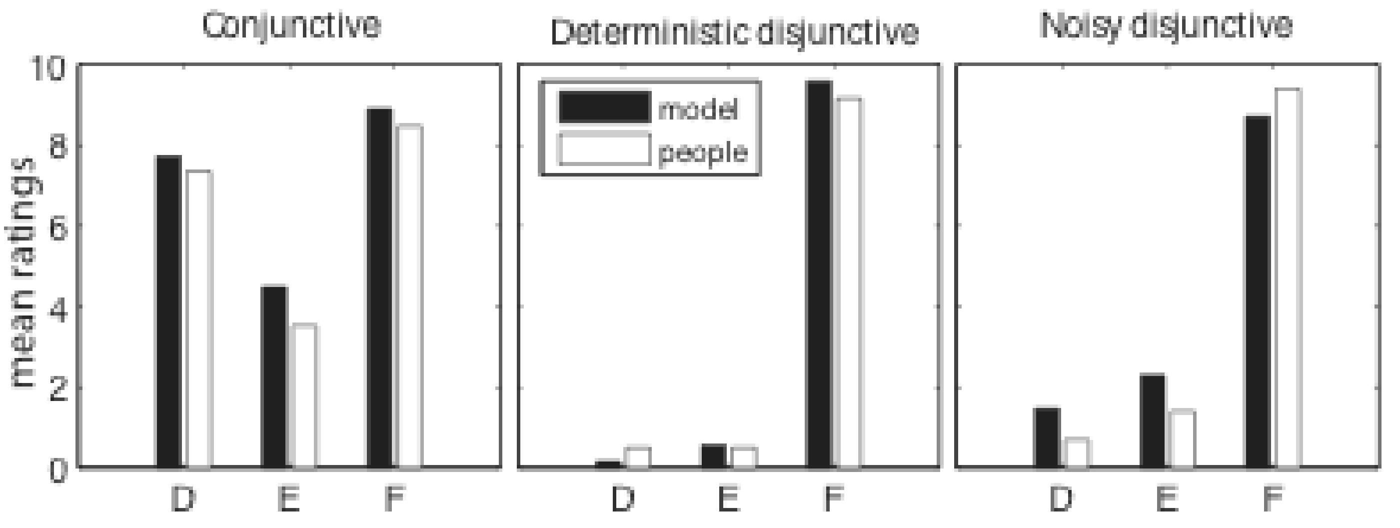


Learning functional form



(Lucas & Griffiths, 2009)

Results



- Model also accounts for fully unsupervised learning of functional form, domain sensitivity
- Compatible with continuous causes

(Lu, Rojas, Beckers & Yuille, 2008)

Learning an ontology

Learning from sparse data requires constraints

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

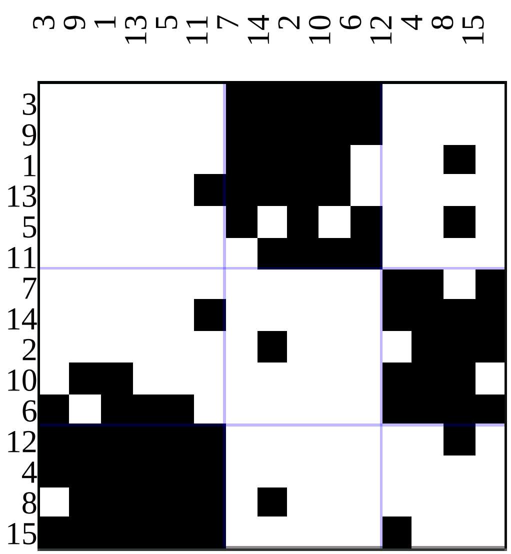
(Segal,Pe'er, Regev, Koller, & Friedman, 2005)

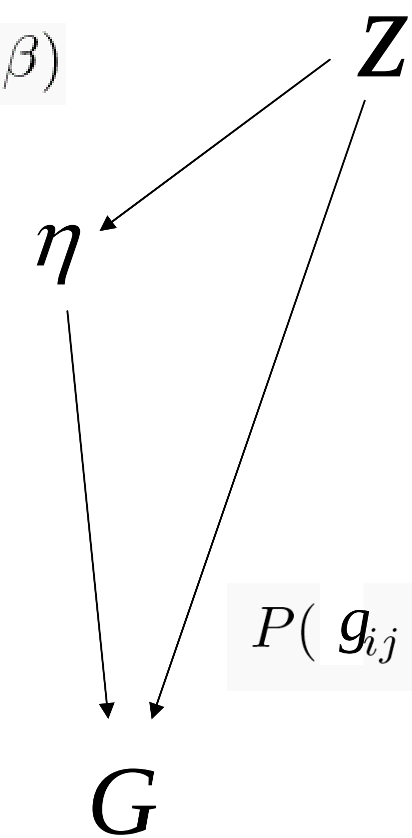
Nonparametric Block Model (NBM)

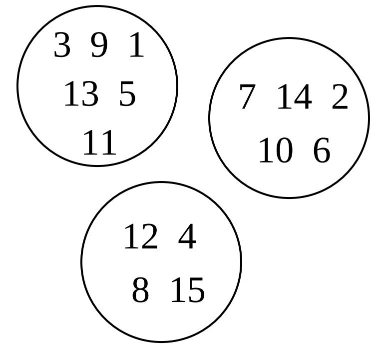
$z \sim \mathrm{CRP}(\alpha)$

$\eta_{ij} \sim \mathrm{Beta}(\beta, \beta)$

0.1	0.9	0.1
0.1	0.1	0.9
0.9	0.1	0.1



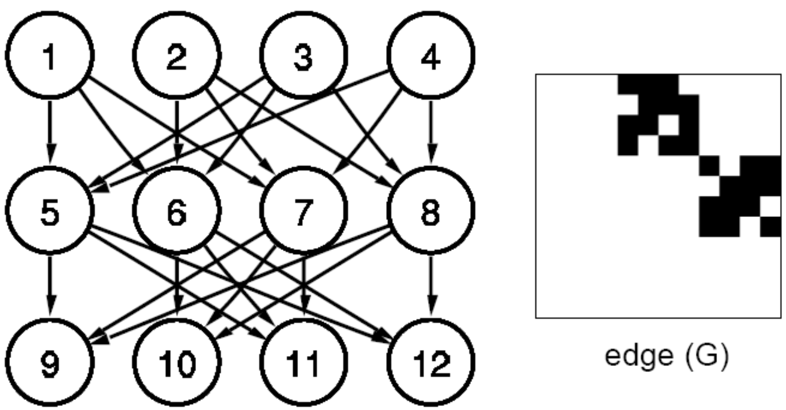


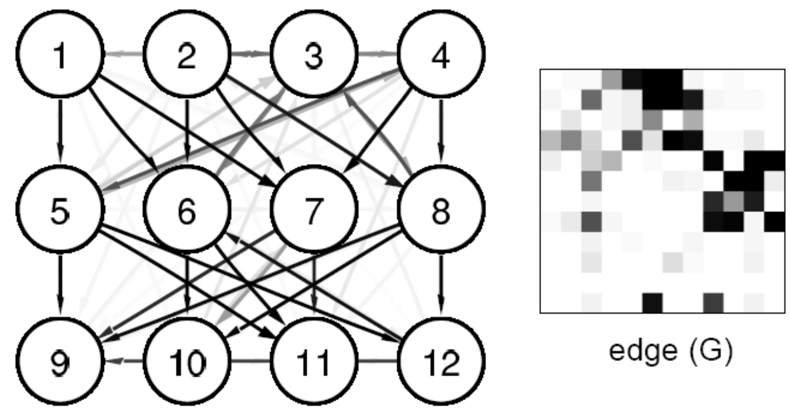


$P(g_{ij} = 1 | z_i, z_j) = \eta_{ij}$

(Mansinghka, Kemp, Tenenbaum & Griffiths, 2006)

Causal learning without a theory:





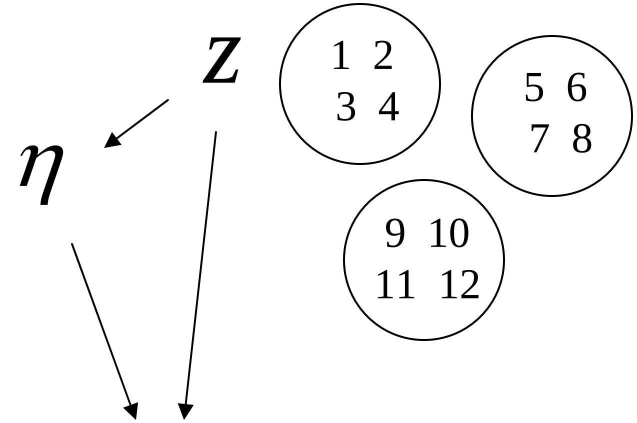
True causal network

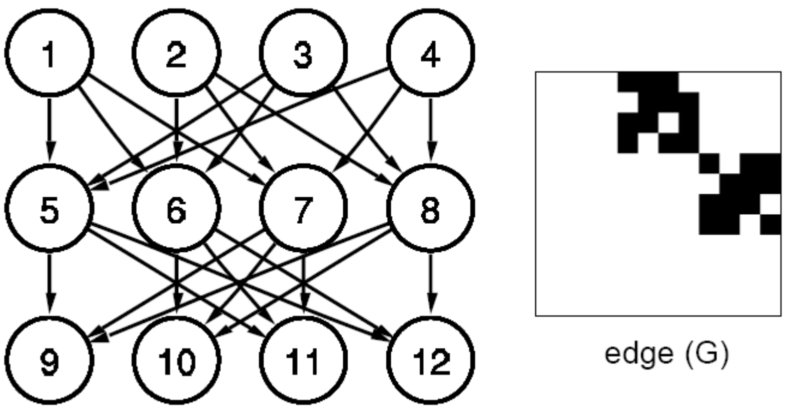
Sample 75 observations…

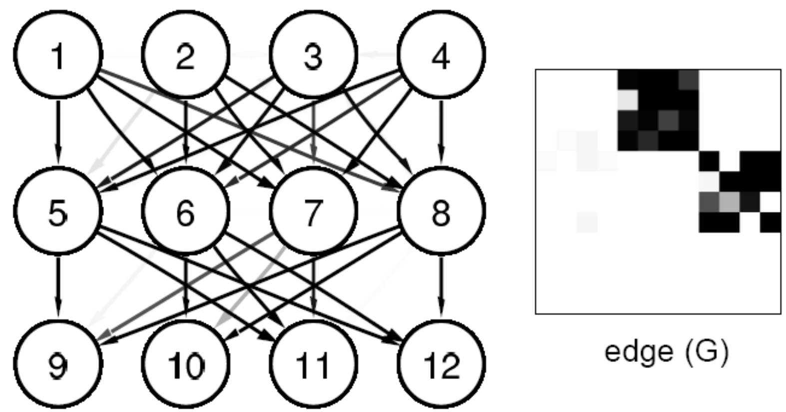
(Mansinghka, Kemp, Tenenbaum & Griffiths, 2006)

Causal learning with a (NBM) theory

0.0	0.8	0.01
0.0	0.0	0.75
0.0	0.0	0.0





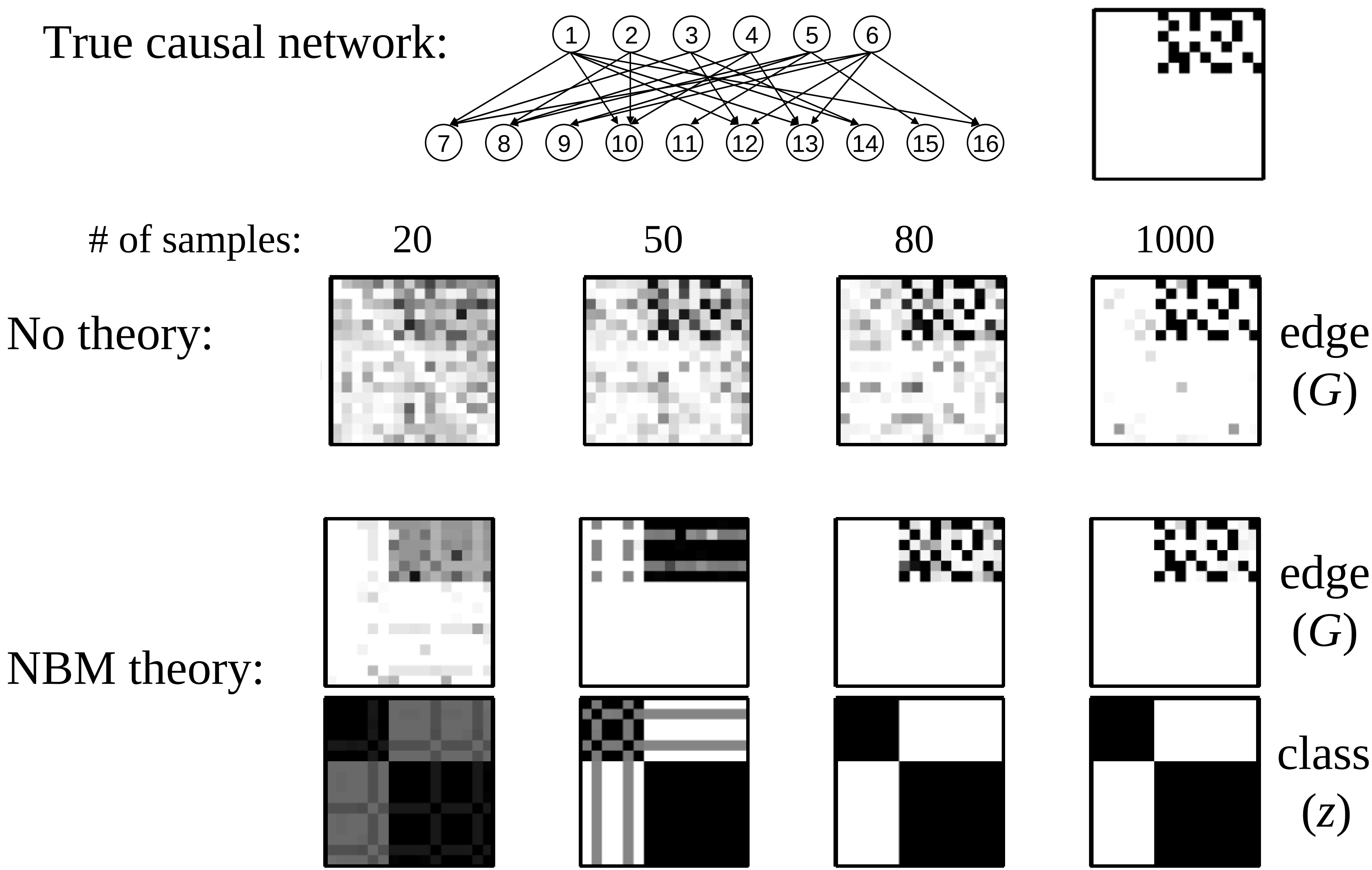


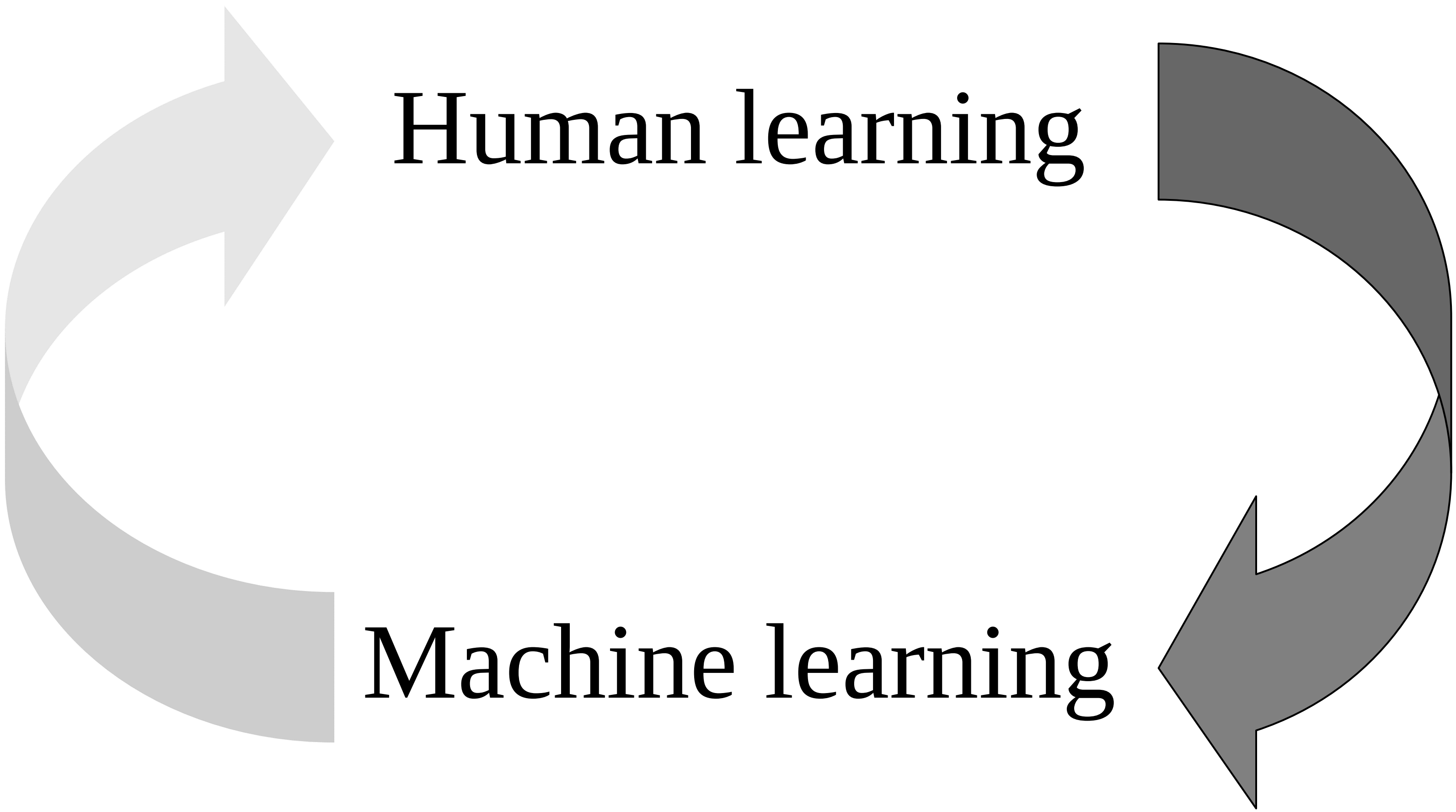
True causal network

Sample 75 observations…

(Mansinghka, Kemp, Tenenbaum & Griffiths, 2006)

The “blessing of abstraction”




Human learning
Machine learning

Causal learning and graphical models

- Tools from machine learning can help to clarify the computational problem of causal learning
- But… human causal learning is guided by strong constraints, which make it possible to learn from small amounts of data
  - e.g., noisy-OR, determinism
- Similar constraints can improve machine learning
  - e.g., nonparametric block model

(Mansinghka, Kemp, Tenenbaum & Griffiths, 2006)