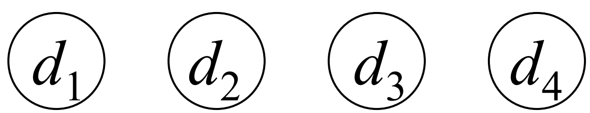
Outline

- Bayesian Ockham’s Razor
- Bayes nets (directed graphical models)
  - Computational motivation: tractable reasoning
  - Cognitive motivation: causal reasoning
  - Sampling methods for approximate inference

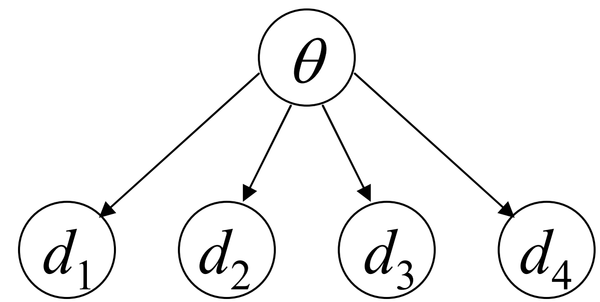
Coin flipping

- Comparing two simple hypotheses
  - $P(\mathsf{H}) = 0.5$ vs. $P(\mathsf{H}) = 1.0$
- Comparing simple and complex hypotheses
  - $P(\mathsf{H}) = 0.5$ vs. $P(\mathsf{H}) = \theta$
- Comparing infinitely many hypotheses
  - $P(\mathsf{H}) = \theta$: Infer θ

Comparing simple and complex hypotheses



vs.



Fair coin, $P(\mathsf{H}) = 0.5$

$P(\mathsf{H}) = \theta$

- Which provides a better account of the data: the simple hypothesis of a fair coin, or the complex hypothesis that $P(\mathsf{H}) = \theta$?

Comparing simple and complex hypotheses

- $P(\mathsf{H}) = \theta$ is more complex than $P(\mathsf{H}) = 0.5$ in two ways:
  - $P(\mathsf{H}) = 0.5$ is a special case of $P(\mathsf{H}) = \theta$
  - for any observed sequence D, we can choose θ such that D is more probable than if $P(\mathsf{H}) = 0.5$

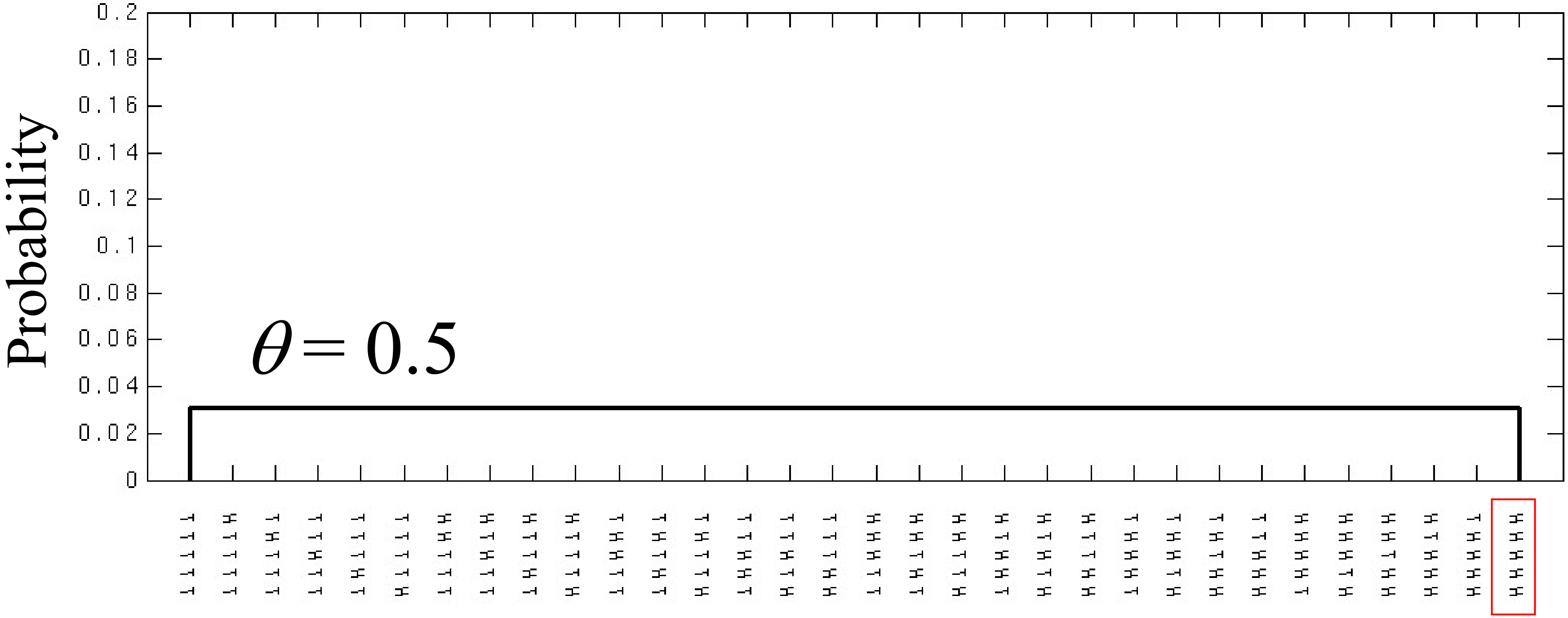
Bernoulli Distribution: $P(D \mid \theta) = \theta^n (1-\theta)^{N-n}$

n = # of heads in D

N = # of flips in D

Comparing simple and complex hypotheses

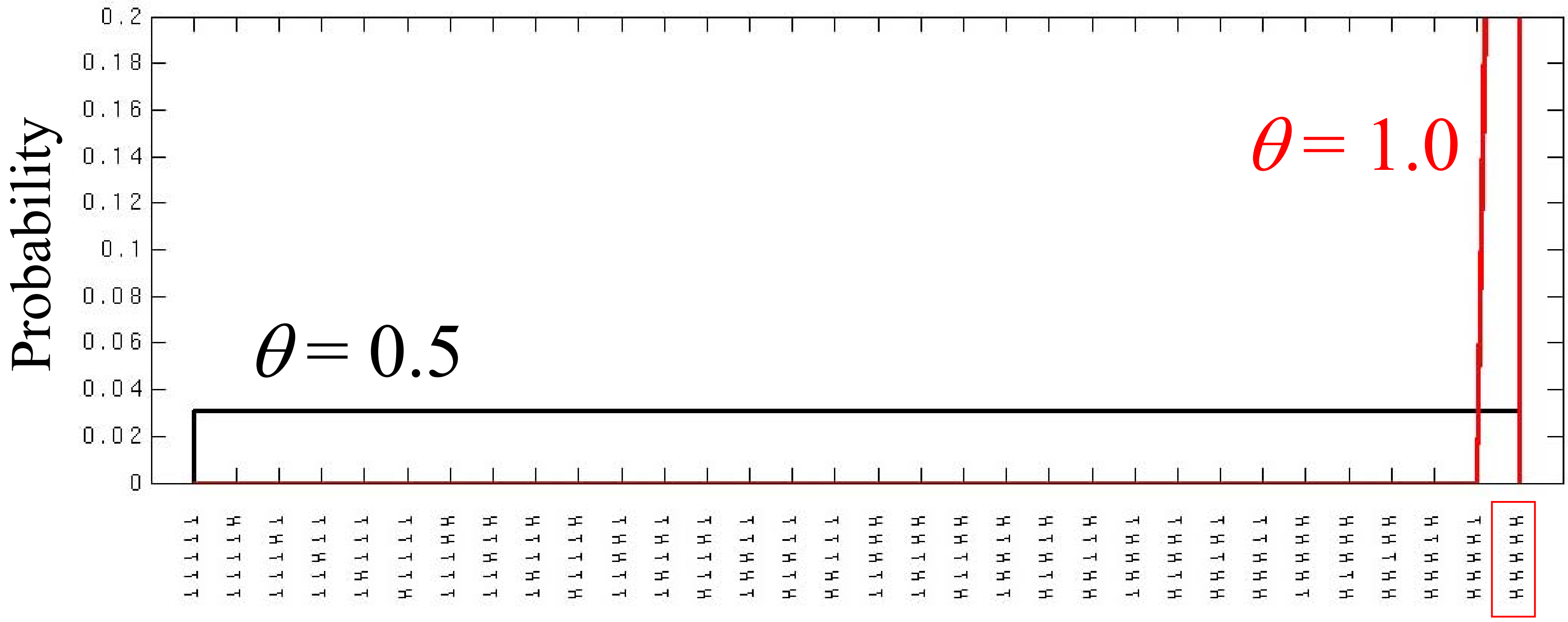
$$P(D\,|\,\theta) = \theta^n (1-\theta)^{N-n}$$



D = HHHHH

Comparing simple and complex hypotheses

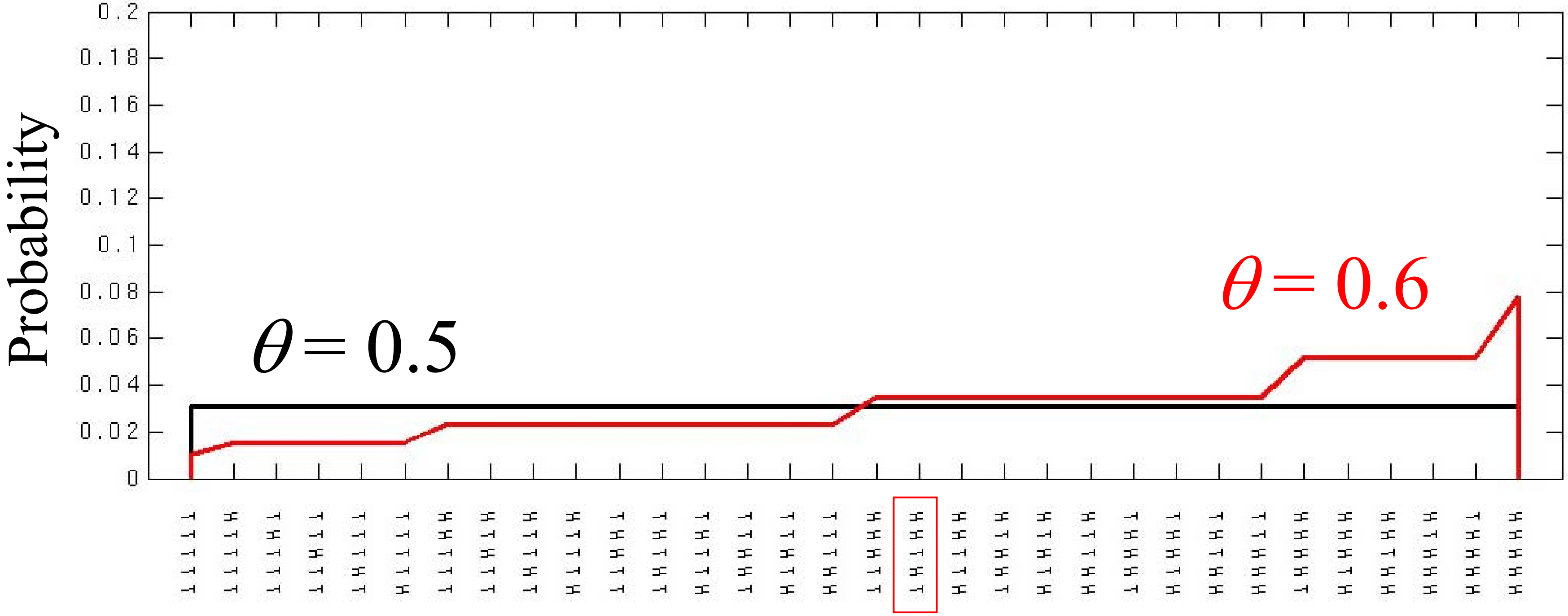
$$P(D \mid \theta) = \theta^{n}(1-\theta)^{N-n}$$



D = HHHHH

Comparing simple and complex hypotheses

$$P(D \mid \theta) = \theta^n (1-\theta)^{N-n}$$



$D =$ HHTHT

Comparing simple and complex hypotheses

- $P(\mathsf{H}) = \theta$ is more complex than $P(\mathsf{H}) = 0.5$ in two ways:
  - $P(\mathsf{H}) = 0.5$ is a special case of $P(\mathsf{H}) = \theta$
  - for any observed sequence X, we can choose θ such that X is more probable than if $P(\mathsf{H}) = 0.5$
- How can we deal with this?
  - Some version of Ockham's razor:?
  - Bayes: just the law of conservation of belief!

Comparing simple and complex hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

Computing $P(D|H_1)$ is easy:

$$P(D \mid H_1) = (1/2)^n (1-1/2)^{N-n} = 1/2^N$$

Compute $P(D|H_2)$ by averaging over θ:

$$P(D \mid H_2) = \int_0^1 P(D \mid \theta)\, p(\theta \mid H_2)\, d\theta$$

Comparing simple and complex hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

Computing $P(D|H_1)$ is easy:

$$P(D \mid H_1) = (1/2)^n (1-1/2)^{N-n} = 1/2^N$$

Compute $P(D|H_2)$ by averaging over θ:

$$P(D \mid H_2) = \int_0^1 P(D \mid \theta) d\theta$$

(assume uniform prior on θ)

Comparing simple and complex hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

Computing $P(D|H_1)$ is easy:

$$P(D \mid H_1) = (1/2)^n (1-1/2)^{N-n} = 1/2^N$$

Compute $P(D|H_2)$ by averaging over θ:

$$P(D \mid H_2) = \int_0^1 \theta^n (1-\theta)^{N-n} \, d\theta$$

Comparing simple and complex hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

Computing $P(D|H_1)$ is easy:

$$P(D \mid H_1) = (1/2)^n (1-1/2)^{N-n} = 1/2^N$$

Compute $P(D|H_2)$ by averaging over θ:

$$P(D \mid H_2) = \int_0^1 \theta^n (1-\theta)^{N-n} \, d\theta = \frac{n!(N-n)!}{(N+1)!}$$

(How is this an average?)

- Consider a discrete approximation with 11 values of θ, from 0 to 1 in steps of 1/10:

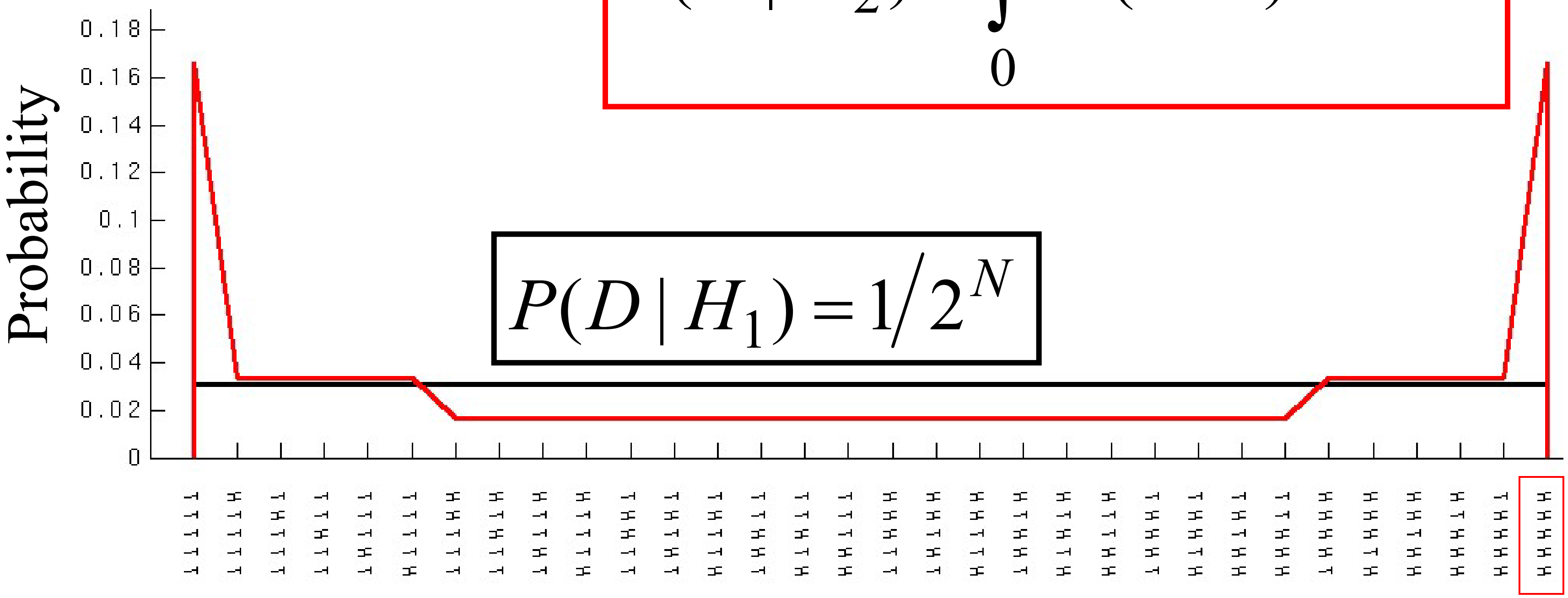
$$P(D \mid H_2) = \sum_{i=0}^{10} P(D \mid \theta = i/10)\, p(\theta = i/10 \mid H_2)$$

$$P(D \mid H_2) = \sum_{i=0}^{10} P(D \mid \theta = i/10)\ (1/11)$$

(c.f., $P(D \mid H_2) = \int_0^1 P(D \mid \theta)\, d\theta$)

Comparing simple and complex hypotheses

$$P(D \mid H_2) = \int_0^1 \theta^n (1-\theta)^{N-n} \, d\theta$$

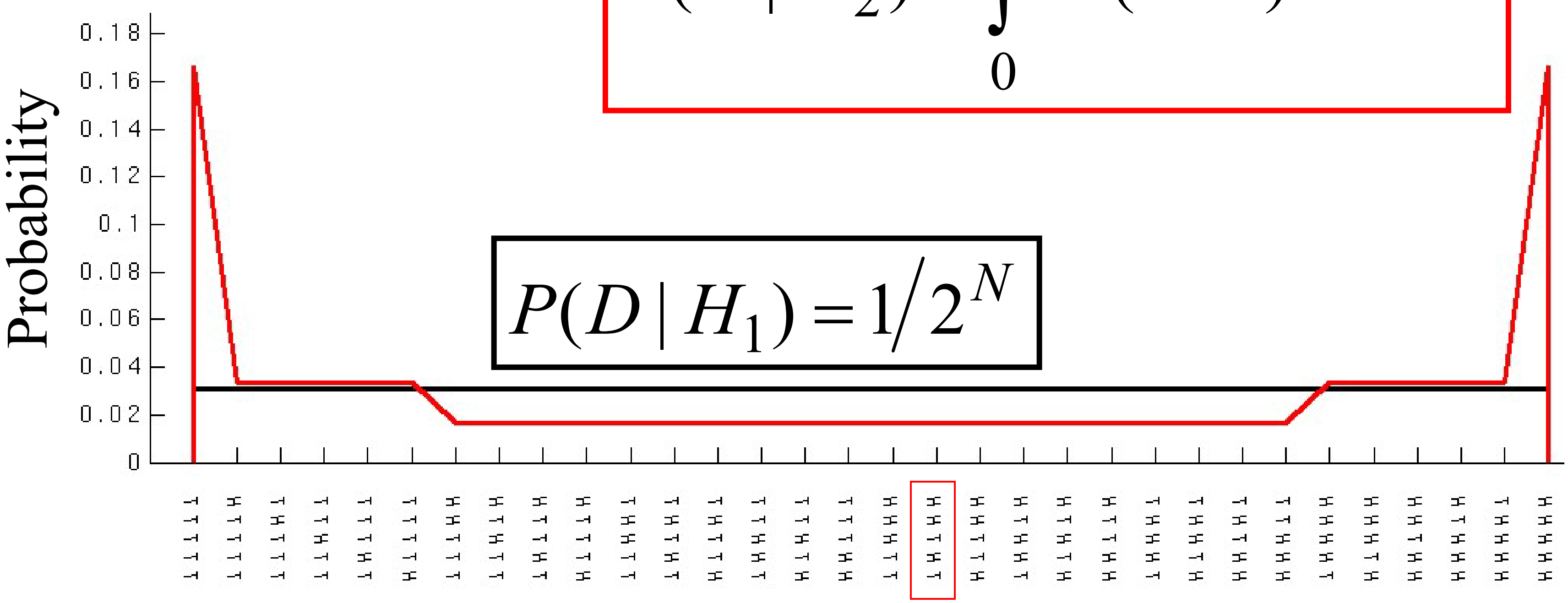


$$P(D \mid H_1) = 1/2^N$$

D = HHHHH

Comparing simple and complex hypotheses

$$P(D \mid H_2) = \int_0^1 \theta^n (1-\theta)^{N-n} d\theta$$



$$P(D \mid H_1) = 1/2^N$$

D = HHHHH

Law of conservation of belief

$$\sum_i P(X = x_i) = 1$$

- Two different stages
  - Prior over model parameter:

$$\int_0^1 p(\theta \mid H_2) d\theta = 1$$

In a model with a wider range of parameter values, each setting of the parameters contributes less to the model predictions.

Law of conservation of belief

$$\sum_i P(X = x_i) = 1$$

- Two different stages
  - Prior over model parameter:

$$\int_0^1 p(\theta \mid H_2) d\theta = 1$$

  - Likelihood (probability over data):

$$\sum_d P(D = d \mid H_2) = \sum_d \int_\theta P(D = d \mid \theta) p(\theta \mid H_2) d\theta = 1$$

A model that predicts some data sets very well must predict others very poorly.

Bayesian Ockham's Razor

Image removed due to copyright considerations.

Two alternative models

- Fudged Newton
  - A new planet: Vulcan?
  - Matter rings around the sun?
  - Sun is slightly lopsided.
  - Exponent in Universal law of gravitation is $2 + \varepsilon$ instead of 2.
  - Each version of this hypothesis has a fudge factor, whose most likely value we can estimate empirically . . . .

Image removed due to copyright considerations.

- Simplifying assumption: predictions of fudged Newton are Gaussian around 0.

More formally….

Image removed due to copyright considerations.

ε : fudge factor

$$p(d \mid M) = \int_{\varepsilon} p(d, \varepsilon \mid M)$$

$$= \int_{\varepsilon} p(d \mid \varepsilon, M)\, p(\varepsilon \mid M)$$

$$\leq \max_{\varepsilon} p(d \mid \varepsilon, M)$$

Two alternative models

- Fudged Newton
- Einstein: General Relativity + experimental error (+/- 2 arc seconds/century).

Comparing the models

Image removed due to copyright considerations.

Where is Occam's razor?

- Why not a more "complex" fudge, in which the Gaussian can vary in both mean and variance?

Image removed due to copyright considerations.

Bayesian Occam's razor

- Recall: predictions of a model are the weighted average over all parameter values.

$$p(d \mid M) = \int_{\mu,\sigma} p(d \mid \mu, \sigma, M)\, p(\mu \mid M)\, p(\sigma \mid M)\, d\mu\, d\sigma$$

Image removed due to copyright considerations.

- Only a small set of parameter values fit the data well, so average fit is poor.

Law of conservation of belief

$$\sum_i P(X = x_i) = 1$$

- Two different stages
  - Priors over model parameters:

$$\int_{\mu,\sigma} p(\mu,\sigma \mid M)\, d\mu\, d\sigma = 1$$

  - Likelihood (probability over data):

$$\int_x p(x \mid M)\, dx = \int_x \int_{\mu,\sigma} p(x \mid \mu,\sigma,M)\, p(\mu,\sigma \mid M)\, d\mu\, d\sigma\, dx = 1$$

A model that can predict many possible data sets must assign each of them low probability.

Bayesian Occam's Razor

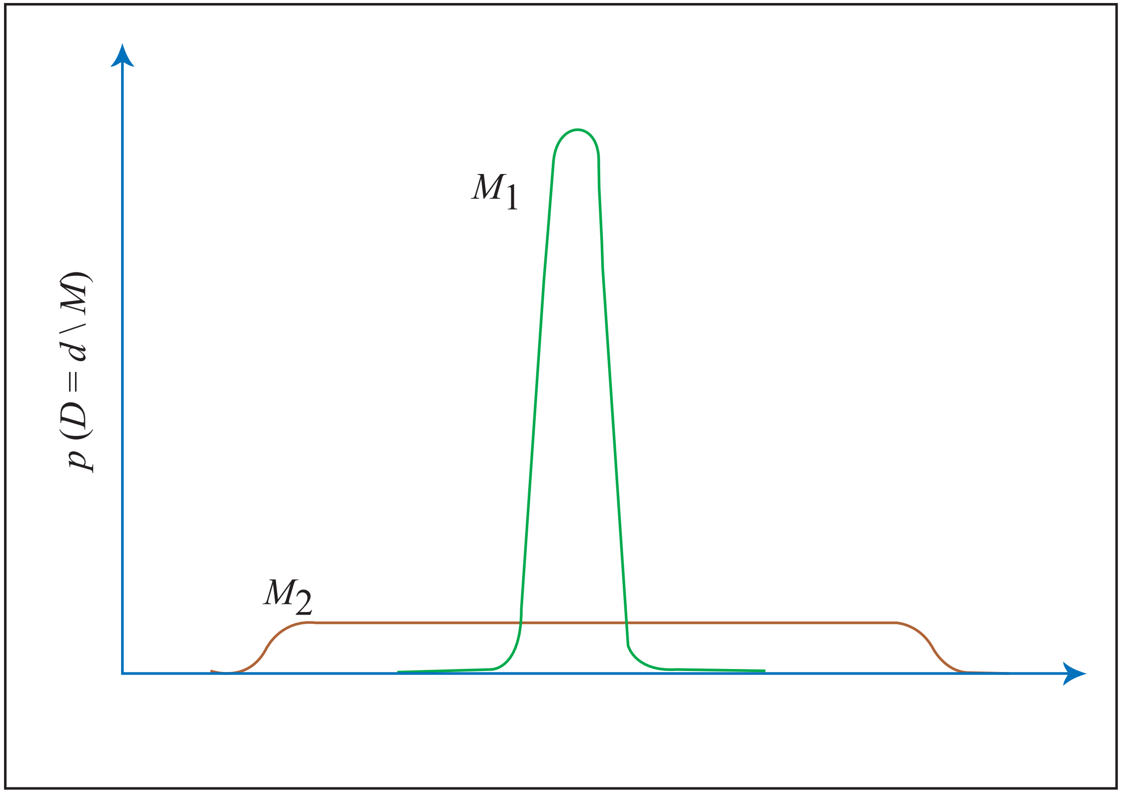


Figure by MIT OCW.

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

Ockham's Razor in curve fitting

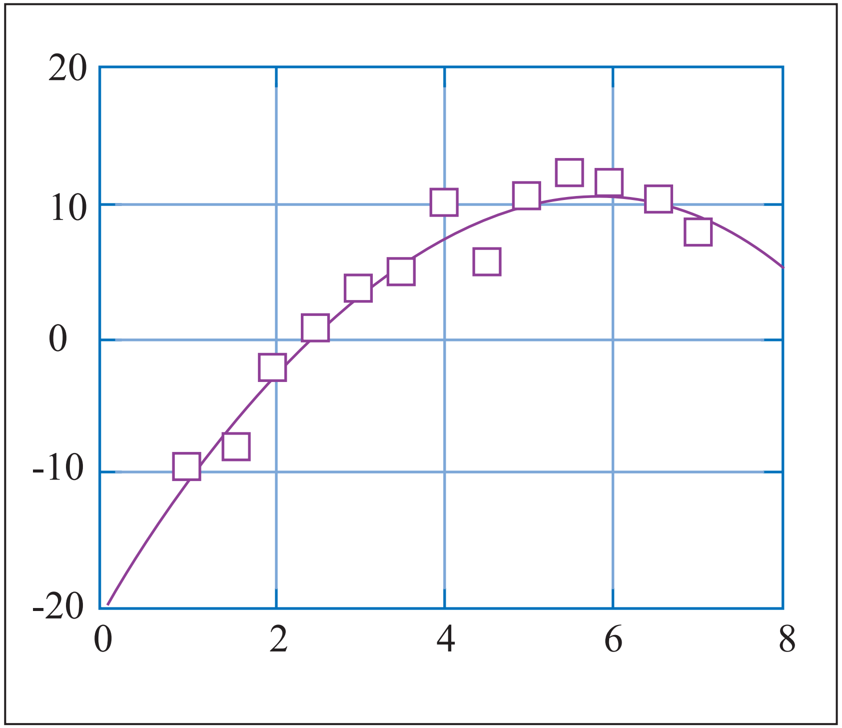


Figure by MIT OCW.

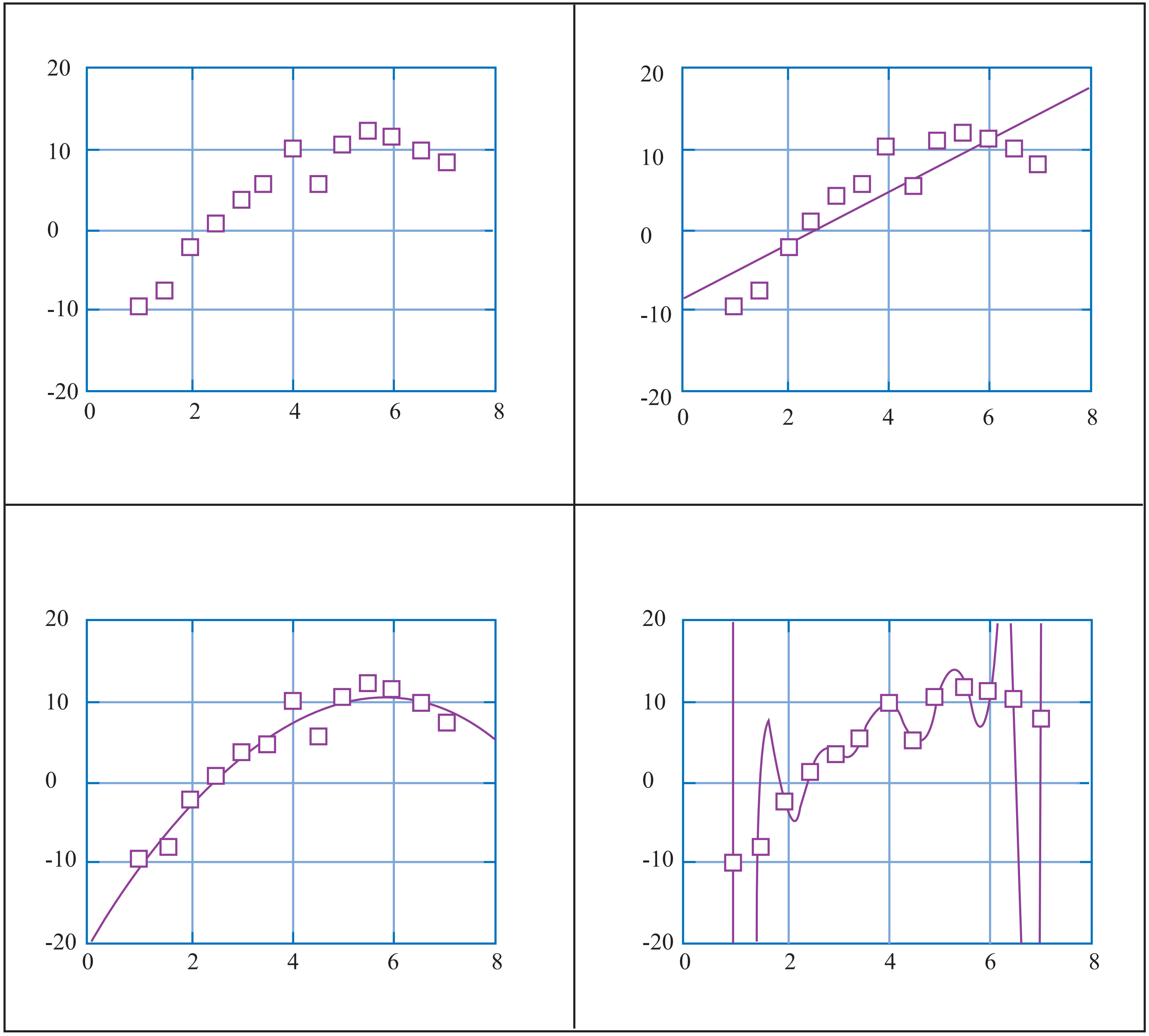


Figure by MIT OCW.

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

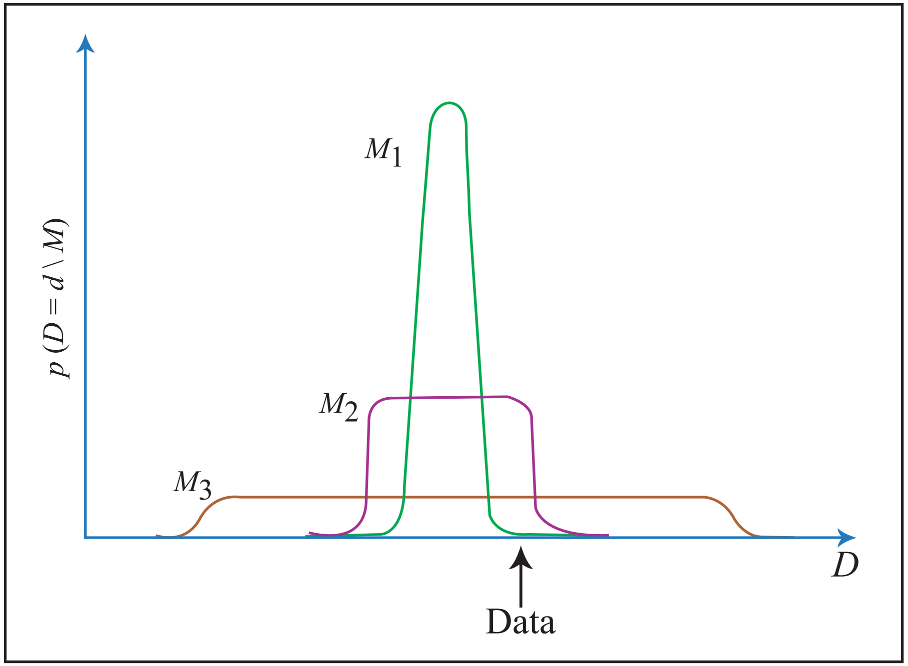


Figure by MIT OCW.

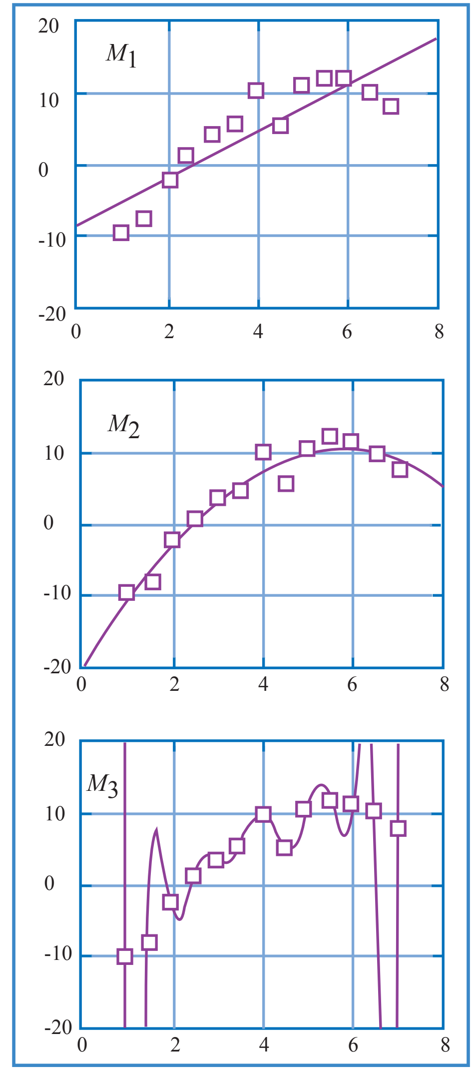
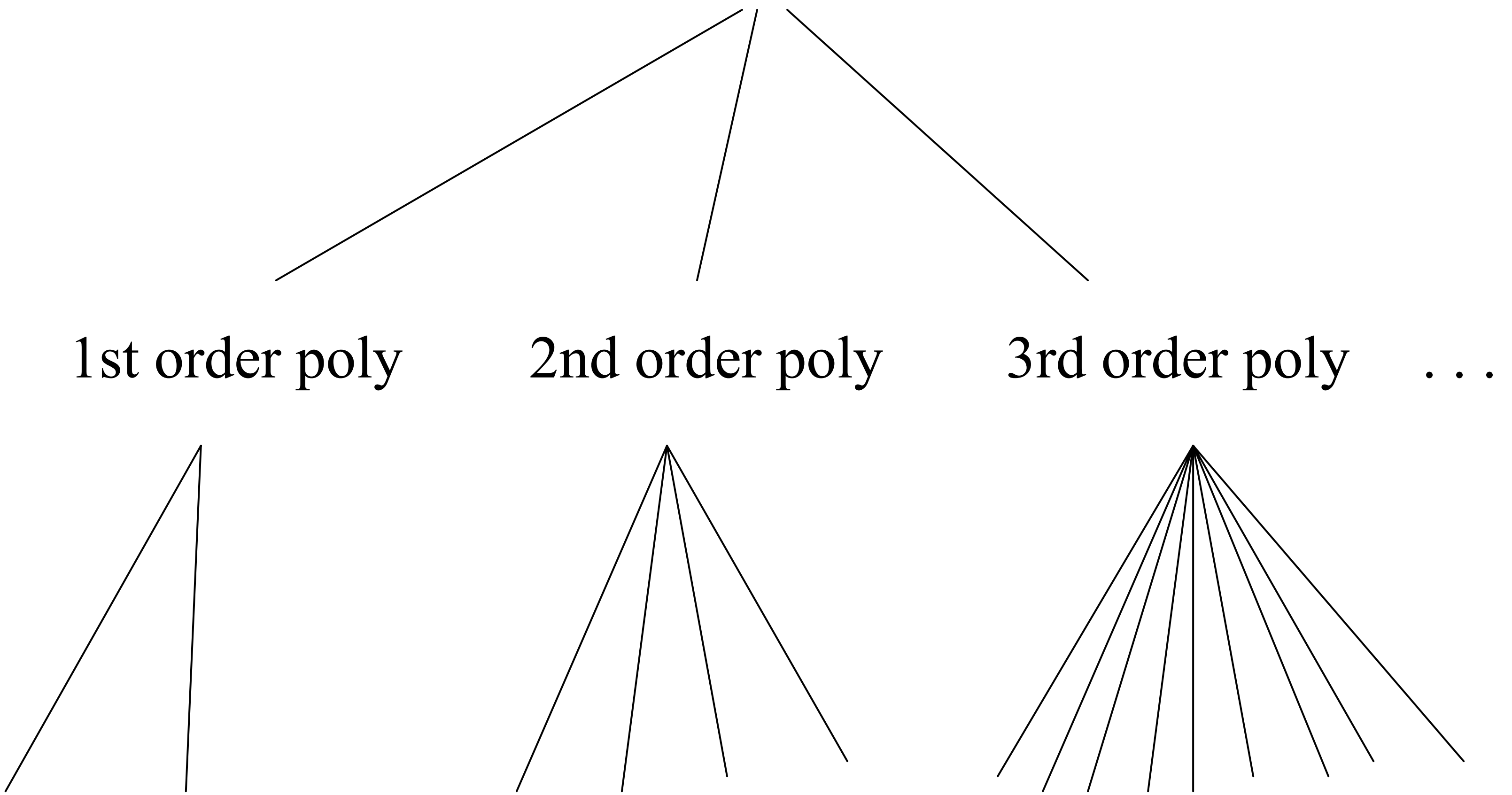


Figure by MIT OCW.

A model that can predict many possible data sets must assign each of them low probability.

Hierarchical prior



Likelihood function for regression

- Assume y is a linear function of x plus Gaussian noise:

Image removed due to copyright considerations.

- Linear regression is maximum likelihood: Find the function f: $x \rightarrow$ y that makes the data most likely.

Likelihood function for regression

- Assume y is a linear function of x plus Gaussian noise:

Image removed due to copyright considerations.

- Linear regression is maximum likelihood: Find the function f: $x \rightarrow$ y that makes the data most likely.

Likelihood function for regression

- Assume *y* is a linear function of *x* plus Gaussian noise:

Image removed due to copyright considerations.

- Not the maximum likelihood function….

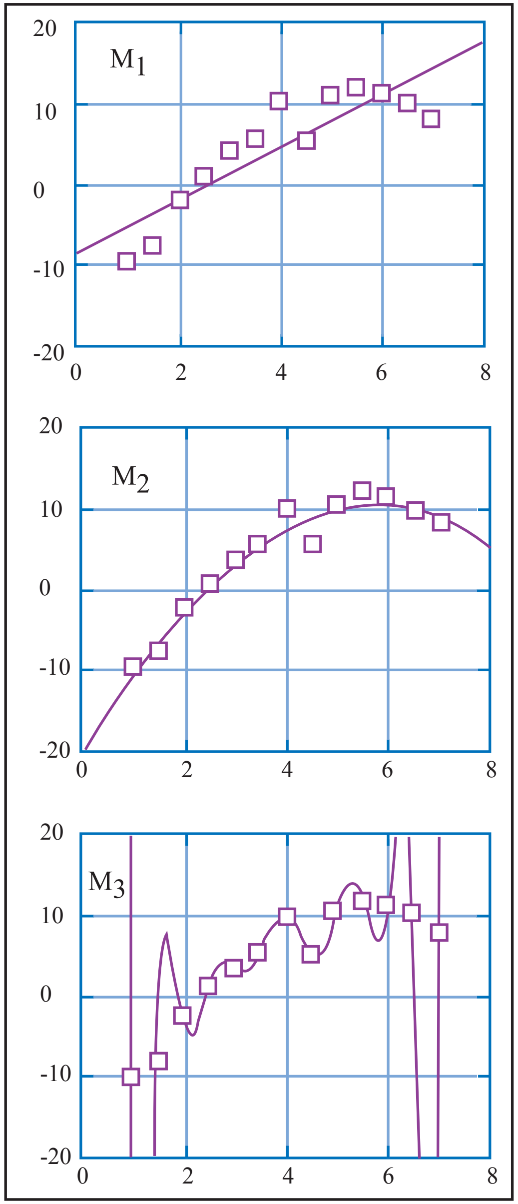


Figure by MIT OCW.

For best fitting version of each model:

Prior	Likelihood
high	low
medium	high
very very very very low	very high

Some questions

- Is the Bayesian Ockham’s razor “purely objective”?

Some questions

- Is the Bayesian Ockham's razor "purely objective"? No.
  - Priors matter. (What about uninformative priors?)
  - Choice of description language/basis functions/hypothesis classes matters.
  - Classes of hypotheses + priors = theory.
    (c.f. Martian grue, coin flipping)

- What do we gain from Bayes over conventional Ockham’s razor?

- What do we gain from Bayes over conventional Ockham's razor?
  - Isolates all the subjectivity in the choice of hypothesis space and priors
  - Gives a canonical way to measure simplicity.
  - A common currency for trading off simplicity and fit to the data: probability.
  - A rigorous basis for the intuition that "the simplest model that fits is most likely to be true".
  - Measure of complexity not just # of parameters.
    - Depends on functional form of the model

Three *one-parameter* models for 10-bit binary sequences

- Model 1:
  - Choose parameter α between 0 and 1.
  - Round(10*α) 0's followed by [10 - Round(10*α)] 1's.
- Model 2:
  - Choose parameter α between 0 and 1.
  - Draw 10 samples from Bernoulli distribution (weighted coin flips) with parameter α.
- Model 3:
  - Choose parameter α between 0 and 1.
  - Convert-to-binary(Round(2^10*α)).

- What do we gain from Bayes over conventional Ockham's razor?
  - Isolates all the subjectivity in the choice of hypothesis space and priors
  - Gives a canonical way to measure simplicity.
  - A common currency for trading off complexity and fit to the data: probability.
  - A rigorous basis for the intuition that "the simplest model that fits is most likely to be true".
  - Measure of complexity not just # of parameters.
    - Depends on functional form of the model
    - Depends on precise shape of priors (e.g., different degrees of smoothness)

Two *infinite-parameter* models for regression

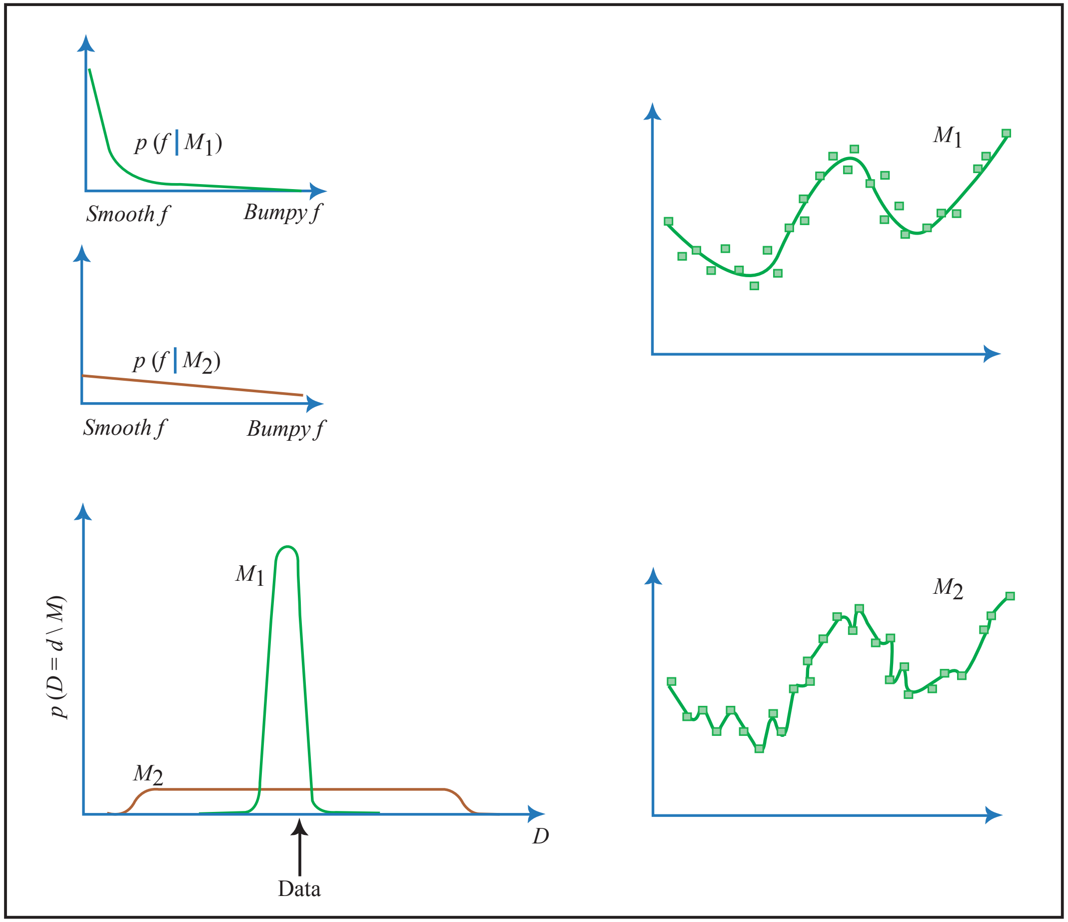
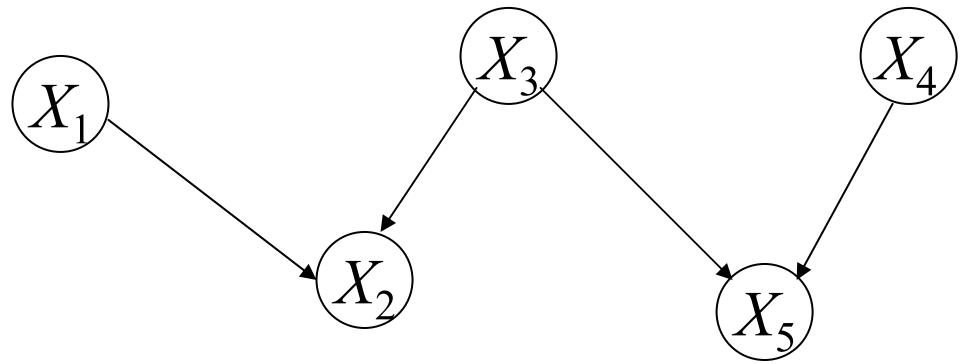


Figure by MIT OCW.

Outline

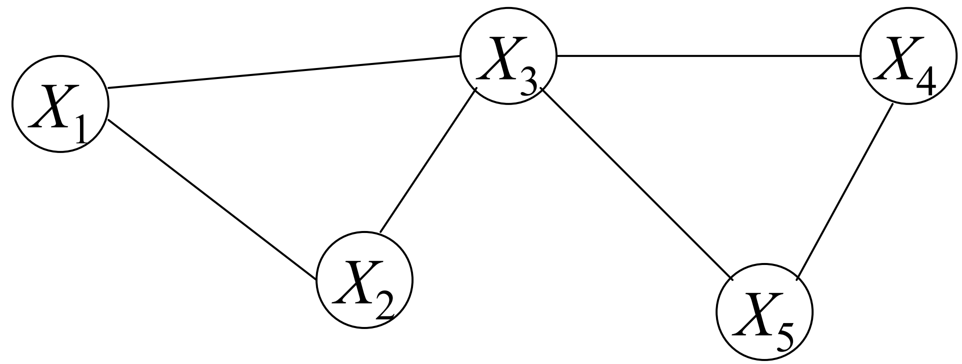
- Bayesian Ockham's Razor
- Bayes nets (directed graphical models)
  - Computational motivation: tractable reasoning
  - Cognitive motivation: causal reasoning
  - Sampling methods for approximate inference

Directed graphical models



- Consist of
  - a set of nodes
  - a set of edges
  - a *conditional probability distribution* for each node, conditioned on its parents, multiplied together to yield the distribution over variables
- Constrained to directed acyclic graphs (DAG)
- AKA: Bayesian networks, Bayes nets

Undirected graphical models



- Consist of
  – a set of nodes
  – a set of edges
  – a *potential* for each *clique*, multiplied together to yield the distribution over variables
- Examples
  – statistical physics: Ising model
  – early neural networks (e.g. Boltzmann machines)
  – low- and mid-level vision

Properties of Bayesian networks

- Efficient representation and inference
  - exploiting dependency structure makes it easier to work with distributions over many variables
- Causal reasoning
  - directed representations elucidates the role of causal structure in learning and reasoning
  - model for non-monotonic reasoning (esp. “explaining away” or causal discounting).
  - reasoning about effects of interventions (exogenous actions on a causal system)

Efficient representation and inference

- Three binary variables: *Cavity*, *Toothache*, *Catch*

Efficient representation and inference

- Three binary variables: *Cavity*, *Toothache*, *Catch*
- Specifying *P*(*Cavity*, *Toothache*, *Catch*) requires 7 parameters.
  - e.g., 1 for each set of values: $P(cav, ache, catch)$, $P(cav, ache, \neg catch)$, ..., minus 1 because it's a probability distribution
  - e.g., chain of conditional probabilities:

$$P(cav), P(ache \mid cav), P(ache \mid \neg cav), P(catch \mid ache, cav),$$
$$P(catch \mid ache, \neg cav), P(catch \mid \neg ache, cav), P(catch \mid \neg ache, \neg cav)$$

Efficient representation and inference

- Three binary variables: *Cavity*, *Toothache*, *Catch*
- Specifying *P*(*Cavity*, *Toothache*, *Catch*) requires 7 parameters.
- With n variables, we need 2^n -1 parameters
  - Here n=3. Realistically, many more: X-ray, diet, oral hygiene, personality, ...
- Problems:
  - Intractable storage, computation, and learning
  - Doesn't really correspond to the world's structure, or what we know of the world's structure.

Conditional independence

- Probabilistically: all three variables are dependent, but *Toothache* and *Catch* are independent given the presence or absence of *Cavity*.
- Causally: *Toothache* and *Catch* are both effects of *Cavity*, via independent causal mechanisms.

Conditional independence

- Probabilistically: all three variables are dependent, but *Toothache* and *Catch* are independent given the presence or absence of *Cavity*.
- Causally: *Toothache* and *Catch* are both effects of *Cavity*, via independent causal mechanisms.
- In probabilistic terms: [Without conditional independence]

$$P(ache \wedge catch \mid cav) = P(ache \mid cav) P(catch \mid ache, cav)$$

$$\begin{aligned} P(\neg ache \wedge catch \mid cav) &= P(\neg ache \mid cav) P(catch \mid \neg ache, cav) \\ &= \left[1 - P(ache \mid cav)\right] P(catch \mid \neg ache, cav) \end{aligned}$$

Conditional independence

- Probabilistically: all three variables are dependent, but *Toothache* and *Catch* are independent given the presence or absence of *Cavity*.
- Causally: *Toothache* and *Catch* are both effects of *Cavity*, via independent causal mechanisms.
- In probabilistic terms: [With conditional independence]

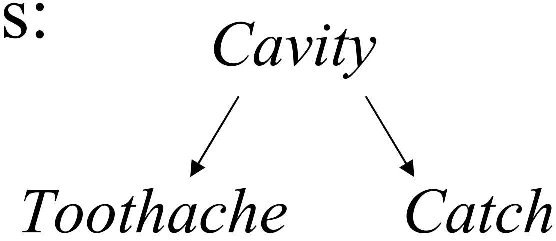
$$P(ache \wedge catch \mid cav) = P(ache \mid cav)P(catch \mid cav)$$

$$\begin{aligned} P(\neg ache \wedge catch \mid cav) &= P(\neg ache \mid cav)P(catch \mid cav) \\ &= \left[1 - P(ache \mid cav)\right]P(catch \mid cav) \end{aligned}$$

- With n pieces of evidence, $x_1, \ldots, x_n$, we need $2n$ conditional probabilities: $P(x_i \mid cav), P(x_i \mid \neg cav)$

A simple Bayes net

- Graphical representation of relations between a set of random variables:

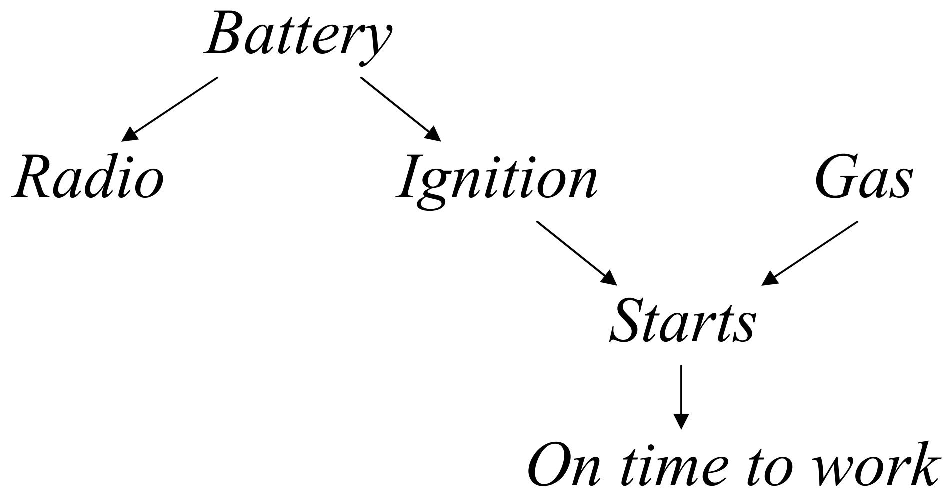


- Causal interpretation: independent local mechanisms
- Probabilistic interpretation: factorizing complex terms

$$P(A,B,C) = \prod_{V \in \{A,B,C\}} P(V \mid \text{parents}[V])$$

$$\begin{aligned} P(Ache, Catch, Cav) &= P(Ache, Catch \mid Cav) P(Cav) \\ &= P(Ache \mid Cav) P(Catch \mid Cav) P(Cav) \end{aligned}$$

A more complex system



- Joint distribution sufficient for any inference:

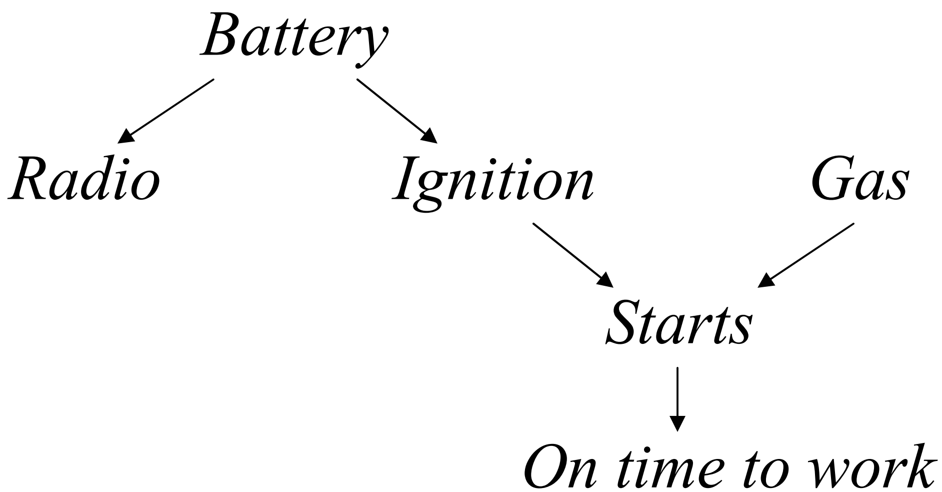
$$P(B,R,I,G,S,O) = P(B)P(R \mid B)P(I \mid B)P(G)P(S \mid I,G)P(O \mid S)$$

$$P(O \mid G) = \frac{P(O,G)}{P(G)} = \frac{\sum_{B,R,I,S} P(B,R,I,G,S,O)}{P(G)}$$

$$P(A) = \sum_{B} P(A,B)$$

"marginalization"

A more complex system

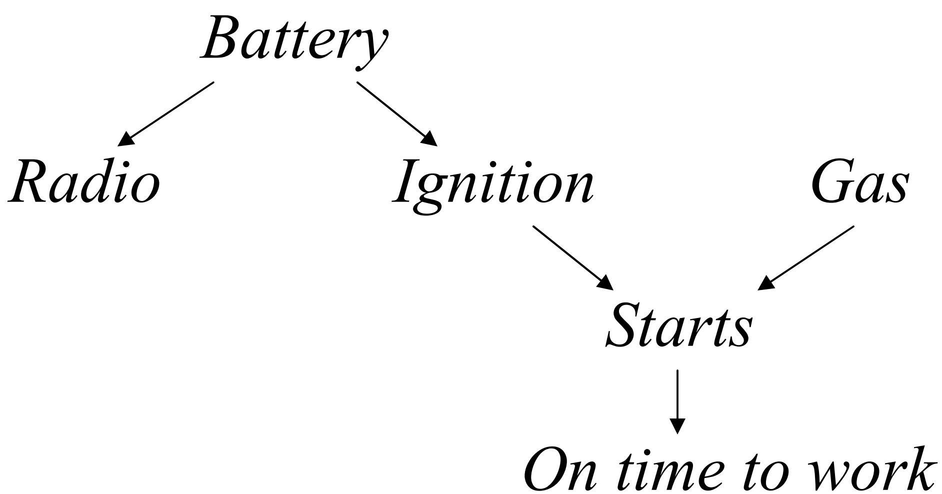


- Joint distribution sufficient for any inference:

$$P(B,R,I,G,S,O) = P(B)P(R \mid B)P(I \mid B)P(G)P(S \mid I,G)P(O \mid S)$$

$$P(O \mid G) = \frac{P(O,G)}{P(G)} = \frac{\sum_{B,R,I,S} P(B)P(R \mid B)P(I \mid B)P(G)P(S \mid I,G)P(O \mid S)}{P(G)}$$

A more complex system

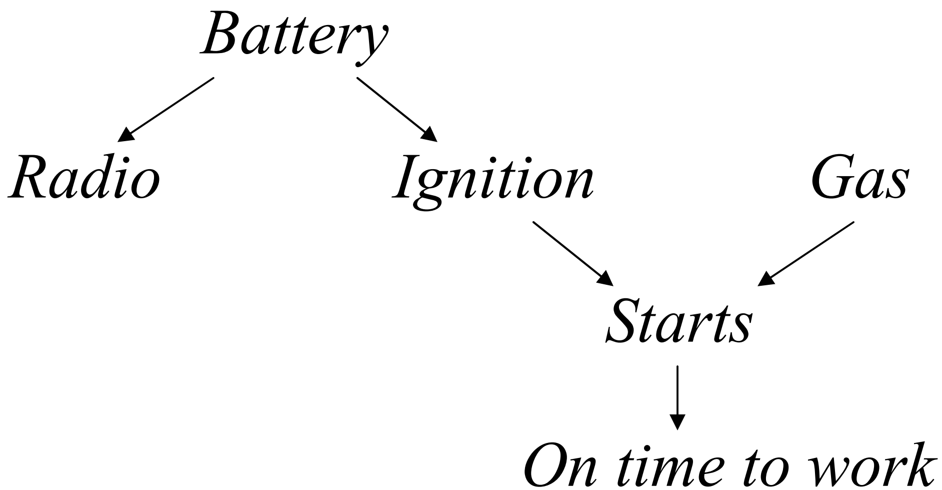


- Joint distribution sufficient for any inference:

$$P(B,R,I,G,S,O) = P(B)P(R \mid B)P(I \mid B)P(G)P(S \mid I,G)P(O \mid S)$$

$$P(O \mid G) = \frac{P(O,G)}{P(G)} = \sum_{S} \left(\sum_{B,I} P(B)P(I \mid B)P(S \mid I,G) \right) P(O \mid S)$$

A more complex system

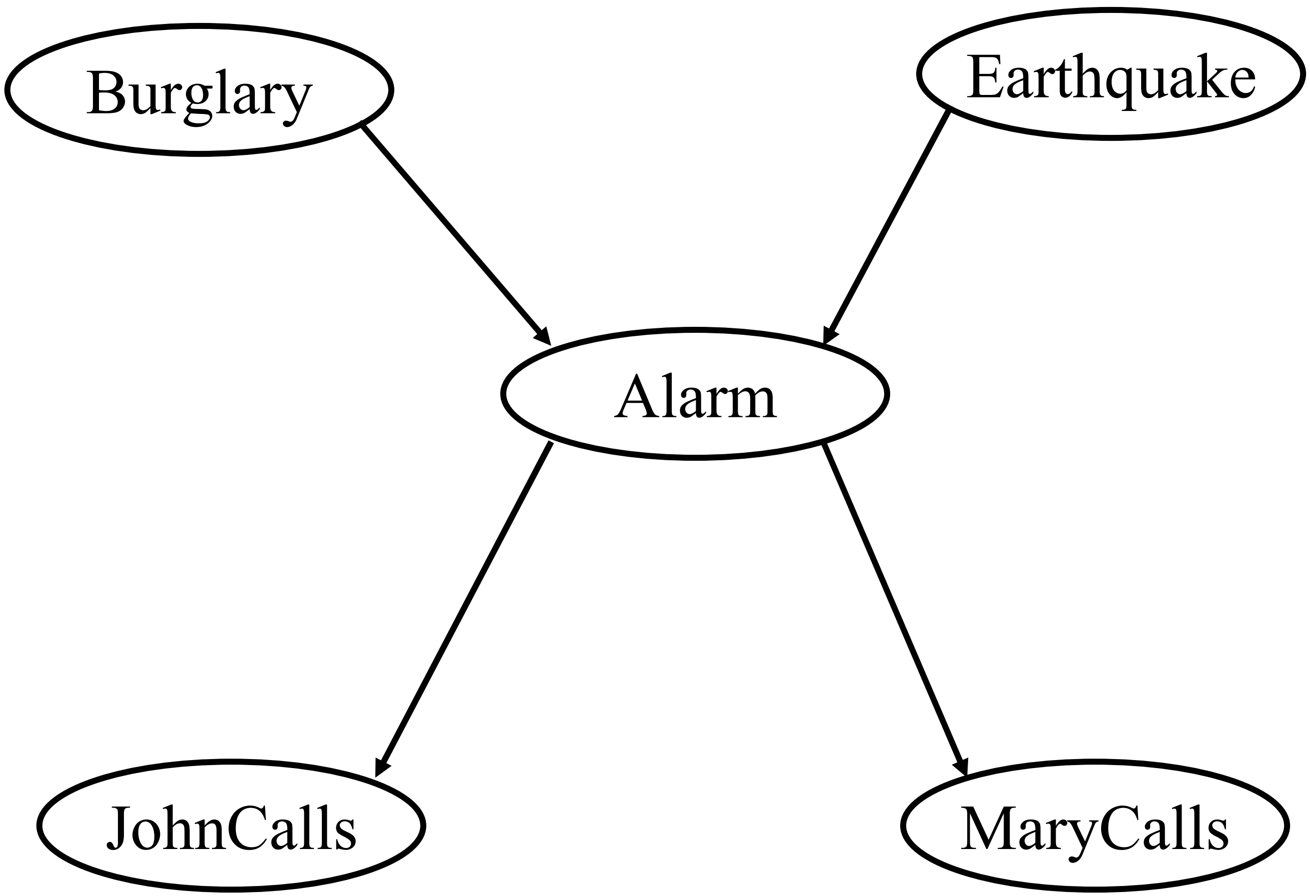


- Joint distribution sufficient for any inference:

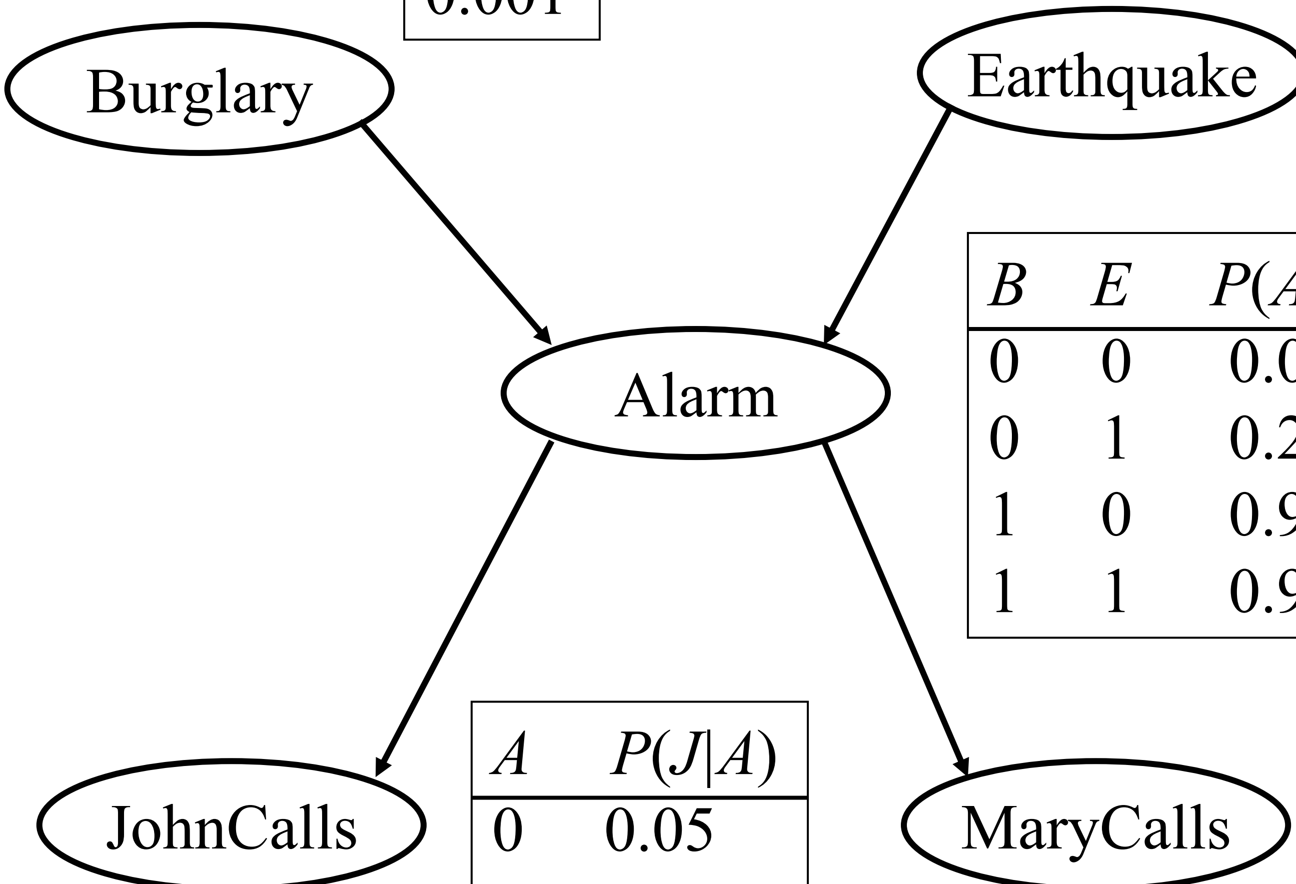
$$P(B,R,I,G,S,O) = P(B)P(R \mid B)P(I \mid B)P(G)P(S \mid I,G)P(O \mid S)$$

- General inference algorithms via local computations
  - for graphs without loops: belief propagation
  - in general: variable elimination, junction tree

More concrete representation



More concrete representation



$P(B)$
0.001

$P(E)$
0.002

B	E	$P(A\|B,E)$
0	0	0.001
0	1	0.29
1	0	0.94
1	1	0.95

"CPT"

A	$P(J\|A)$
0	0.05
1	0.90

A	$P(M\|A)$
0	0.01
1	0.70

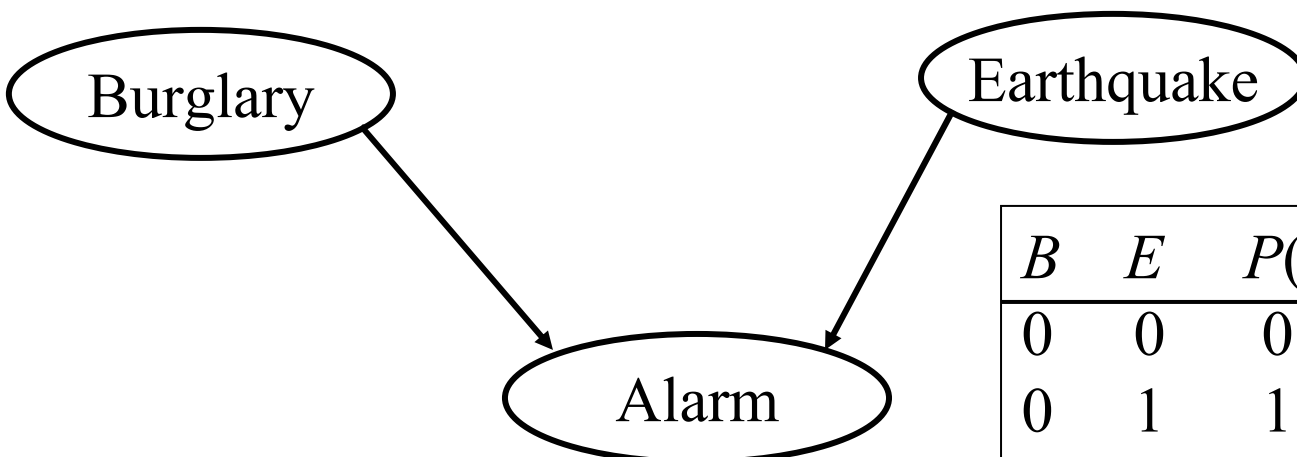
Parameterizing the CPT

Size of CPT is exponential in number of parents. Often use a simpler parameterization based on knowledge of how causes interact.

Parameterizing the CPT

Size of CPT is exponential in number of parents. Often use a simpler parameterization based on knowledge of how causes interact.

- Logical OR: Independent deterministic causes

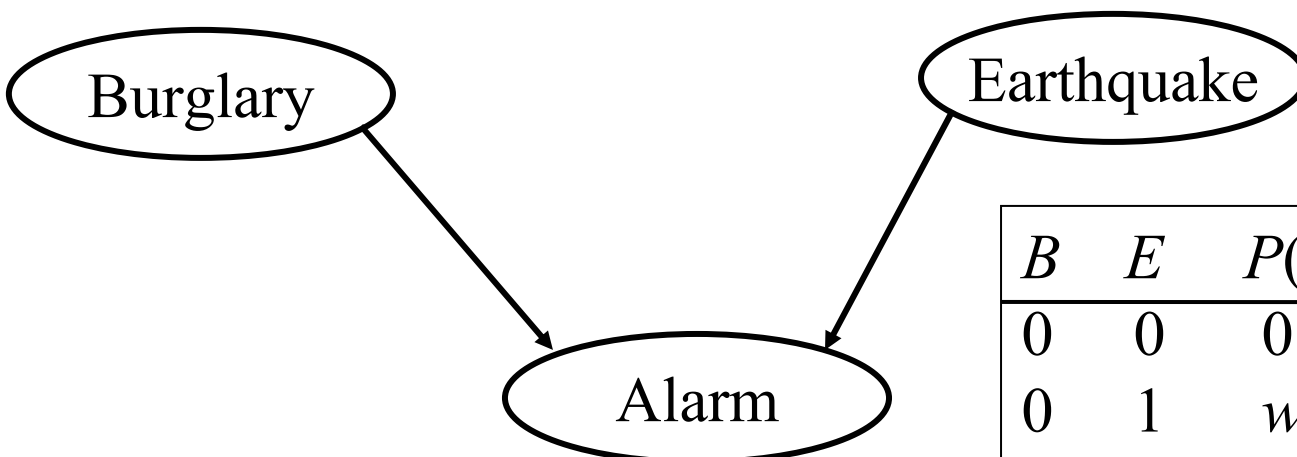


B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

Parameterizing the CPT

Size of CPT is exponential in number of parents. Often use a simpler parameterization based on knowledge of how causes interact.

- Noisy OR: Independent probabilistic causes

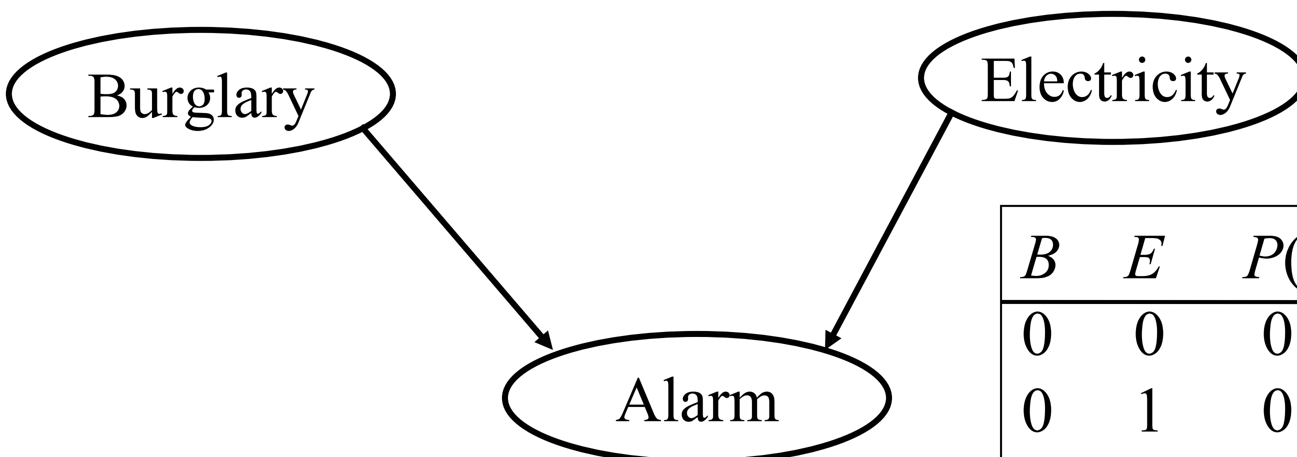


B	E	$P(A \mid B,E)$
0	0	0
0	1	w_B
1	0	w_E
1	1	$w_B + (1 - w_B) w_E$

Parameterizing the CPT

Size of CPT is exponential in number of parents. Often use a simpler parameterization based on knowledge of how causes interact.

- AND: cause + enabling condition

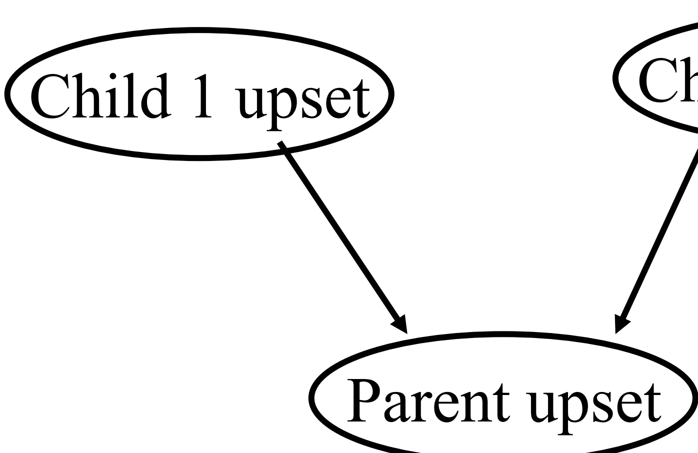


B	E	$P(A\|B,E)$
0	0	0
0	1	0
1	0	0
1	1	1 (or w_B)

Parameterizing the CPT

Size of CPT is exponential in number of parents. Often use a simpler parameterization based on knowledge of how causes interact.

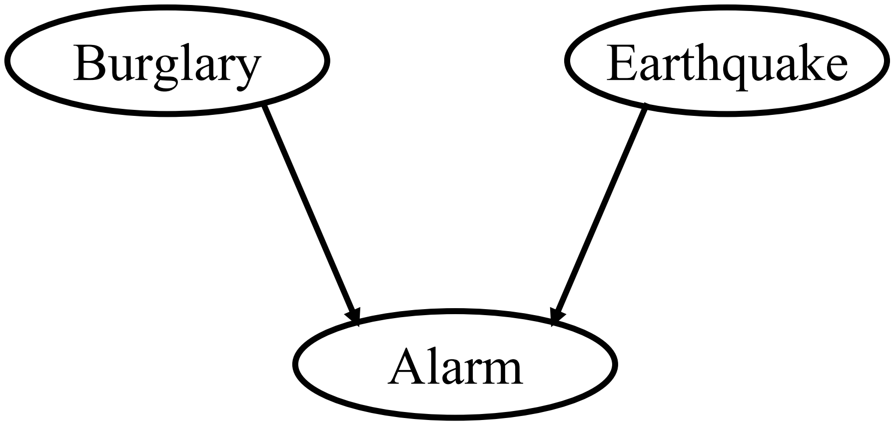
- Logistic: Independent probabilistic causes with varying strengths w_i and a threshold θ



$C1$	$C2$	$P(Pa \mid C1, C2)$
0	0	$1/[1+\exp(\theta)]$
0	1	$1/[1+\exp(\theta - w_1)]$
1	0	$1/[1+\exp(\theta - w_2)]$
1	1	$1/[1+\exp(\theta - w_1 - w_2)]$

Explaining away

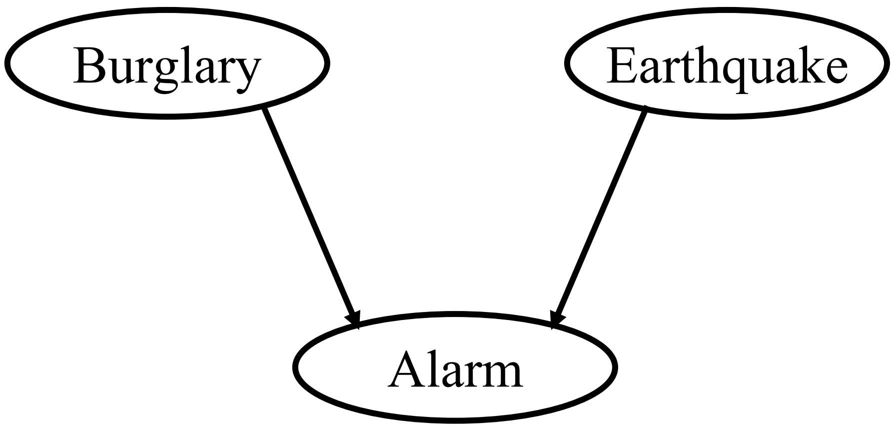
- Logical OR: Independent deterministic causes



B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

Explaining away

- Logical OR: Independent deterministic causes



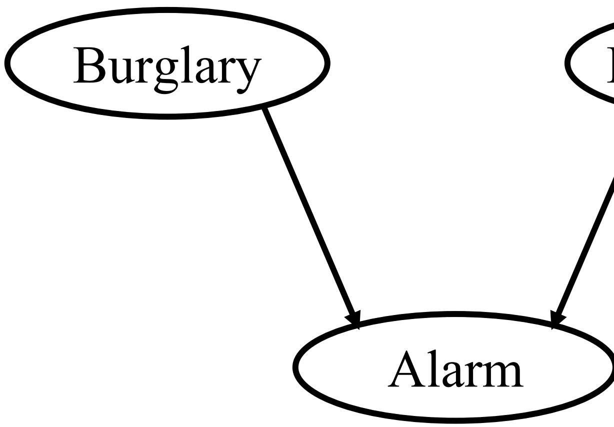
B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

A priori, no correlation between *B* and *E*:

$$P(B,E) = \sum_A P(B,E,A)$$

Explaining away

- Logical OR: Independent deterministic causes



B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

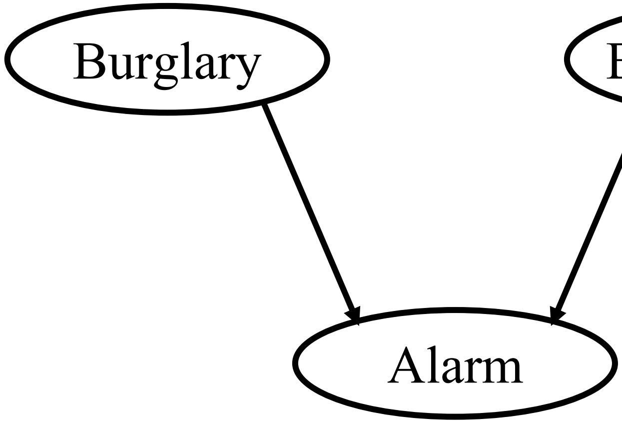
A priori, no correlation between *B* and *E*:

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

$$P(A,B,C) = \prod_{V \in \{A,B,C\}} P(V \mid \text{parents}[V])$$

Explaining away

- Logical OR: Independent deterministic causes



B	E	$P(A \mid B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

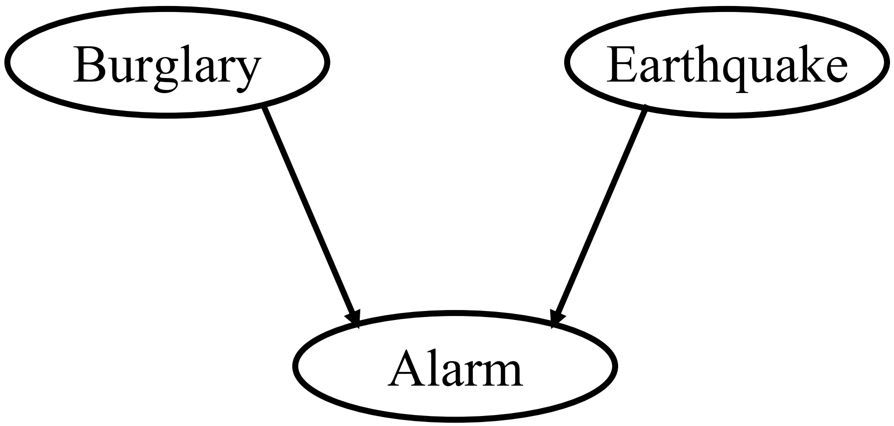
A priori, no correlation between *B* and *E*:

$$P(B,E) = \boxed{\sum_A P(A \mid B,E)} P(B)\, P(E)$$

=1, for any values of *B* and *E*

Explaining away

- Logical OR: Independent deterministic causes



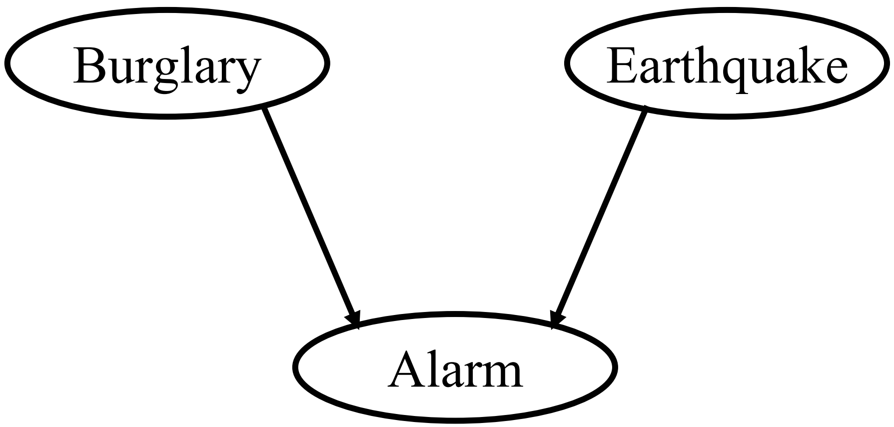
B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

A priori, no correlation between *B* and *E*:

$$P(B,E) = P(B)\,P(E)$$

Explaining away

- Logical OR: Independent deterministic causes



B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

After observing A=1 …

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

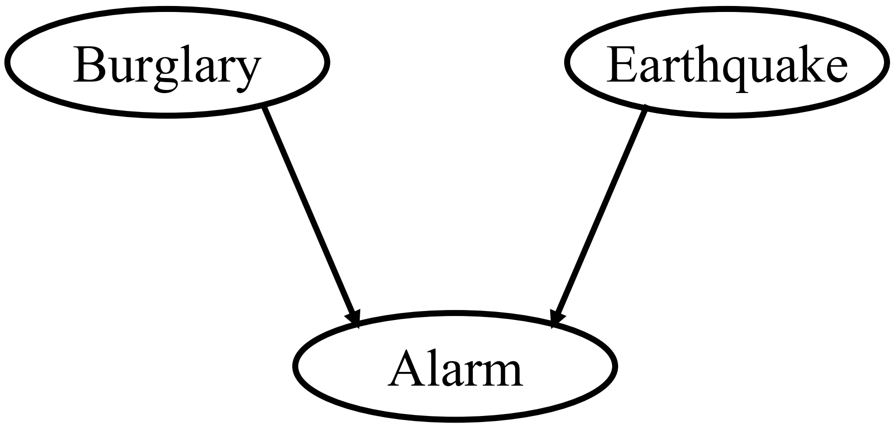
$$\propto P(A=1 \mid B,E)P(B)P(E)$$

Assume

$$P(B) = P(E) = 1/2$$

Explaining away

- Logical OR: Independent deterministic causes



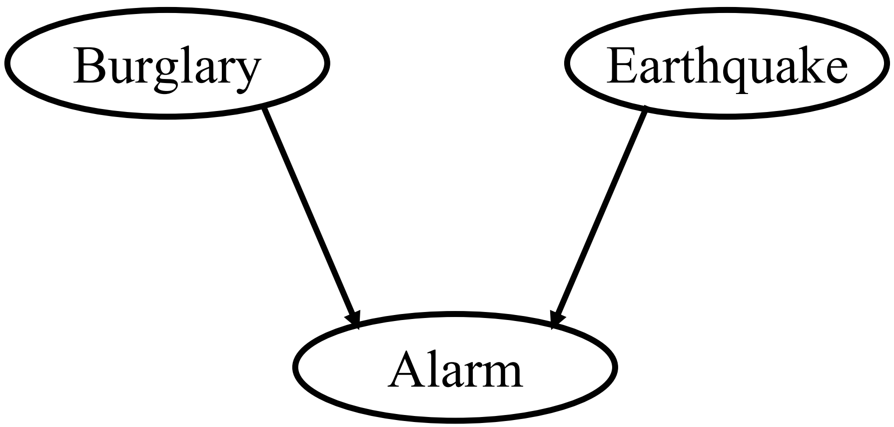
B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

After observing A=1 …

$$P(B, E \mid A = 1) \propto P(A = 1 \mid B, E)$$

Explaining away

- Logical OR: Independent deterministic causes



B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

After observing A=1 …

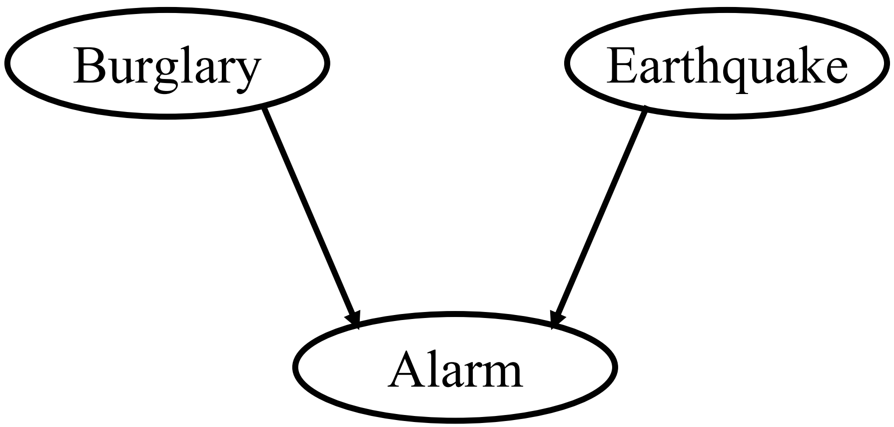
$$P(B, E \mid A = 1) \propto P(A = 1 \mid B, E)$$

… $P(B|A{=}1) = 2/3$

B and E are anti-correlated

Explaining away

- Logical OR: Independent deterministic causes



B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

After observing A=1, E=1 …

$$P(B \mid A=1, E=1) \propto P(A=1 \mid B, E=1)$$

… $P(B|A=1) = 1/2$

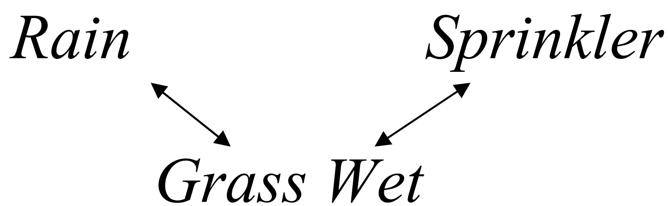
Back to $P(B)$.

"Explaining away" or
"Causal discounting"

Explaining away

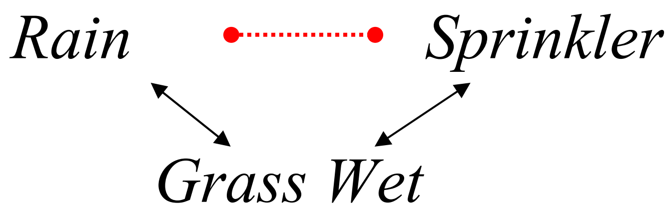
- Depends on the functional form (the parameterization) of the CPT
  - OR or Noisy-OR: Discounting
  - AND: No Discounting
  - Logistic: Discounting or Augmenting

Spreading activation or recurrent neural networks



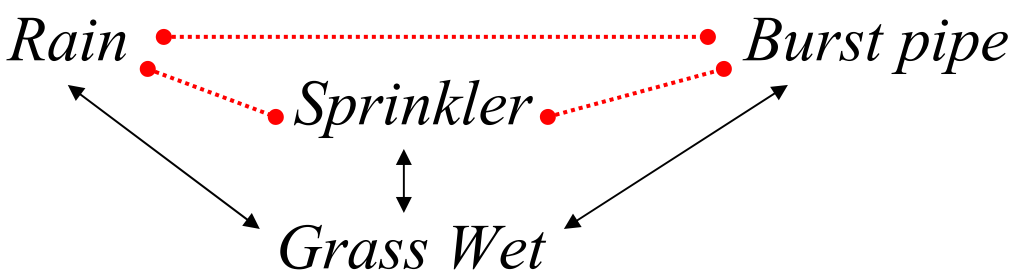
- Excitatory links: *Rain* $\longleftrightarrow$ *Wet*, *Sprinkler* $\longleftrightarrow$ *Wet*
- Observing rain, *Wet* becomes more active.
- Observing grass wet, *Rain* and *Sprinkler* become more active.
- Observing grass wet and sprinkler, *Rain* cannot become less active. No explaining away!

Spreading activation or recurrent neural networks



- Excitatory links: *Rain* $\leftrightarrow$ *Wet*, *Sprinkler* $\leftrightarrow$ *Wet*
- Inhibitory link: *Rain* •·······• *Sprinkler*
- Observing grass wet, *Rain* and *Sprinkler* become more active.
- Observing grass wet and sprinkler, *Rain* becomes less active: explaining away.

Spreading activation or recurrent neural networks



- Each new variable requires more inhibitory connections.
- Interactions between variables are not causal.
- Not modular.
  - Whether a connection exists depends on what other connections exist, in non-transparent ways.
  - Combinatorial explosion.

Summary

Bayes nets, or directed graphical models, offer a powerful representation for large probability distributions:

– Ensure tractable storage, inference, and learning
– Capture causal structure in the world and canonical patterns of causal reasoning.
– This combination is not a coincidence.