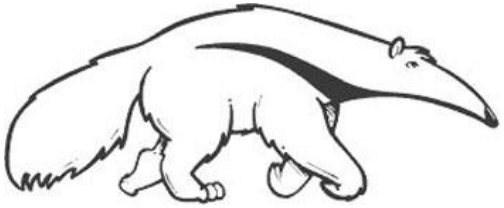
COMPSCI 276: Causal and Probabilistic Reasoning

Rina Dechter, UCI

Lecture 1: Introduction



Darwiche chapters 1,3
Dechter-Morgan&claypool book: Chapters 1-2
Causal Inference in Statistics, A primer, J. Pearl, M Glymur and N. Jewell Ch1,
Why, ch1

Class Information

Course Topics

Probabilistic Graphical Models, Structural causal models,The Causal Hierarchy.

1.Representing independencies by graphs. d-seperation.
2.Algorithms (Bucket-elimination, Join-trees, The induced-width.).
3.Sampling schemes for graphical models (MCMC, IS)
4.Structural Causal Models; Identification of Causal Effect;
5.The Back-Door and Front-Door Criteria and the Do-Calculus.
6.Linear Causal Models.
7.Counterfactuals.
8.Algorithms for identification. The ID algorithm.
9.Learning Bayesian networks and Causal graphs (causal discovery).

Class page

Grading

- **Four or five homeworks**
- Project: Class presentation and a report: Students will present a paper and write a report

Textbooks

[P] Judea Pearl, Madelyn Glymour, Nicholas P. Jewell,
Causal Inference in Statistics: A Primer,
Cambridge Press, 2016.

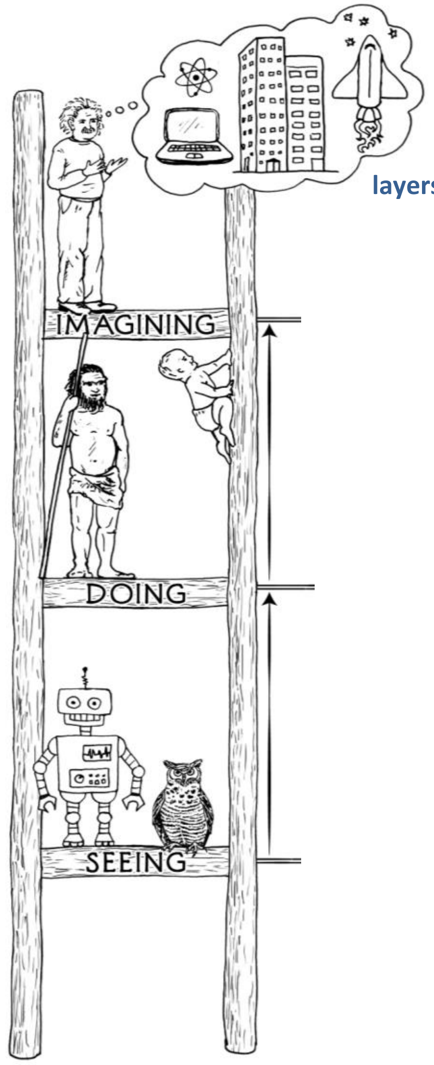
[C] Judea Pearl,
Causality: Models, Reasoning, and Inference
Cambridge Press, 2009.

[W] Judea Pearl, Dana Mackenzie,
The Book of Why,
Basic books, 2018.

•[Darwiche] Adnan Darwiche, "Modeling and Reasonin with Bayesian Networks“

•[Dechter] Rina Dechter, "Reasoning with Probabilistic and Deterministic Graphical Models: Exact Algorithms"

Why Causality?



layers/rungs of the causal hierarchy

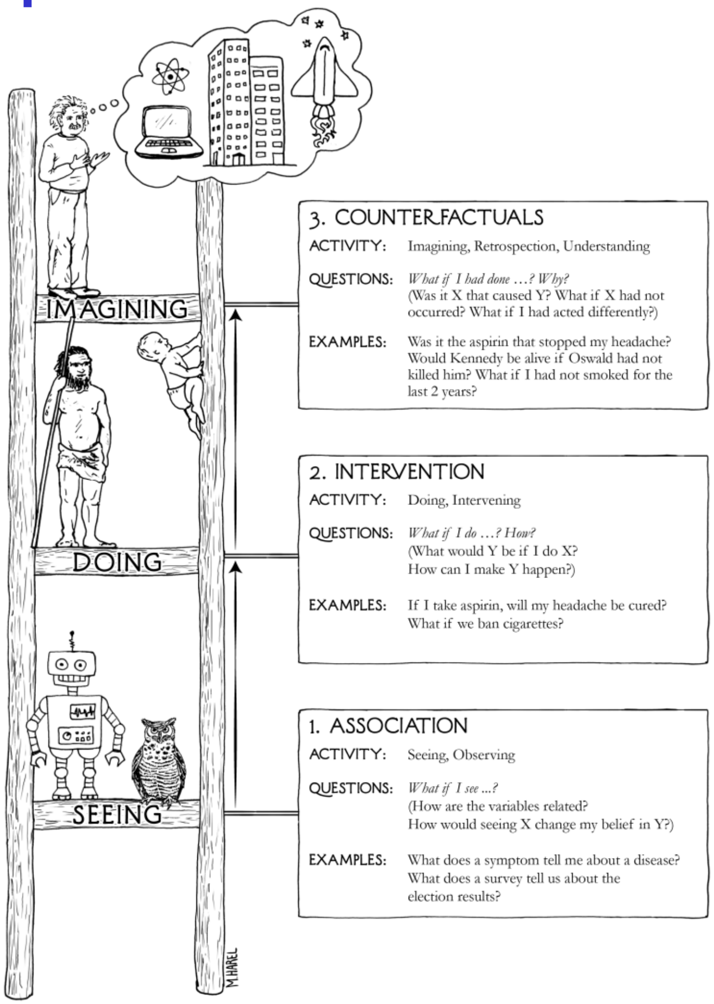
This course last part

This course main part

This course, first part

There are three distinct levels of cognitive ability: seeing, doing and imagining

Ladder of Causation`



seeing, doing, and imagining.

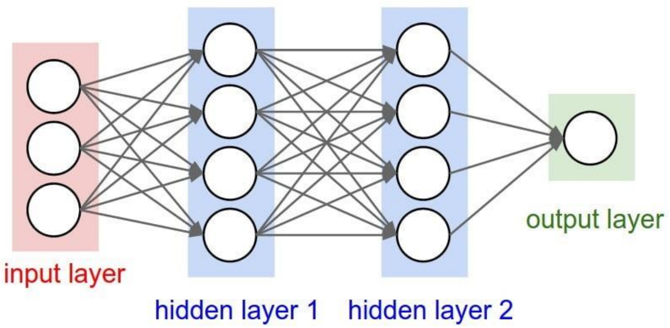
- Most animals, learning machines are on the first rung, learning from association.
- Tool users, such as early humans, are on the second rung, if they act by planning and not merely by imitation. We can also use experiments to learn the effects of interventions, and presumably this is how babies acquire much of their causal knowledge.
- On the top rung, counterfactual learners can imagine worlds that do not exist and infer reasons for observed phenomena.

The Primary AI Challenges

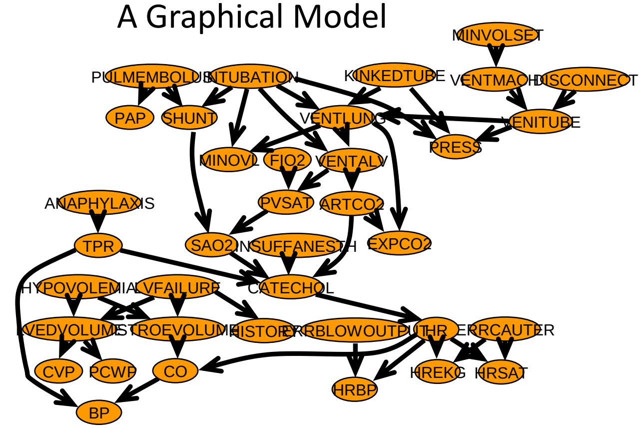
- **Machine Learning** focuses on replicating humans learning

- **Automated reasoning** focuses on replicating how people reason.

A neural network



A Graphical Model



Automated Reasoning

Medical Doctor

Lawyer

Policy Maker

Queries:

- Prediction: what will happen?
- Diagnosis: what had happened?
- Situation assessment: What is going on?
- Planning, decision making: what to do?
- Explanation: need causal models
- Counterfactuals: What if? need Structural causal models

Automated Reasoning

Causal Models **Data**

answers

Queries:

- Prediction
- Diagnosis
- Situation assessment
- Planning, decision making
- Explanation, causal effect
- Counterfactuals

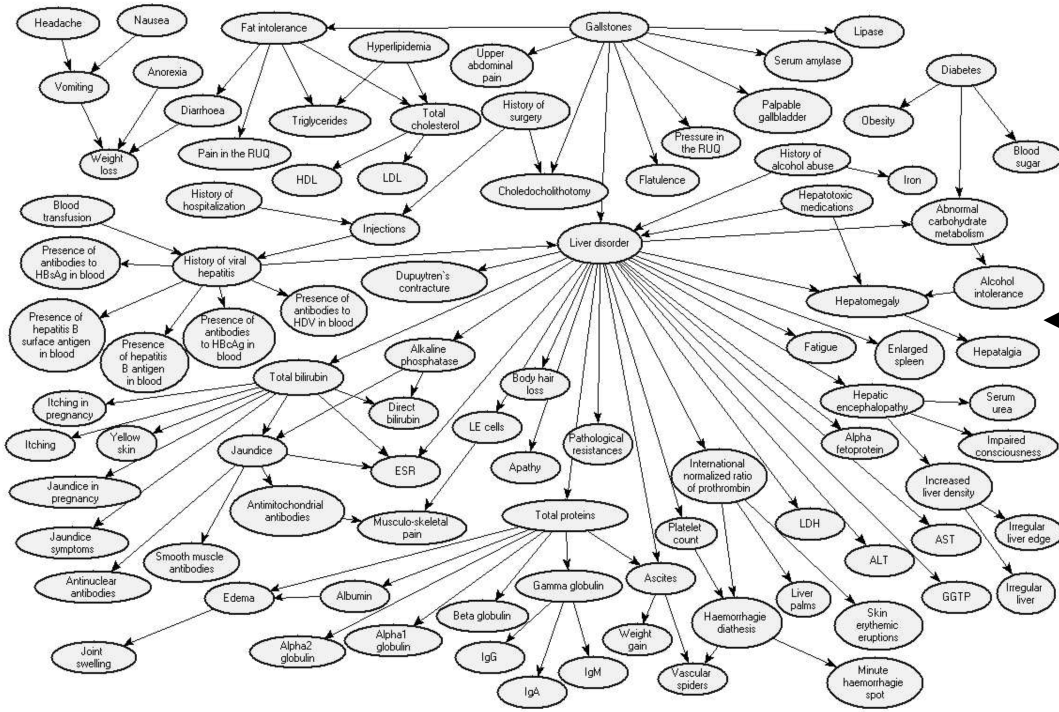
Knowledge is huge, so How to identify what's relevant? → **Causal Graphical Models**

The field of Automated Reasoning developing general purpose formalisms (languages, models) that enable us to represent knowledge in such a way that we can exploit the relevance and causal relationship quickly.

Answer query in the 3 levels of the causal heirarchy

Graphical Models

Example: diagnosing liver disease (Onisko et al., 1999)



Queries:

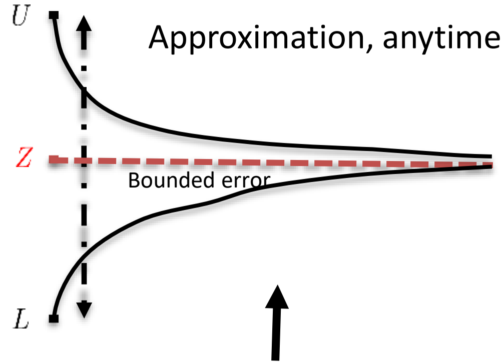
- Prediction
- Diagnosis, explanation
- Situation assessment
- Planning, decision making
- Counterfactual reasoning

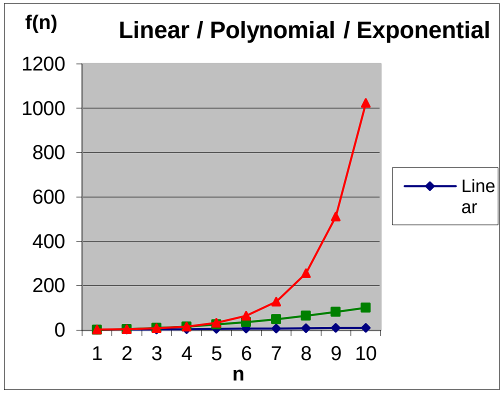
Automated Reasoning:

- Develop methods to answer these questions.
- Learning the models: from experts and data.

Complexity of Automated Reasoning

- Prediction
- Diagnosis
- Planning and scheduling
- Probabilistic Inference
- Explanation
- Decision-making
- Causal reasoning

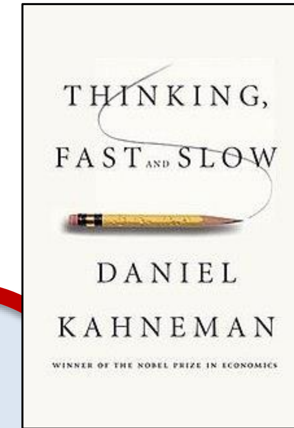


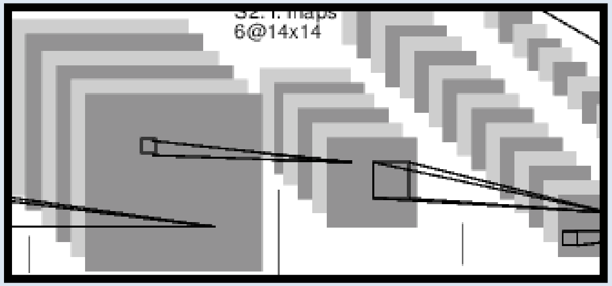


Reasoning is computationally hard

Complexity is exponential

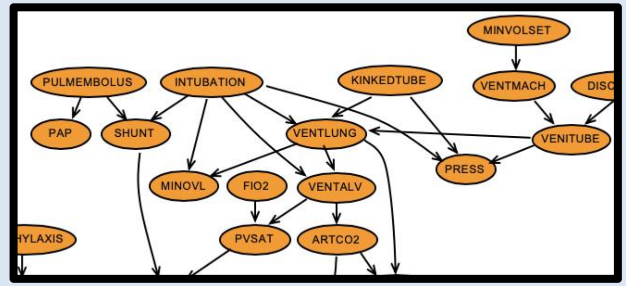
AI Renaissance





- Deep learning
  - Fast predictions
  - "Instinctive"

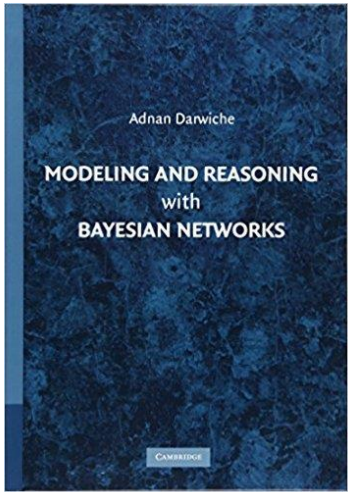
Tools:
 Tensorflow, PyTorch, …

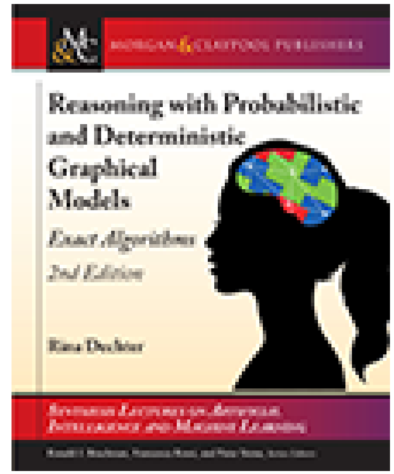


- Probabilistic models
  - Slow reasoning
  - "Logical / deliberative"

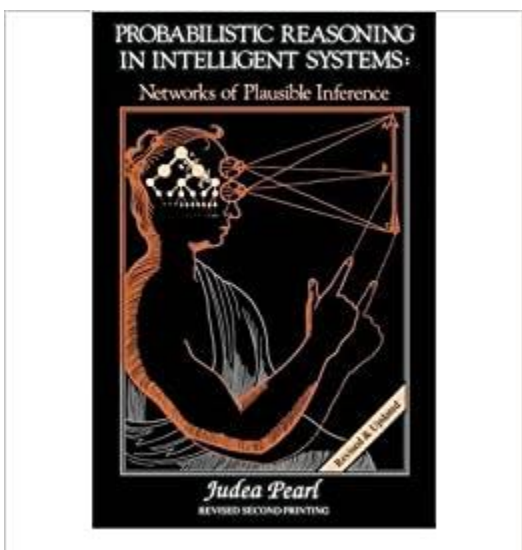
Tools:
 Graphical Models,
Probabilistic programming,
 Markov Logic, …

Text Books

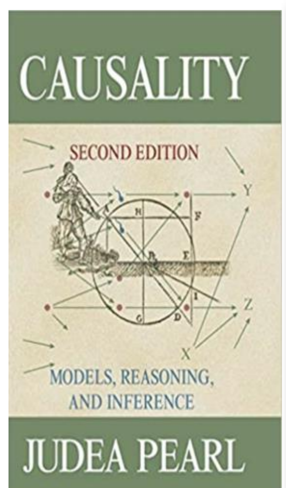




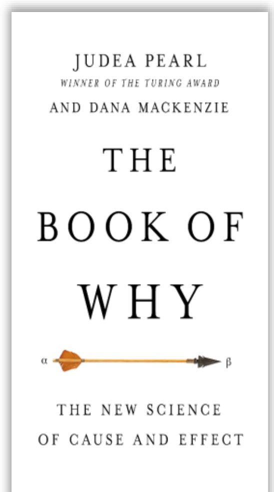
[Class page]()

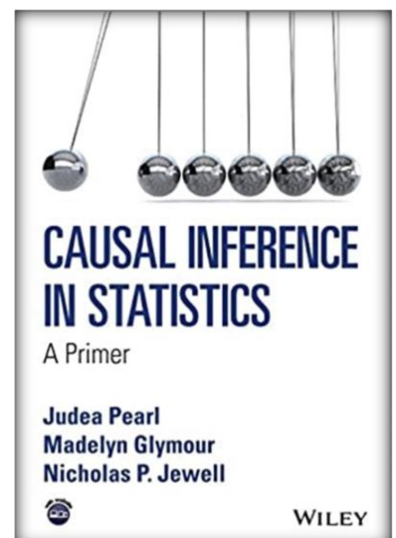


2009



2018





Probabilistic Graphical models

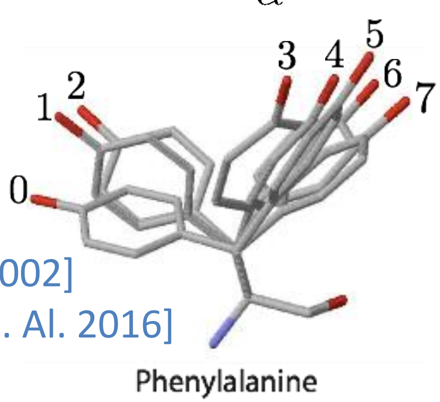
- Describe structure in large problems
  - Large complex system $F(X)$
  - Made of "smaller", "local" interactions $f_\alpha(x_\alpha)$
  - Complexity emerges through interdependence

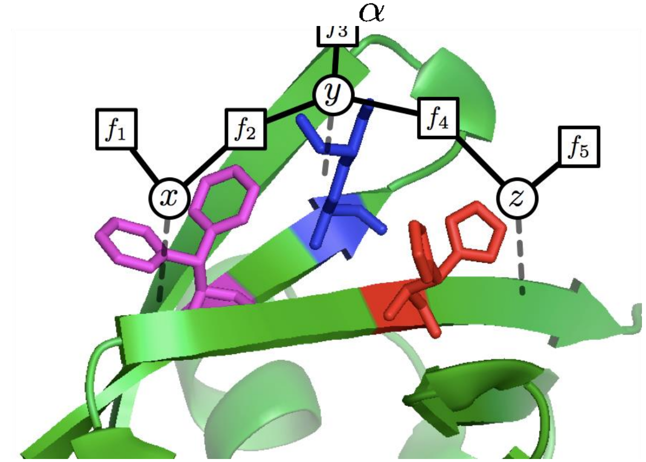
Probabilistic Graphical models

- Describe structure in large problems
  - Large complex system $F(X)$

- Protein Structure **prediction**: predicting the 3d structure from given sequences
- CPD: Computational Protein **design** (backbone) algorithms enumerate a combinatorial number of candidate structures to compute the Global Minimum Energy Conformation (GMEC).

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha) \qquad f(\mathbf{x}^*) = \max_{\mathbf{x}} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha)$$





[Yanover & Weiss 2002]
[Bruce R. Donald et. Al. 2016]

We can model the compatibility of two parts of the protein with a reward for positions that are compatible, and penalty for incompatible ones.

Probabilistic Graphical models

- Describe structure in large problems
  - Large complex system $F(X)$
  - Made of “smaller”, “local” interactions $f_\alpha(x_\alpha)$
  - Complexity emerges through interdependence
- Examples & Tasks
  - Summation & marginalization

“partition function”

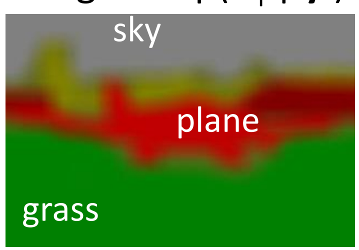
$$p(x_i) = \frac{1}{Z} \sum_{\mathbf{x} \setminus x_i} \prod_\alpha f_\alpha(\mathbf{x}_\alpha) \quad \text{and} \quad Z = \sum_{\mathbf{x}} \prod_\alpha f_\alpha(\mathbf{x}_\alpha)$$

Image segmentation and classification:

Observation **y**

$\Rightarrow$

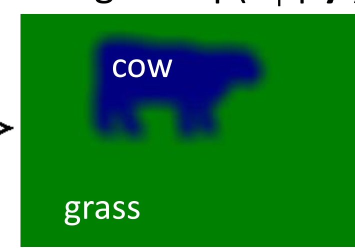
Marginals p(x_i | **y**)



Observation **y**

$\Rightarrow$

Marginals p(x_i | **y**)



e.g., [Plath et al. 2009]

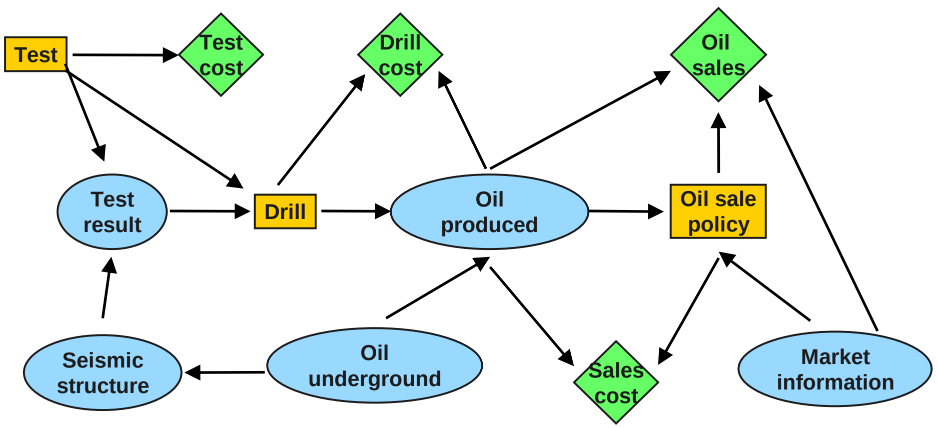
Probabilistic (Causal) Graphical models

- Describe structure in large problems
  - Large complex system $F(X)$
  - Made of "smaller", "local" interactions $f_\alpha(x_\alpha)$
  - Complexity emerges through interdependence
- Examples & Tasks
  - Mixed inference (marginal MAP, MEU, …)

$$f(\mathbf{x}_M^*) = \max_{\mathbf{x}_M} \sum_{\mathbf{x}_S} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha)$$

Influence diagrams & optimal decision-making

(the "oil wildcatter" problem)



e.g., [Raiffa 1968; Shachter 1986]

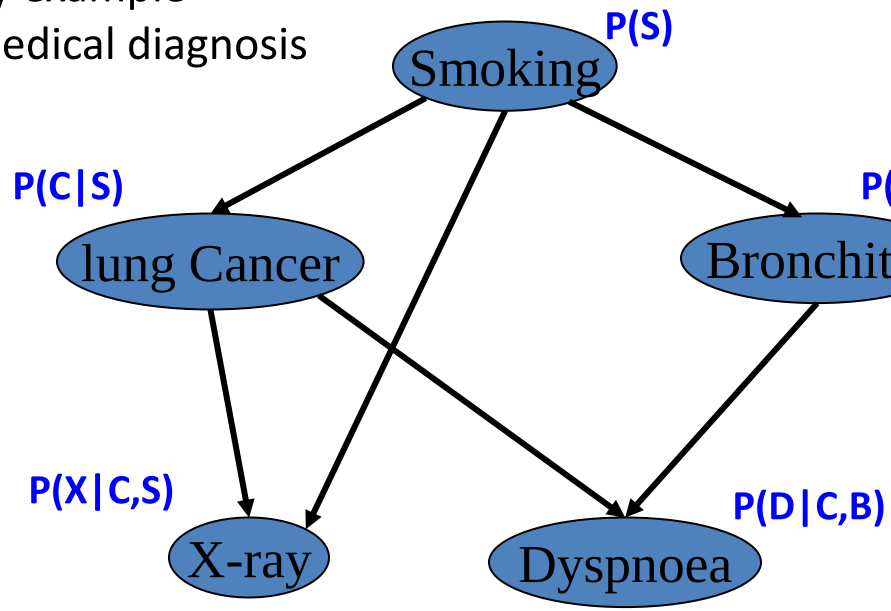
Example domains for graphical models

- Natural Language processing
  - Information extraction, semantic parsing, translation, topic models, …
- Computer vision
  - Object recognition, scene analysis, segmentation, tracking, …
- Computational biology
  - Pedigree analysis, protein folding and binding, sequence matching, …
- Networks
  - Webpage link analysis, social networks, communications, citations, ….
- Robotics
  - Planning & decision making
- Social sciences, man-machine interaction requires causality

In more details…

Bayesian Networks (Pearl 1988)

An early example
From medical diagnosis



$\mathrm{BN} = (G, \Theta)$

CPD:

C	B	P(D|C,B)	
0	0	0.1	0.9
0	1	0.7	0.3
1	0	0.8	0.2
1	1	0.9	0.1

P(S, C, B, X, D) = ***P(S) P(C|S) P(B|S) P(X|C,S) P(D|C,B)***

Combination: Product
Marginalization: sum/max

- Posterior marginals, probability of evidence, MPE

Is this a causal model?

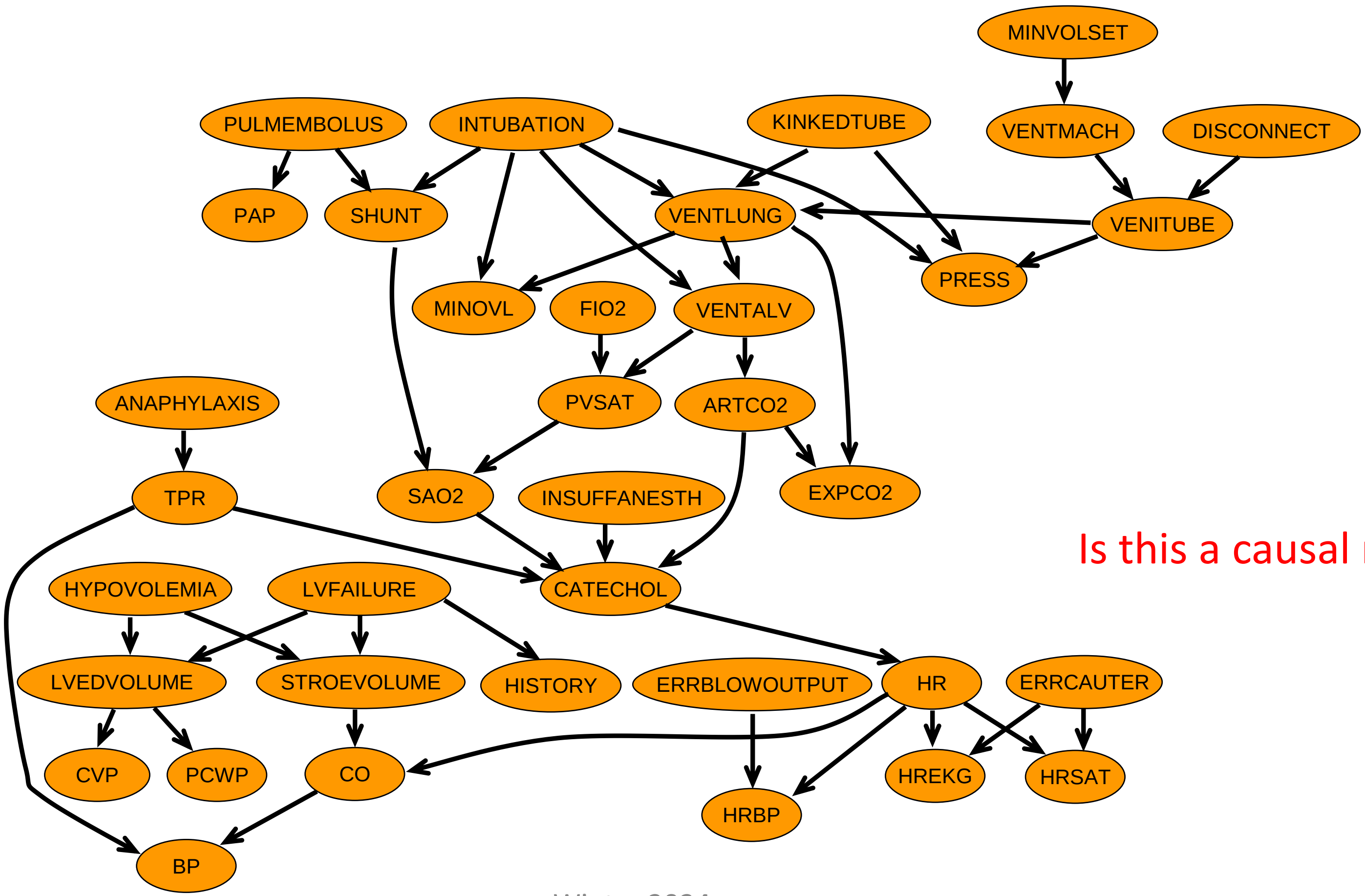
- P(D= 0) = $\sum_{S,L,B,X}$ P(S)· P(C|S)· P(B|S)· P(X|C,S)· P(D|C,B

MAP(P)= $max_{S,L,B,X}$ P(S)· P(C|S)· P(B|S)· P(X|C,S)· P(D|C,B)

Alarm network [Beinlich et al., 1989]

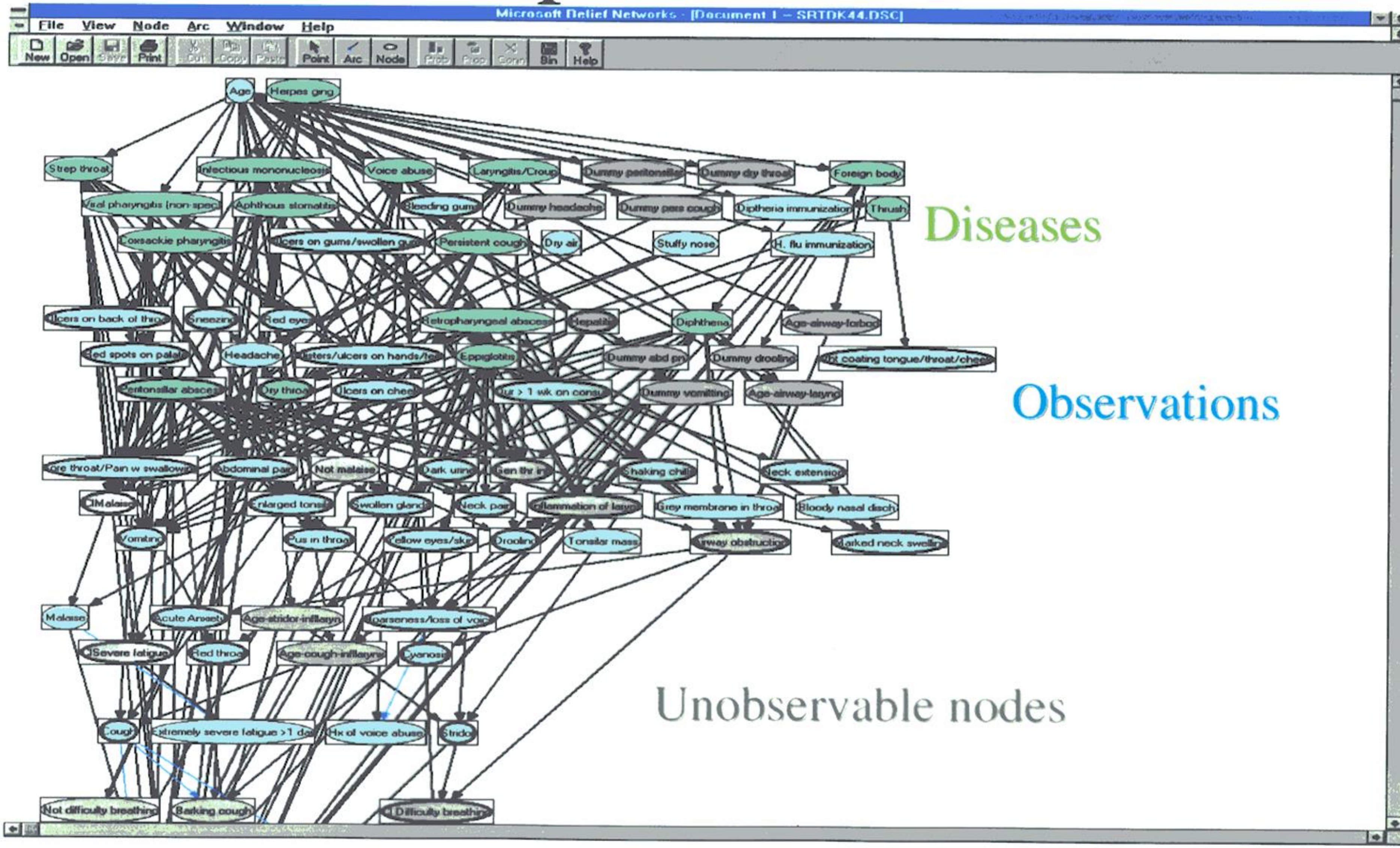
- Bayes nets: compact representation of large joint distributions

The "alarm" network: 37 variables, 509 parameters (rather than $2^{37} = 10^{11}$!)



Is this a causal model?

Chief Complaint: Sore Throat



Constraint Networks

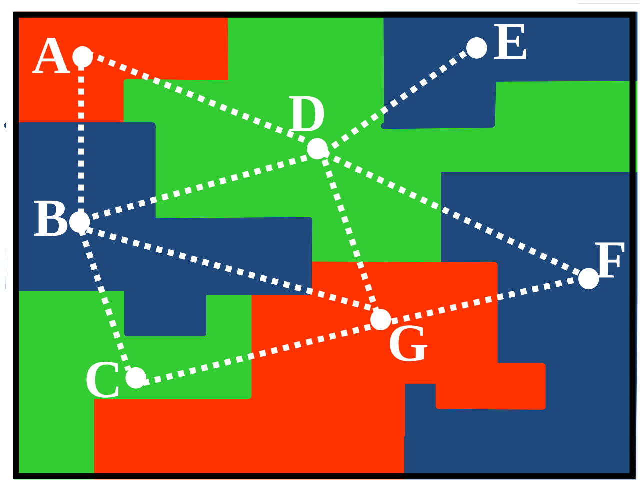
Example: map coloring

Variables - countries (A,B,C,etc.)

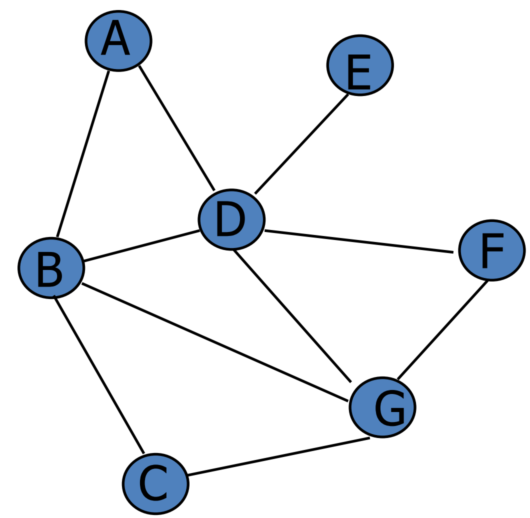
Values - colors (red, green, blue)

Constraints: A ≠ B, A ≠ D, D ≠ E, etc.

A	B
red	green
red	yellow
green	red
green	yellow
yellow	green
yellow	red



Constraint graph

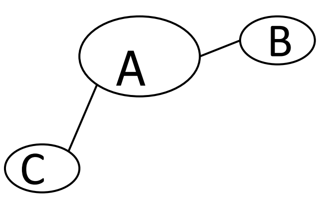


Propositional Reasoning

Example: party problem

- If Alex goes, then Becky goes: $A \rightarrow B$
  (Alex goes = A, Becky goes = B)
- If Chris goes, then Alex goes: $C \rightarrow A$
  (Chris goes = C, Alex goes = A)

- **Question:**
  Is it possible that Chris goes to the party but Becky does not?

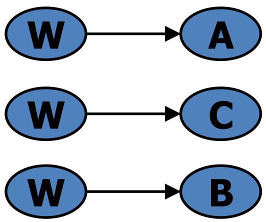


Is the *propositional theory* $\phi = \{A \rightarrow B, C \rightarrow A, \neg B, C\}$ satisfiable?

Probabilistic reasoning (directed)

Party example: the weather effect

- Alex is-likely-to-go in bad weather
- Chris rarely-goes in bad weather
- Becky is indifferent but unpredictable



P(A|W=bad)=.9

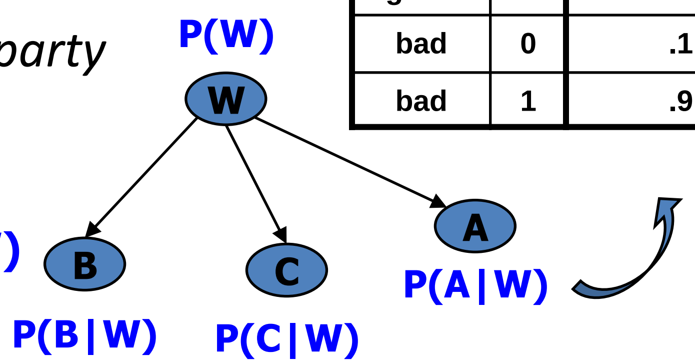
P(C|W=bad)=.1

P(B|W=bad)=.5

Questions:

- *Given bad weather, which group of individuals is most likely to show up at the party?*
- *What is the probability that Chris goes to the party but Becky does not?*

W	A	P(A|W)
good	0	.01
good	1	.99
bad	0	.1
bad	1	.9

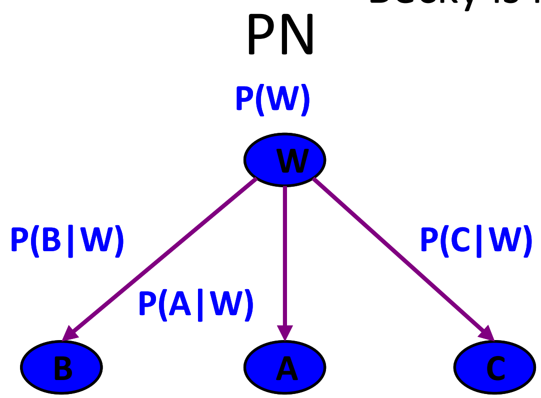


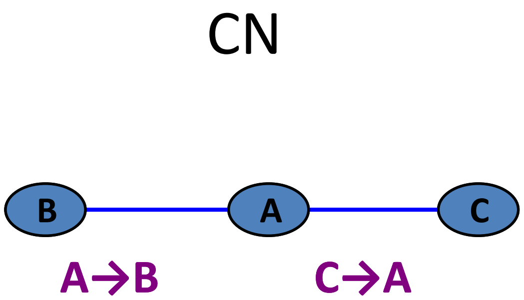
$P(W,A,C,B) = P(B|W) \cdot P(C|W) \cdot P(A|W) \cdot P(W)$

$P(A,C,B|W=bad) = 0.9 \cdot 0.1 \cdot 0.5$

Mixed Probabilistic and Deterministic networks

Alex is-likely-to-go in bad weather
Chris rarely-goes in bad weather
Becky is indifferent but unpredictable





Query:
Is it likely that Chris goes to the party if Becky does not but the weather is bad?

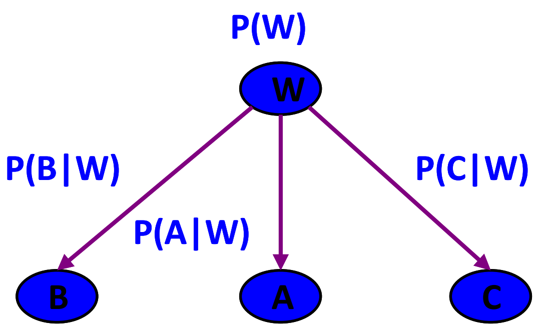
$$P(C, \neg B | w = bad, A \rightarrow B, C \rightarrow A)$$

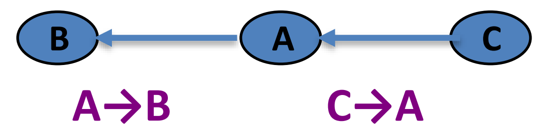
Causal Probabilistic and Deterministic networks

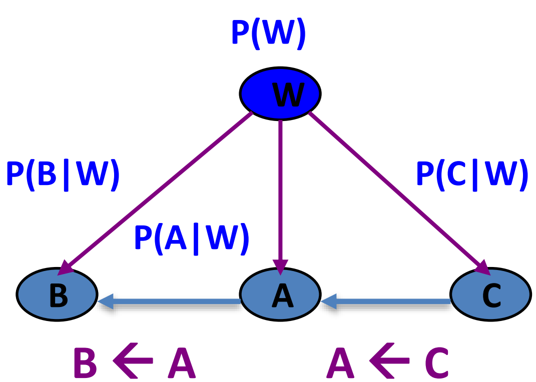
$$P(C, \neg B | w = bad, A \to B, C \to A)$$

Alex is-likely-to-go in bad weather
Chris rarely-goes in bad weather
Becky is indifferent but unpredictable

PN







Causal effect query vs obs query:

- *Is it likely that Chris goes to the party if Becky does not?*
- *Is it likely that Chris goes to the party if* ***we force Becky to not go.***

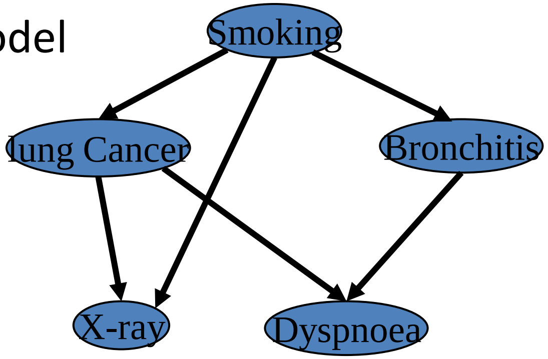
$$P(C \,|do(B = notgo)\,, w = bad)$$

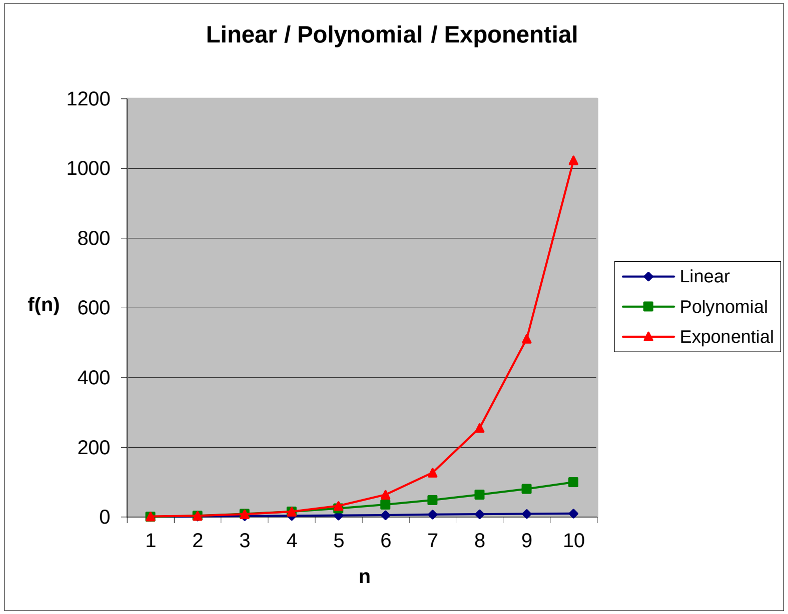
$$P(C \,|B = notgo.\ w = bad)$$

Complexity of Reasoning Tasks

- Constraint satisfaction
- Counting solutions
- Combinatorial optimization
- Belief updating
- Most probable explanation
- Decision-theoretic planning

Given a **full** model





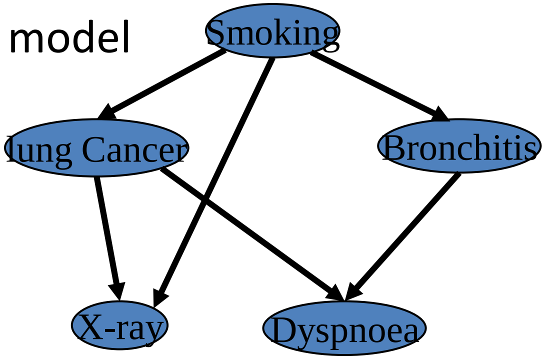
Reasoning is
computationally hard

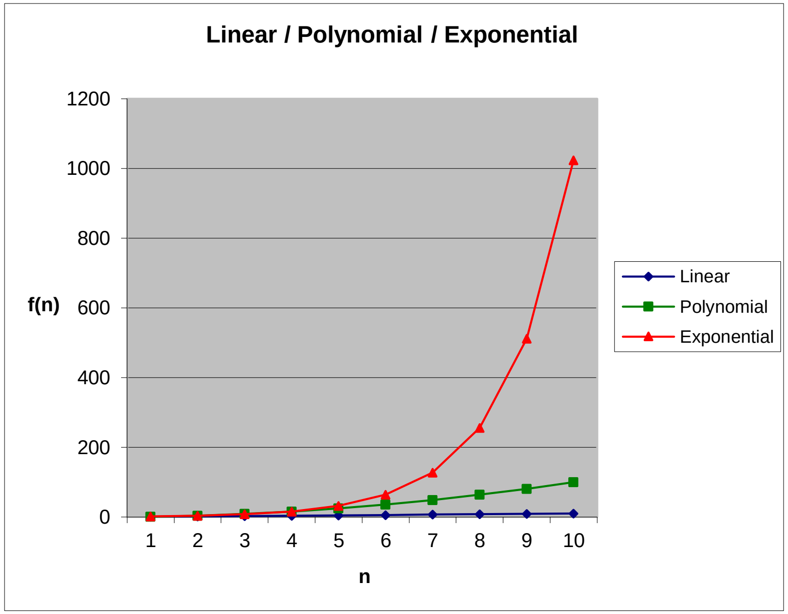
Complexity is
Time and space(memory)

Complexity of Causal Tasks

- Constraint satisfaction
- Counting solutions
- Combinatorial optimization
- Belief updating
- Most probable explanation
- Decision-theoretic planning

Given a **partial** model
And **data**…





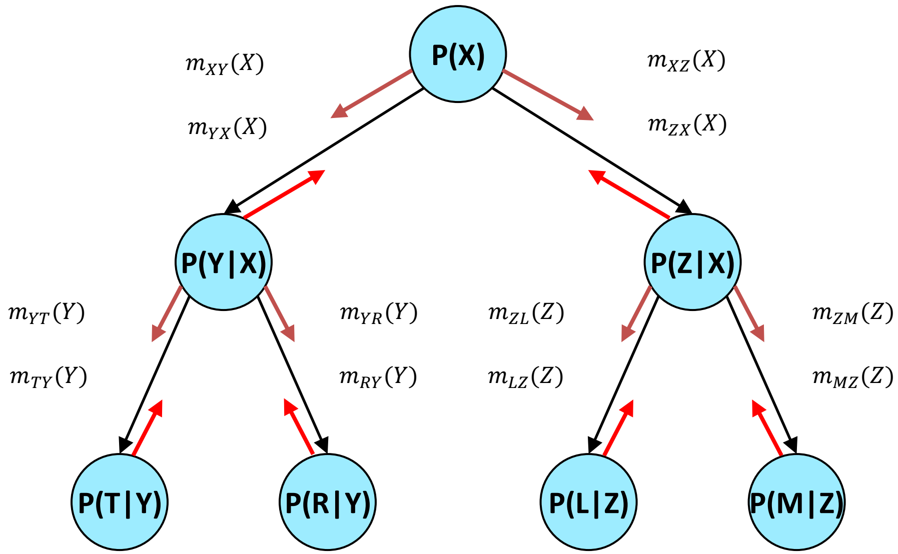
Reasoning is computationally hard

Complexity is
Time and space(memory)

Tree-solving is easy

Belief updating (sum-prod)

CSP – consistency (projection-join)



MPE (max-prod)

#CSP (sum-prod)

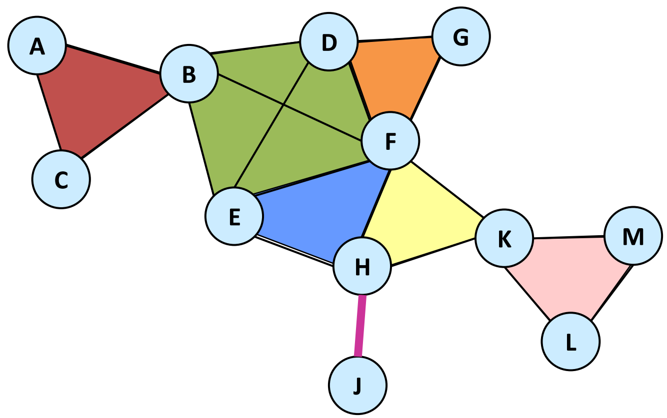
Trees are processed in linear time and memory

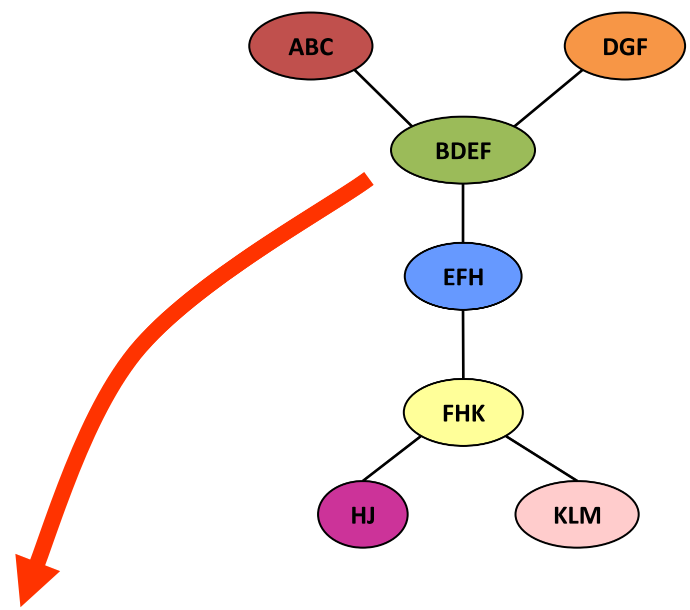
Transforming into a Tree

- **By Inference (thinking)**
  - Transform into a single, equivalent tree of sub-problems

- **By Conditioning (guessing)**
  - Transform into many tree-like sub-problems.

Inference and Treewidth

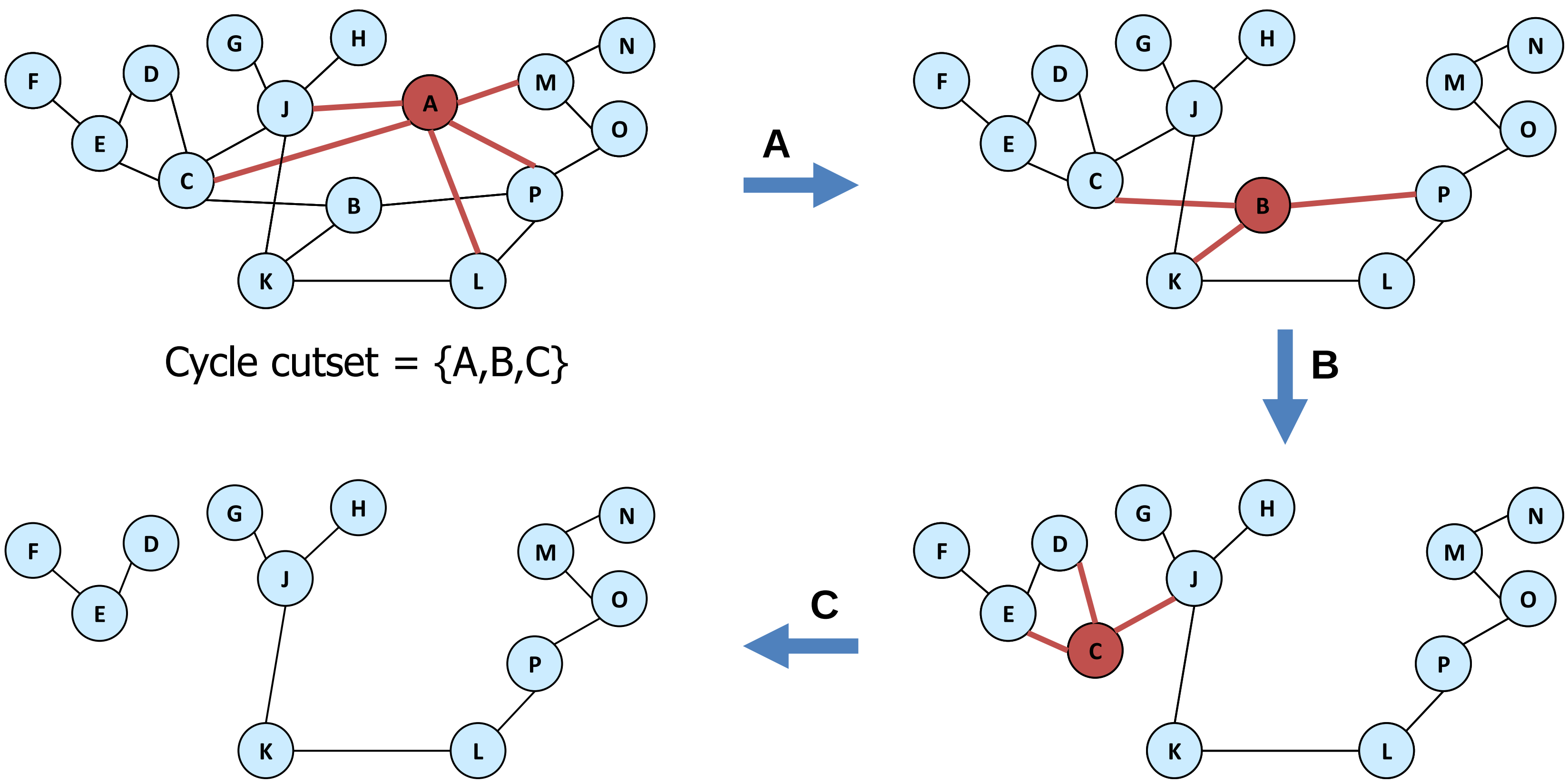




Inference algorithm:
Time: exp(tree-width)
Space: exp(tree-width)

treewidth = 4 - 1 = 3
treewidth = (maximum cluster size) - 1

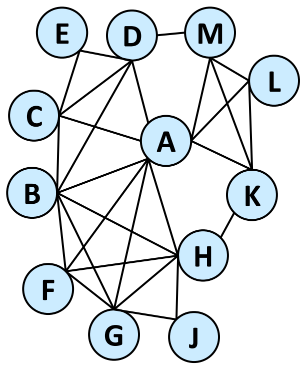
Conditioning and Cycle cutset



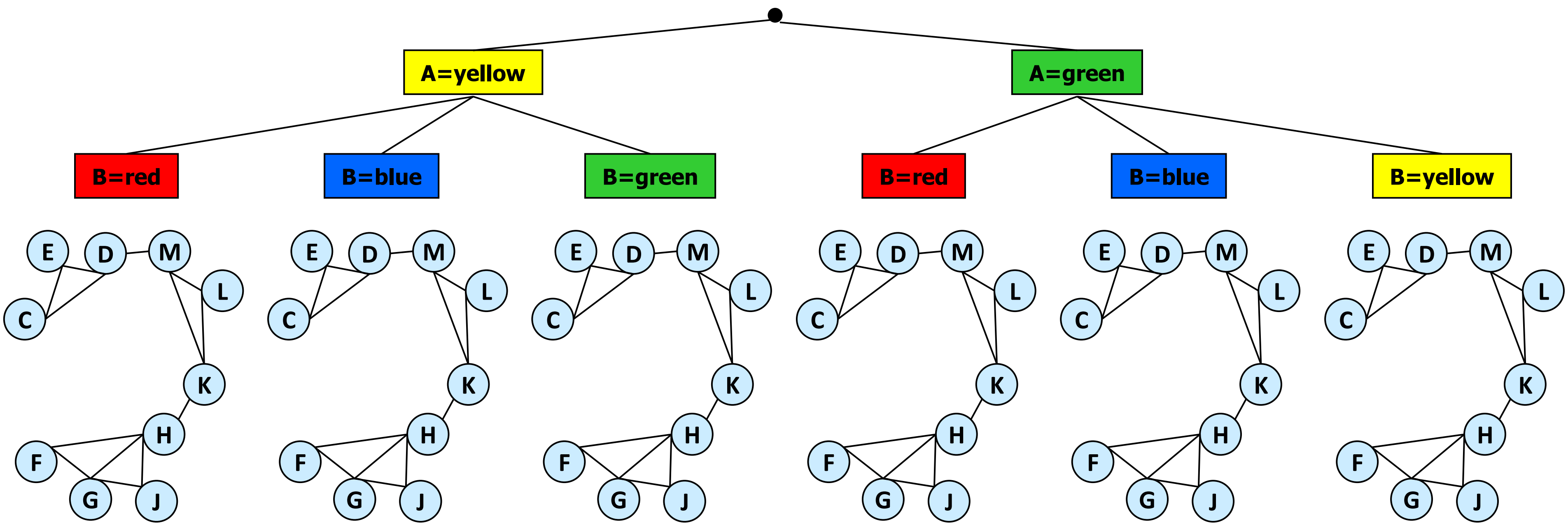
Cycle cutset = {A,B,C}

Search over the Cutset

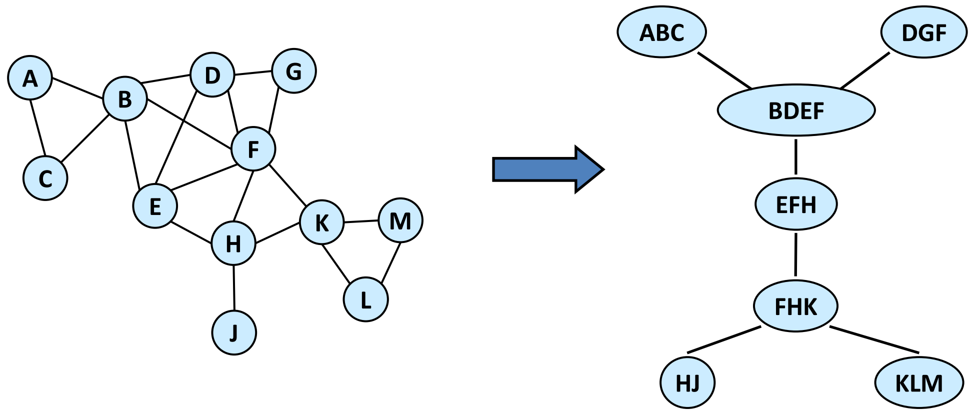
Graph Coloring problem



- Inference may require too much memory
- **Condition** on some of the variables

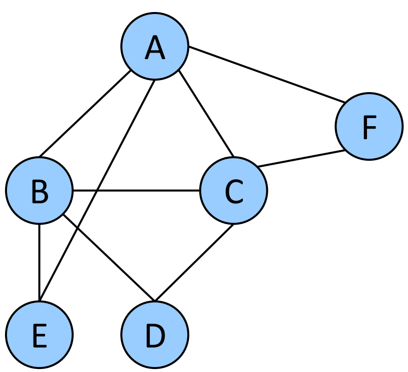


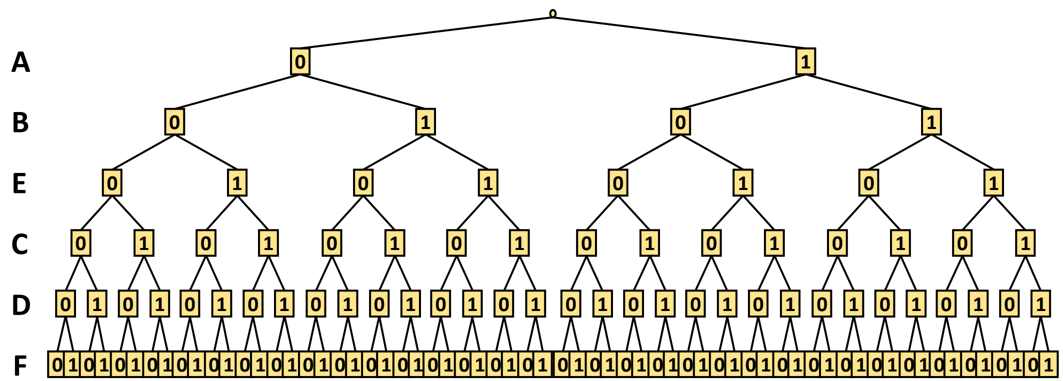
Bird's-eye View of Exact Algorithms



Inference

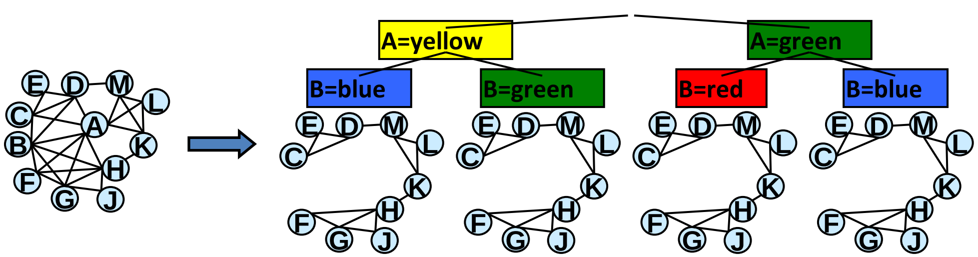
exp(w*) time/space





Search

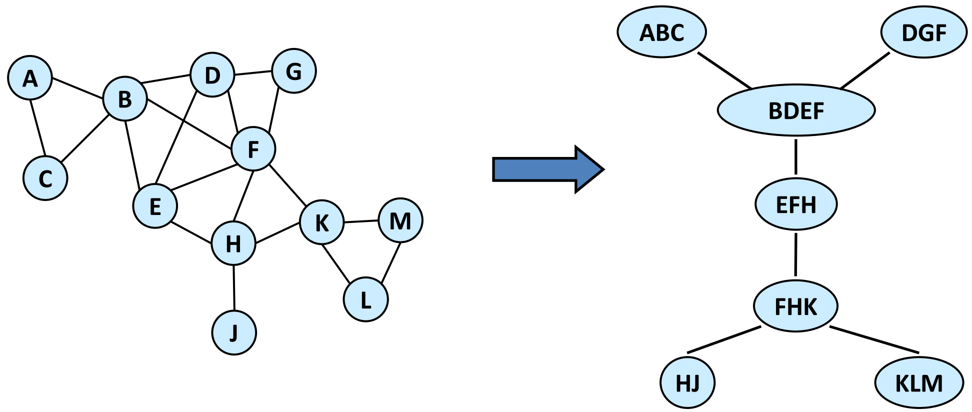
Exp(w*) time
O(w*) space



Search+inference:
Space: exp(q)
Time: exp(q+c(q))

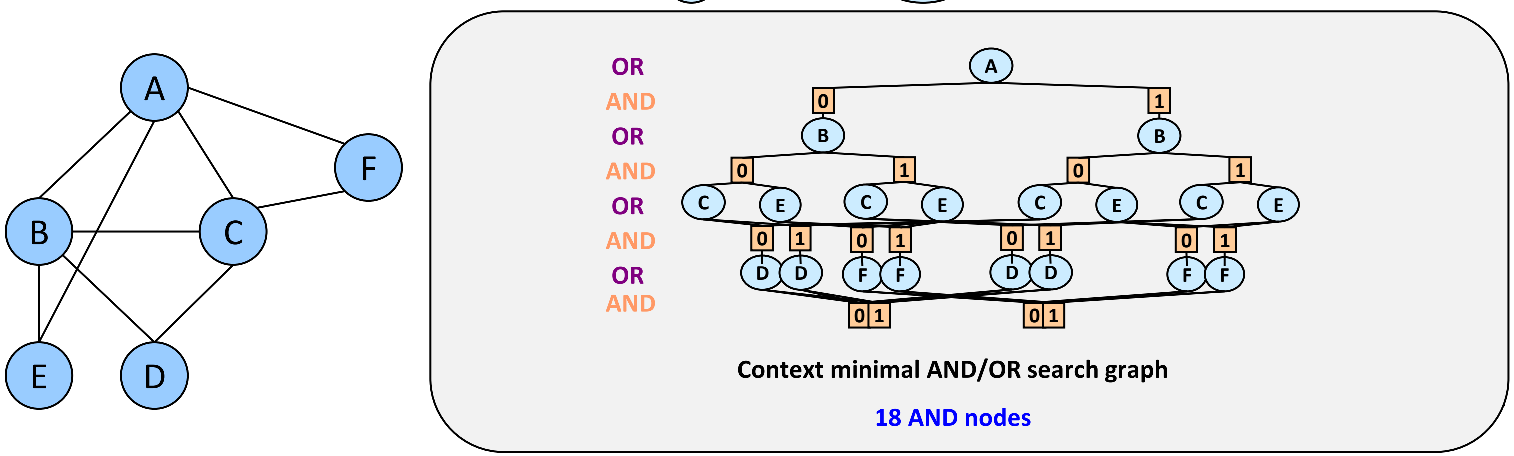
q: user controlled

Bird's-eye View of Exact Algorithms



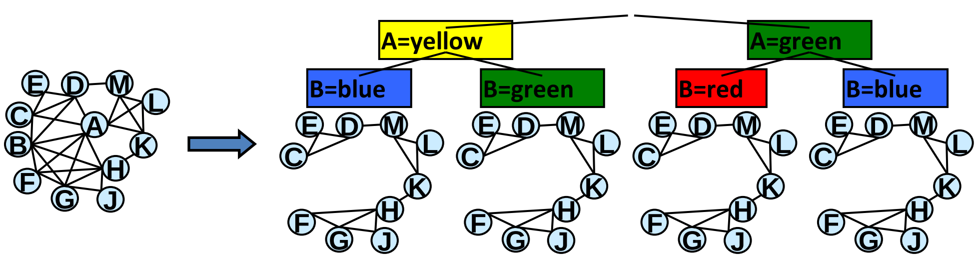
Inference

exp(w*) time/space



Search

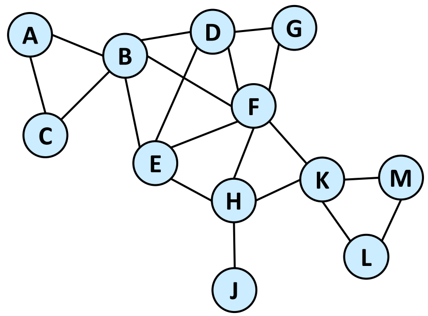
Exp(w*) time
O(w*) space

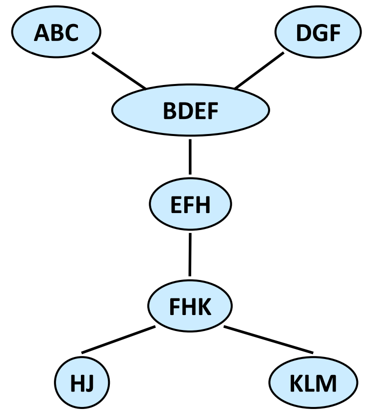


Search+inference:
Space: exp(q)
Time: exp(q+c(q))

q: user controlled

Bird's-eye View of Approximate Algorithms

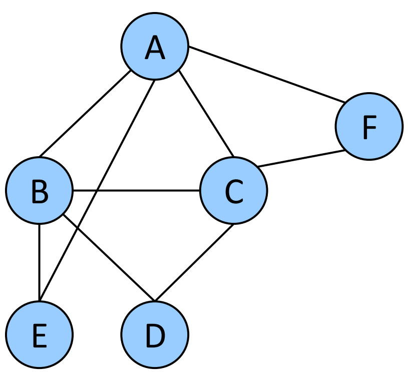


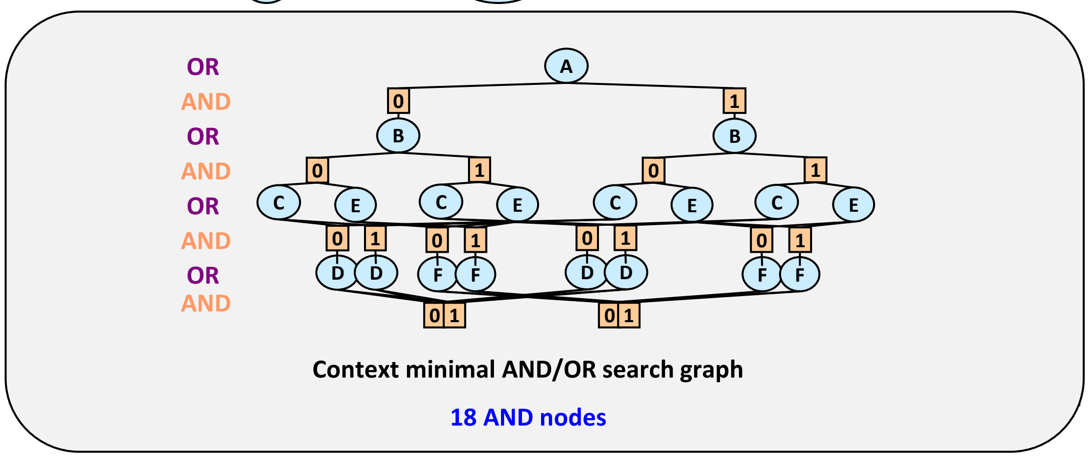


Inference

Bounded Inference

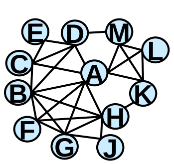
And what about causal-effect? Counterfactuals? Confounding?

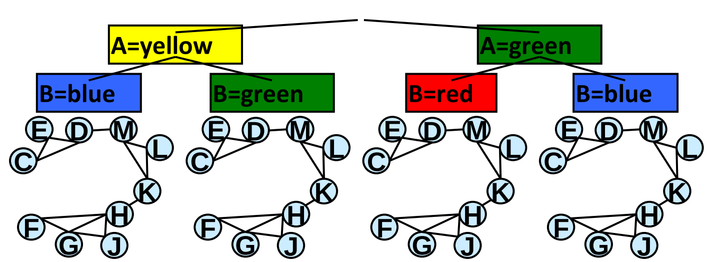




Search

Sampling





Search + inference:

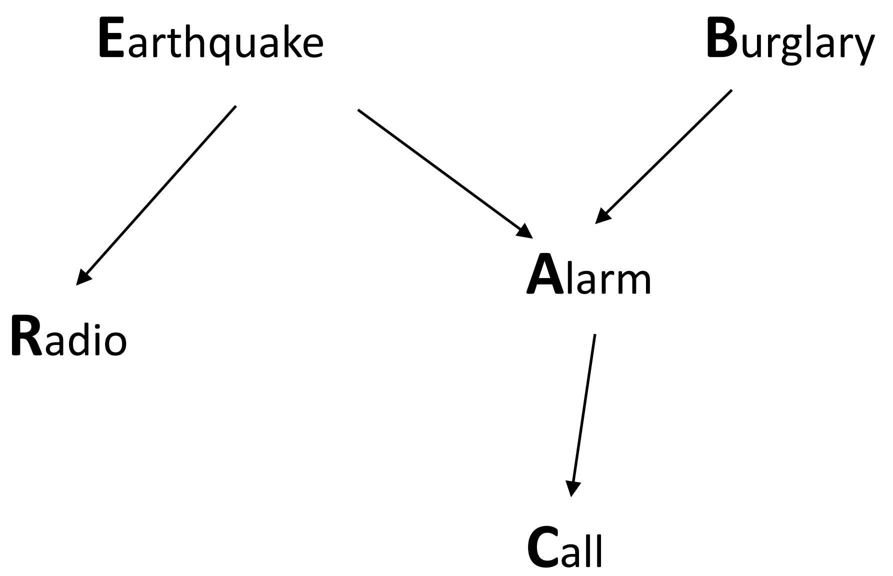
Sampling + bounded inference

Why/What/How Uncertainty?

- Why Uncertainty?
  - Answer: It is abundant
- What formalism to use?
  - Answer: Probability theory
- How to overcome exponential representation?
  - Answer: Graphs, graphs, graphs… to capture irrelevance, independence, causality
- Why Causality?
  - Because it is everywhere (what would have happened in January 6th had the Capitol been better protected?)
  - If we seek strong AI, AGI (Artificial General Intelligence) we must have causal models.

Basics of Probabilistic Calculus (Chapter 3)

The Burglary Example



Degrees of Belief

- Assign a degree of belief or probability in $[0, 1]$ to each world ω and denote it by $\Pr(\omega)$.
- The belief in, or probability of, a sentence α:

$$\Pr(\alpha) \stackrel{def}{=} \sum_{\omega \models \alpha} \Pr(\omega).$$

world	Earthquake	Burglary	Alarm	$\Pr(.)$
ω_1	true	true	true	.0190
ω_2	true	true	false	.0010
ω_3	true	false	true	.0560
ω_4	true	false	false	.0240
ω_5	false	true	true	.1620
ω_6	false	true	false	.0180
ω_7	false	false	true	.0072
ω_8	false	false	false	.7128

Properties of Beliefs

- A bound on the belief in any sentence:

$$0 \leq \Pr(\alpha) \leq 1 \qquad \text{for any sentence } \alpha.$$

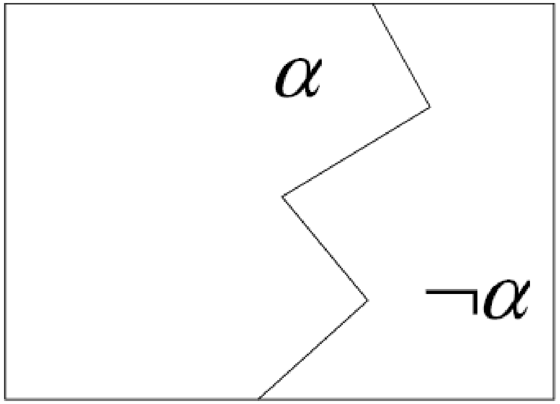
- A baseline for inconsistent sentences:

$$\Pr(\alpha) = 0 \qquad \text{when } \alpha \text{ is inconsistent.}$$

- A baseline for valid sentences:

$$\Pr(\alpha) = 1 \qquad \text{when } \alpha \text{ is valid.}$$

Properties of Beliefs



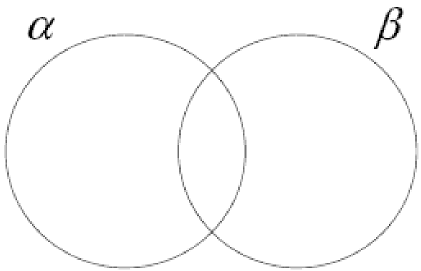
- The belief in a sentence given the belief in its negation:

$$\Pr(\alpha) + \Pr(\neg\alpha) = 1.$$

Example

$$\begin{aligned} \Pr(\text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_5) + \Pr(\omega_6) = .2 \\ \Pr(\neg\text{Burglary}) &= \Pr(\omega_3) + \Pr(\omega_4) + \Pr(\omega_7) + \Pr(\omega_8) = .8 \end{aligned}$$

Properties of Beliefs



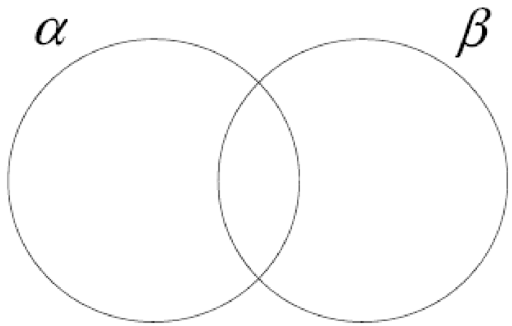
- The belief in a disjunction:

$$\Pr(\alpha \vee \beta) = \Pr(\alpha) + \Pr(\beta) - \Pr(\alpha \wedge \beta).$$

- Example:

$$\begin{aligned}
\Pr(\text{Earthquake}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_3) + \Pr(\omega_4) = .1 \\
\Pr(\text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_5) + \Pr(\omega_6) = .2 \\
\Pr(\text{Earthquake} \wedge \text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) = .02 \\
\Pr(\text{Earthquake} \vee \text{Burglary}) &= .1 + .2 - .02 = .28
\end{aligned}$$

Properties of Beliefs



- The belief in a disjunction:

$\Pr(\alpha \vee \beta) = \Pr(\alpha) + \Pr(\beta)$ when α and β are mutually exclusive.

Entropy

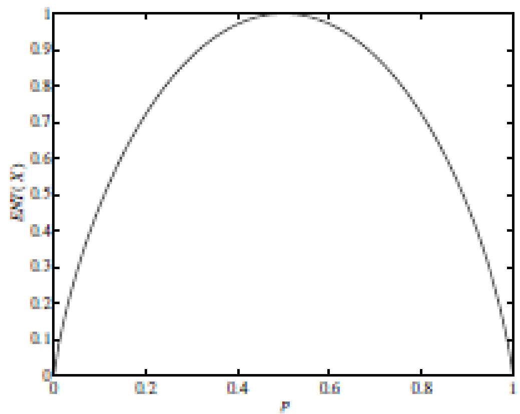
Quantify uncertainty about a variable X using the notion of entropy:

$$\mathrm{ENT}(X) \stackrel{def}{=} -\sum_{x} \Pr(x) \log_2 \Pr(x),$$

where $0 \log 0 = 0$ by convention.

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558
ENT(.)	.469	.722	.802

Entropy



- The entropy for a binary variable X and varying $p = \Pr(X)$.
- Entropy is non-negative.
- When $p = 0$ or $p = 1$, the entropy of X is zero and at a minimum, indicating no uncertainty about the value of X.
- When $p = \frac{1}{2}$, we have $\Pr(X) = \Pr(\neg X)$ and the entropy is at a maximum (indicating complete uncertainty).

Bayes Conditioning

Alpha and beta are events

Closed form for Bayes conditioning:

$$\Pr(\alpha|\beta) = \frac{\Pr(\alpha \wedge \beta)}{\Pr(\beta)}.$$

Defined only when $\Pr(\beta) \neq 0$.

Degrees of Belief

world	Earthquake	Burglary	Alarm	$\Pr(.)$
ω_1	true	true	true	.0190
ω_2	true	true	false	.0010
ω_3	true	false	true	.0560
ω_4	true	false	false	.0240
ω_5	false	true	true	.1620
ω_6	false	true	false	.0180
ω_7	false	false	true	.0072
ω_8	false	false	false	.7128

$$\begin{aligned}
\Pr(\text{Earthquake}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_3) + \Pr(\omega_4) = .1 \\
\Pr(\text{Burglary}) &= .2 \\
\Pr(\neg\text{Burglary}) &= .8 \\
\Pr(\text{Alarm}) &= .2442
\end{aligned}$$

Belief Change

Burglary is independent of Earthquake

Conditioning on evidence Earthquake:

$$\begin{aligned} \Pr(\text{Burglary}) &= .2 \\ \Pr(\text{Burglary}|\text{Earthquake}) &= .2 \\ \\ \Pr(\text{Alarm}) &= .2442 \\ \Pr(\text{Alarm}|\text{Earthquake}) &\approx .75 \uparrow \end{aligned}$$

The belief in Burglary is not changed, but the belief in Alarm increases.

Belief Change

Earthquake is independent of burglary

Conditioning on evidence Burglary:

$$
\begin{aligned}
\Pr(\text{Alarm}) &= .2442 \\
\Pr(\text{Alarm}|\text{Burglary}) &\approx .905 \uparrow \\
\\
\Pr(\text{Earthquake}) &= .1 \\
\Pr(\text{Earthquake}|\text{Burglary}) &= .1
\end{aligned}
$$

The belief in Alarm increases in this case, but the belief in Earthquake stays the same.

Belief Change

The belief in Burglary increases when accepting the evidence Alarm. How would such a belief change further upon obtaining more evidence?

- Confirming that an Earthquake took place:

$$\begin{aligned} \Pr(\text{Burglary}|\text{Alarm}) &\approx .741 \\ \Pr(\text{Burglary}|\text{Alarm} \wedge \text{Earthquake}) &\approx .253 \downarrow \end{aligned}$$

  We now have an explanation of Alarm.

- Confirming that there was no Earthquake:

$$\begin{aligned} \Pr(\text{Burglary}|\text{Alarm}) &\approx .741 \\ \Pr(\text{Burglary}|\text{Alarm} \wedge \neg\text{Earthquake}) &\approx .957 \uparrow \end{aligned}$$

  New evidence will further establish burglary as an explanation.

Conditional Independence

$\Pr$ finds α conditionally independent of β given γ iff

$$\Pr(\alpha|\beta \wedge \gamma) = \Pr(\alpha|\gamma) \quad \text{or } \Pr(\beta \wedge \gamma) = 0.$$

Another definition

$$\Pr(\alpha \wedge \beta|\gamma) = \Pr(\alpha|\gamma)\Pr(\beta|\gamma) \quad \text{or } \Pr(\gamma) = 0.$$

Variable Independence

$\Pr$ finds **X** independent of **Y** given **Z**, denoted $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, means that $\Pr$ finds **x** independent of **y** given **z** for all instantiations **x**, **y** and **z**.

Example

$\mathbf{X} = \{A, B\}$, $\mathbf{Y} = \{C\}$ and $\mathbf{Z} = \{D, E\}$, where A, B, C, D and E are all propositional variables. The statement $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is then a compact notation for a number of statements about independence:

$A \wedge B$ is independent of C given $D \wedge E$;

$A \wedge \neg B$ is independent of C given $D \wedge E$;

$\vdots$

$\neg A \wedge \neg B$ is independent of $\neg C$ given $\neg D \wedge \neg E$;

That is, $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is a compact notation for $4 \times 2 \times 4 = 32$ independence statements of the above form.

Further Properties of Beliefs

Chain rule

$$\begin{aligned}
&\Pr(\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_n) \\
&\quad = \Pr(\alpha_1|\alpha_2 \wedge \ldots \wedge \alpha_n)\Pr(\alpha_2|\alpha_3 \wedge \ldots \wedge \alpha_n)\ldots\Pr(\alpha_n).
\end{aligned}$$

Case analysis (law of total probability)

$$\Pr(\alpha) = \sum_{i=1}^{n} \Pr(\alpha \wedge \beta_i),$$

where the events $\beta_1, \ldots, \beta_n$ are mutually exclusive and exhaustive.

Further Properties of Beliefs

Another version of case analysis

$$\Pr(\alpha) = \sum_{i=1}^{n} \Pr(\alpha|\beta_i)\Pr(\beta_i),$$

where the events $\beta_1, \ldots, \beta_n$ are mutually exclusive and exhaustive.

Two simple and useful forms of case analysis are these:

$$\begin{aligned}\Pr(\alpha) &= \Pr(\alpha \wedge \beta) + \Pr(\alpha \wedge \neg\beta) \\ \Pr(\alpha) &= \Pr(\alpha|\beta)\Pr(\beta) + \Pr(\alpha|\neg\beta)\Pr(\neg\beta).\end{aligned}$$

The main value of case analysis is that, in many situations, computing our beliefs in the cases is easier than computing our beliefs in α. We shall see many examples of this phenomena in later chapters.

Further Properties of Beliefs

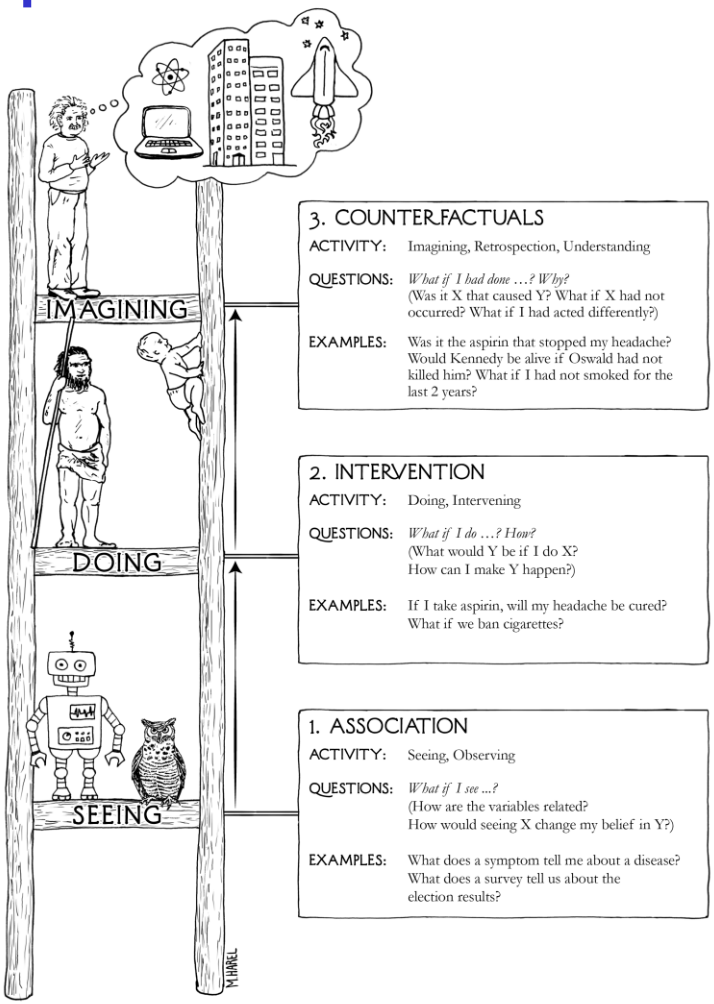
Bayes rule

$$\Pr(\alpha|\beta) = \frac{\Pr(\beta|\alpha)\Pr(\alpha)}{\Pr(\beta)}.$$

- Classical usage: α is perceived to be a cause of β.
- Example: α is a disease and β is a symptom–
- Assess our belief in the cause given the effect.
- Belief in an effect given its cause, $\Pr(\beta|\alpha)$, is usually more readily available than the belief in a cause given one of its effects, $\Pr(\alpha|\beta)$.

Ladder of Causation (PCH)

seeing, doing, and imagining.



- Most animals, learning machines are on the first rung, learning from association.
- Tool users, such as early humans, are on the second rung, if they act by planning and not merely by imitation. We can also use experiments to learn the effects of interventions, and presumably this is how babies acquire much of their causal knowledge.
- On the top rung, counterfactual learners can imagine worlds that do not exist and infer reasons for observed phenomena.

Darwiche 2017: “Human-Level Intelligence or Animal-Like Abilities?”