Introduction to Bayesian Belief Nets

Russ Greiner

Dep't of Computing Science

Alberta Ingenuity Centre for Machine Learning

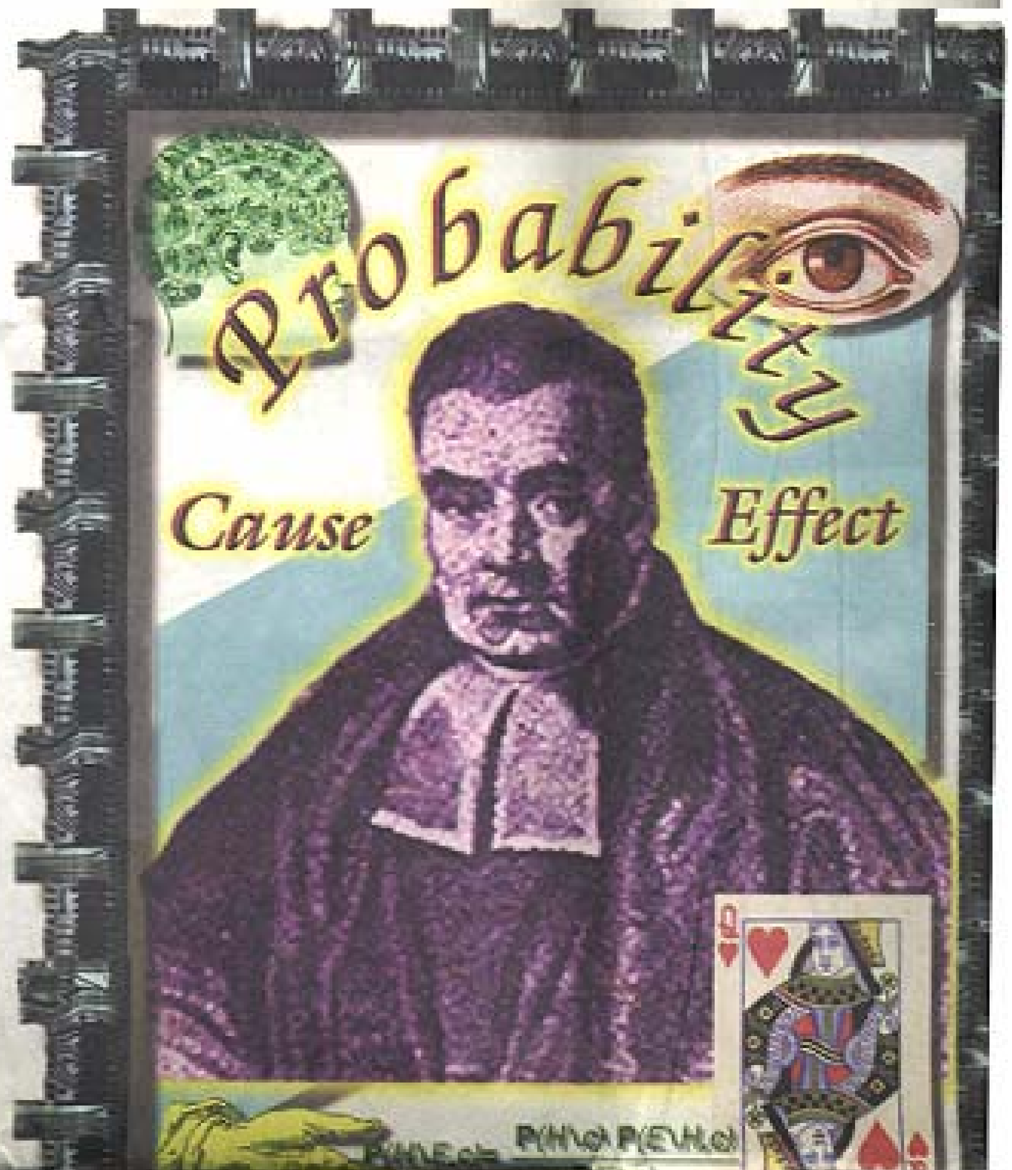
University of Alberta

http://www.cs.ualberta.ca/~greiner/bn.html

BUSINESS

MONDAY TECHNOLOGY SPECIAL

The future of software may lie in the obscure theories of an 18th century cleric named Thomas Bayes.



Motivation

- Gates says *[LATimes, 28/Oct/96]:*

  Microsoft's competitive advantages is its expertise in "Bayesian networks"

- *Current Products*
  - *Microsoft Pregnancy and Child Care (MSN)*
  - *Answer Wizard (Office, …)*
  - *Print Troubleshooter*

    Excel Workbook Troubleshooter

    Office 95 Setup Media Troubleshooter

    Windows NT 4.0 Video Troubleshooter

    Word Mail Merge Troubleshooter

Motivation (II)

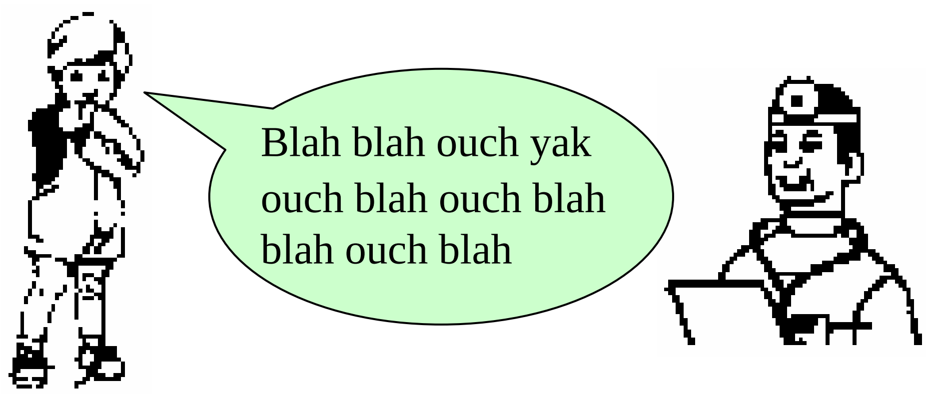
- US Army: **SAIP** (Battalion Detection from SAR, IR… GulfWar)
- NASA: **Vista** (DSS for Space Shuttle)
- GE: **Gems** (real-time monitor for utility generators)
- Intel: (infer possible processing problems from end-of-line tests on semiconductor chips)
- KIC:
  - medical: sleep disorders, pathology, trauma care, hand and wrist evaluations, dermatology, home-based health evaluations
  - DSS for capital equipment: locomotives, gas-turbine engines, office equipment

Motivation (III)

- Lymph-node pathology diagnosis
- Manufacturing control
- Software diagnosis
- Information retrieval
- *Types of tasks*
  - *Classification/Regression*
  - *Sensor Fusion*
  - *Prediction/Forecasting*

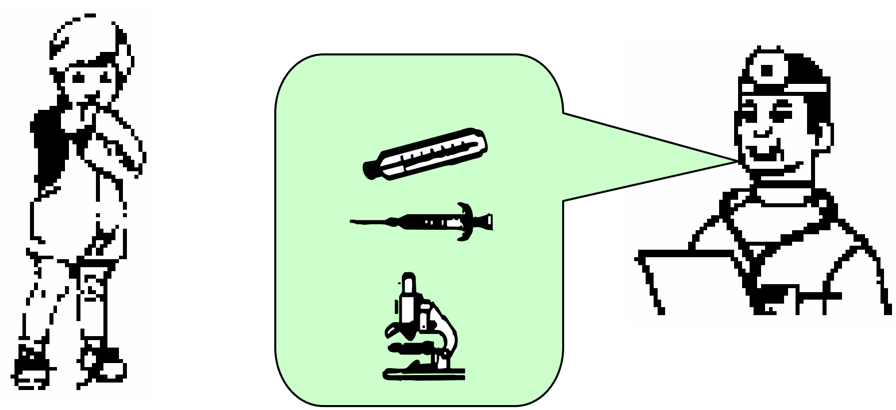
Outline

- Existing uses of Belief Nets (BNs)
- What is a BN ?
- Specific Examples of BNs
- Contrast with Rules, Neural Nets, ...
- Possible applications of BNs
- Challenges
  - How to reason efficiently
  - How to *learn* BNs



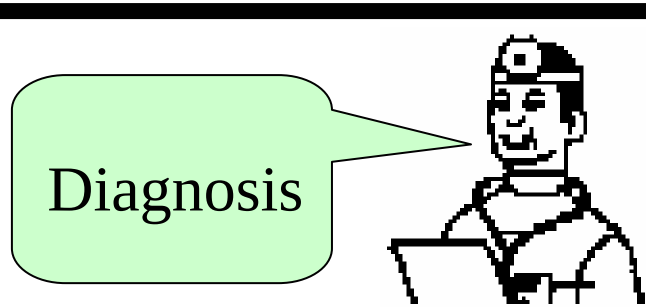
Symptoms

Chief complaint
History, …



Signs

Physical Exam
Test results, …



Plan

Treatment, …

Objectives: Decision Support System

- Determine
  - which *tests* to perform
  - which *repair* to suggest

  based on *costs, sensitivity/specificity*, ...
- Use all sources of information
  - *symbolic* (discrete observations, history, ...)
  - *signal* (from sensors)
- Handle *partial* information
- *Adapt* to track fault distribution

Underlying Task

- *Situation*: Given observations $\{O_1=v_1, \ldots\ O_k=v_k\}$

  (symptoms, history, test results, ...)

  what is best DIAGNOSIS Dx_i for patient?

- *Approach1*: Use set of $\boxed{obs_1\ \&\ \ldots\ \&\ obs_m \rightarrow Dx_i}$ rules

  but… *Need rule for each situation*

  - for each diagnosis Dx_r
  - for each set of possible values v_j for O_j
  - for each subset of obs. $\{O_{x1}, O_{x2}, \ldots\} \subset \{O_j\}$

  Can't use

  $\boxed{\text{If Temp>100 \& BP = High \& Cough = Yes} \rightarrow \text{DiseaseX}}$

  if only know Temp and BP

- *Seldom Completely Certain*

Underlying Task

- *Situation*: Given observations $\{O_1=v_1, \ldots O_k=v_k\}$
  (symptoms, history, test results, ...)
  what is best DIAGNOSIS Dx_i for patient?

- *Approach 2*: Compute Probabilities of Dx_i
  given observations $\{ obs_j \}$

$$P(\text{Dx} = u \mid O_1 = v_1, \ldots, O_k = v_k)$$

- *Challenge:* How to express Probabilities?

How to deal with Probabilities

- Sufficient: "atomic events":

$$P(\ Dx = u,\ O_1 = v_1, \ldots,\ O_k = v_k, \ldots,\ O_N = v_N\ )$$

for all 2^{1+N} values $u \in \{T, F\}$, $v_j \in \{T, F\}$

$$\Rightarrow \quad \begin{array}{l} P(\ Dx=T,\ O_1=T,\ O_2=T,\ \ldots,\ O_N=T\ ) = 0.03 \\ P(\ Dx=T,\ O_1=T,\ O_2=T,\ \ldots,\ O_N=F\ ) = 0.4 \\ \quad\quad\ldots \\ P(\ Dx=T,\ O_1=F,\ O_2=F,\ \ldots,\ O_N=T\ ) = 0 \\ \quad\quad\ldots \\ P(\ Dx=F,\ O_1=F,\ O_2=F,\ \ldots,\ O_N=F\ ) = 0.01 \end{array}$$

- *Then: Marginalize:*

$$P(Dx = u, O_1 = v_1, \ldots O_7 = v_7) = \sum_{v_8, \ldots v_N} P(Dx = u,\ O_1 = v_1, \ldots O_7 = v_7, \ldots O_N = v_N)$$

Conditionalize:

$$P(Dx = u \mid O_1 = v_1, \ldots O_7 = v_7) = \frac{P(Dx = u,\ O_1 = v_1, \ldots O_7 = v_7)}{P(\ O_1 = v_1, \ldots O_7 = v_7)}$$

- But… even if binary Dx, 20 binary obs.'s. $\Rightarrow$ *>2,097,000* numbers!

Problems with "Atomic Events"

- Representation *is not intuitive*

  ⇒ Should make "*connections*" explicit

  use "local information"

P(Jaundice | Hepatitis), P(LightDim | BadBattery),…

- Too many numbers – $O(2^N)$
  - Hard to store
  - Hard to use

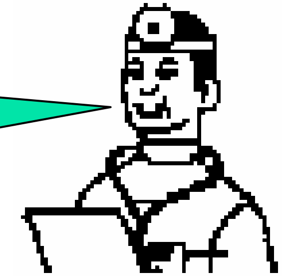
    [Must add 2^r values to marginalize r variables]
  - Hard to learn

    [Takes $O(2^N)$ samples to learn 2^N parameters]

  ⇒ Include only *necessary* "connections"

⇒ Belief Nets

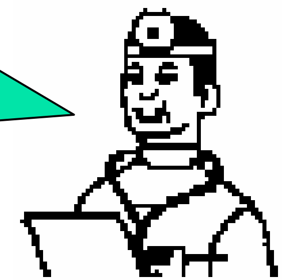
? Hepatitis?



Jaunticed

BloodTest

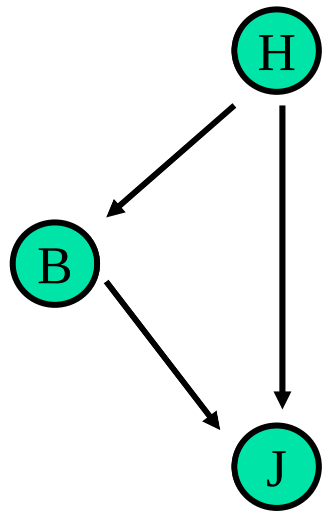
? Hepatitis,
not Jaunticed
but +BloodTest
?



Encoding Causal Links

- Simple Belief Net:

P(H=1)	P(H=0)
0.05	0.95



h	P(B=1 | H=h)	P(B=0 | H=h)
1	0.95	0.05
0	0.03	0.97

h	b	P(J=1|h,b)	P(J=0|h,b)
1	1	0.8	0.2
1	0	0.8	0.2
0	1	0.3	0.7
0	0	0.3	0.7

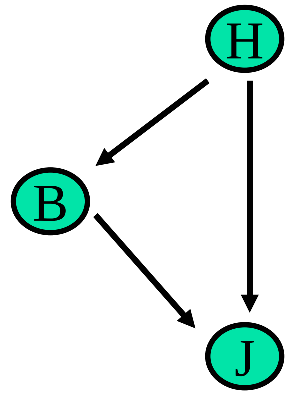
- Node ~ Variable

  Link ~ "Causal dependency"
- "CPTable" ~ P(child | parents)

Encoding Causal Links

P(H=1)
0.05

h	P(B=1 | H=h)
1	0.95
0	0.03



h	b	P(J=1|h , b)
1	1	0.8
1	0	0.8
0	1	0.3
0	0	0.3

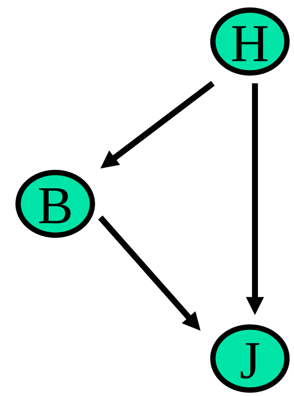
- $P(J \mid H, B=0) \; = \; P(J \mid H, B=1) \;\; \forall\, J,\; H\,!$

  $\Rightarrow \quad \mathbf{P(\, J \mid H, B) \; = \; P(J \mid H)}$
- *J* is <u>INDEPENDENT</u> of *B*, once we know *H*
- Don't need $B \rightarrow J$ arc!

Encoding Causal Links

P(H=1)
0.05

h	P(B=1 | H=h)
1	0.95
0	0.03



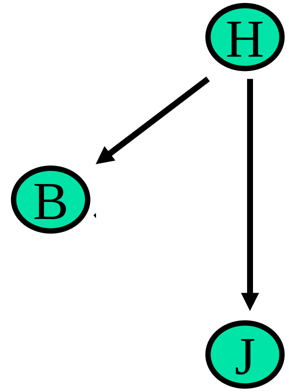
h		P(J=1|h)
1		0.8
1		
0		0.3
0		

- $P(J \mid H, B{=}0) \;=\; P(J \mid H, B{=}1) \;\; \forall\, J,\; H\,!$

  $\Rightarrow$ $\mathbf{P(\,J\;\mid H, B) \;=\; P(J\;\mid H)}$
- *J* is <u>INDEPENDENT</u> of *B*, once we know *H*
- Don't need $B \rightarrow J$ arc!

Encoding Causal Links

h	P(B=1 | H=h)
1	0.95
0	0.03



P(H=1)
0.05

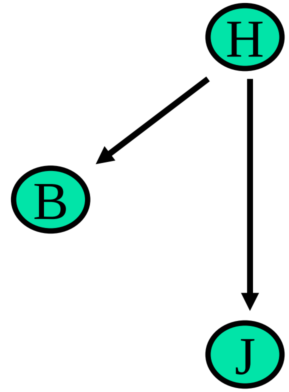
h	P(J=1|h)
1	0.8
0	0.3

- $P(J \mid H, B{=}0) \;=\; P(J \mid H, B{=}1) \quad \forall\, J,\; H\,!$

  $\Rightarrow \quad \mathbf{P(\,J \mid H, B) \;=\; P(J \mid H)}$
- *J* is <u>INDEPENDENT</u> of *B*, once we know *H*
- Don’t need $B \rightarrow J$ arc!

Sufficient Belief Net

h	P(B=1 | H=h)
1	0.95
0	0.03



P(H=1)
0.05

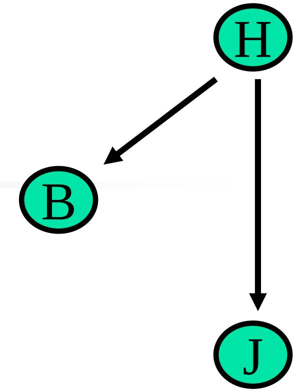
h	P(J=1|h)
1	0.8
0	0.3

- Requires: *P(H=1)* known
  P(J=1 | H=1) known
  P(B=1 | H=1) known

  (Only 5 parameters, not 7)

Hence: $P(\mathrm{H}=1 \mid \mathrm{B}=1, \mathrm{J}=0) = \frac{1}{\alpha} P(\mathrm{H}=1)\, P(\mathrm{B}=1 \mid \mathrm{H}=1)\ \overset{P(\mathrm{J}=0 \mid \mathrm{H}=1)}{\cancel{\mathrm{P}(\mathrm{J}=0 \mid \mathrm{B}=1,\mathrm{H}=1)}}$

"Factoring"



- B *does* depend on J:

  If J=1, then likely that H=1 $\Rightarrow$ *B =1*

- *but…* *ONLY THROUGH* H:
  - If know H=1, then likely that B=1
  - … doesn't matter whether J=1 or J=0 !

$$\Rightarrow \quad \boxed{\mathbf{P(J=0 \mid B=1,\ H=1) = P(J=0 \mid H=1)}}$$

N.b., B *and* J ARE *correlated* a priori *P(J | B) ≠ P(J)*
GIVEN H, *they become uncorrelated* *P(J | B, H) = P(J | H)*

Factored Distribution

- Symptoms *independent,* given Disease

H	Hepatitis
J	Jaundice
B	(positive) Blood test

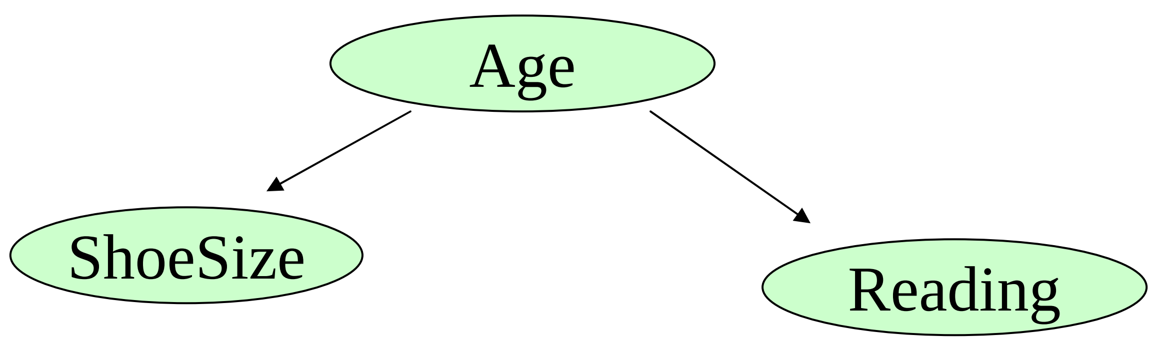
$P(\text{B} \mid \text{J}) \neq P(\text{B})$ but
$P(\text{B} \mid \text{J,H}) = P(\text{B} \mid \text{H})$

- **ReadingAbility** and **ShoeSize** are dependent,

$P(\text{ReadAbility} \mid \text{ShoeSize}) \neq P(\text{ReadAbility})$

but become independent, given Age

$P(\text{ReadAbility} \mid \text{ShoeSize, Age}) = P(\text{ReadAbility} \mid \text{Age})$



"Naïve Bayes"

- Classification Task:

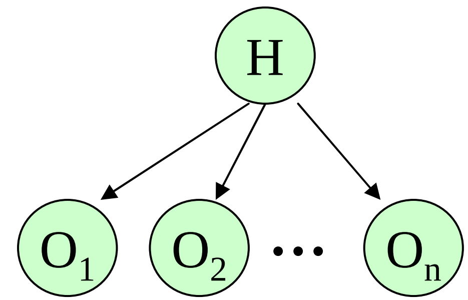
  Given $\{ O_1 = v_1, \ldots, O_n = v_n \}$

  Find h_i that maximizes $(H = h_i \mid O_1 = v_1, \ldots, O_n = v_n)$

- Given

  $P(H = h_i)$

  $P(O_j = v_j \mid H = h_i)$

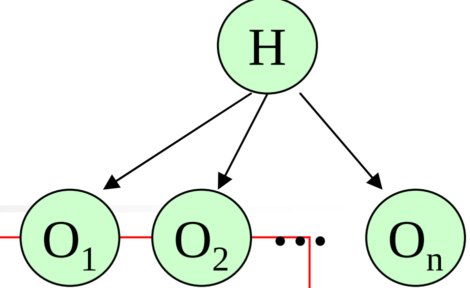


  Independent: $P(O_j \mid H, O_k, \ldots) = P(O_j \mid H)$

$$P(H = h_i \mid O_1 = v_1 \ldots, O_n = v_n) = \frac{1}{\alpha} P(H = h_i) \prod_j P(O_j = v_j \mid H = h_i)$$

- Find argmax $\{h_i\}$

Naïve Bayes (con't)



$$P(H = h_i \mid O_1 = v_1, \ldots, O_n = v_n) = \frac{1}{\alpha} P(H = h_i) \prod_j P(O_j = v_j \mid H = h_i)$$

- *Normalizing term* $\alpha = P(O_1 = v_1, \ldots, O_n = v_n) = \sum_i P(H = h_i) \prod_j P(O_j = v_j \mid H = h_i)$

  (No need to compute, as same for all h_i)

- Easy to use for Classification
- Can use even if some v_js not specified
- If k Dx's and n O_is,

  requires only k priors, $n * k$ pairwise-conditionals

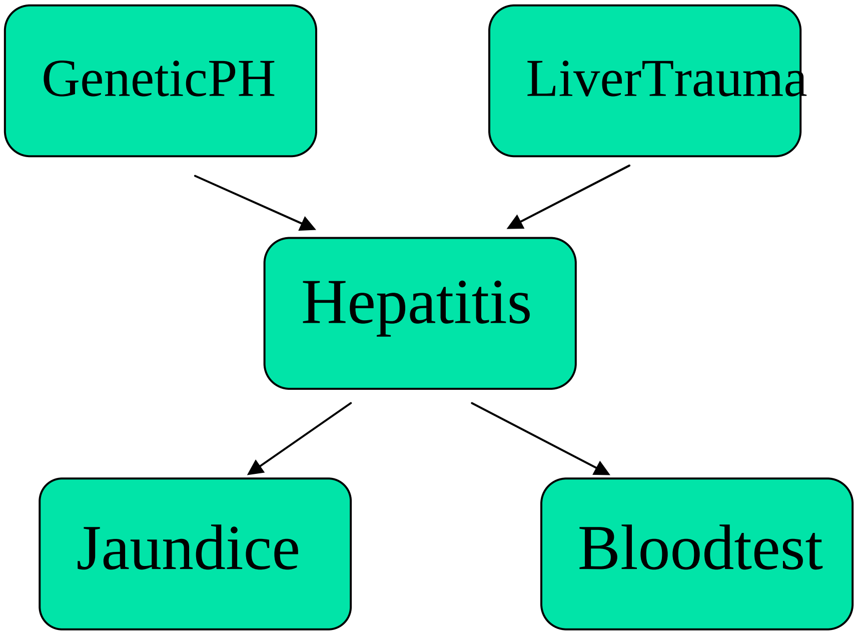
  (Not 2^{n+k}… relatively easy to learn)

n	1+2n	$2^{n+1} - 1$
10	21	2,047
30	61	2,147,438,647

Bigger Networks

P(H=1)
0.32

P(I=1)
0.20



g	lt	P(H=1|g ,lt)
1	1	0.82
1	0	0.10
0	1	0.45
0	0	0.04

h	P(J=1| h)
1	0.8
0	0.3

h	P(B=1| h)
1	0.98
0	0.01

- Intuition: Show *CAUSAL* connections:
  GeneticPH CAUSES Hepatitis; Hepatitis CAUSES Jaundice
- If GeneticPH, then expect Jaundice: $\boxed{\text{GeneticPH} \Rightarrow \text{Hepatitis} \Rightarrow \text{Jaundice}}$

But only via Hepatitis: $\boxed{\text{GeneticPH and not Hepatitis} \not\Rightarrow \text{Jaundice}}$

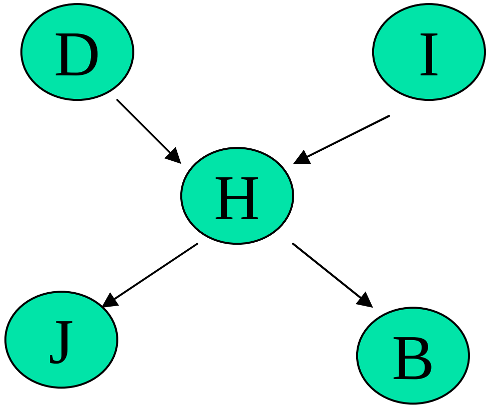
$$P(\text{ J} \mid \text{G }) \neq P(\text{ J }) \quad \text{but}$$
$$P(\text{ J} \mid \text{G,H }) = P(\text{ J} \mid \text{H})$$

Belief Nets

- DAG structure
  - Each node ≡ Variable *v*
  - *v* depends (only) on its parents

  \+ conditional prob: $P(v_i \mid \text{parent}_i = \langle 0,1,\ldots \rangle)$

- *v is INDEPENDENT* of non-descendants, given assignments to its parents



Given H = 1,

- D has no influence on J
- J has no influence on B
- etc.

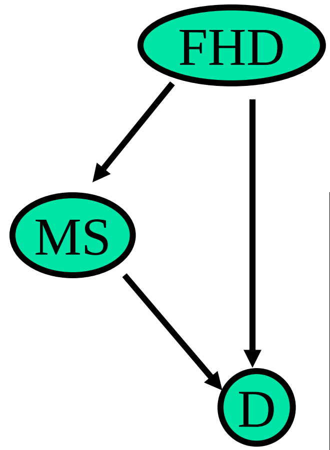
Less Trivial Situations

- *N.b.*, obs_1 is *not* always independent of obs_2 given H
- *Eg,* **FamilyHistoryDepression** 'causes' **MotherSuicide** and **Depression**

  MotherSuicide causes **Depression** (w/ or w/o **F.H.Depression)**

P(FHD=1)
0.001

f	P(MS=1 | FHD=f)
1	0.10
0	0.03



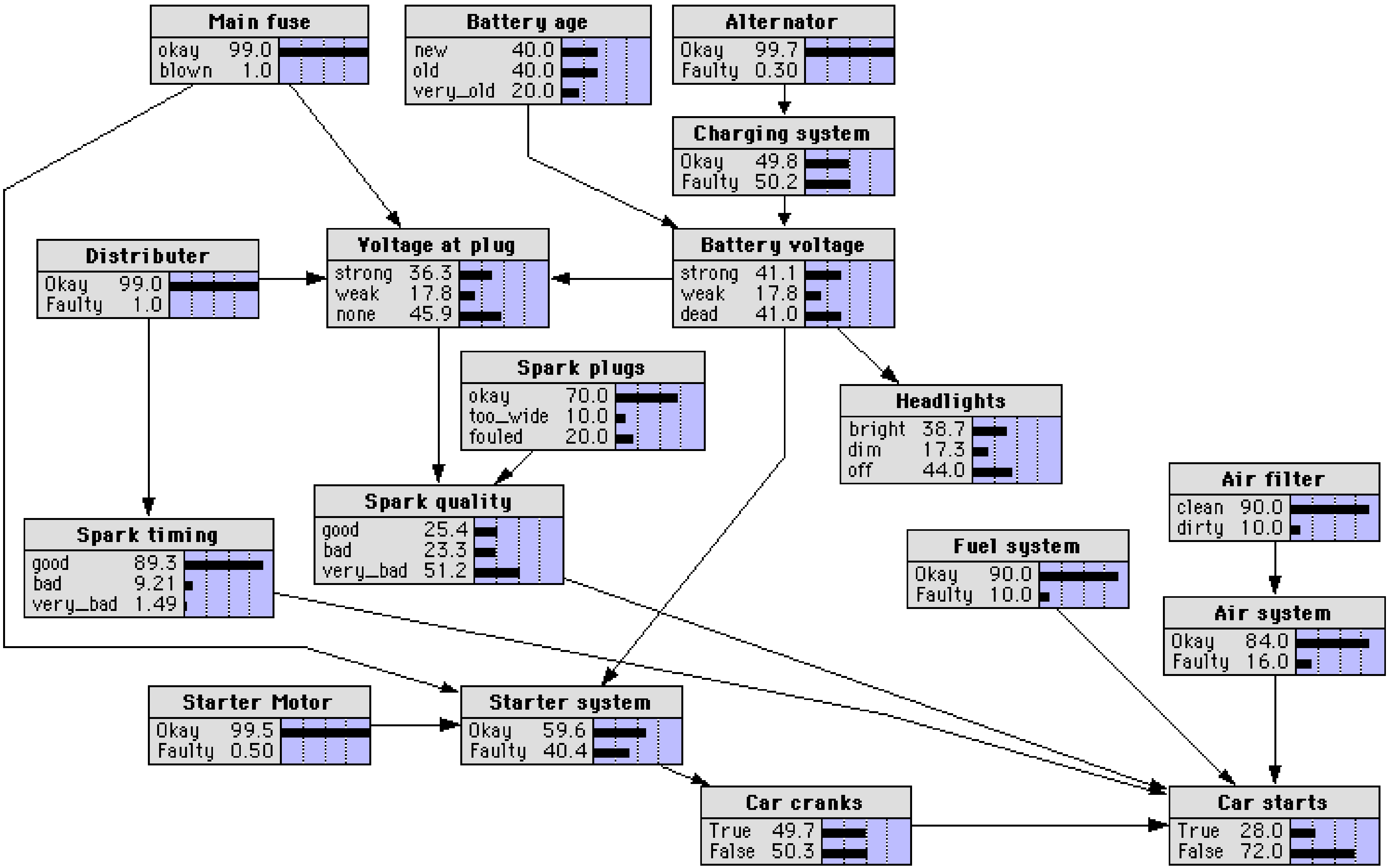
f	m	P(D=1 | FHD=f, MS=m)
1	1	0.97
1	0	0.90
0	1	0.08
0	0	0.04

- Here, $P(\text{ D} \mid \text{MS, FHD }) \neq P(\text{ D} \mid \text{FHD })$!
- Can be done using Belief Network,

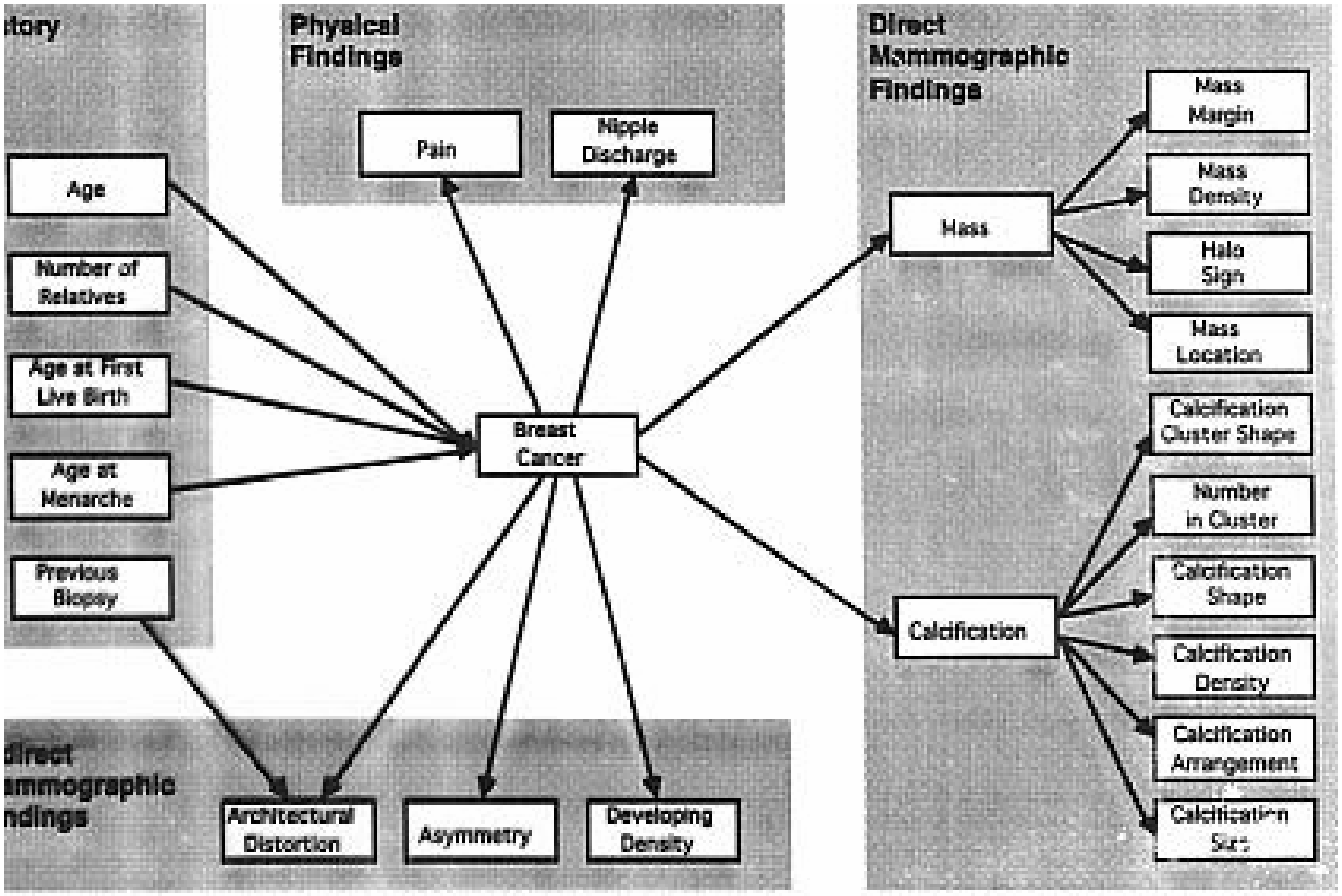
  but need to specify:

$P(FHD)$	1
$P(MS \mid FHD)$	2
$P(D \mid MS, FHD)$	4

Example: Car Diagnosis

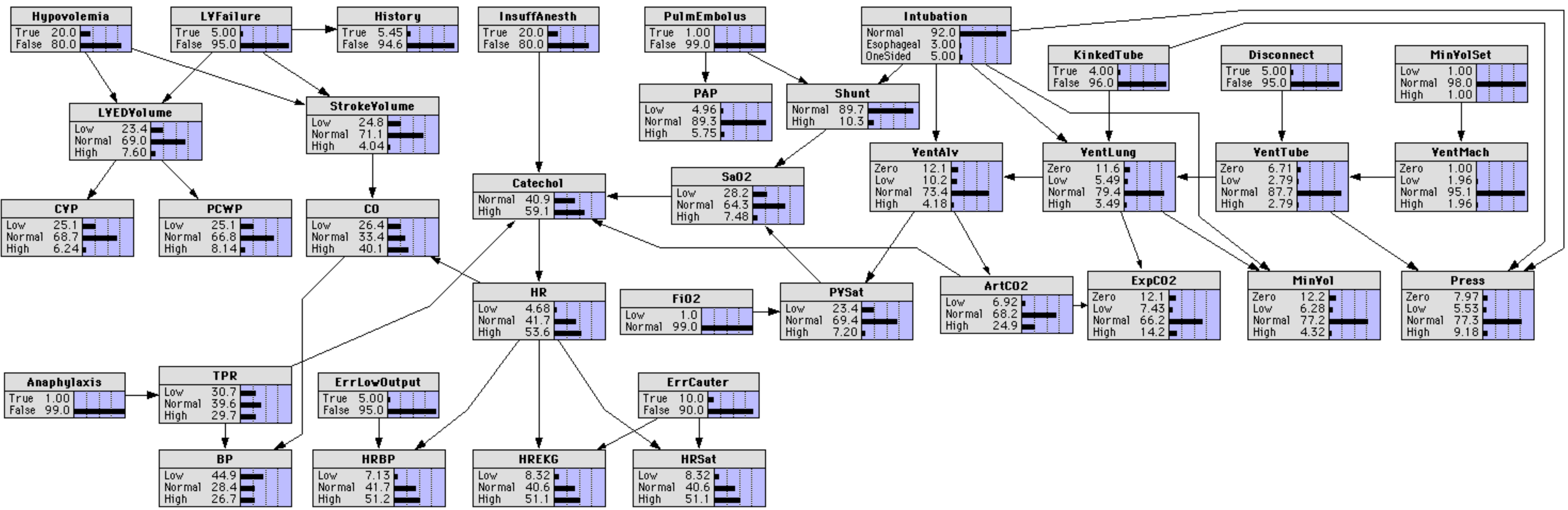


MammoNet



tory
Age
Number of Relatives
Age at First Live Birth
Age at Menarche
Previous Biopsy
Physical Findings
Pain
Nipple Discharge
Breast Cancer
Direct Mammographic Findings
Mass
Mass Margin
Mass Density
Halo Sign
Mass Location
Calcification
Calcification Cluster Shape
Number in Cluster
Calcification Shape
Calcification Density
Calcification Arrangement
direct ammographic ndings
Architectural Distortion
Asymmetry
Developing Density

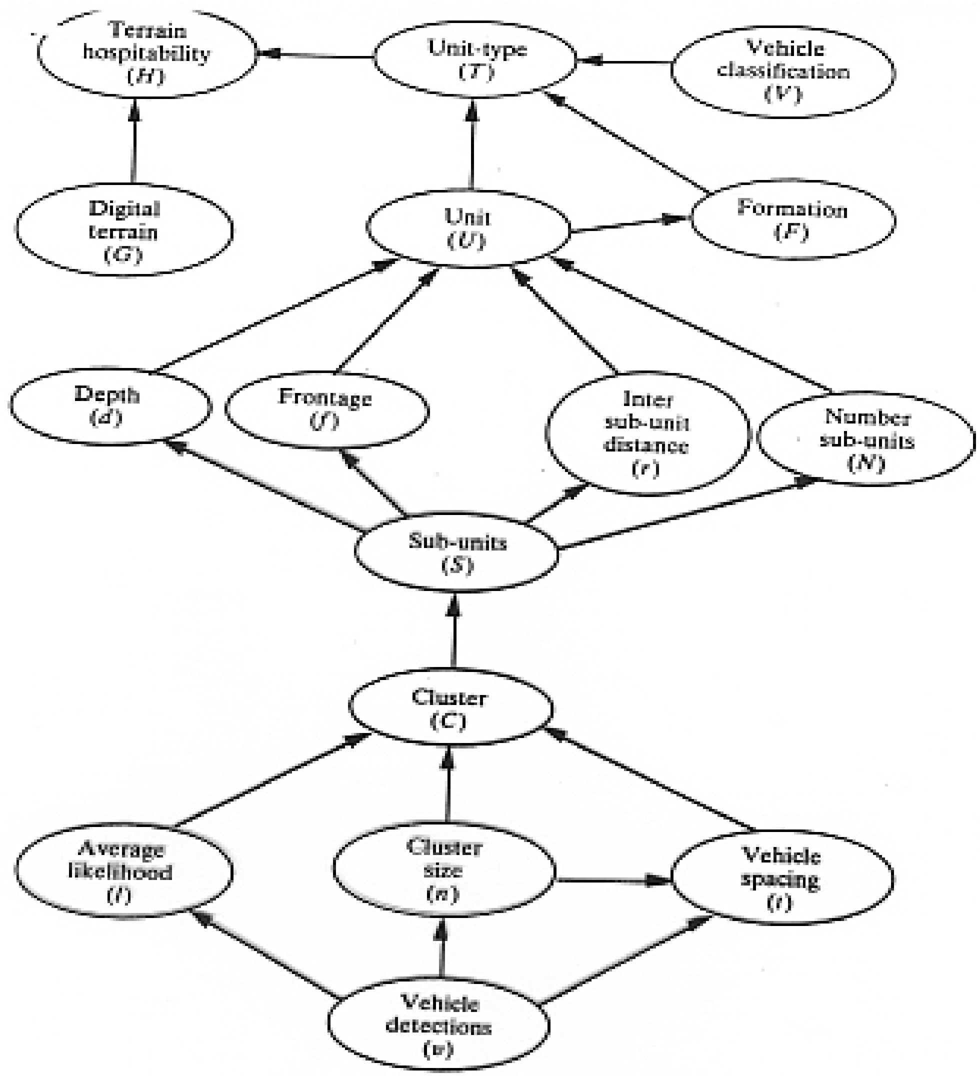
ALARM



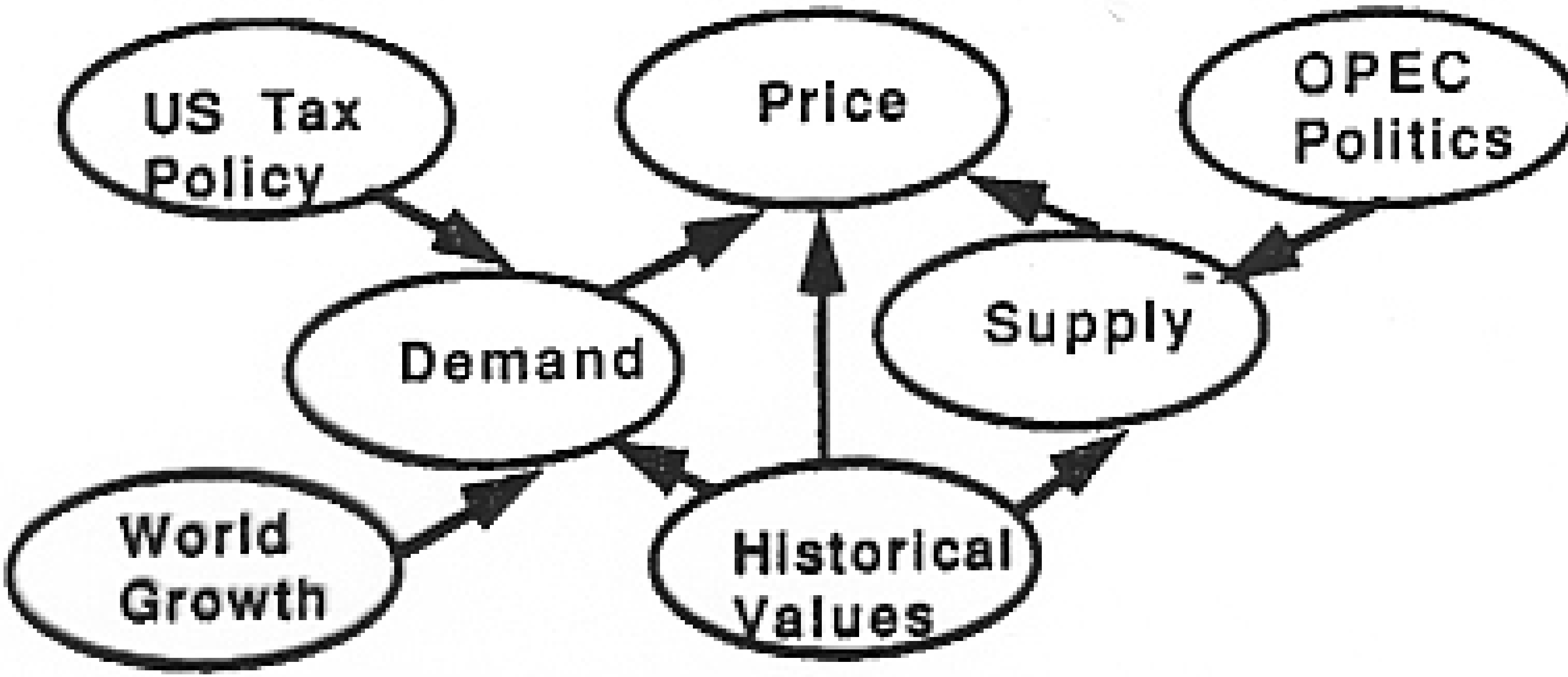
A Logical Alarm Reduction Mechanism

- 8 diagnoses, 16 findings, …

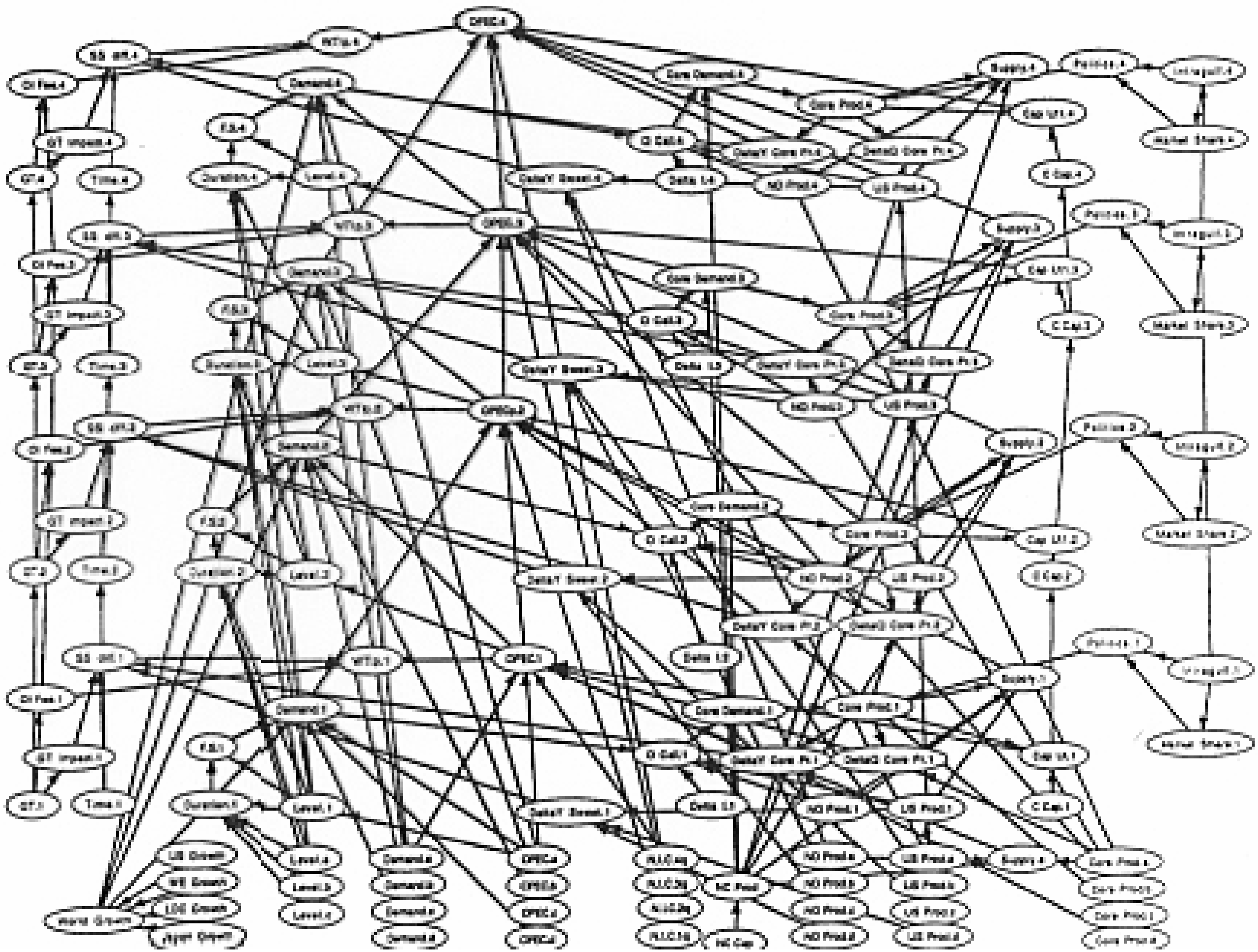
Troup Detection



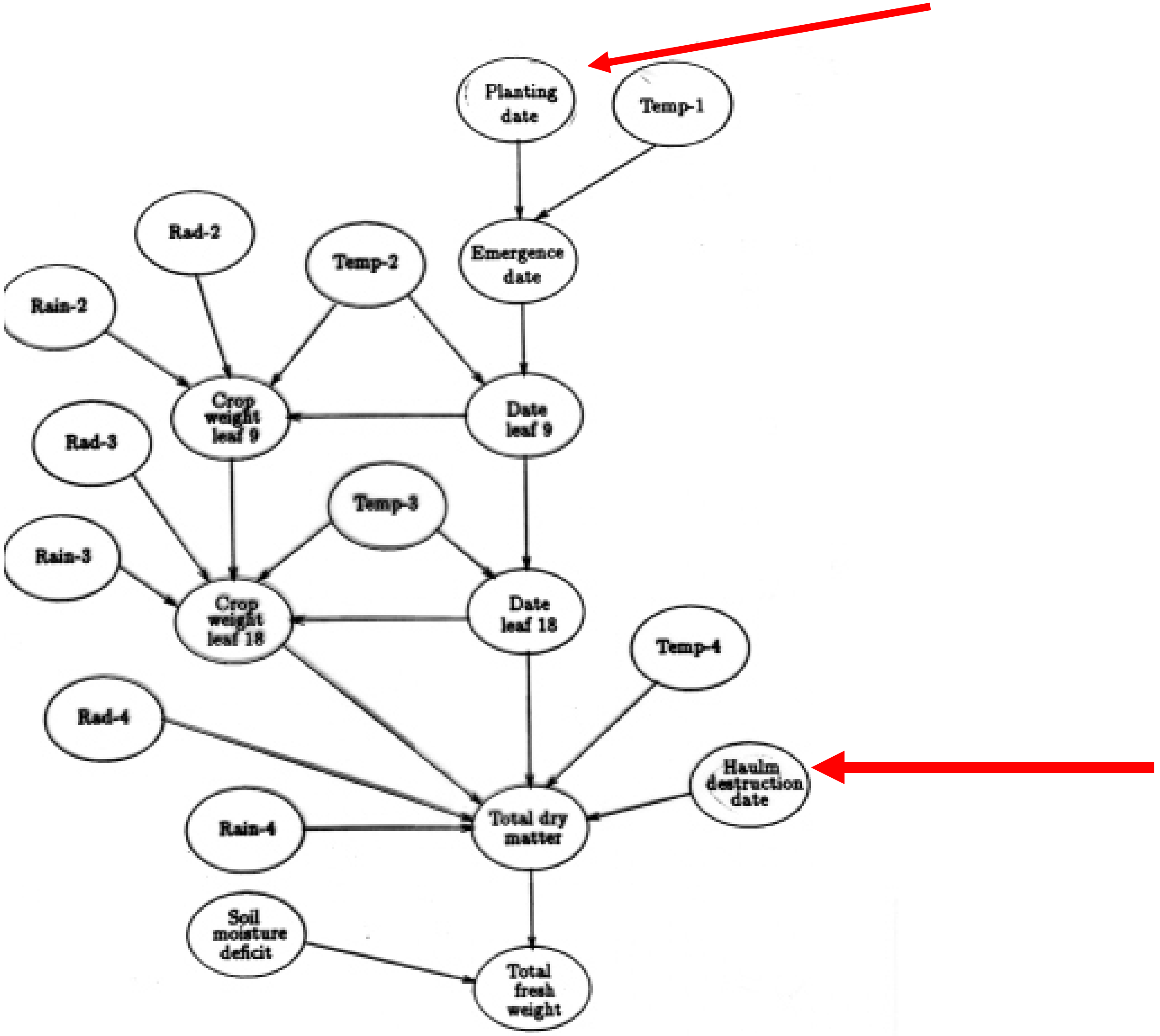
ARCO1: Forecasting Oil Prices



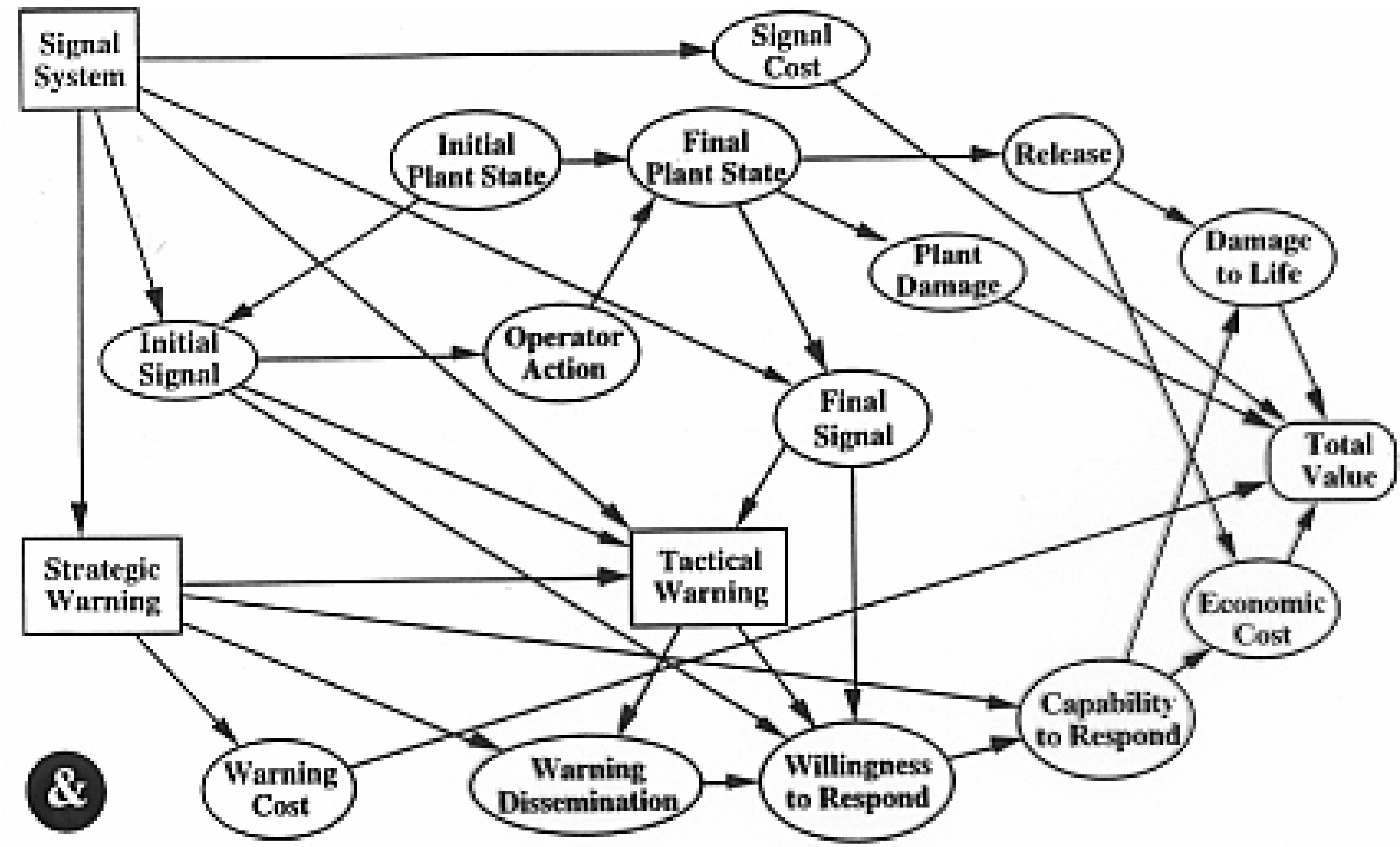
ARCO1: Forecasting Oil Prices



Forecasting Potato Production



Warning System



Extensions

- Find best values (posterior distr.) for
  SEVERAL (> 1) "output" variables
- *Partial* specification of "input" values
  - only subset of variables
  - only "distribution" of each input variable
- General Variables
  - Discrete, but domain > 2
  - Continuous (Gaussian: $x = \Sigma_i b_i y_i$ for parents $\{Y\}$)
- Decision Theory $\Rightarrow$ *Decision Nets* (Influence Diagrams)
  Making Decisions, not just assigning prob's
- Storing $P(v \mid p_1, p_2,\ldots,p_k)$
  General "CP Tables" $0(2^k)$
  Noisy-Or, Noisy-And, Noisy-Max
  "Decision Trees"

Outline

- Existing uses of Belief Nets (BNs)
- What is a BN ?
- Specific Examples of BNs
- Contrast with Rules, Neural Nets, ...
- Possible applications of BNs
- Challenges
  - How to reason efficiently
  - How to *learn* BNs

Belief Nets vs Rules

- Both have "*Locality*"
  Specific clusters (rules / connected nodes)
- Often *same nodes* (rep'ning Propositions) but

BN:	Cause	⇒	Effect	
	"Hep	⇒	Jaundice"	$P(J \mid H)$
Rule:	Effect	⇒	Cause	
	"Jaundice	⇒	Hep"	

*WHY?: Easier for people to reason **CAUSALLY***
*even if use is **DIAGNOSTIC***

- BN *provide OPTIMAL way to deal with*
  + ***Uncertainty***
  + ***Vagueness*** ***(var not given, or only dist)***
  + ***Error***

  …Signals meeting Symbols …

- BN *permits* different "direction"s of inference

Belief Nets vs Neural Nets

- Both have "*graph structure*" but

BN:	Nodes have SEMANTICs Combination Rules: Sound Probability
NN:	Nodes: arbitrary Combination Rules: Arbitrary

- So harder to
  - *Initialize NN*
  - *Explain NN*

  (But perhaps easier to learn NN from examples only?)

- BNs can deal with
  - *Partial Information*
  - *Different "direction"s of inference*

Belief Nets vs Markov Nets

- Each uses "*graph structure*"

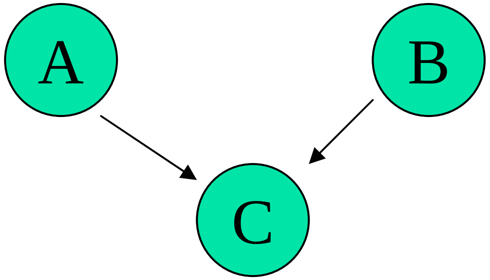
  to FACTOR a distribution

  ... explicitly specify dependencies, implicitly independencies...

- but subtle differences…
  - BNs capture "causality", "hierarchies"
  - MNs capture "temporality"

Technical: BNs use DIRECTRED arcs

$\Rightarrow$ allow "induced dependencies"

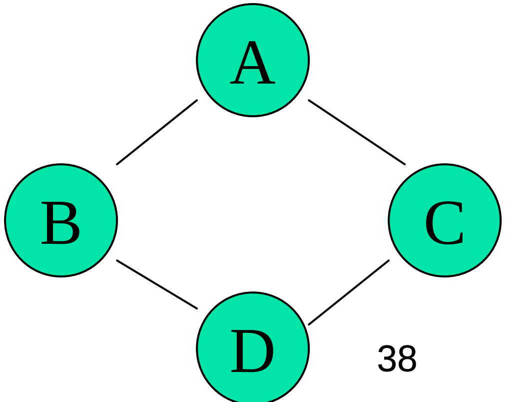


$I\,(A, \{\}, B)$ "*A* independent of *B*, given {}"

$\neg\, I\,(A, C, B)$ "*A* dependent on *B*, given *C*"

MNs use UNDIRECTED arcs

$\Rightarrow$ allow other independencies



$I(A, BC, D)$ *A* independent of *D*, given *B*, *C*

$I(B, AD, C)$ *B* independent of *C*, given *A*, *D*

Uses of Belief Nets #1

- Medical Diagnosis: "Assist/Critique" MD
  - identify diseases not ruled-out
  - specify additional tests to perform
  - suggest treatments appropriate/cost-effective
  - react to MD's proposed treatment

- Decision Support: Find/repair faults in complex machines

  [Device, or Manufacturing Plant, or …]

  … based on sensors, recorded info, history,…

- Preventative Maintenance:

  Anticipate problems in complex machines

  [Device, or Manufacturing Plant, or …]

  …based on sensors, statistics, recorded info, device history,…

Uses (con't)

- Logistics Support: Stock warehouses appropriately
  …based on (estimated) freq. of needs, costs,
- Diagnose Software:

  Find most probable bugs, given
  program behavior, core dump, source code, ...
- Part Inspection/Classification:
  … based on multiple sensors, background, model of production,…
- Information Retrieval:

  Combine information from various sources,
  based on info from various "agents",…

General: Partial Info, Sensor fusion
-Classification **-Interpretation**
-Prediction **-…**

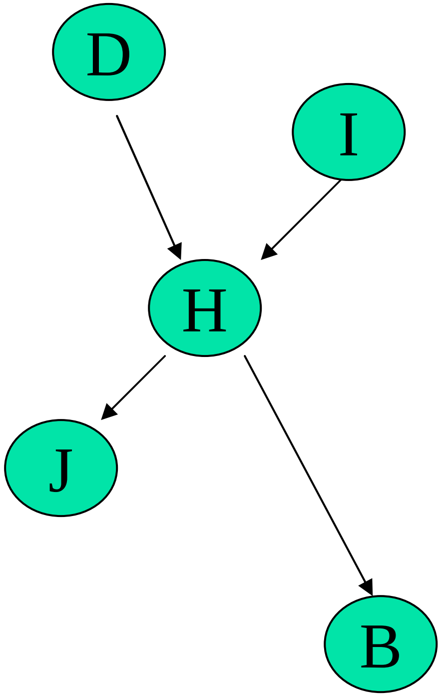
Challenge #1
Computational Efficiency

For given BN:

General problem is

Given $\boxed{O_1 = v_1, \ldots, O_n = v_n}$

Compute $\boxed{P(H \mid O_1 = v_1, \ldots, O_n = v_n)}$



+ If BN is "poly tree", ∃ efficient alg.

- If BN is gen'l DAG (>1 path from *X* to *Y*)
  - NP-hard in theory
  - slow in practice

Tricks: Get *approximate* answer (quickly)

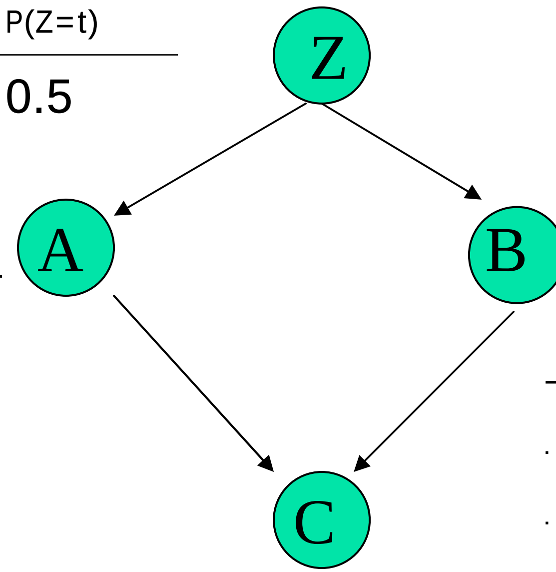
+ Use abstraction of BN
+ Use "abstraction" of query (range)

Why Reasoning is Hard

- BN reasoning may look easy:

  Just "propagate" information from node to node

P(Z=t)
0.5



z	P(A=t|Z=z)
t	1.0
f	0.0

z	P(B=t|Z=z)
t	0.0
f	1.0

a	b	P(C=t|a,b)
t	t	1.0
t	f	0.0
f	t	0.0
f	f	0.0

- Challenge: What is P(C=t) ?

$A = Z = \neg B \quad P(A = t) = P(B = f) = ½$

So… ? $P(C = t) = P(A = t, B = t)$

$= P(A = t) \times P(B = t) = ½ \times ½ = ¼$

- Wrong: $P(C = t) = 0$!

Need to maintain dependencies! $P(A = t, B = t) = P(A = t) * P(B = t \mid A = t)$

2a:Obtaining Accurate BN

✓ BN encodes distribution over *n* variables

Not $O(2^n)$ *values, but "only"* $\sum_i 2^{k\_i}$

(Node n_i binary, with k_i parents)

Still *lots* of values! …structure ..

⇒ ***Qualitative Information***

Structure: "What depends on what?"

- Easy *for people (background knowledge)*
- *But NP-hard to learn from samples…*

Knowledge acquisition: from human experts

⇒ ***Quantitative Information***

Actual CP-tables

- *Easy* to learn, given lots of examples.
- But people have hard time…

Simple learning algorithm

Notes on Learning

- **Mixed Sources:** Person provides structure; Algorithm fills-in numbers.

- **Just Learning Algorithm:** ∃ algorithms that

$$\text{learn} \left\{ \begin{array}{c} \text{structure} \\ \text{values} \end{array} \right\} \text{from sample}$$

- **Just Human Expert:** People produce CP-table, as well as structure

  Relatively few values really required

  Esp. if NoisyOr, NoisyAnd, NaiveBayes, …

  Actual values not *that* important

  …Sensitivity studies

2b: Maintaining Accurate BN

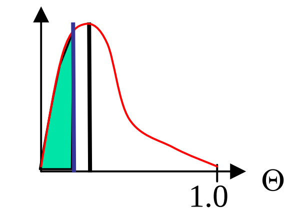
- The world changes.

  Information in BN^* may be
  - perfect at time t
  - sub-optimal at time t + 20
  - worthless at time t + 200
- Need to *MAINTAIN* a BN over time

  using *on-going* human consultant
- Adaptive BN
  - Dirichlet distribution (variables)
  - Priors over BNs

My Results Related to Belief Nets

- Quantifying Uncertainty in BN Response



  - $\Pr_{\Theta}($ C=true | D = false $) = 0.3 \pm 0.05$
  - Uses: Good Decision, Bad Outcome Bias^2+Variance; Mixture using Variance
- Learning Structure – Generatively
  - BDe, 2-foldCV work well (not MDL)
- Learning Structure – Discriminatively
  - Bias^2+Variance works well (not MDL)
- Learning Parameters – Discriminately
  - NaïveBayes : Logistic Regression :: Belief Nets : ***ELR***

Conclusions

- Belief Nets *are PROVEN TECHNOLOGY*
  - Medical Diagnosis
  - DSS for complex machines
  - Forecasting, Modeling, InfoRetrieval…
- Provide effective way to
  - Represent complicated, inter-related events
  - Reason about such situations
    - Diagnosis, Explanation, ValueOfInfo
    - Explain conclusions
    - Mix Symbolic and Numeric observations
- Challenges
  - Efficient ways to use BNs
  - How to create accurate/effective BNs
  - How to maintain BNs
  - Reason about time…