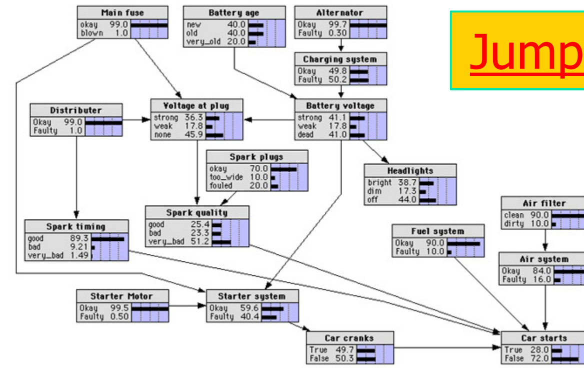
Cmput 466 / 551



Jump

Introduction to Bayesian Belief Nets

Readings: ≈HTF 17

\+ Bayesian Networks without the Tears (Charniak)

R Greiner

University of Alberta

BUSINESS

MONDAY TECHNOLOGY SPECIAL

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

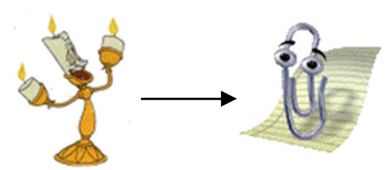
LATimes, 28/Oct/96

Motivation

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

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

- *Microsoft Products*



  - *Answer Wizard (Office, …)*
  - *Microsoft Pregnancy and Child Care (MSN)*
  - *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

Outline

Jump

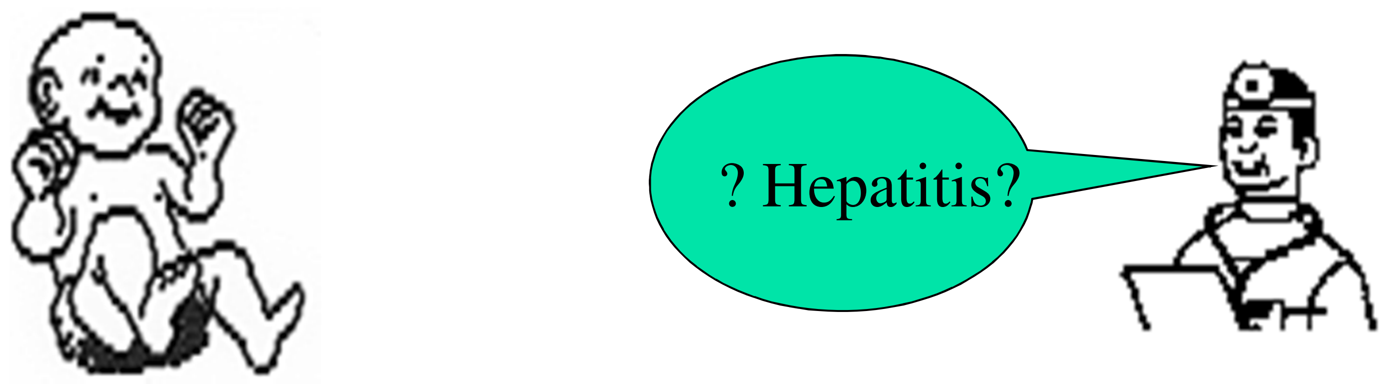
- Motivation
- What is a Belief Net?
  - Factored Distribution
    - ... use *connections*... just *some* connections
    - (Conditional) Independence
  - Semantics
    - Engineering a structure
- Inference (Reasoning)
- Applications
- Relation to other Models
- Learning a Belief Net

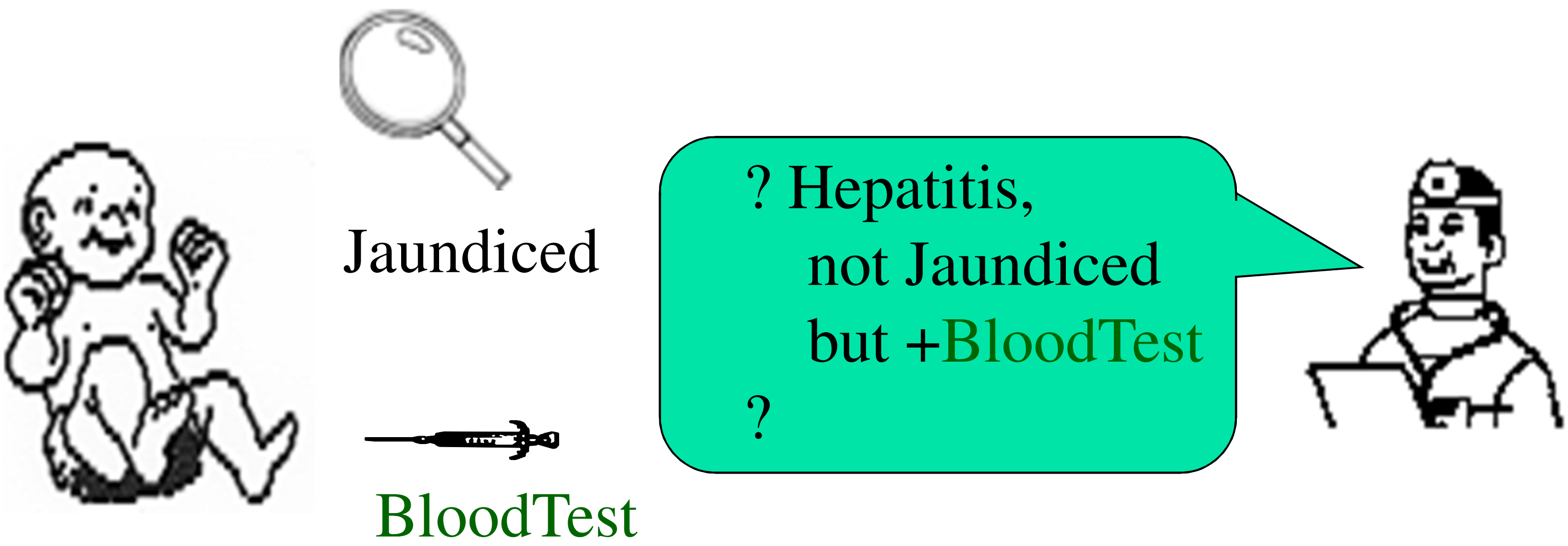
Assume everyone knows …

- Random Variable
  - Domain(X)
- Probability
  - P(X=x): Probability that x holds
  - P(x | y): Probability that X=x holds, given that Y=y holds
  - Joint, Marginal, … Distribution
  - (Conditional) Independence
- Graphs
  - Node, (Directed) Arc, …

- Complexity … exponential, NP-hard, …

Repeat





What is P(+h | - j, +b) ?

Inference by Enumeration

Repeat

- Using only joint probability distribution:

H	Hepatitis
J	Jaundice
B	(positive) Blood test

J	H	B	P(j,b,h)
0	0	0	0.03395
0	0	1	0.0095
0	1	0	0.0003
0	1	1	0.1805
1	0	0	0.01455
1	0	1	0.038
1	1	0	0.00045
1	1	1	0.722

- Can compute *conditional probabilities*:

$$P(-h \mid +j) = \frac{P(-h \wedge +j)}{P(+j)} = \frac{0.01455 + 0.038}{0.01455 + 0.038 + 0.00045 + 0.722}$$

$$\approx 0.0678$$

Just use Table ?

- Just need single table!! But…
- Unnatural:
  - Easier to think about CORRELATIONS
    - P(Jaudice | Hepatitis)
    - P(DimLight | BadBattery), …

  ⇒ better to use **CONDITIONAL** EVENTS
- Too MANY NUMBERS!!
  - Exponential size to store
    $O(2^N)$ numbers…
  - Exponential cost for inference

  ⇒ only use **some** connections

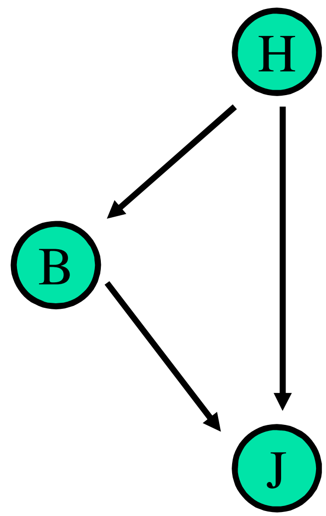
⇒ **Bayesian Belief Net**

J	B	H	P(j,b,h)
0	0	0	0.03395
0	0	1	0.0095
0	1	0	0.0003
0	1	1	0.1805
1	0	0	0.01455
1	0	1	0.038
1	1	0	0.00045
1	1	1	0.722

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



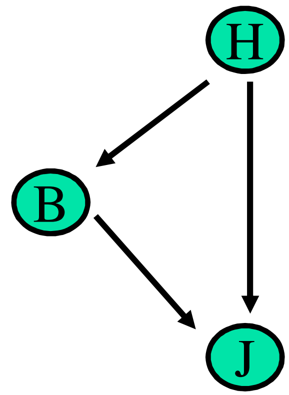
h	b	P(J=1|h,b)	P(J=0|h,b)
1	1	0.8	0.2
1	0	0.8	0.2
0	1	0.3	0.7
0	0	0.3	0.7

- Node ~ Variable
  Link ~ “Causal dependency”
- “CPTable” ~ P(child | parents)

Encoding Direct 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

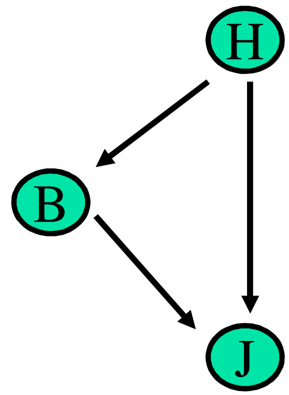
- $P(J \mid H, B{=}0) \;=\; P(J \mid H, B{=}1) \quad \forall\; J,\; H\,!$

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

Encoding Direct Links

P(H=1)
0.05

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



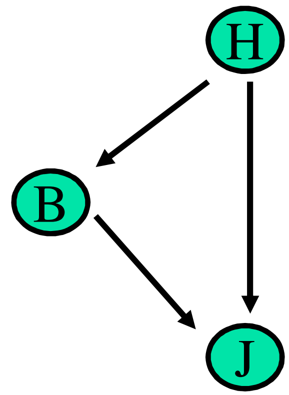
h		P(J=1|h)
1		0.8
1		
0		0.3
0		

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

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

Encoding Direct Links

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



P(H=1)
0.05

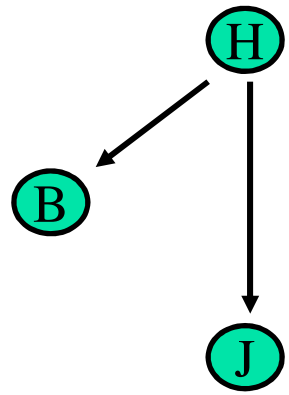
h	P(J=1|h)
1	0.8
0	0.3

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

  $\Rightarrow \quad \mathbf{P(\,J \mid H, B) \;=\; P(J \mid H)}$
- 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: $$P(\mathrm{H}=1 \mid \mathrm{B}=1, \mathrm{J}=0) = \frac{1}{\alpha} P(\mathrm{H}=1)\, P(\mathrm{B}=1 \mid \mathrm{H}=1)\ \overset{P(\mathrm{J}=0 \mid \mathrm{H}=1)}{\cancel{P(\mathrm{J}=0 \mid \mathrm{B}=1, \mathrm{H}=1)}}$$

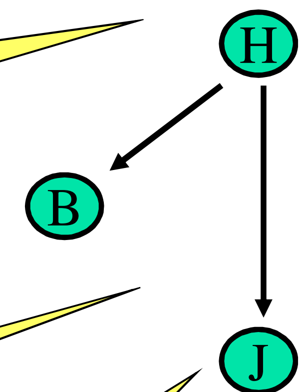
What is probability that Fred is …
Jaundiced, given {}?

$P(+j) = 0.325$

… Jaundiced, given –BloodTest ?

$P(+j \mid -b) = 0.301$

… Jaundiced, given +Hepatitis?



$P(+j \mid +h) = 0.8$

… Jaundiced, given +Hepatitis,
and also -BloodTest ?

Same: $P(+j \mid +h, -b) = 0.8$

So Jaundice DOES depend on BloodTest, initially

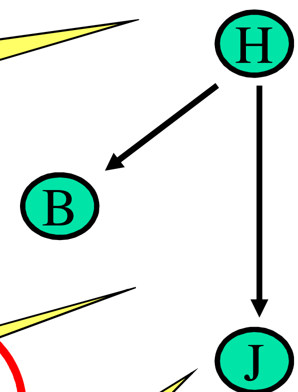
$P(+j) = 0.325$

$P(+j \mid -b) = 0.301$

But Jaundice does NOT depend on BloodTest, given Hepatitis

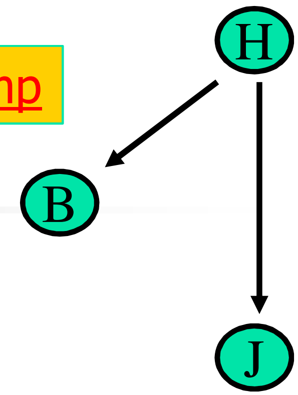
$P(+j \mid +h) = 0.8$

Same: $P(+j \mid +h, -b) = 0.8$



Dependencies…

Jump



- B *does* depend on J:

  If J=1, then likely that H=1 ⇒ 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 $P(J \mid B) \neq P(J)$
GIVEN H, *they become uncorrelated* $P(J \mid B, H) = P(J \mid H)$

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

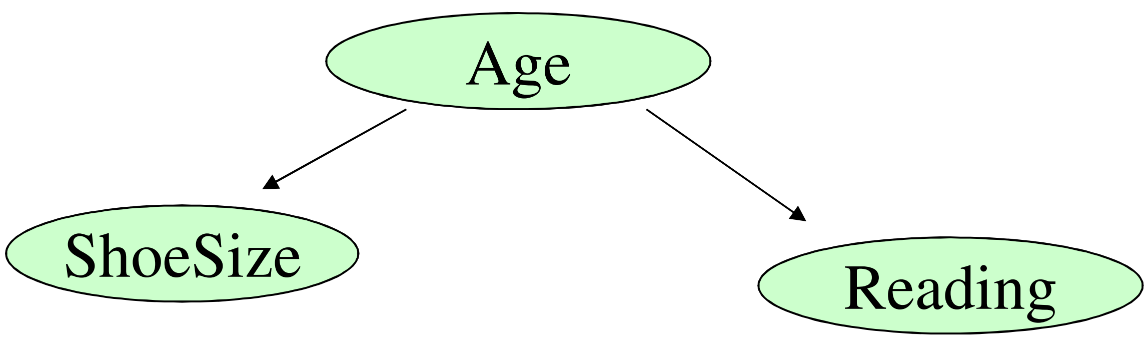
$P(\text{ B} \mid \text{J,H })\ =\ P(\text{ B} \mid \text{H })$

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

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

but become independent, given Age

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



(a) Independence

- Coin tosses:
  - T_1: the first toss is a head; T_2: the second toss is a tail
  - $P(T_2 \mid T_1) = P(T_2)$

- α and β ***independent*** iff $P(\beta|\alpha)=P(\beta)$
  - $\mathbf{P} \models (\alpha \perp \beta)$
  - … distr'n **P** entails α independent of β

- **Definition:** α and β *independent*
  if and only if
  $\mathbf{P}(\alpha,\beta) = \mathbf{P}(\alpha)\ \mathbf{P}(\beta)$

(a) Independence

- Events α and B are independent *iff*
  - $P(\alpha, \beta) = P(\alpha) P(\beta)$
  - $P(\alpha \mid \beta) = P(\alpha)$
  - $P(\alpha \vee \beta) = 1 - (1 - P(\alpha))(1 - P(\beta))$

- Variables independent
  $\Leftrightarrow$ independent for all values

$\forall a, b \quad P(A = a, B = b) = P(A = a)\ P(B = b)$

(b) Conditional independence

- *Unconditional* independence is rare,
  but *conditional* independence is common…
  - Shoe size is NOT independent of Reading Ability
  - But is independent, given AGE…

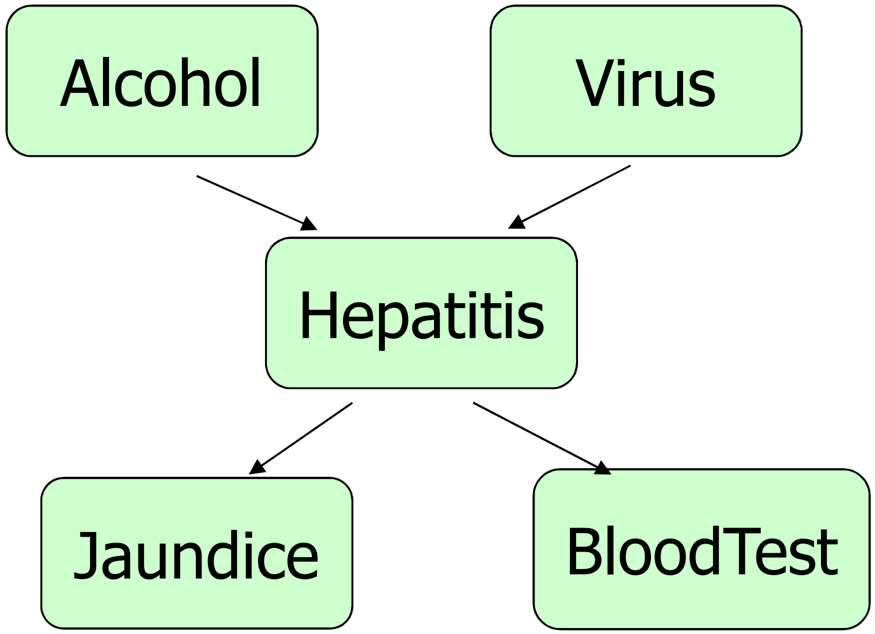
- α and β ***conditionally independent*** given γ
  iff
  given γ, knowing α does not change the probability of β

  - $P(\beta \mid \alpha, \gamma) = P(\beta \mid \gamma)$
  - $P(\alpha \mid \beta, \gamma) = P(\alpha \mid \gamma)$
  - $P(\alpha, \beta \mid \gamma) = P(\alpha \mid \gamma)\ P(\beta \mid \gamma)$

Bigger Networks

P(A=1)
0.32

P(V=1)
0.20



a	v	P(H=1|a ,v)
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:
  Alcohol CAUSES Hepatitis; Hepatitis CAUSES Jaundice
- If Alcohol, then expect Jaundice: $\boxed{\text{Alcohol} \Rightarrow \text{Hepatitis} \Rightarrow \text{Jaundice}}$

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

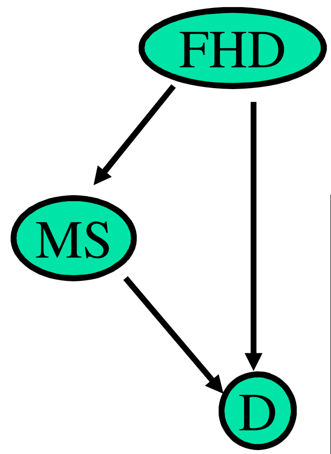
$$P(\,J \mid A\,) \neq P(\,J\,) \quad \text{but}$$
$$P(\,J \mid A, H\,) = P(\,J \mid H)$$

Less Trivial Situation

- N.b., obs_1 is *not* always independent of obs_2 given H
  - 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(D | MS, FHD) ≠ P(D | FHD)
- Can be done using Belief Network… but need to specify

P(FHD) 1
P(MS | FHD) 2
P(D | MS, FHD) 4

Advantages of Belief Net

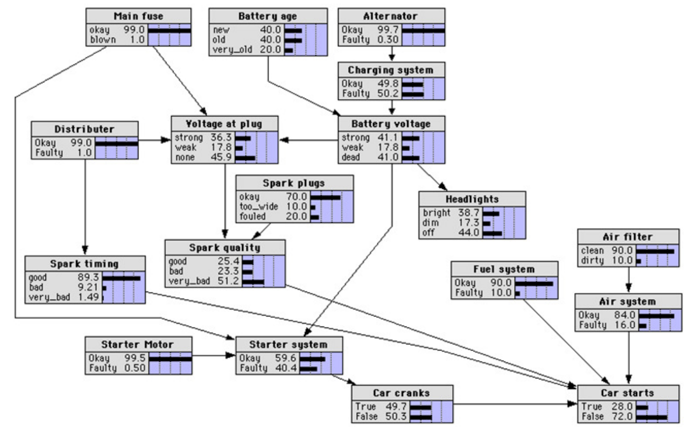
- All of advantages of *Probability Theory*
  - Not CertaintyFactor, Fuzzy, Dempster-Schaeffer, …
  - Formal understanding of how things relate
  - Well-defined inference
- Explanatory power
  - What is related to what? … and how strongly?
- Efficient encoding
  - 5 nodes $\Rightarrow 2^5 = 32$ values ... But needed only 10 values
  - 422 node "CPCS Network" $\Rightarrow 2^{422}$
    - Modeling disease/symptom for internal medicine

    Needed only 8,254 values
- Effective learning…

What to do with a Belief Net?

Jump

- Examine its connections
  - What depends on what?



Semantics

- Get answers to specific questions
  - What is $P(\text{Cancer} \mid G_3{=}+, \text{Age}{>}52)$?
  - What is most likely cause of symptoms?

Inference

Outline

- Motivation
- What is a Belief Net?
  - Factored Distribution
  - Semantics
    - Engineering a structure
- Inference
- Applications
- Relation to other Models
- Learning a Belief Net

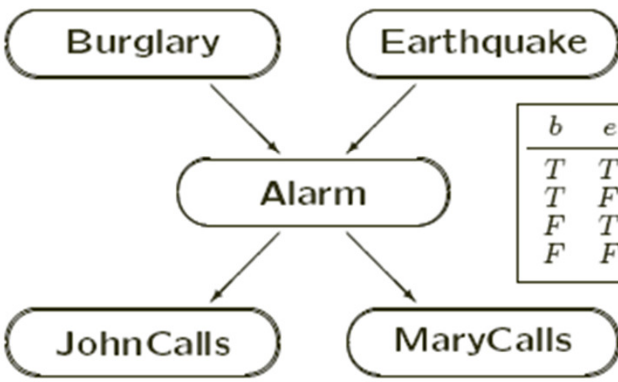
Components of a Bayesian Net

Directed Acyclic Graph:

$$\mathcal{BN} = \left\{ \begin{array}{lll} \mathcal{N} & \text{Nodes} & \equiv \text{Variables} \\ \mathcal{A} & \text{Arcs} & \equiv \text{Dependencies} \\ \mathcal{C} & \text{CPTables} & \equiv \text{“weights”} \end{array} \right\}$$

$P(B)$
0.001

$P(E)$
0.002



b	e	$P(A \mid B=b, E=e)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

a	$P(J \mid A=a)$
T	0.90
F	0.05

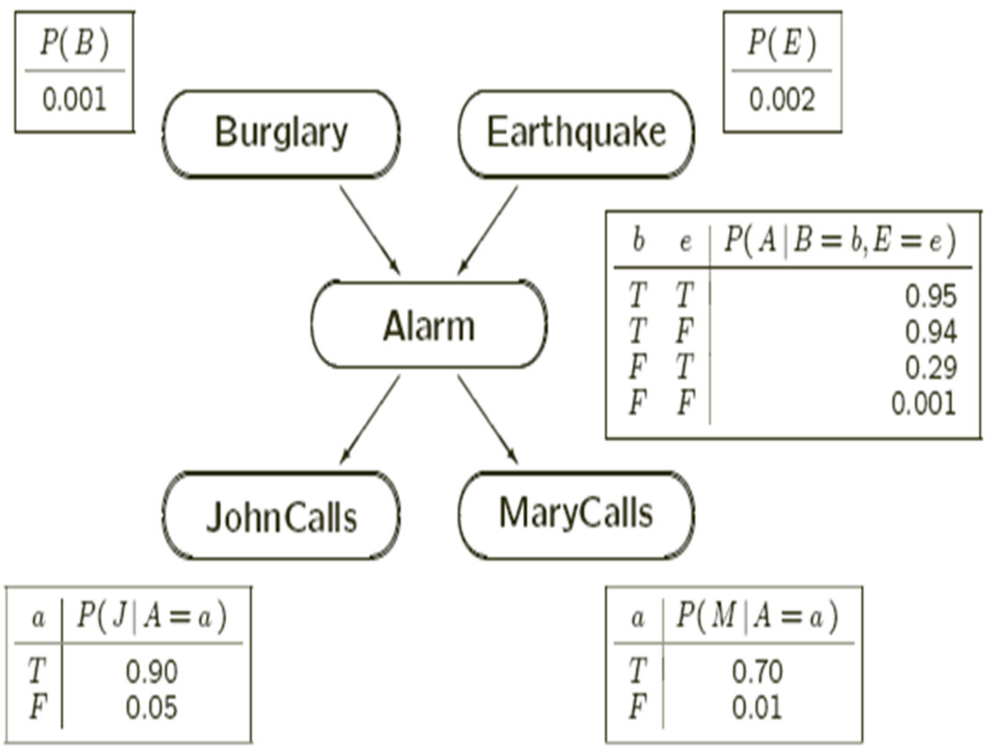
a	$P(M \mid A=a)$
T	0.70
F	0.01

- **Nodes**: one for each random variable
- **Arcs**: one for each direct influence between two random variables
- **CPT**: each node stores a conditional probability table

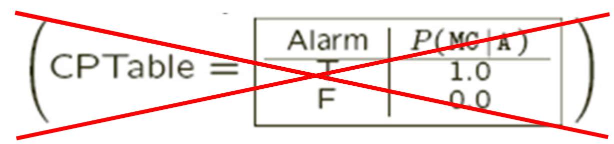
  P(Node | Parents(Node))

  to quantify effects of “parents” on child

Causes, and Bayesian Net

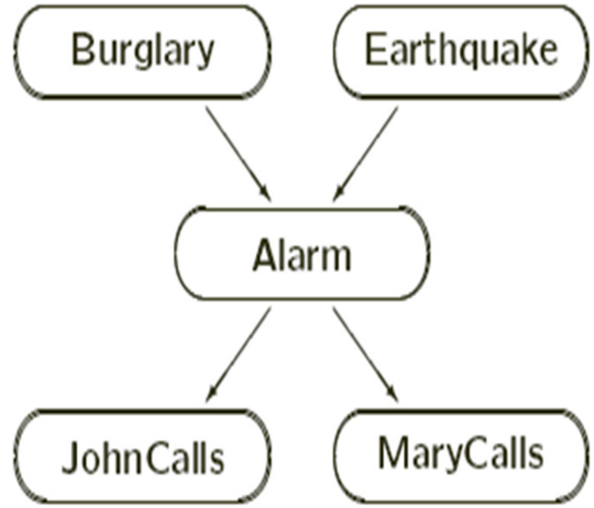


- What "causes" Alarm?
  A: Burglary, Earthquake
- What "causes" JohnCall?
  A: Alarm
  N.b., NOT Burglary, ...
- Why not Alarm $\Rightarrow$ MaryCalls?



A: Mary not always home
... phone may be broken
...

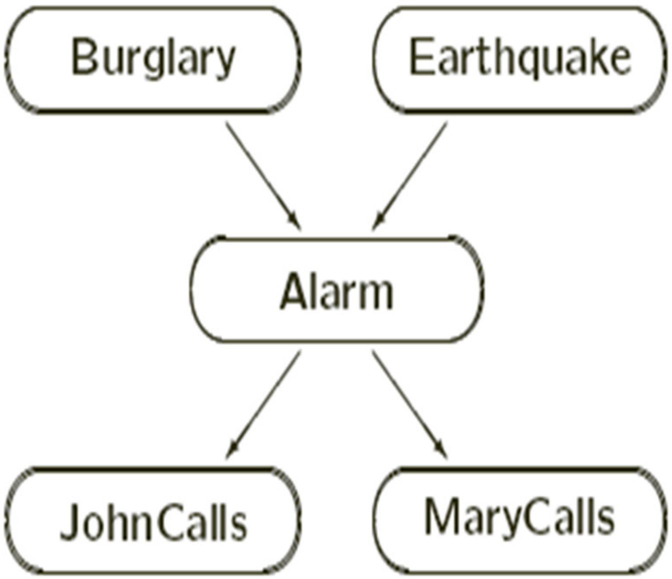
Independence in a Belief Net



- Burglary, Earthquake independent
  - $B \perp E$
- Given Alarm, JohnCalls and MaryCalls independent
  - $J \perp M \mid A$
  - JohnCalls is correlated with MaryCalls $\neg(J \perp M)$

    as JohnCalls suggest Alarm which suggests MaryCalls
  - But given Alarm, JohnCalls gives no NEW evidence wrt MaryCalls

The Independence Assumption

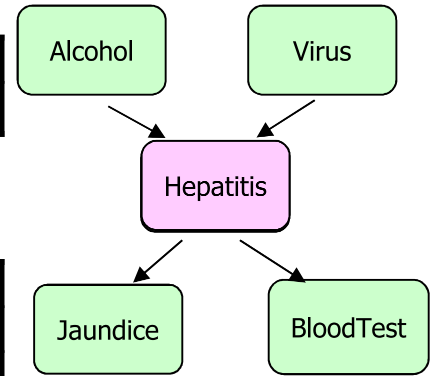


Local Markov Assumption:
A variable X is independent of its non-descendants given its parents
($X_i \perp NonDescendants_{Xi} \mid Pa_{Xi}$)

- $B \perp E \mid \{\}$ $\qquad (B \perp E)$
- $M \perp \{ B, E, J \} \mid A$
- Given graph G,
  $I_{LM}(G) = \{ (X_i \perp \mathbf{NonDescendants}_{Xi} \mid Pa_{Xi}) \}$

Belief Nets

P(A=1)
0.32



P(V=1)
0.20

a	v	P(H=1|a ,v)
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

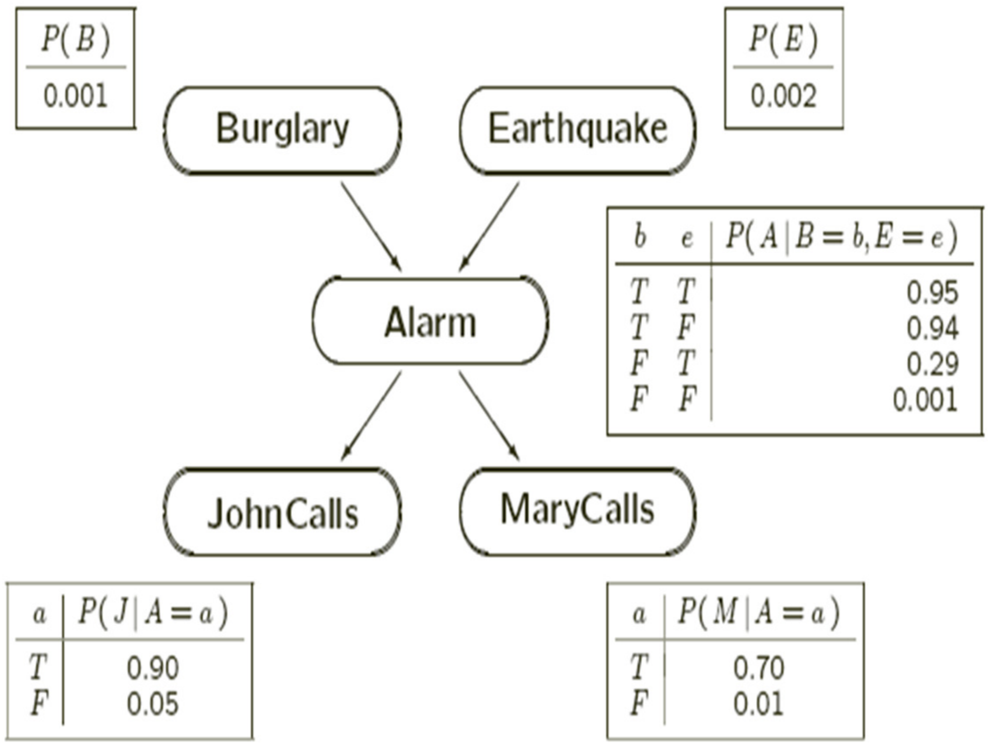
- DAG structure
  - Each node ≡ Variable *v*
  - *v* depends (only) on its parents

  + 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 = true,
  - A has no influence on J
  - J has no influence on B
  - etc.

What about probabilities? Conditional probability tables (CPTs)



- Each CPTable is called a “Factor”

Factoid…

- $P(A,B,C) = P(A \mid B,C)\ \boxed{P(B,C)}$

  $= P(A \mid B,C)\ \boxed{P(B|C)\ P(C)}$
- In general:

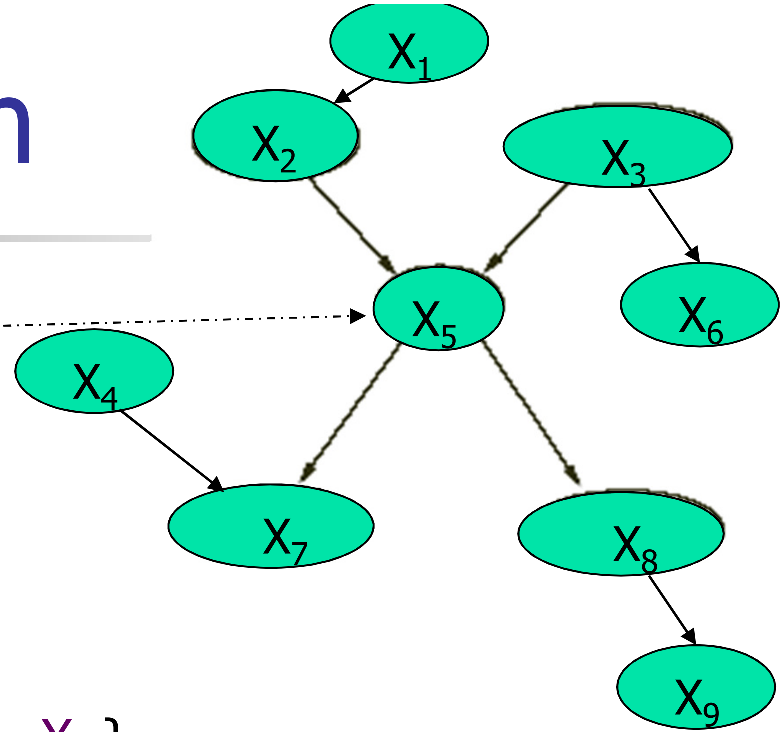
  $P(X_m, X_{m-1}, \ldots, X_1) =$

  $P(X_m \mid X_{m-1}, \ldots, X_1)\ P(X_{m-1}, \ldots, X_1) =$

  $P(X_m \mid X_{m-1}, \ldots, X_1)\ P(X_{m-1}, | X_{m-2} \ldots, X_1)\ P(X_{m-2}, \ldots, X_1)$

  $= \prod_i P(X_i \mid X_{i-1}, \ldots, X_1)$
- … or other order: eg, from 1..m, or …

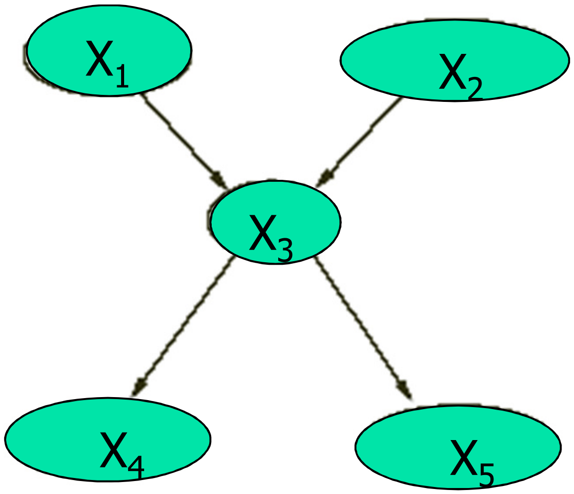
Joint Distribution



- Wrt X_5
  - Parents(X_5) = { X_2, X_3 }
    - Note 2,3 < 5
    - So Parents(X_5) ⊆ $\{ X_1 , \ldots , X_4 \}$
  - Descendants(X_5) = { X_7, X_8 , X_9 }
    - Note 7,8,9 > 5
  - Non-Descendants(X_5) = { X_1, …, X_4 , X_6 }
    - Note $\{ X_1 , \ldots , X_4 \}$ ⊆ Non-Descendants(X_5)
- Parents(X_5) ⊆ $\{ X_1 , \ldots , X_4 \}$ ⊆ Non-Descendants(X_5)
  - As node independent of non-descendants, given parents:
  - $P(X_5 | X_1, \ldots, X_{5-1}) = P(X_i | \text{Parents}(X_i))$
- In general

$$P(X_i | X_1, \ldots, X_{i-1}) = P(X_i | \text{Parents}(X_i))$$

Joint Distribution



- In gen'l,

$$P(X_1, X_2, \ldots, X_m) = \prod_i P(X_i \mid X_1, \ldots, X_{i-1})$$

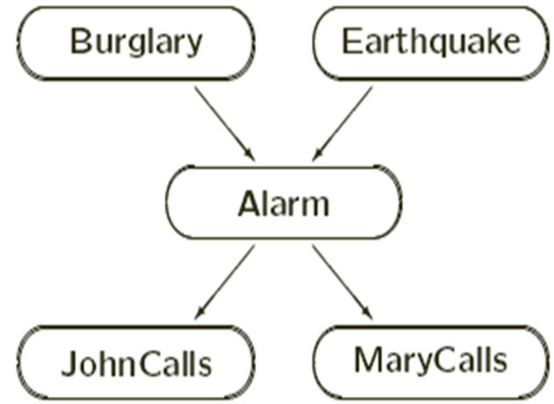
- Local Markov Assumption means…

$$P(X_i \mid X_1, \ldots, X_{i-1}) \quad = \quad P(X_i \mid \text{Parents}(X_i))$$

  Node independent of other non-descendants, given parents

- So... $P(X_1, X_2, \ldots, X_m) \quad = \quad \prod_i P(X_i \mid \text{Parents}(X_i))$

Joint Distribution



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

$$P(+j, +m, +a, -b, -e)$$

$$= \cancel{P(+j \mid +m, +a, -b, -e)} \xrightarrow{J \perp \{M,B,E\} \mid A} P(+j \mid +a)$$

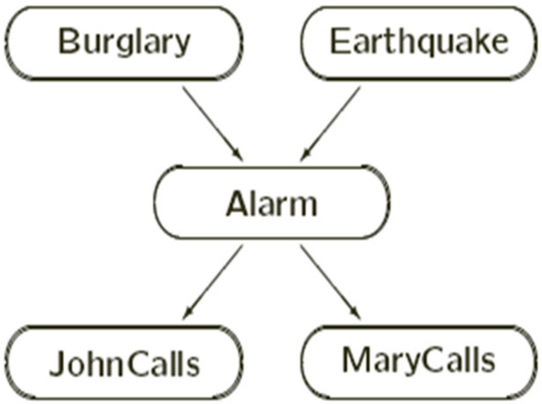
$$\cancel{P(+m \mid +a, -b, -e)} \xrightarrow{M \perp \{B,E\} \mid A} P(+m \mid +a)$$

$$\cancel{P(+a \mid -b, -e)} \longrightarrow P(+a \mid -b, -e)$$

$$\cancel{P(-b \mid -e)} \xrightarrow{B \perp E} P(-b)$$

$$\cancel{P(-e)} \longrightarrow P(-e)$$

Joint Distribution



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

P(+j, +m, +a, -b, -e)

= P(+j | +a)

P(+m | +a)

P(+a | -b, -e)

P(-b)

P(-e)

Recovering Joint

[Jump]

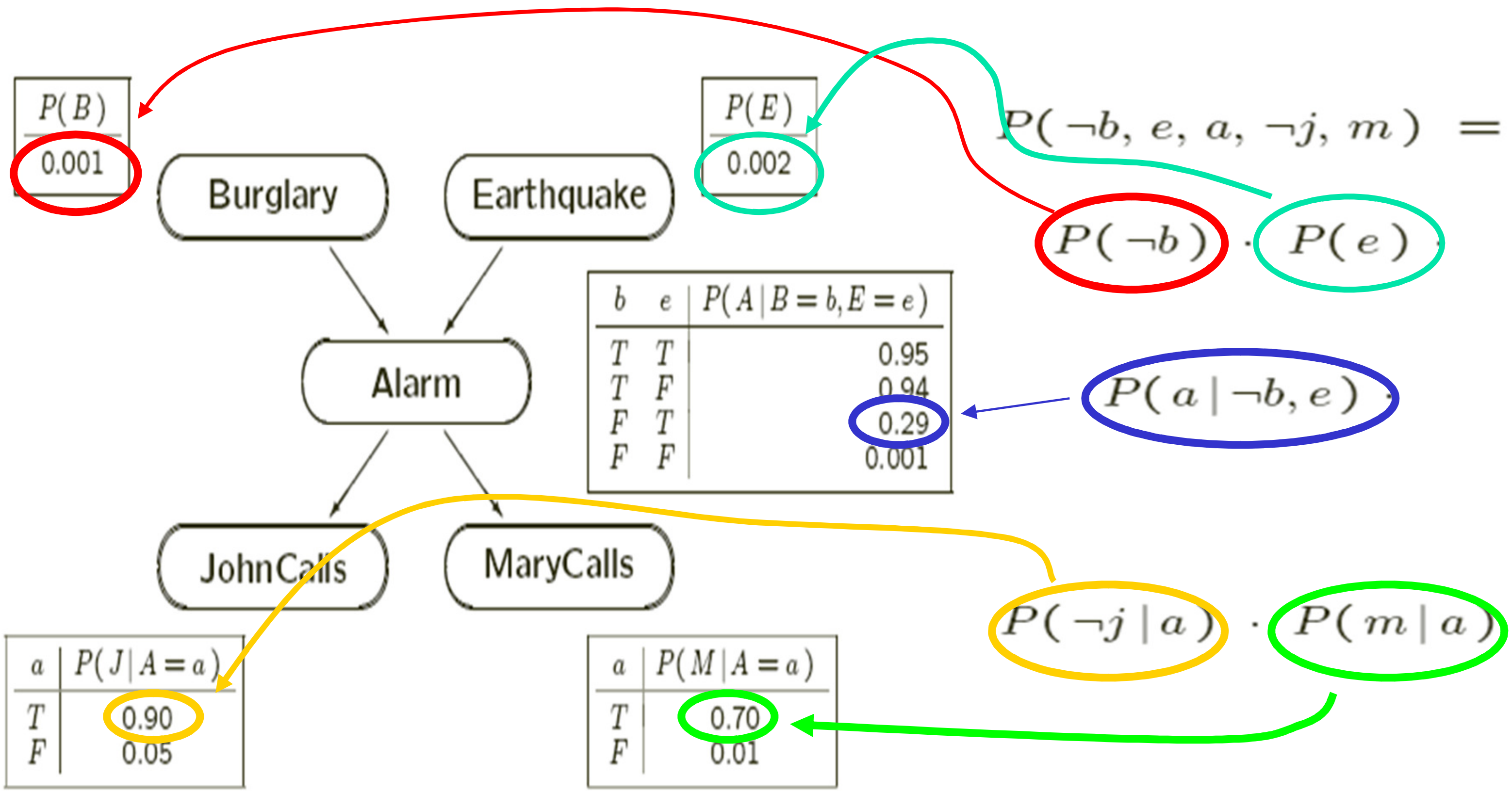
$$P(\neg b, e, a, \neg j, m) =$$

$$P(\neg b)\ P(e \mid \neg b)\ P(a \mid e, \neg b)\ P(\neg j \mid a, e, \neg b)\ P(m \mid \neg j, a, e, \neg b)$$

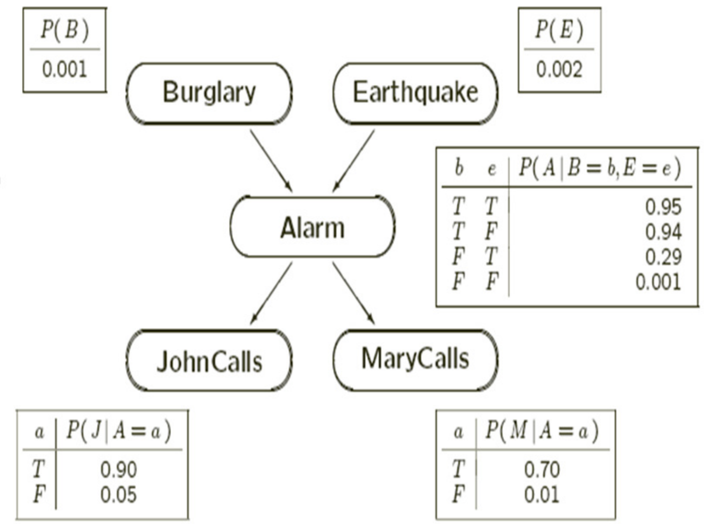
$$P(\neg b)\ P(e)\ P(a \mid e, \neg b)\ P(\neg j \mid a)\ P(m \mid a)$$

$$0.99 \times 0.02 \times 0.29 \times 0.1 \times 0.70$$

Node independent of predecessors, given parents



Comments

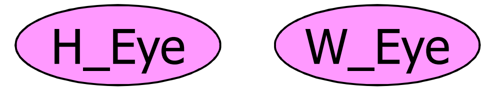


- BN used 10 entries
  ... can recover full joint
  (2^5 entries)
  (Given structure,
  other 2^5 – 10 entries are REDUNDANT)

$\Rightarrow$ Can compute
  P(+burglary | +johnCalls, -maryCalls) :
  Get joint, then marginalize, conditionalize, ...
  $\exists$ ***better ways*** ...

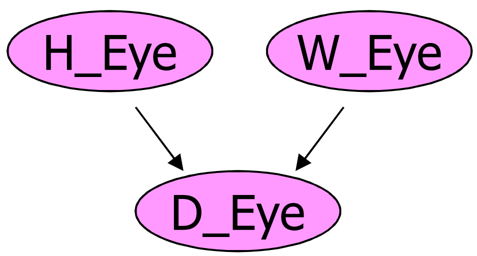
- Note: Given structure, ANY CPT is consistent.
  $\nexists$ redundancies in BN...

"V"-Connections



- What color are my wife's eyes?
- Would it help to know MY eye color?
  NO! H_Eye and W_Eye are independent!

- We have a DAUGHTER, who has BROWN eyes
  Now do you want to know my eye_color?



h	w	P(D= bl | h , w)
bl	bl	1.0
bl	br	0.5
br	bl	0.5
br	br	0.25

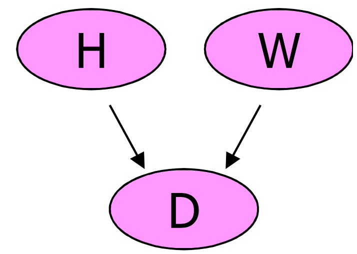
- H_Eye and W_Eye became dependent!

What color is W?

Prior is $P(W = br) = 0.8$?

But I know H!
Should I tell you?

Don't bother; it doesn't matter
$P(W = br \mid H= bl) = 0.8$
$P(W = br \mid H= br) = 0.8$

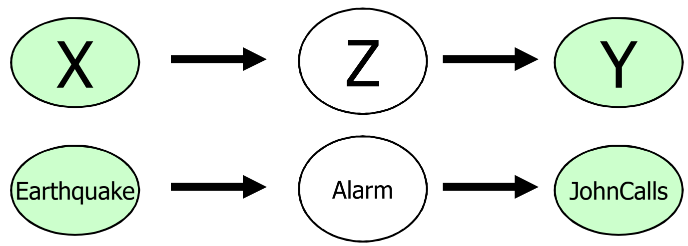


I also know $D = br$. Now do you care?

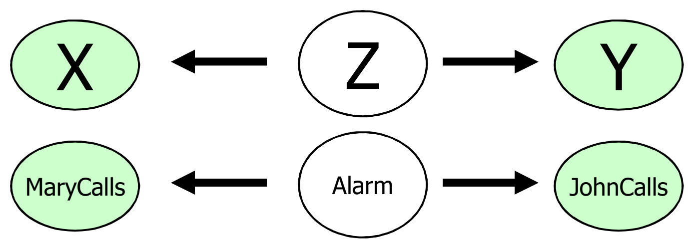
Yes, yes!!! Now tell me H!
$P(W = br \mid H= bl, D=br) = 0.50$
$P(W = br \mid H= br, D=br) = 0.22$

d-separation Conditions

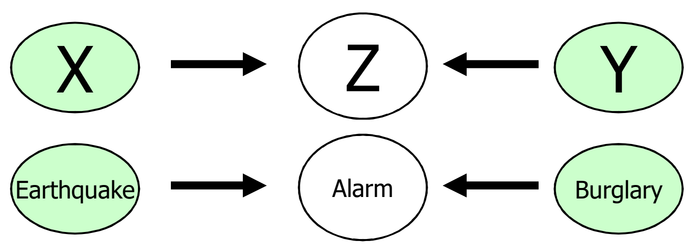
$\neg(X \perp Y)$



$\neg(X \perp Y)$

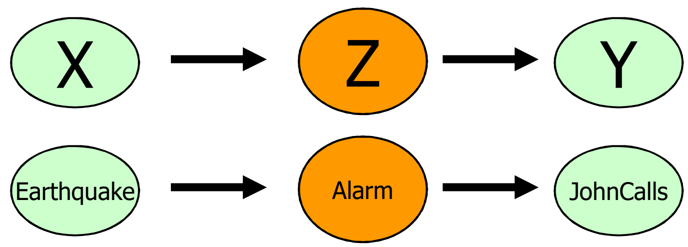


$X \perp Y$

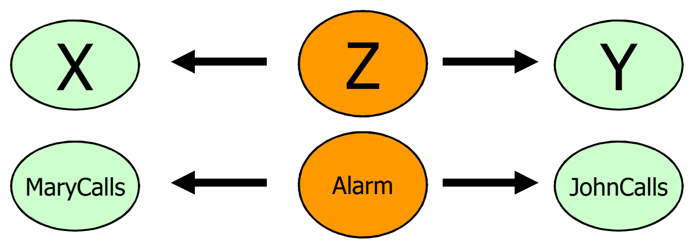


d-separation Conditions

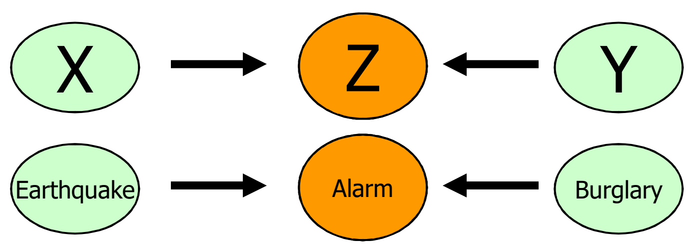
Jump



$$X \perp Y \mid Z$$



$$X \perp Y \mid Z$$



$$\neg(X \perp Y \mid Z)$$

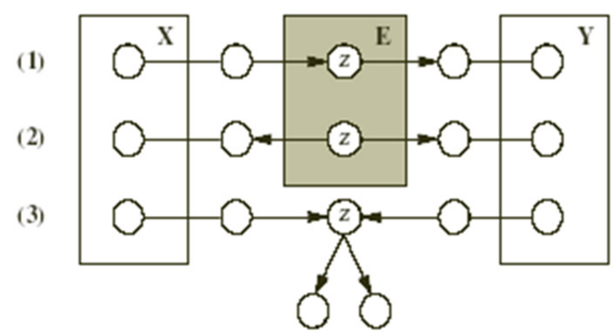
Conditional Independence

- Node X is independent of its non-descendants given assignment to immediate parents parents(X)

- **General** question: "**X** $\perp$ **Y** | **E**"
  - Are nodes **X** independent of nodes **Y**, given assignments to (evidence) nodes **E**?

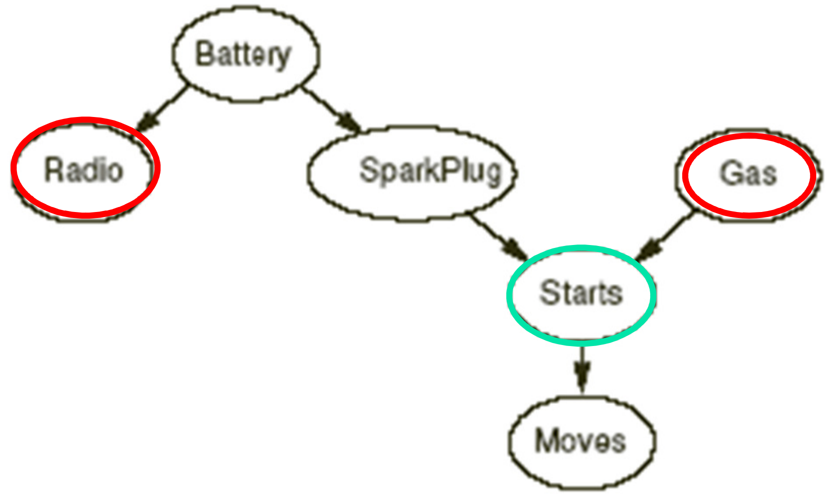
- **Answer**: If every undirected path from **X** to **Y** is d-separated by **E**, then **X** $\perp$ **Y** | **E**

- *d-separated* if every path from **X** to **Y** is blocked by **E**
  ... if $\exists$ node Z on path s.t.
  1. Z $\in$ **E**, and Z has 1 out-link (on path)
  2. Z $\in$ **E**, and Z has 2 out-link, *or*
  3. Z has 2 in-links, Z $\notin$ **E**, no child of Z in **E**



Example of *d*-separation, II

d-separated if
every path from X to Y is blocked by **E**



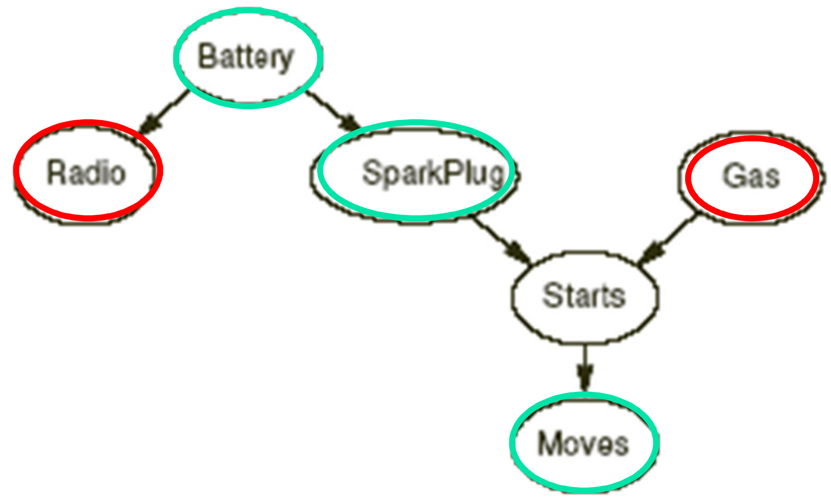
Is Radio *d*-separated from Gas given . . .

1. **E** = {} ?
   YES: P(R | G) = P(R)
   Starts ∉ **E**, and Starts has 2 in-links
2. **E** = Starts ?
   NO!! P(R | G, S) ≠ P(R| S)
   Starts ∈ **E**, and Starts has 2 in-links

If car does not start,
expect radio to NOT work.
… unless you see it is out of gas!

Example of *d*-separation, II

d-separated if
every path from X to Y is blocked by **E**



Is Radio *d*-separated from Gas given . . .

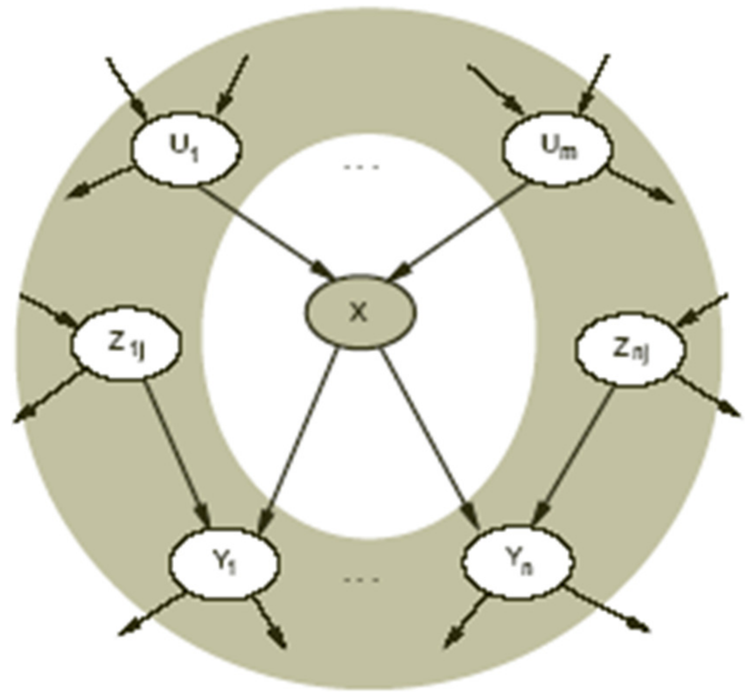
1. **E** = {} ?
   YES: P(R | G) = P(R)
   Starts ∉ **E**, and Starts has 2 in-links
2. **E** = Starts ?
   NO!! P(R | G, S) ≠ P(R| S)
   Starts ∈ **E**, and Starts has 2 in-links
3. **E** = Moves ?
   NO!! P(R | G, M) ≠ P(R| M)
   Moves ∈ **E**, Moves child-of Starts, and Starts has 2 in-links (on path)
4. **E** = SparkPlug ?
   YES: P(R | G, Sp) = P(R| Sp)
   SparkPlug ∈ **E**, and SparkPlug has 1 out-link
5. **E** = Battery ?
   YES: P(R | G, B) = P(R| B)
   Battery ∈ **E**, and Battery has 2 out-links

If car does not MOVE,
expect radio to NOT work.
… unless you see it is out of gas!

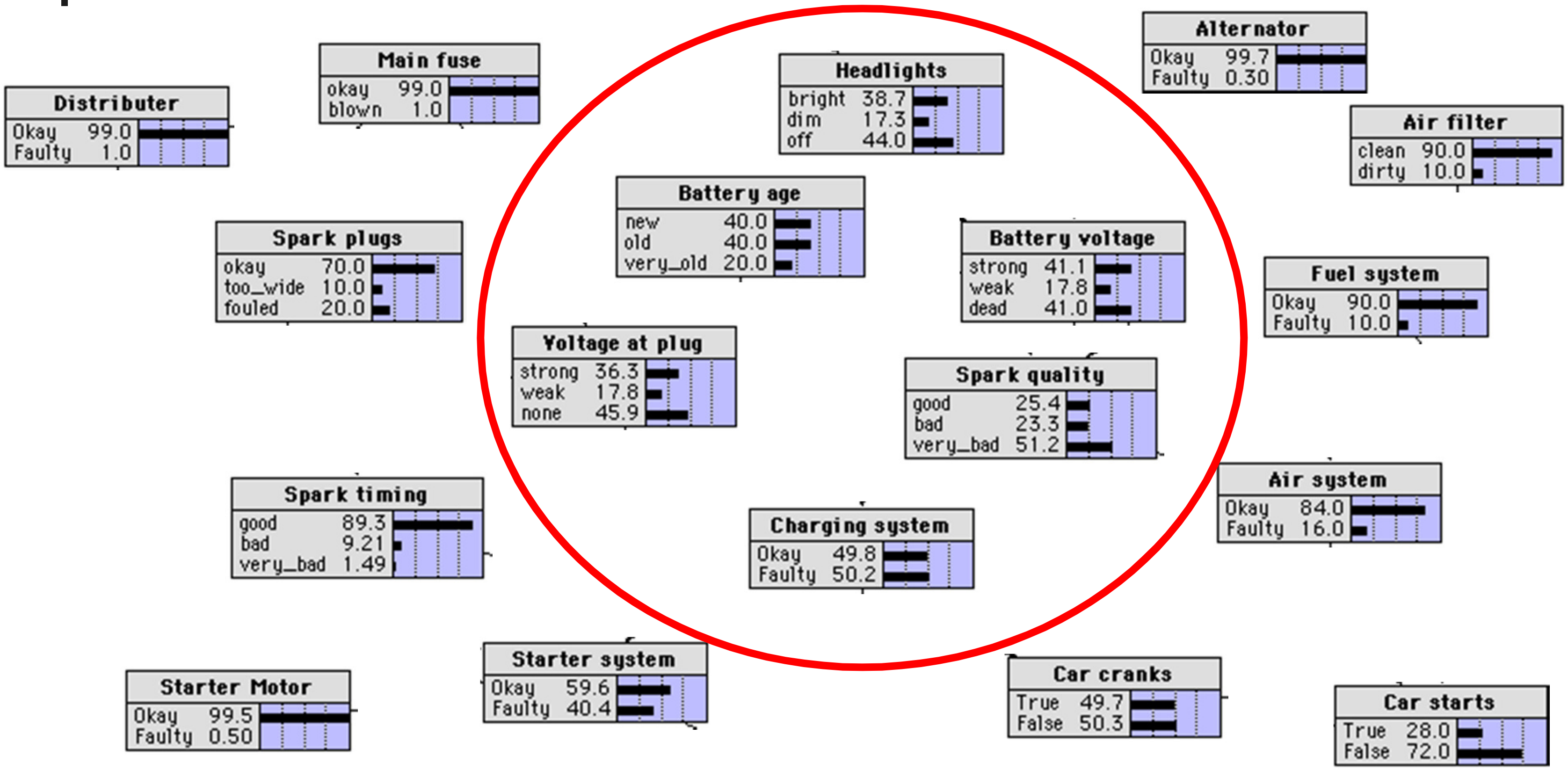
Markov Blanket

Each node is
conditionally independent of all others
given its *Markov blanket:*

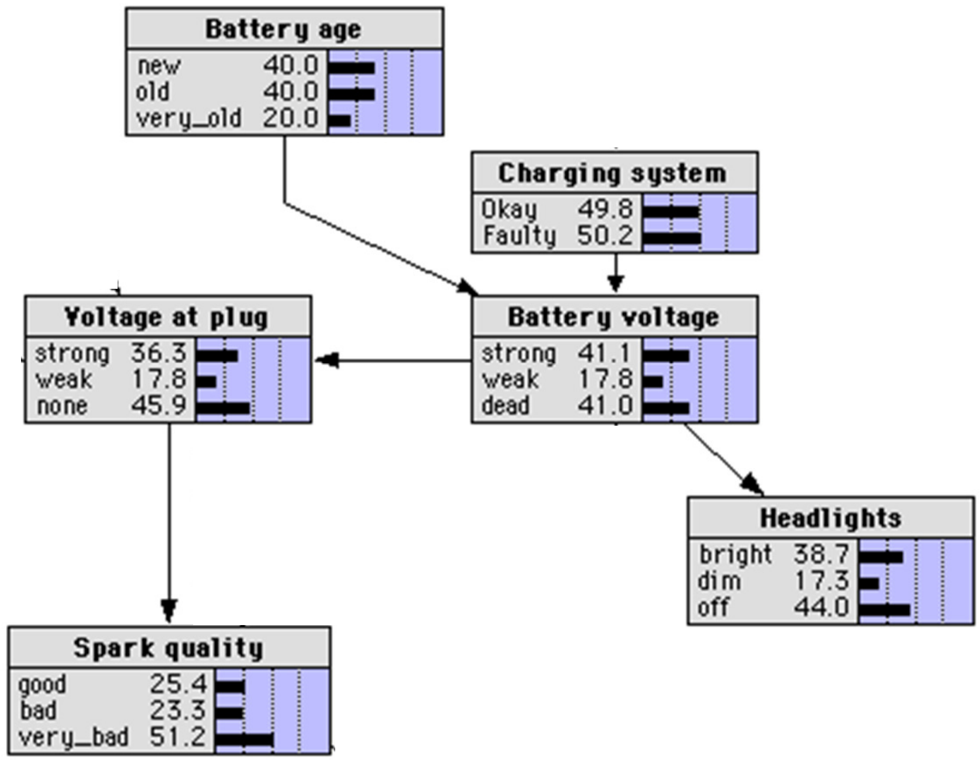
- parents
- children
- children's parents



Engineer a Belief Net



Example: Car Diagnosis



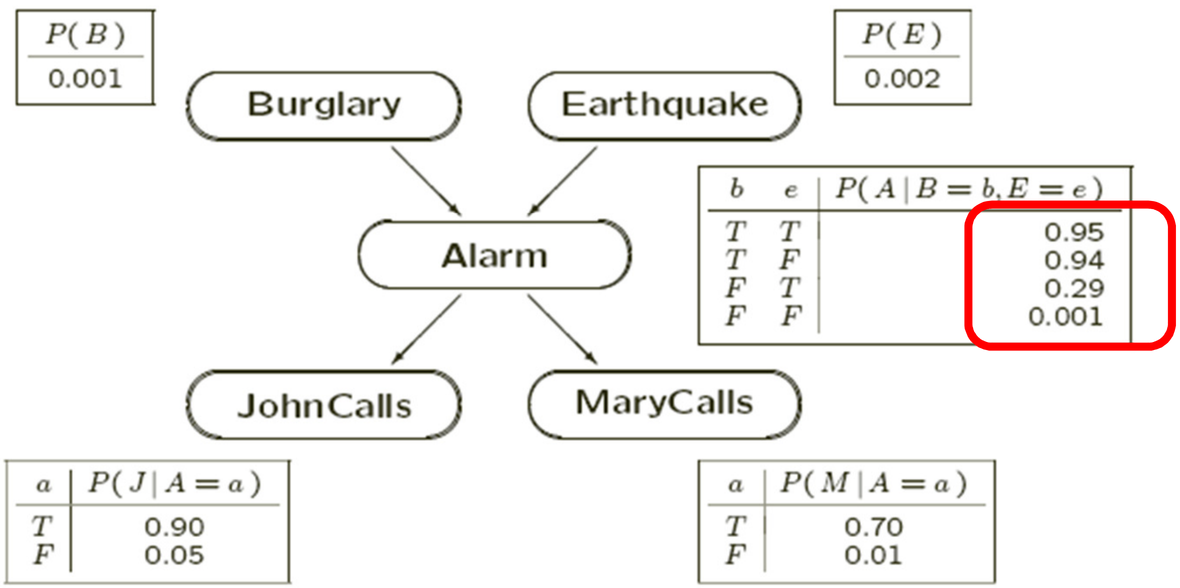
Alternative Conditional Probabilities

- Special forms for Discrete variables
  - Determistic
  - Noisy Or
  - Sigmoid
  - Context-Specific Independencies
- Continuous variables – Gaussians
  - Linear Gaussian Model
  - Hybrid Versions

Example of a Belief Net

Directed Acyclic Graph:

$$\mathcal{BN} = \left\{ \begin{array}{lll} \mathcal{N} & \text{Nodes} & \equiv \text{Variables} \\ \mathcal{A} & \text{Arcs} & \equiv \text{Dependencies} \\ \mathcal{C} & \text{CPTables} & \equiv \text{“weights”} \end{array} \right\}$$



$P(B)$
0.001

$P(E)$
0.002

b	e	$P(A\mid B=b, E=e)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

a	$P(J\mid A=a)$
T	0.90
F	0.05

a	$P(M\mid A=a)$
T	0.70
F	0.01

- Discrete variables
- Explicit table

- **Nodes**: one for each random variable
- **Arcs**: one for each direct influence between two r.v.s
- **CPT**: each node stores a conditional probability table

  P(Node | Parents(Node))

  to quantify effects of “parents” on child

Simple forms of CPTable

- In gen'l: CPTable is function mapping *values of parents* to *distribution over child*

$$f: \left[\prod_{U \in Parents(X)} Dom(U) \right] \times Dom(X) \;\mapsto\; [0.1]$$

(Actually, $f': \prod_{U \in Parents(X)} Dom(U) \;\mapsto$ dist over X)

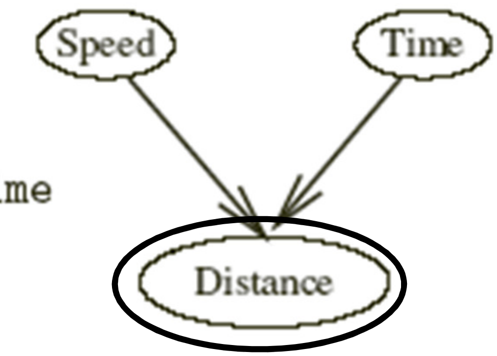
Cold	Flu	Malaria	P(Fever | C,F,M)	P(¬Fever | C,F,M)
F	F	F	0.0	1.0
F	F	T	0.9	0.1
F	T	F	0.8	0.2
F	T	T	0.98	0.02
T	F	F	0.4	0.6
T	F	T	0.94	0.06
T	T	F	0.88	0.12
T	T	T	0.988	0.012

- Standard: Include $\prod_{U \in Parents(X)} |Dom(U)|$ rows, each with $|Dom(X)| - 1$ entries
- But... can be structure within CPTable: Deterministic, Noisy-Or, Decision Tree, ...

Skip

Deterministic Node

- Given value of parent(s), specify unique value for child (logical, functional)



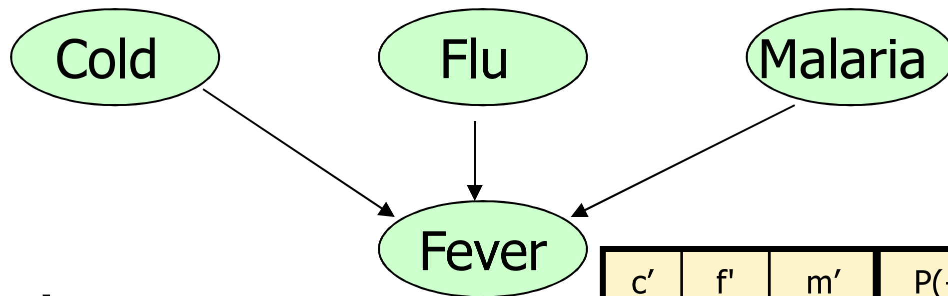
$$P(\texttt{Distance} \mid \texttt{Rate}, \texttt{Time}) = \begin{cases} 1.0 & \text{if } \texttt{Distance} = \texttt{Rate} \cdot \texttt{Time} \\ 0.0 & \text{otherwise} \end{cases}$$

As if each row has just one 1, rest 0s:

Rate	Time	$P(\texttt{Dist=0} \mid \texttt{R,T})$	$P(\texttt{Dist=1} \mid \texttt{R,T})$	$P(\texttt{Dist=2} \mid \texttt{R,T})$
0	1	1.0	0.0	0.0
1	0	1.0	0.0	0.0
1	1	1.0	1.0	0.0
1	2	0.0	0.0	1.0
2	1	0.0	0.0	1.0
⋮		⋮		

Noisy Or… from simple "Or"

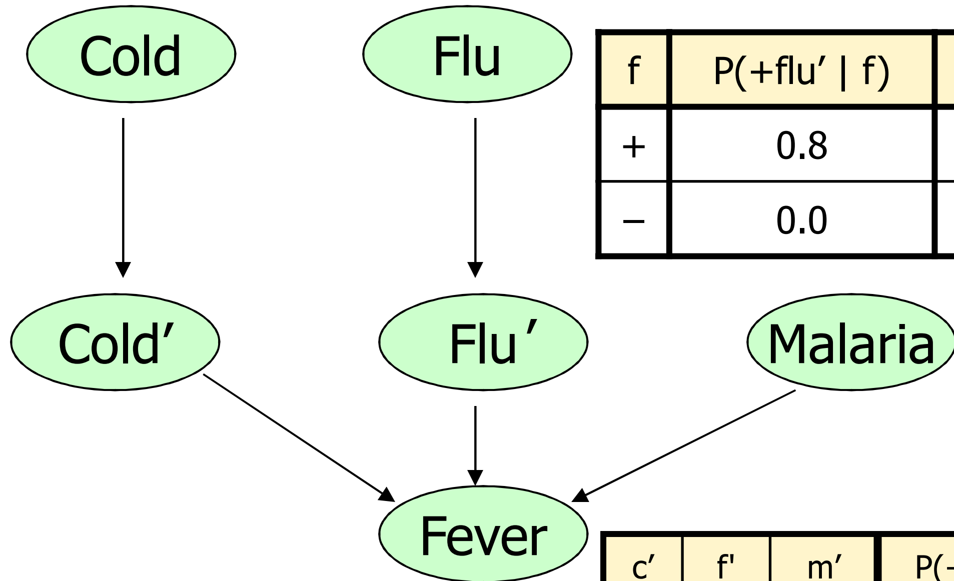
- Suppose person has Fever if have ANY of { Cold, Flu, Malaria }



- But ∃ chance that someone
  - has Cold
  - but not Fever
- Think of
  - Cold′ ALWAYS causing Fever
  - Cold′ occurring in 60% of Cold

c′	f'	m′	P(+Fever | c′, f′, m′)
+	+	+	1.0
+	+	–	1.0
+	–	+	1.0
+	–	–	1.0
+	+	+	1.0
–	+	–	1.0
–	–	+	1.0
–	–	–	0.0

Towards Noisy-Or …



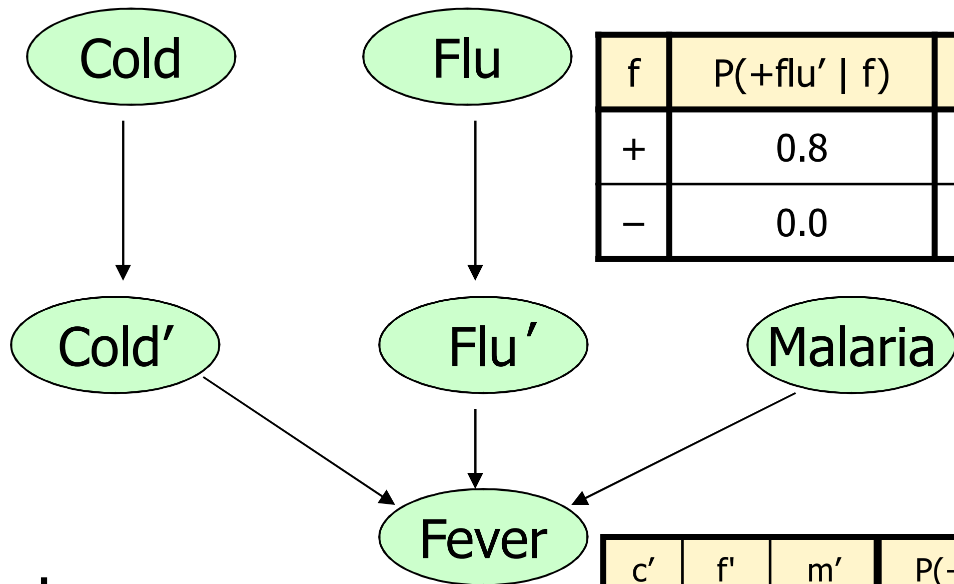
f	P(+flu’ | f)	P(-flu’ | f)
+	0.8	0.2
–	0.0	1.0

c	P(+cold’ | c)	P(-cold’ | c)
+	0.4	0.6
–	0.0	1.0

c’	f'	m’	P(+Fever | c’, f’, m’)
+	+	+	1.0
+	+	–	1.0
+	–	+	1.0
+	–	–	1.0
+	+	+	1.0
–	+	–	1.0
–	–	+	1.0
–	–	–	0.0

- Now … what about Flu?
- Perhaps
  - Flu’ ALWAYS causes Fever
  - Flu’ occurring in 20% of Flu

Towards Noisy-Or …



f	P(+flu′ | f)	P(-flu′ | f)
+	0.8	0.2
–	0.0	1.0

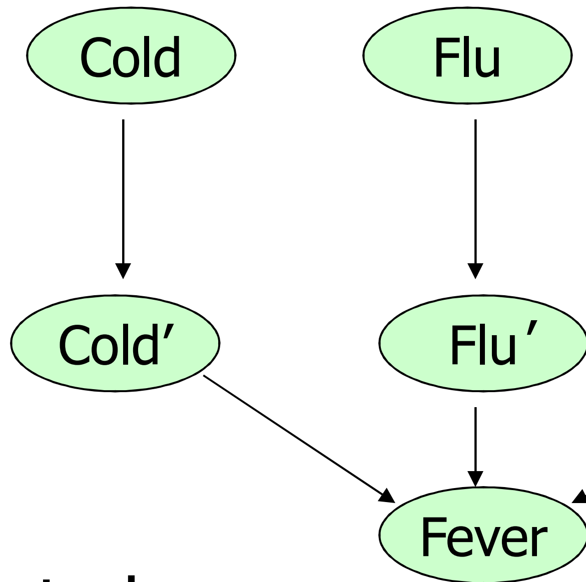
c	P(+cold′ | c)	P(-cold′ | c)
+	0.4	0.6
–	0.0	1.0

c′	f'	m′	P(+Fever | c′, f′, m′)
+	+	+	1.0
+	+	–	1.0
+	–	+	1.0
+	–	–	1.0
+	+	+	1.0
–	+	–	1.0
–	–	+	1.0

- Whatever prevented
  Cold $\not\Rightarrow$ Fever
  is independent of
  whatever prevented
  Flu $\not\Rightarrow$ Fever

$$P(\neg\text{Fev} \mid +\text{Cold}, +\text{Flu}) \approx P(\neg\text{Fev} \mid +\text{Cold}) \times P(\neg\text{Fev} \mid +\text{Flu})$$

Towards Noisy-Or …



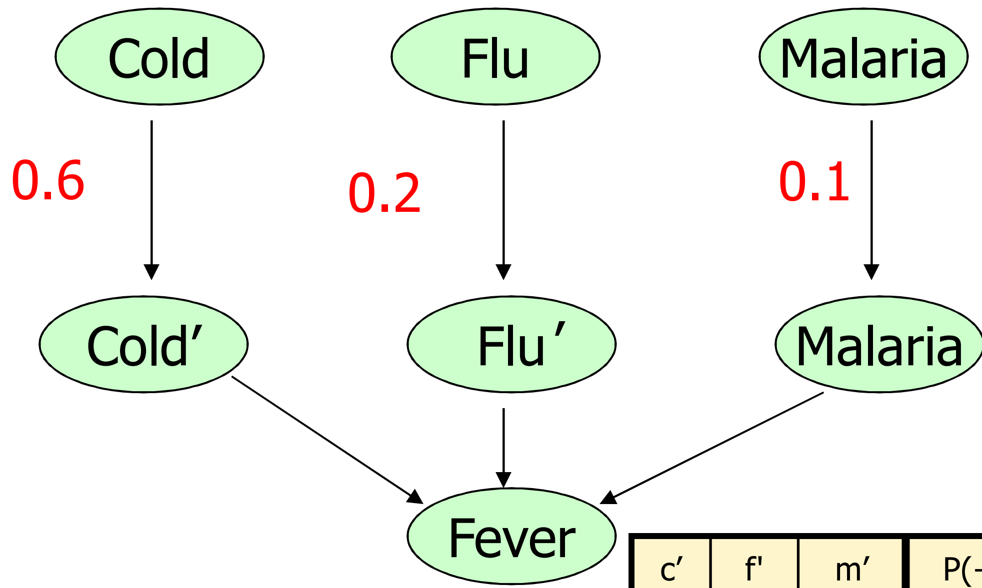
f	P(+flu′ | f)	P(-flu′ | f)
+	0.8	0.2
–	0.0	1.0

c	P(+cold′ | c)	P(-cold′ | c)
+	0.4	0.6
–	0.0	1.0

c′	f'	m′	P(+Fever | c′, f′, m′)
+	+	+	1.0
+	+	–	1.0
+	–	+	1.0
+	–	–	1.0
+	+	+	1.0
–	+	–	1.0
–	–	+	1.0
–	–	–	0.0

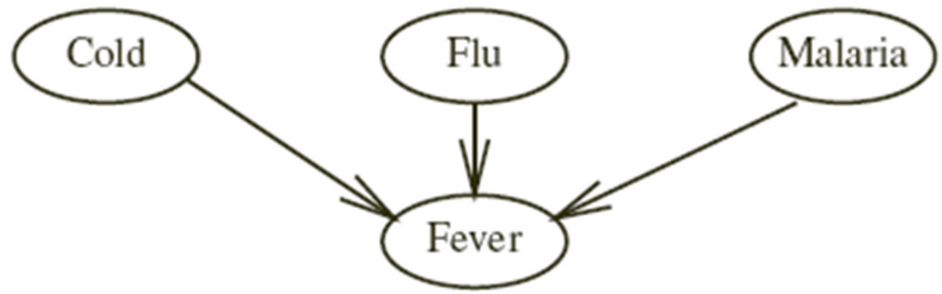
- Whatever prevented
  Cold $\not\Rightarrow$ Fever
  is independent of
  whatever prevented
  Flu $\not\Rightarrow$ Fever
- Just one parameter for each feature

Noisy-Or …



c′	f'	m′	P(+Fever | c′, f′, m′)
+	+	+	1.0
+	+	−	1.0
+	−	+	1.0
+	−	−	1.0
+	+	+	1.0
−	+	−	1.0
−	−	+	1.0
−	−	−	0.0

Noisy-OR "CPTable" (2)



- $P(\texttt{Fev} \mid \neg\texttt{Col}, \neg\texttt{Flu}, \neg\texttt{Mal}) = 0$

$$P(\neg\texttt{Fev} \mid \texttt{Col}) \approx q_{col} = 0.6$$
$$P(\neg\texttt{Fev} \mid \texttt{Flu}) \approx q_{flu} = 0.2$$
$$P(\neg\texttt{Fev} \mid \texttt{Mal}) \approx q_{mal} = 0.1$$

- Independent inhibiters:
$P(\neg\texttt{Fev} \mid \texttt{Col}, \texttt{Flu}) \approx P(\neg\texttt{Fev} \mid \texttt{Col}) \times P(\neg\texttt{Fev} \mid \texttt{Flu})$

$$P(\neg\texttt{Fever} \mid \pm_i d_i) = \prod_{i:+d_i} q_i$$

Cold	Flu	Malaria	$P(\neg\texttt{Fever} \mid \texttt{c,f,m})$	$P(\texttt{Fever} \mid \texttt{c,f,m})$
F	F	F	1.0	0.0
F	F	T	0.1	0.9
F	T	F	0.2	0.8
F	T	T	$\boxed{0.02 = 0.2 \times 0.1}$	0.98
T	F	F	0.6	0.4
T	F	T	$0.06 = 0.6 \times 0.1$	0.94
T	T	F	$0.12 = 0.6 \times 0.2$	0.88
T	T	T	$0.012 = 0.6 \times 0.2 \times 0.1$	0.988

Noisy-Or (Ge

CPCS Network:

- Modeling disease/symptom for internal medicine
- Using Noisy-Or & Noisy-Max
- 448 nodes, 906 links
- Required 8,254 values (not 13,931,430) !

- Fever if Cold, Flu or Malaria

$$\text{Want} \begin{cases} P(\texttt{Fev} \mid \neg\texttt{Col}, \neg\texttt{Flu}, \neg\texttt{Mal}) & = 0 \\ P(\neg\texttt{Fev} \mid \texttt{Col}) & \approx q_{col} = 0.6 \\ P(\neg\texttt{Fev} \mid \texttt{Flu}) & \approx q_{flu} = 0.2 \\ P(\neg\texttt{Fev} \mid \texttt{Mal}) & \approx q_{mal} = 0.1 \end{cases}$$

- Define 3 numbers ("noise" parameters)

  $q_{col}, \ q_{flu}, \ q_{mal}$

Def'n: $P(\neg\texttt{Fev} \mid \pm S_1, \pm S_2, \pm S_3) = \prod_{i:+S_i} q_i$

$$P(\neg\texttt{Fev} \mid \neg\texttt{Col}, \neg\texttt{Flu}, \neg\texttt{Mal}) = 1.0$$

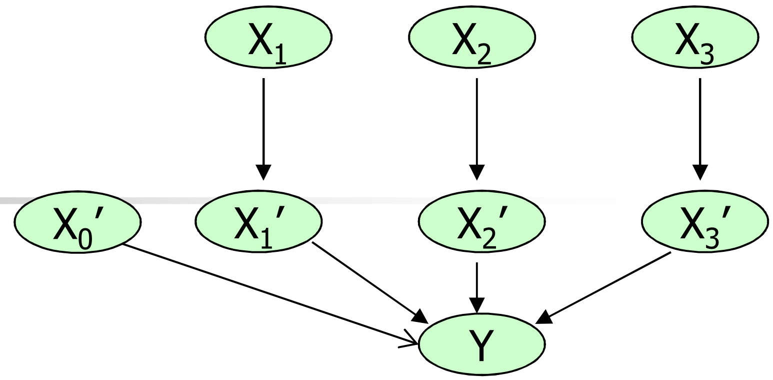
$$P(\neg\texttt{Fev} \mid \neg\texttt{Col}, \texttt{Flu}, \neg\texttt{Mal}) = q_{flu} = 0.2$$

$$P(\texttt{Fev} \mid \neg\texttt{Col}, \texttt{Flu}, \neg\texttt{Mal}) = 0.8$$

Assumes:
- each cause has independent chance of causing effect
- all causes listed
  (Leak node, to handle ALL OTHER CAUSES. . .)
- inhibiting factors independent

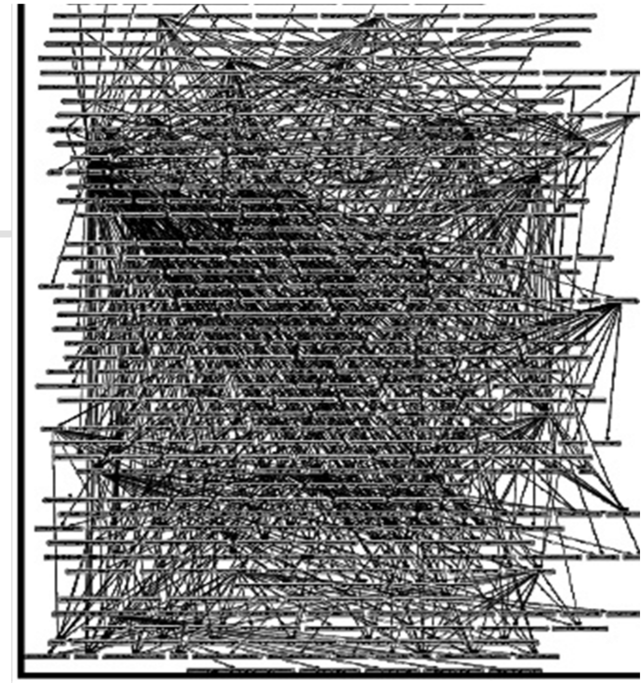
Note: Only k parameters, not 2^k

Extensions



- Can have a LEAK variable
  - Chance of +outcome,
    even if NONE of the variables hold
- Noisy **And**
- Noisy **Max** (not just binary variables)
- …

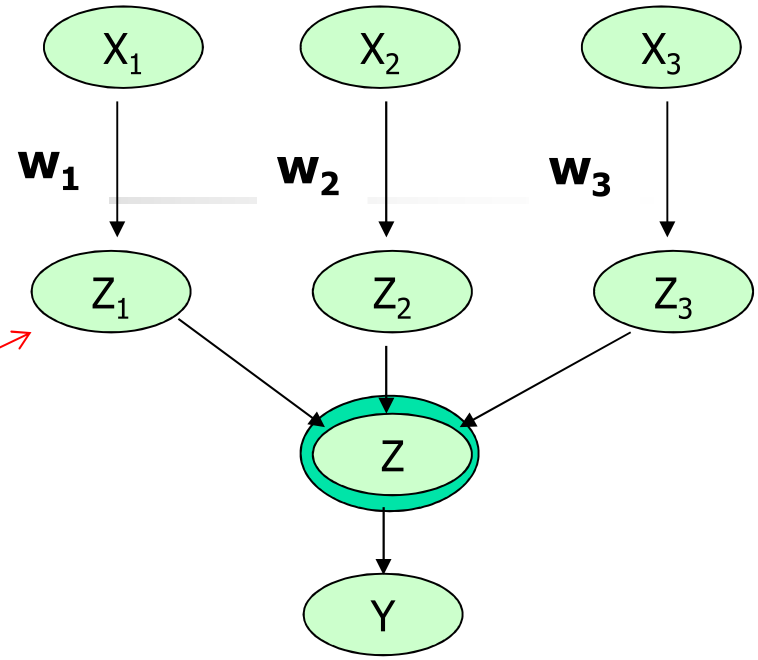
CPCS network



- 422 nodes
  - 14 describe diseases
  - 33 risk factors
  - 375 findings
- each variable with ≈ 4 values
- Naïvely … $\approx 4^{422}$ parameters
- When factored… ≈ 134 million
- Using NoisyMax: $\approx 8{,}000$

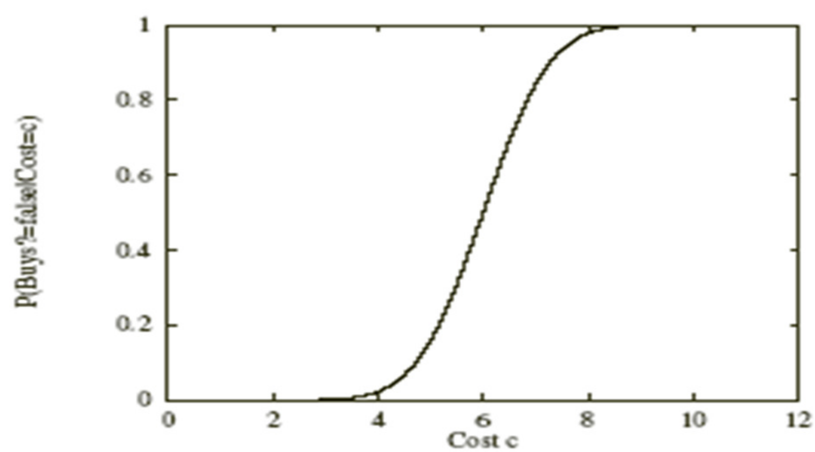
Sigmoid CPD

$$Z = w_0 + \sum_{i=1}^{k} \boxed{w_i X_i}$$



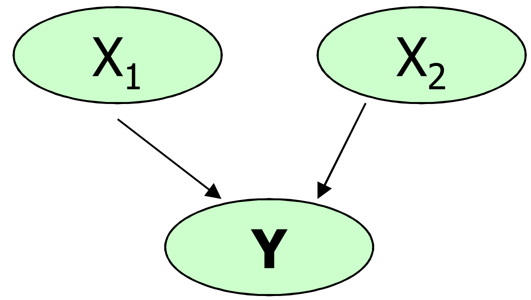
Jump

- $\mathrm{P}(+y \mid X_1, \ldots, X_k) = \mathrm{sigmoid}(Z)$
  - $\mathrm{sigmoid}(z) = \frac{e^z}{1+e^z}$



Structural CP Distributions

$Y = X_1 \vee X_2$



- Notice:
  - $\neg(X_1 \perp X_2 \mid Y)$
  - $\neg(Y \perp X_1 \mid X_2)$

But…

- If Y = 0, then both X_1 =0 and X_2 = 0, so $X_1 \perp X_2 \mid \neg y$

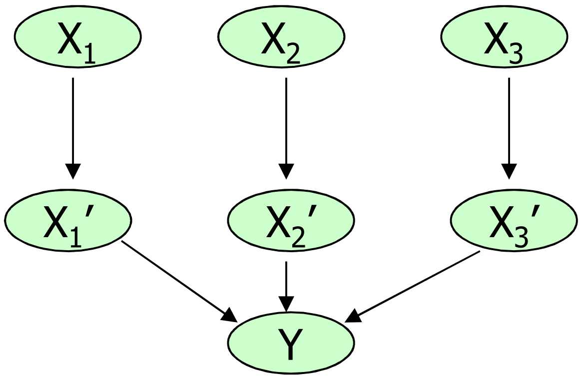
$$\boxed{X_1 \perp_c X_2 \mid Y, (Y = -)}$$

- If $X_2 = 1$, then Y is true… so $Y \perp X_1 \mid +x_2$

$$\boxed{Y \perp_c X_1 \mid X_2, (X_2 = +)}$$

Context Specific Independence

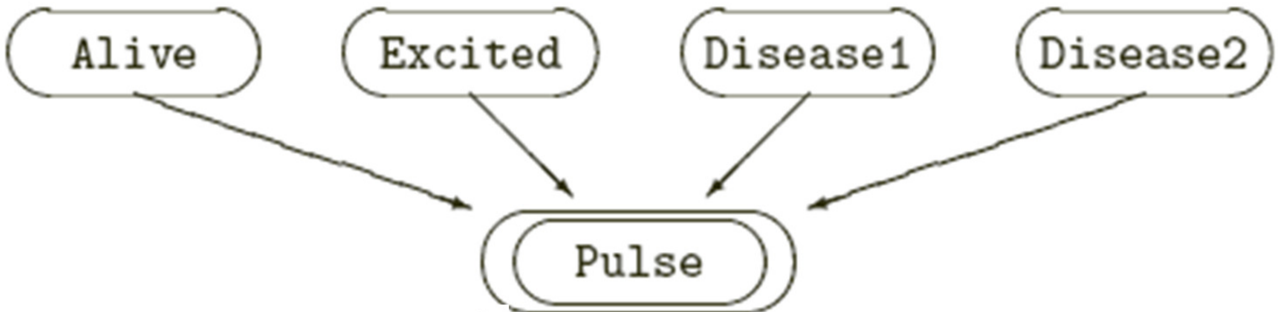
- For NoisyOr:
  $X'_i \perp X'_j \mid$ -y



- $X \perp_c Y \mid Z, c$
  - assignment c
- P(X, Y | Z, c) = P(X | Z, c) P(Y | Z, c)
- P(X | Y, Z, c) = P(X | Z, c)
- P(Y | X, Z, c) = P(Y | Z, c)

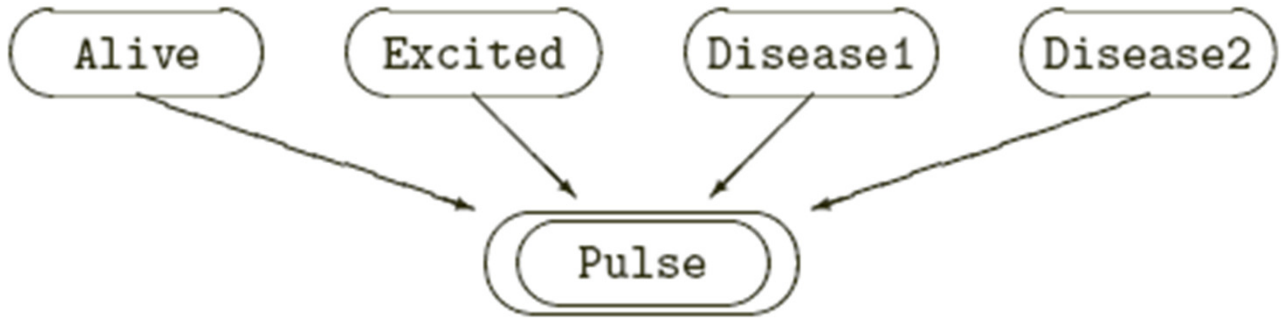
DecisionTree CPTable

Jump



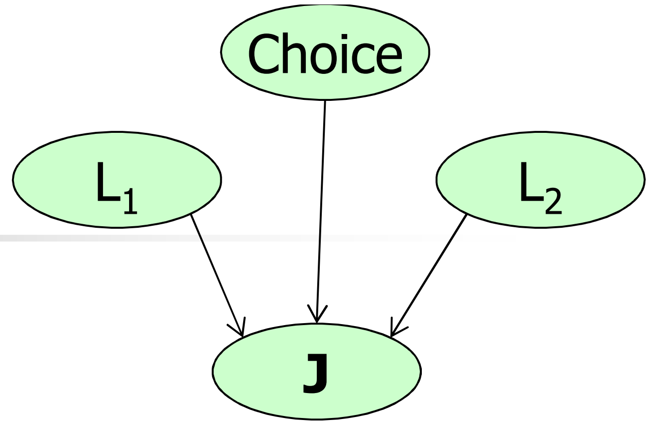
A	E	D1	D2	χ s.t. $P(\texttt{Pulse}=\chi \mid \texttt{A,E,D1,D2}) = 1.0$
Y	Y	Y	Y	vhigh
Y	Y	Y	N	vhigh
Y	Y	N	Y	vhigh
Y	Y	N	N	vhigh
Y	N	Y	Y	high
Y	N	Y	N	med
Y	N	N	Y	med
Y	N	N	N	ok
N	Y	Y	Y	none
N	Y	Y	N	none
N	Y	N	Y	none
N	Y	N	N	none
N	N	Y	Y	none
N	N	Y	N	none
N	N	N	Y	none
N	N	N	N	none

DecisionTree CPTable



- $P \perp E \ | \ A = no$
- $P \perp E, D1, D2 \ | \ A = no$
- $P \perp D1 \ | \ E = vhigh$

Multiplexer CPD

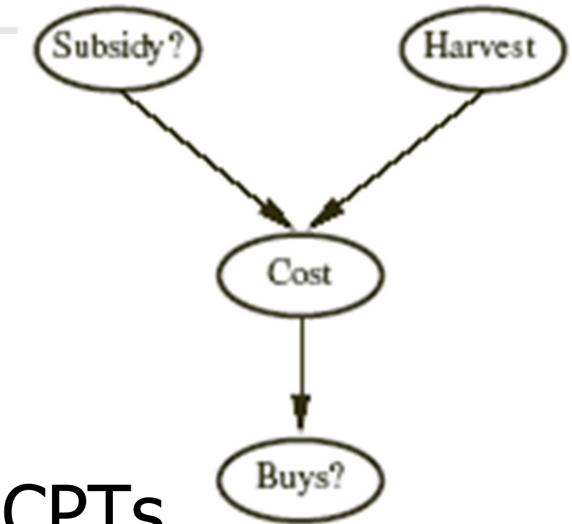


- Job depends on
  Choice of LetterWriter#1
  vs LetterWriter#2
- Note that

$$L_1 \perp L_2 \mid J, C$$

- Why?
  - $L_1 \perp_c L_2 \mid J, c = 1$
  - $L_1 \perp_c L_2 \mid J, c = 2$
- General: $P(Y \mid A, Z_1, \ldots Z_k) = 1$ iff $Y = Z_A$

Hybrid (discrete+continuous) Networks



- **Discrete**: Subsidy?, Buys?

  Continuous: Harvest, Cost

Option 1: Discretization

but possibly large errors, large CPTs

Option 2: Finitely parameterized canonical families

Problematic cases to consider...

- Continuous variable, discrete+continuous parents

  Cost

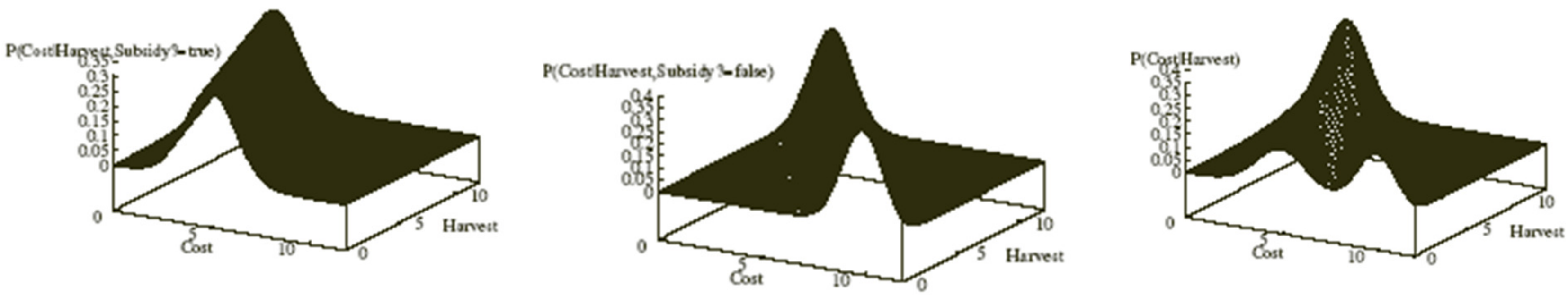
- Discrete variable, continuous parents

  Buys?

Skip

If everything continuous is Gaussian...

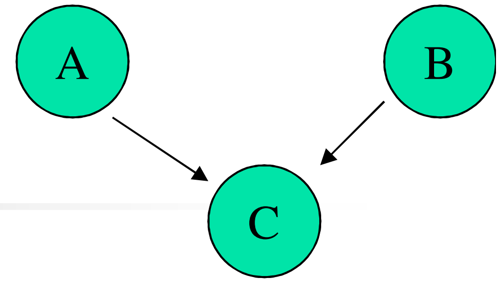
- All nodes continuous w/ LinearGaussian dist'ns
  $\Rightarrow$ full joint is a multivariate Gaussian



- Discrete+continuous LG network
  $\Rightarrow$ conditional Gaussian network

multivariate Gaussian over all continuous variables for each combination of discrete variable values

Linear Gaussian Model

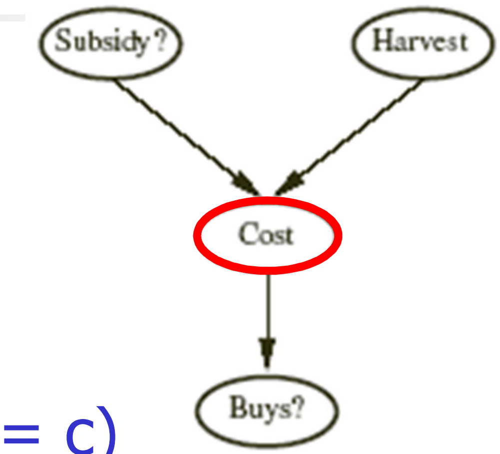


- $x_i \mid pa_i \sim \mathcal{N}(x_i \mid b_i + \sum_{j \in pa\_i} w_{ij}\, x_j,\ \ v_i)$
- So…
  - $P(x_A) \sim \mathcal{N}(x_A \mid b_A,\ \ v_A)$
  - $P(x_B) \sim \mathcal{N}(x_B \mid b_B,\ \ v_B)$
  - $P(x_C \mid x_A, x_B) \sim \mathcal{N}(x_C \mid b_C + w_{AC}\, x_A + w_{BC}\, x_B,\ \ v_C)$

  … eg, $\mathcal{N}(x_C \mid 2.9 + 1.3\, x_A + -21\, x_B,\ \ 0.5)$
- $\ln p(\mathbf{x}) = \sum_i \ln p(x_i \mid pa_i) =$

$$-\sum_i \frac{1}{2v_i}\left(x_i - \sum_{j \in pa_i} w_{ij} x_i - b_i\right)^2 + const.$$

Continuous Child Variables



- For each "continuous" child E,
  - with continuous parents C
  - with discrete parents D
- Need conditional density function

  P(E = e | C = c, D = d) = $P_{D=d}$(E = e | C = c)

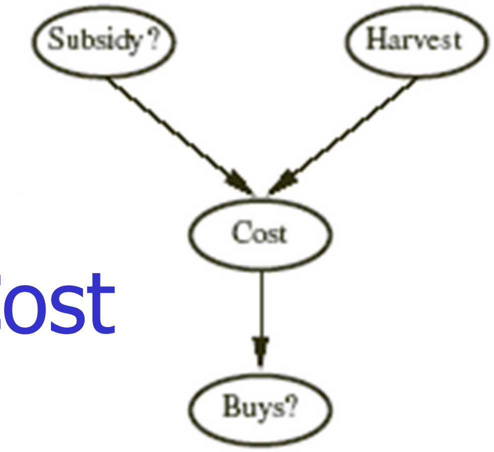
  for each assignment to discrete parents D=d
- Common: linear Gaussian model

  f(Harvest, Subsidy?) = "dist'n over Cost"

$$P(\texttt{Cost} = \texttt{c} \mid \texttt{Harvest} = \texttt{h},\ \texttt{Subsidy?} = \texttt{true})$$
$$= \mathcal{N}[a_t h + b_t,\ \sigma_t](c)$$
$$= \frac{1}{\sigma_t\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{c-(a_t h + b_t)}{\sigma_t}\right)^2\right)$$
$$P(\texttt{Cost} = \texttt{c} \mid \texttt{Harvest} = \texttt{h},\ \texttt{Subsidy?} = \texttt{false})$$
$$= \mathcal{N}[a_f h + b_f,\ \sigma_f](c)$$

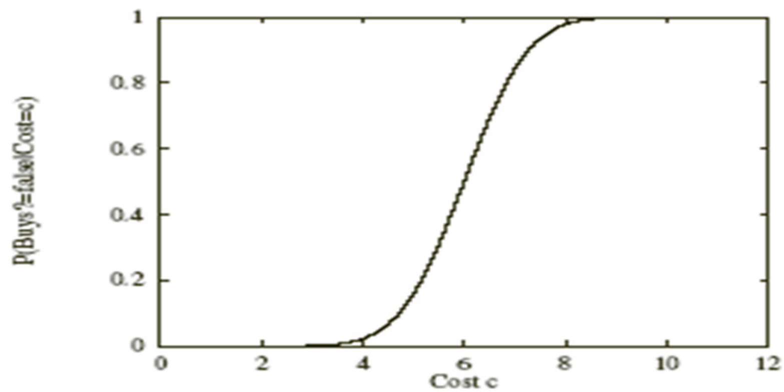
Need parameters:
$\sigma_t \quad a_t \quad b_t$
$\sigma_f \quad a_f \quad b_f$

Discrete variable w/ Continuous Parents



- Probability of Buys? given Cost

  $\approx^{?}$ "soft" threshold:



- Probit distribution uses integral of Gaussian:

$$\Phi(x) = \int_{-\infty}^{x} N(x; \mu = 0, \sigma^2 = 1)\, dx$$

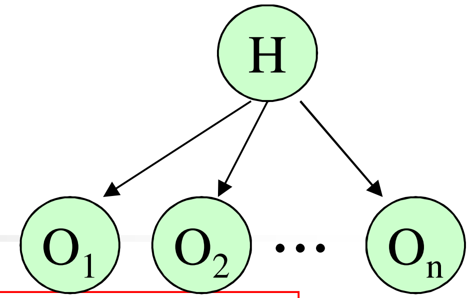
P(Buy? = true | Cost=c) $= \Phi\left(\frac{\mu - c}{\sigma}\right)$

$\approx$ hard threshold, whose location is subject to noise

Outline

- Motivation
- What is a Belief Net?
  - Factored Distribution
  - Semantics
- Inference (Reasoning)
  - Examples
  - Complexity
  - Maximize Expected Utility
- Applications
- Relation to other Models
- Learning a Belief Net

Naïve Bayes



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

- *Normalizing term*

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

  - Needed to computing posteriors
  - Not needed for computing MOST-Likely: as same for all h_i's

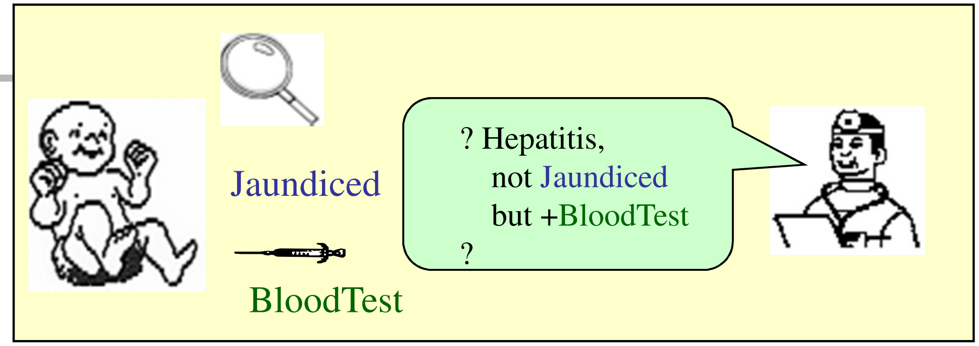
- Easy to use for Classification
- Can use even if some v_js not specified
- If k h_i's and n O_js,

  requires only k priors, $n \times k$ 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

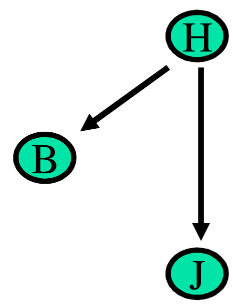
Classification



- Which is more likely: $+h$ vs $-h$?

- Given independencies:
  + values:

h	P(+b | h)	P(-b | h)
1	0.95	0.05
0	0.03	0.93



P(+h)	P(-h)
0.05	0.95

h	P(+j |h)	P(-j |h)
1	0.8	0.2
0	0.3	0.7

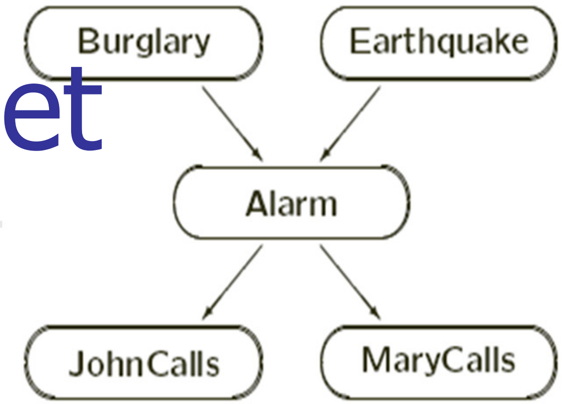
- $\mathrm{argmax}_h\ P(\ h \mid +b, -j\ )$

  $= \mathrm{argmax}_h\ P(\ h\ ) \times P(+b \mid h\ ) \times P(-j \mid h)$

  $= \mathrm{argmax}_h\ \{\ +:\ 0.05 \times 0.95 \times 0.2,\ \ -\ :\ 0.95 \times 0.03 \times 0.7\ \}$

  $\boxed{-h \quad \text{as} \quad 0.0095 \ < \ 0.01995}$

Inference in Belief Net



- Given: Fixed Belief Net

$$BN = \begin{cases} \mathbf{N} & \text{Nodes} = \text{Variables} \\ \mathbf{A} & \text{Arcs} = \text{Dependencies} \\ \mathrm{C} & \text{CPtable} = \text{"weights"} \end{cases}$$

- Belief Assessment: Compute P(H | **E**)
  - H, **E** arbitrary

Relatively easy

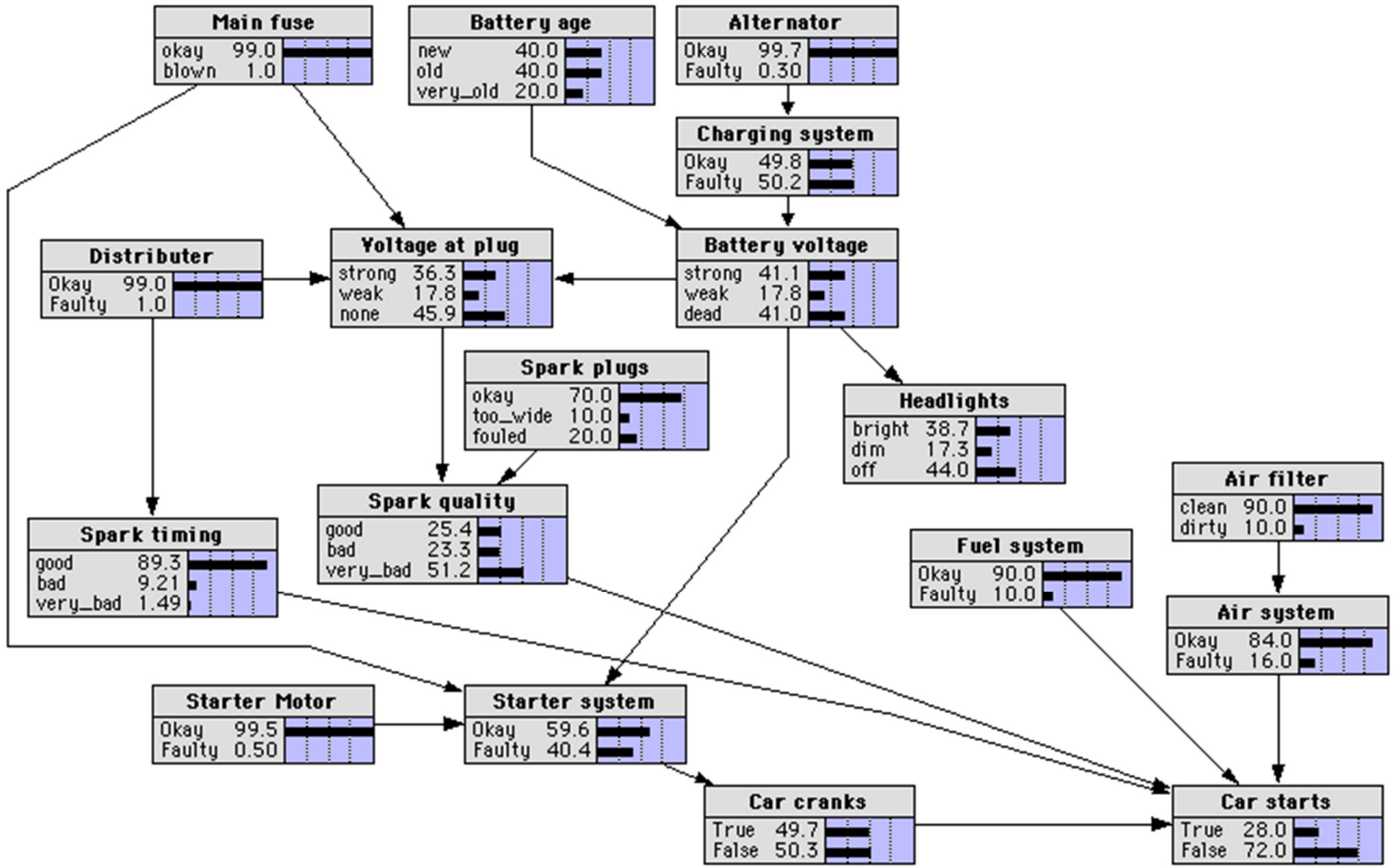
- MAP (Maximum a Posteriori)
  Compute argmax_h P(H=h | **E**)
  - H is single node (eg, *Meningitis*)
  - **E** = **N** – {H} is set of evidence nodes (everything except H)

    BloodTest = +, Jaundice = –, Smoker = –, Temp = 99.8, ...

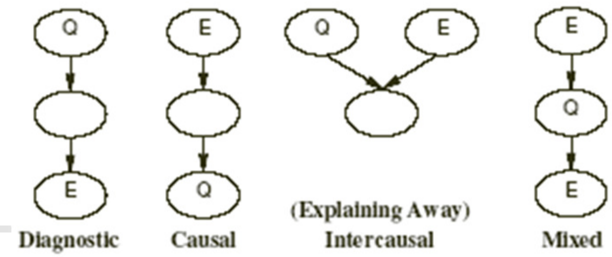
Relatively hard

- MPE (Most Probable Explanation)
  Compute argmax_h P(H=h | **E**)
  - H, **E** arbitrary

Example: Car Diagnosis



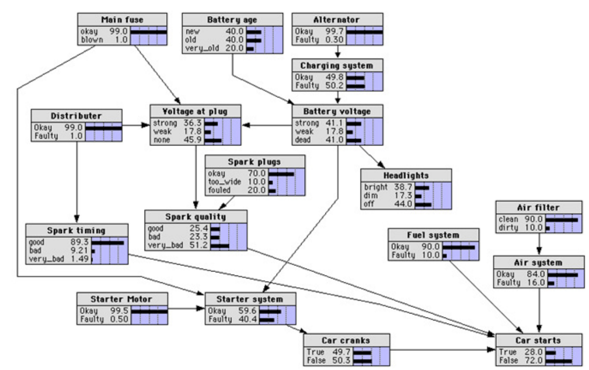
Types of Reasoning



- Typical case: P(QueryVar | EvidenceVars = vals)
  - Eg: P(+starts | +fuel, -voltage)
- **Diagnostic**: from effect to (possible) causes
  - P(-fuse | -starts) = 0.016
- **Causal**: from cause to effects
  - P(-starts | -fuse) = 0.86
- **InterCausal**: between causes of common effect
  - P(-fuel | -starts) = 0.376
  - P(-fuel | -starts, -filter) = 0.003

  Bad_Filter EXPLAINS no_start, and so

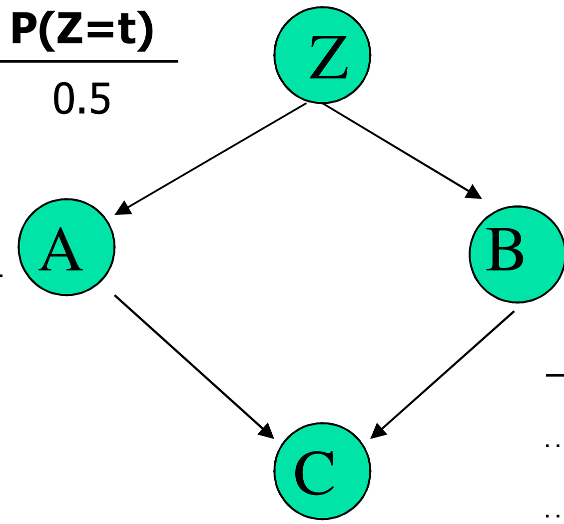
  Bad_Filter EXPLAINS AWAY low-fuel
- **Mixed**: combinations of . . .
  - P(+headlights | +voltage, -starts) = 0.03



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) ?

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

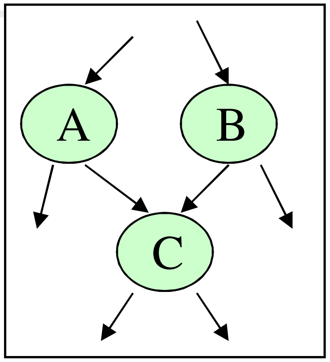
So… ? ~~P(C = t) = P(A = t, B = t)~~

~~= P(A = t) × P(B = t) = ½ × ½ = ¼~~

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

Need to maintain dependencies! $P(A = t, B = t) = P(A = t) \times \boxed{P(B = t \mid A = t)}$

Inference not "Detachable"



- Reasoning is "detached" if, given can compute

  $P(\ C \mid \mathbf{E}\ ) = f(\alpha, \beta)$

  from $\alpha = P(\ A \mid \mathbf{E}\ )$ $\beta = P(\ B \mid \mathbf{E}\ )$
- True IF A, B independent (ie, if poly-tree)
- But FALSE in general...

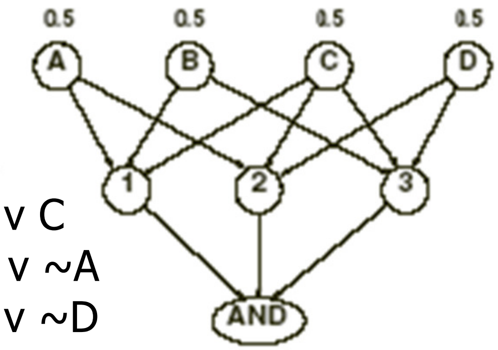
  if uninstantiated (undirected) cycle !

$\Rightarrow$ Need to preserve "justifications" for

$P(B \mid \mathbf{E}\ )$ $P(A \mid \mathbf{E}\ )$

(to see if they interact)

Inherent Complexity



1. A v B v C
2. C v D v ~A
3. B v C v ~D

- Worst case:
  - NP-hard to get exact answer (#P-complete)
  - NP-hard to get answer within 0.5
  - Cannot get **relative error** within $2^{n^{1-\varepsilon}}$ unless P = NP
  - Cannot **stochastically** approximate 1-bit, unless P=RP
- Efficient algorithm ...
  - for "PolyTree": Poly time
    - $\leq$ 1 path between any two nodes
  - if CPtable "bounded" (sub-exp time) wrt λ = M/m
    M = largest CPtable entry; m = smallest

Approaches to Belief Assessment

- Exact, Guaranteed
  - PolyTree Algorithm
  - Clustering Approach
  - Bucket Elimination
  - CutSet Approach

HMM !
Kalman Filter !

- Approximate, Guaranteed
  - Algorithm Modification
  - Value Merging
  - Node Merging
  - Arc Removal

- Approximate, Probabilistic
  - Logic Sampling
  - Likelihood Sampling

Interested?
Take Cmput499 / 659
(Probabilistic Graphical Models)

Outline

- Motivation
- What is a Belief Net?
- Inference
  - Examples
  - Complexity
  - Maximize Expected Utility
- Real-World Applications
- Relation to other Models
- Learning a Belief Net

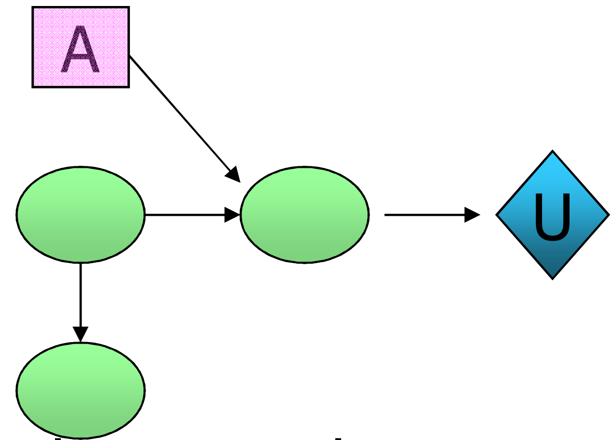
Utility-Based Agents

- MEU Principle:
  Agent should act to maximize expected utility
- Choose action $A^* = \boxed{\text{argmax}_A \{ E[U(A|O)] \}}$
  that maximizes
  expected utility of state after A,
  given prior observations O:

$$E[U(A | O)] =$$
$$= \sum_{S'} \boxed{P(S' | A, O)} \; U(S')$$
$$= \sum_{S'} \boxed{\sum_S P(S | O) P(S' | S, A)} \; U(S')$$

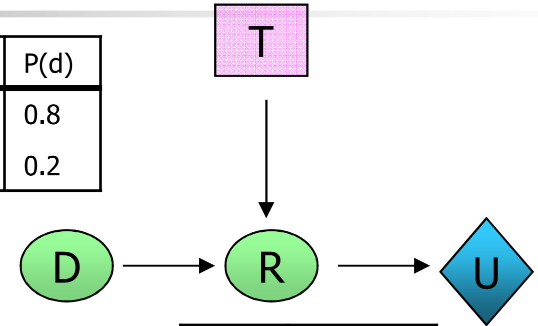
- Given simple assumptions, this is best possible action!
  (Average of utility, not of utility², not minimaxing...)
- Good decision, bad outcome

Decision Network



- Chance Nodes: S, O, S′
  - Belief Net ≈ Decision Network w/ only chance nodes
  - Specify: P(S), P(O | S), P(S′ | S, A)
  - Here: S ≈ Current State O ≈ Observation
    S′ ≈ Resulting State
- Decision Nodes: A
  - represents decision/action to make.
  - Specify: set of possible actions $a \in \mathrm{Dom}(A)$
- Utility Node(s): U
  - represents utility of each value-set of its parent chance variables
  - Specify: U(s′) for each $s' \in \mathrm{Dom}(S')$

Perform a Medical Treatment?



d	P(d)
0	0.8
1	0.2

d	t	P(+r|d,t)
0	0	.001
0	1	.001
1	0	.950
1	1	.010

r	U(r)
0	0
1	-1000

- $EU(\ T = 1\ ) =$
  $\sum_r P(R = r \mid T = 1\ )\ U(R = r)$

  $EU(\ T = 0\ ) =$
  $\sum_r P(R = r \mid T = 0\ )\ U(R = r)$

- $P(\ R = 1 \mid T = 1\ ) =$
  $\sum_d\ P\ (\ R = 1, D = d \mid T = 1\ )$

  $= \sum_d P(\ R = 1 \mid D = d, T = 1\ )\ P(\ D = d\ )$

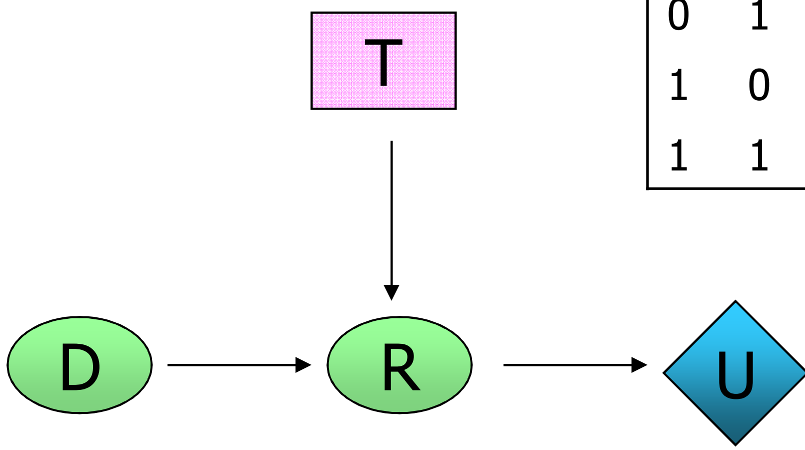
  $= \boxed{P(\ R = 1 \mid D = 0, T = 1\ )}\ P(D = 0) + P(R = 1 \mid D = 1, T = 1)\ P(D = 1)$
  $= \ (0.001 \times 0.8) + (0.01 \times 0.2) = 0.0028$

- $P(\ R = 0 \mid T = 1\ ) = 1 - P(\ R = 1 \mid T = 1\ ) = 0.9972$
- Similarly:
  - $P(\ R = 1 \mid \mathbf{T = 0}\ ) = 0.1908$
  - $P(\ R = 0 \mid \mathbf{T = 0}\ ) = 0.8092$

Medical Treatment (con't)

d	t	P(+r|d,t)
0	0	.001
0	1	.001
1	0	.950
1	1	.010

d	P(d)
0	0.8
1	0.2

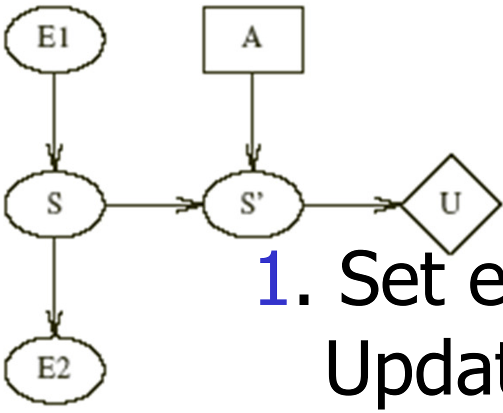


r	U(r)
0	0
1	-1000

	$P(R\|T)$		$U(R)$		
T	0	1	0	1	$EU(T)$
0	.8092	.1908	0	-1000	-190.8
1	.9972	.0028	0	-1000	-2.8

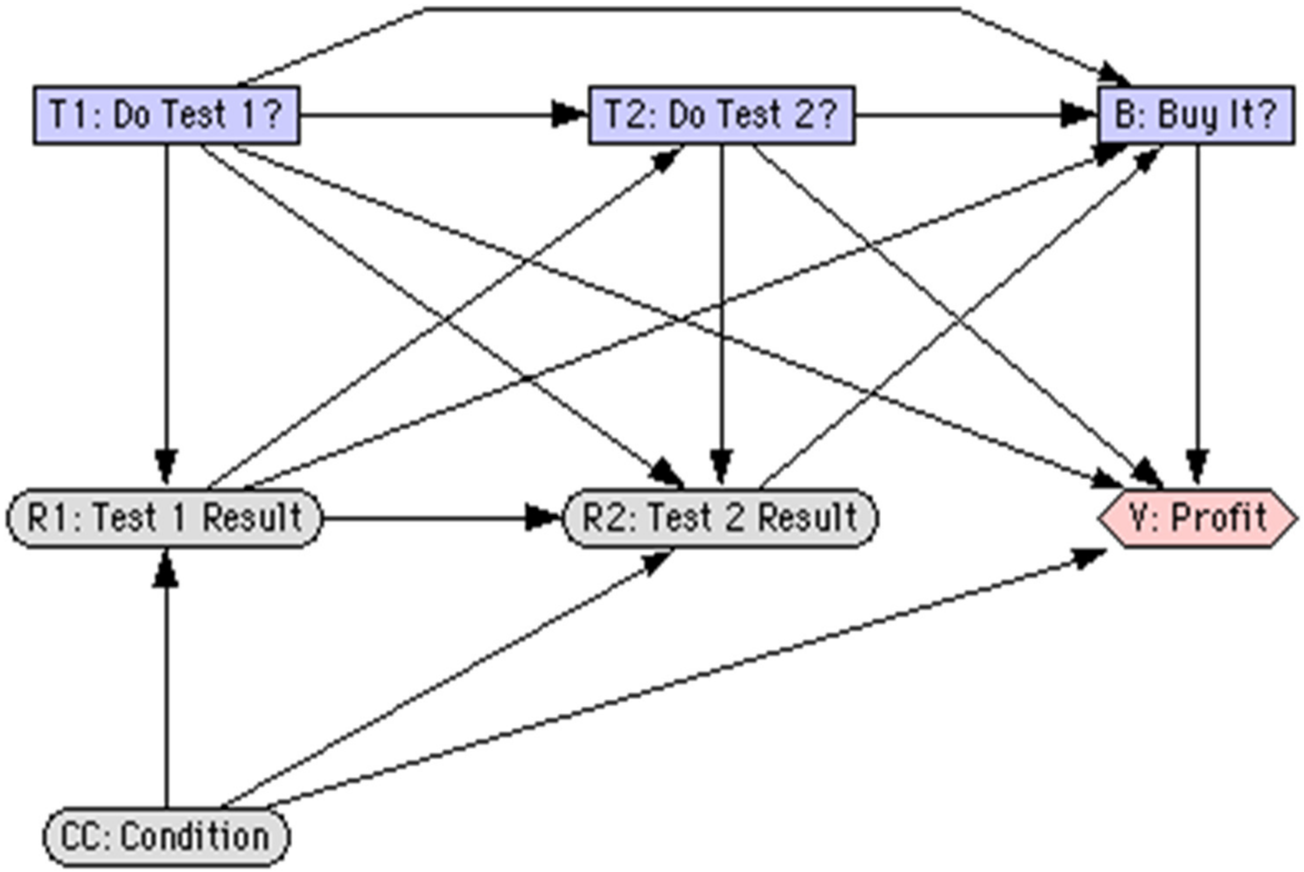
$\Leftarrow$ chosen action

Evaluating a Decision Network



1. Set evidence variables E_1, E_2
   Update distribution over current state S
2. For each possible action a of decision node A
   (a) Set decision node A to a
   (b) For each parent $\{ S' \}$ of utility node U:
      Calculate posterior probability of S
      Here, just $P(S' \mid E_1, E_2, A = a)$
   (c) Calculate expected utility for action a:
      $EU(A \mid E_1, E_2) = \sum_{S'} P(S' \mid E_1, E_2, a) \, U(S')$
3. Choose action $a^* = \arg\max_a \{ EU(a \mid \ldots) \}$
   with highest expected utility

Decision Net: Test/Buy a Car



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 $\{Y_i\}$)
- 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

- Motivation
- What is a Belief Net?
- Inference
- Real-World Applications
- Relation to other Models
- Learning a Belief Net

Deployed Applications (I, II)

Repeat!

- Microsoft: Troubleshooters, Wizards, …
- 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

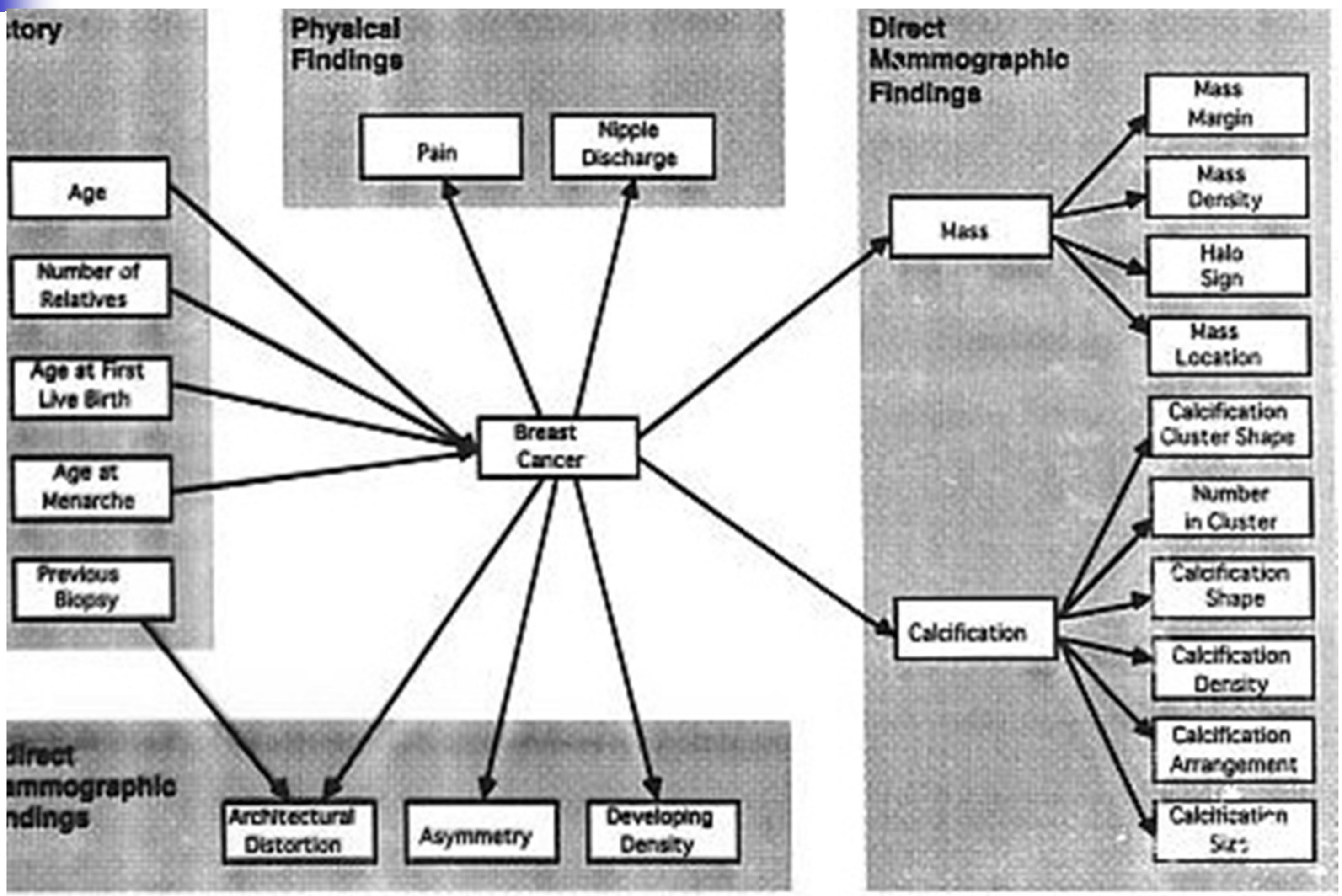
Deployed Applications (III)

- Speech recognition
- Human genome analysis
- Robot mapping
- Identify meteorites to study
- Modeling fMRI data
- Anomaly detection
- Fault diagnosis
- Modeling sensor network data

Deployed Applications (IV)

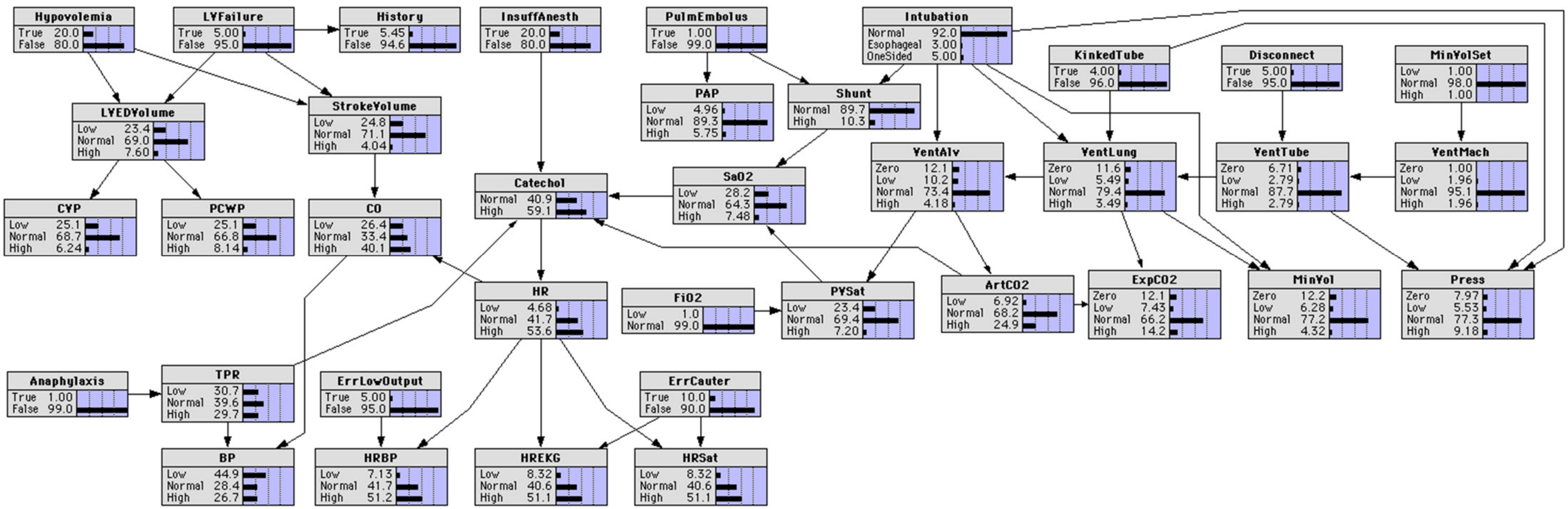
- Lymph-node pathology diagnosis
- Manufacturing control
- Software diagnosis
- Information retrieval
- *Types of tasks*
  - *Classification/Regression*
  - *Sensor Fusion*
  - *Prediction/Forecasting*

MammoNet



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

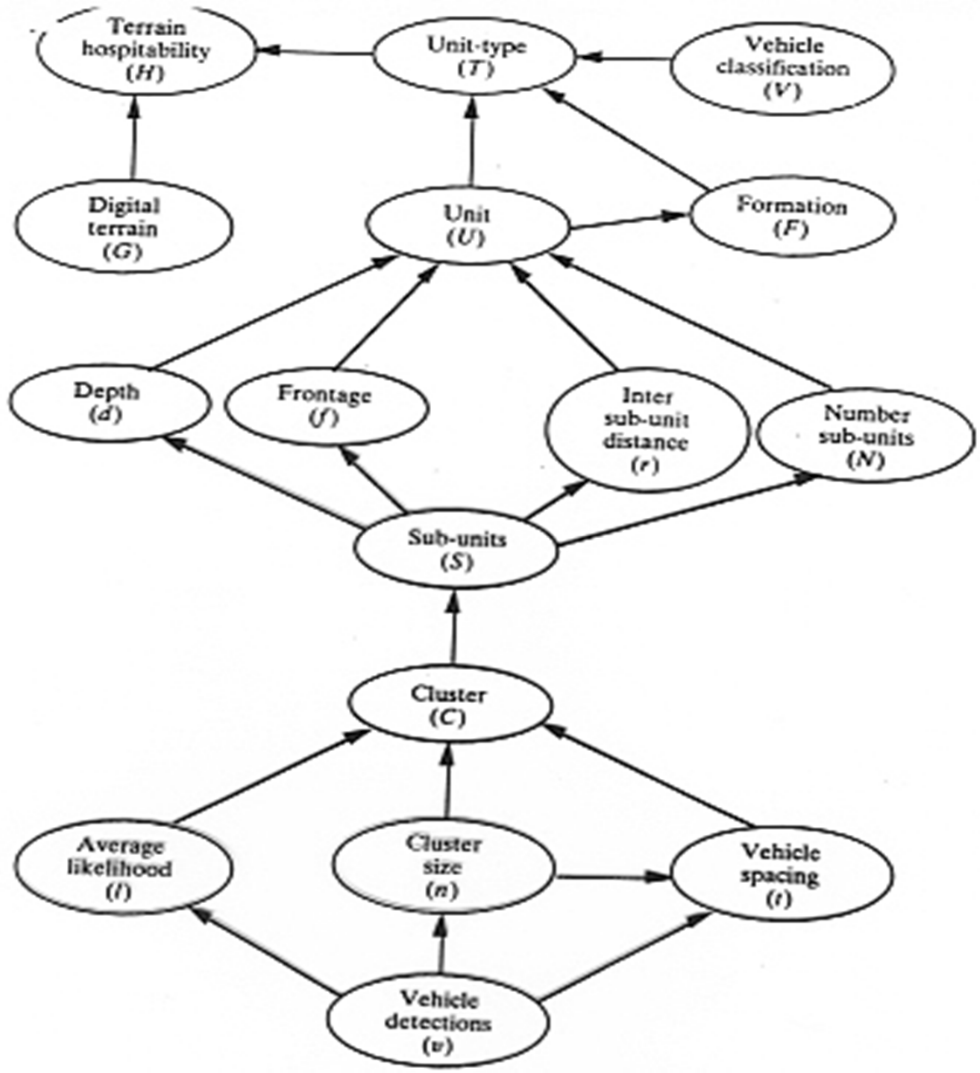
ALARM



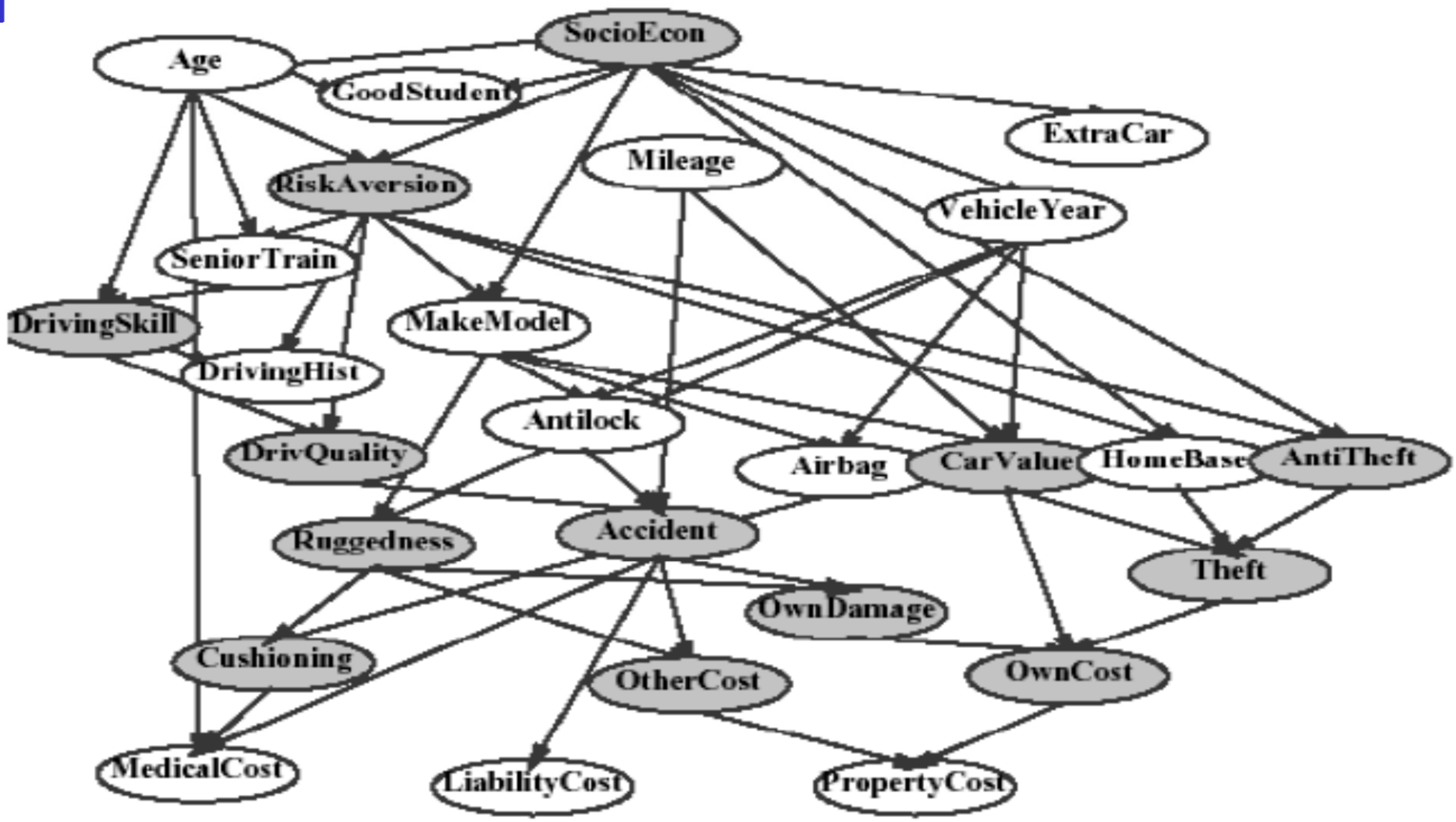
A Logical Alarm Reduction Mechanism

- 8 diagnoses, 16 findings, …

Troup Detection



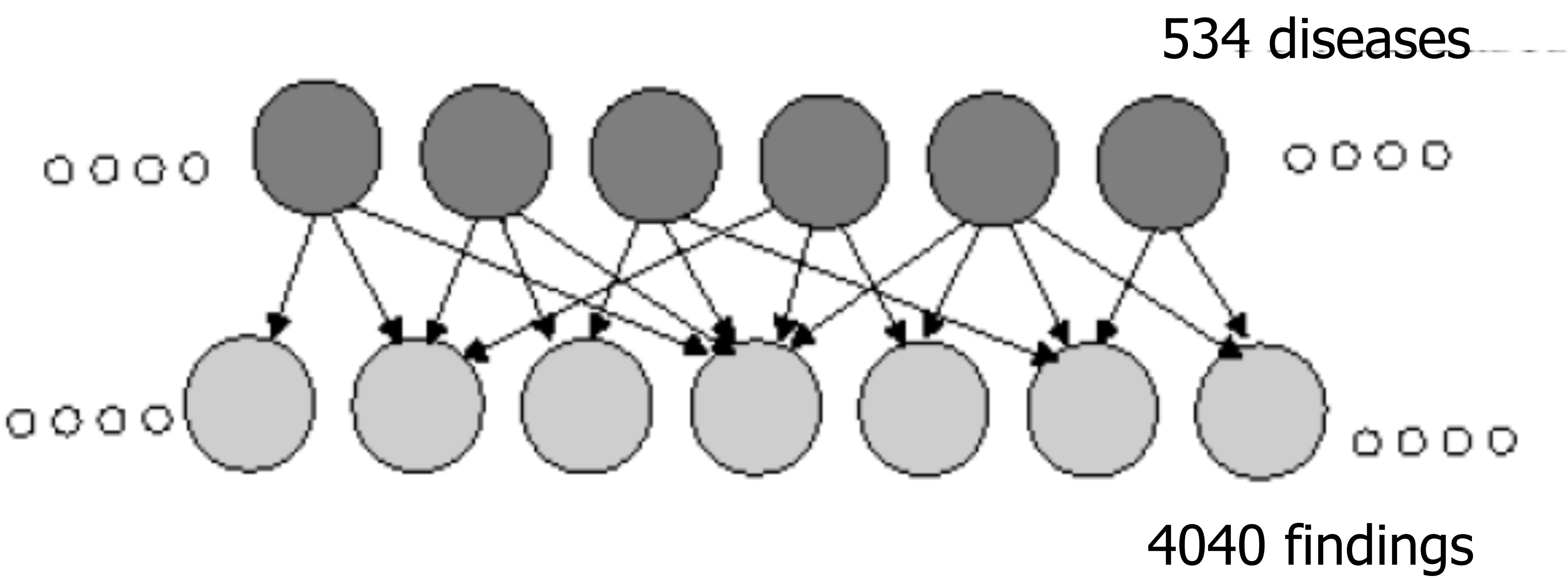
Car Insurance



Predict claim costs (medical, liability) based on application data

QMR-DT

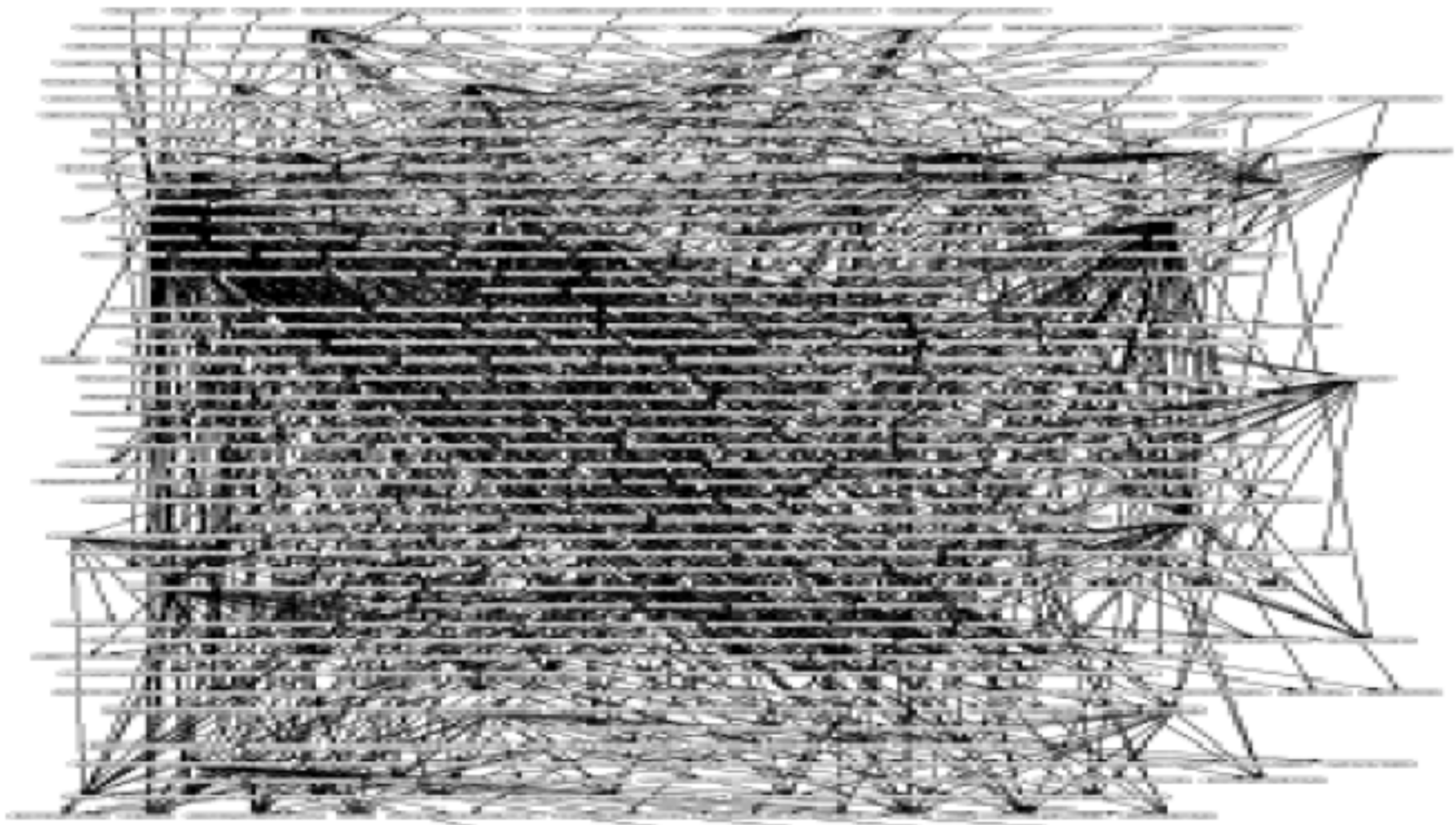
- **Medical diagnosis in internal medicine**
- Bipartite network of disease/findings relations



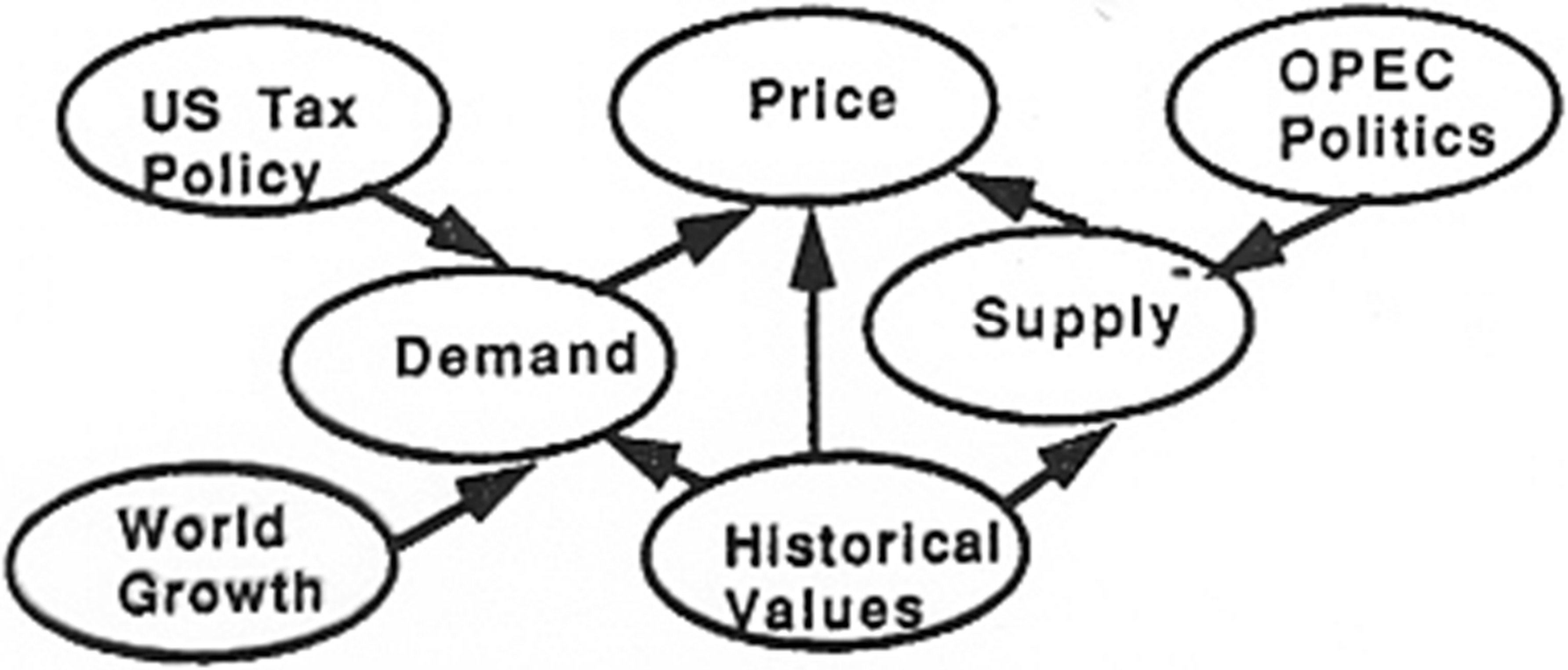
40,740 arcs

CPCS

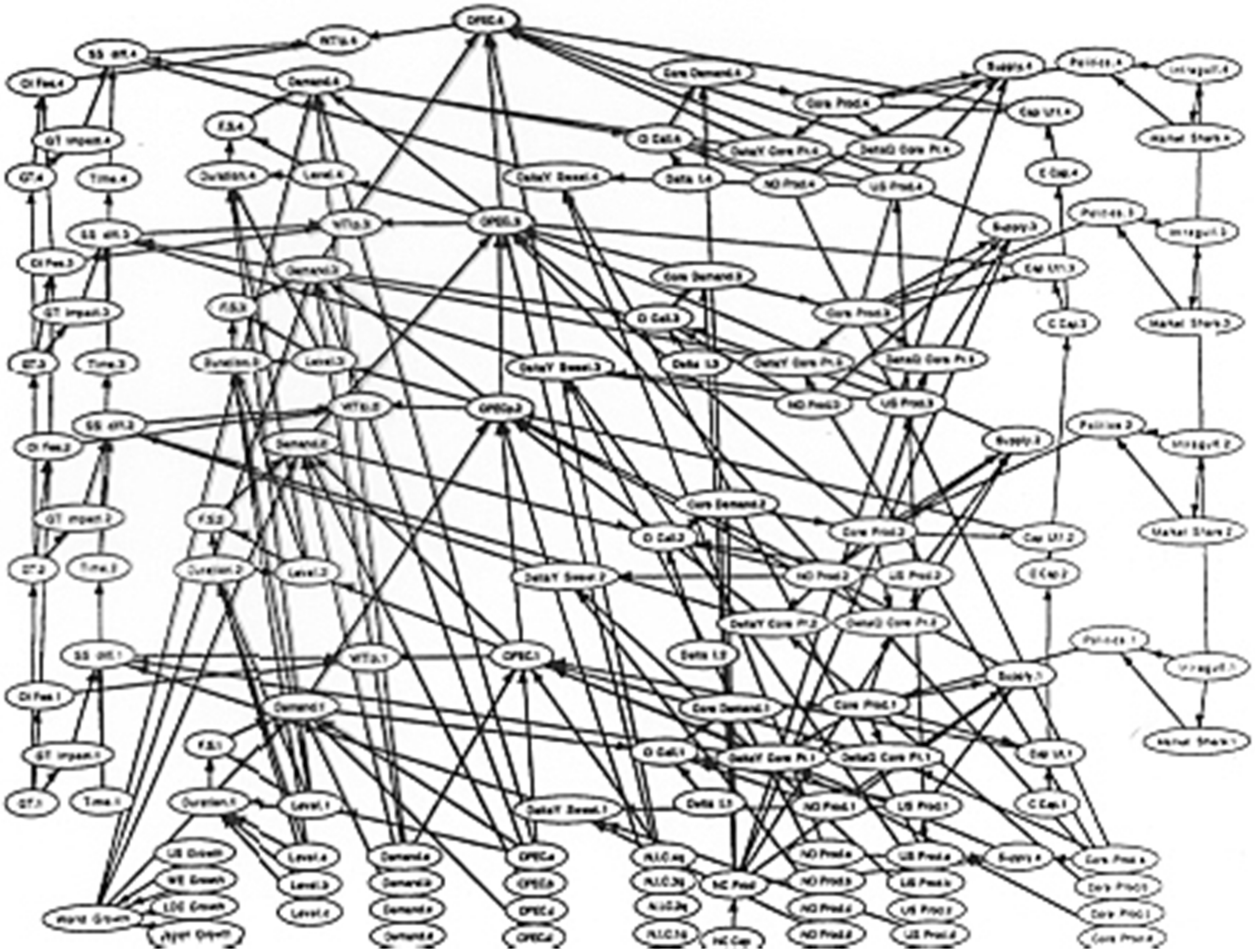
- **C**omputer-based **P**atient **C**ase **S**imulation system
- 422 nodes; 867 arcs



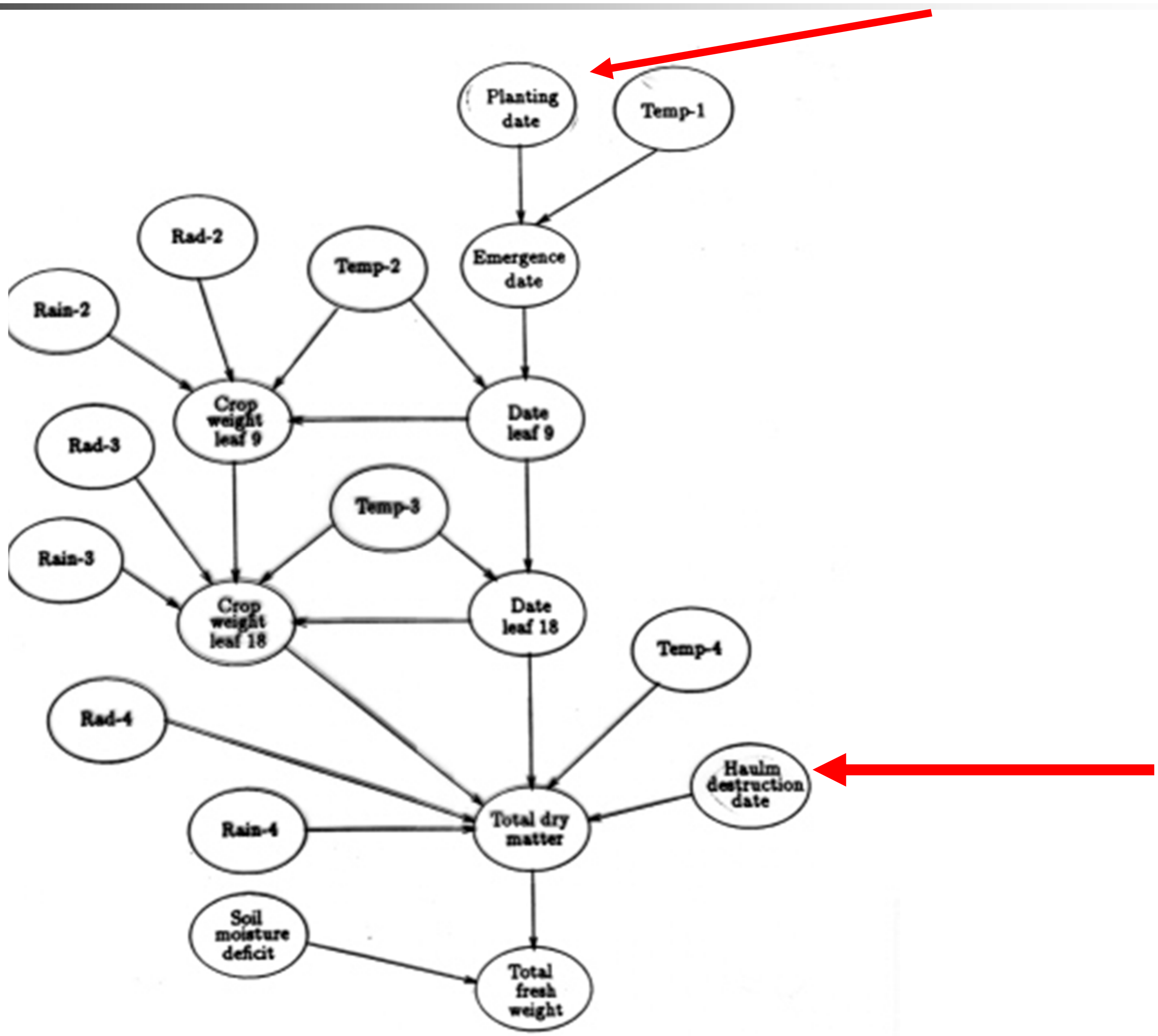
ARCO1: Forecasting Oil Prices



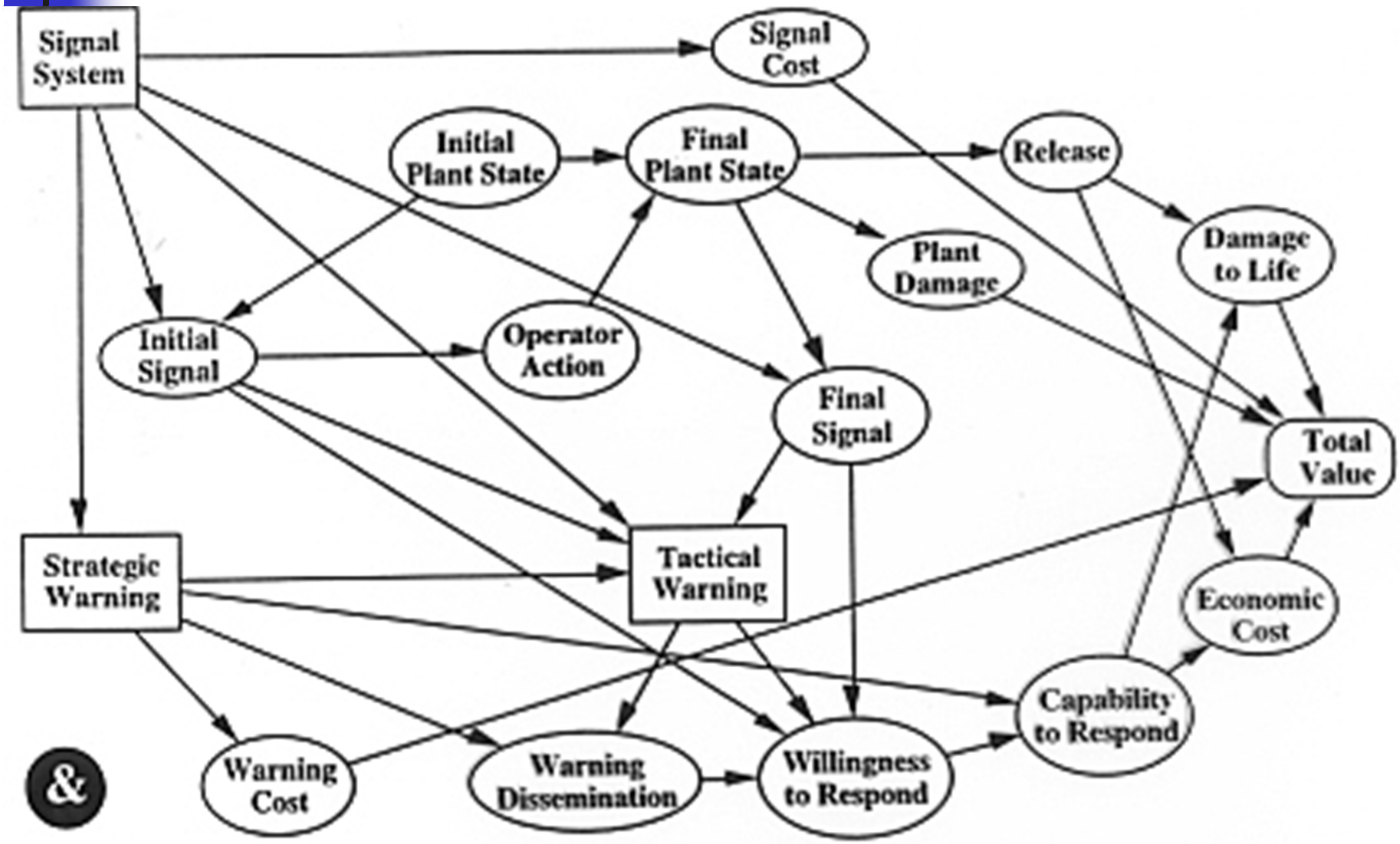
ARCO1: Forecasting Oil Prices



Forecasting Potato Production



Warning System



Outline

- Motivation
- What is a Belief Net?
- Inference
- Real-World Applications
- Relation to other Models
  - Rules
  - Neural Nets
  - Markov Nets
- Learning a Belief Net

Belief Nets vs Rules

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

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

WHY? Easier for people to "input" information **CAUSALLY**
even if use is **DIAGNOSTIC**

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

… 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*"

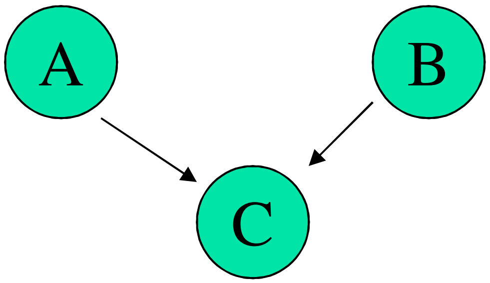
  to FACTOR a distribution

  … explicitly specify dependencies, implicitly independencies…

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

Technical: BNs use DIRECTED arcs

$\Rightarrow$ allow "induced dependencies"

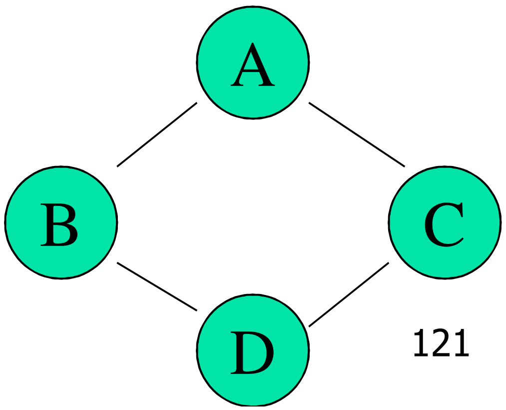


$I(A, \{\}, B)$ "A independent of B, given $\{\}$"

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

MNs use UNDIRECTED arcs

$\Rightarrow$ allow other independencies

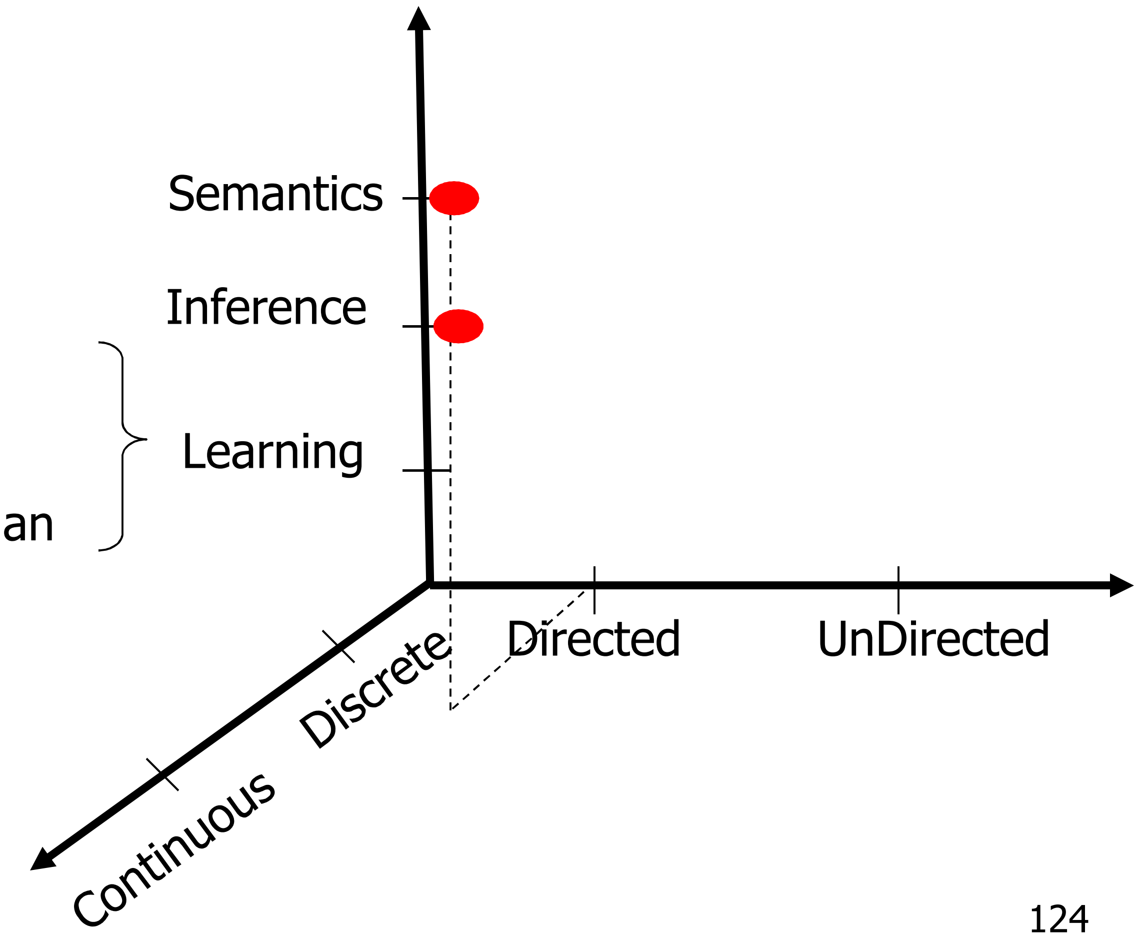


$I(A, BC, D)$ A independent of D, given B, C

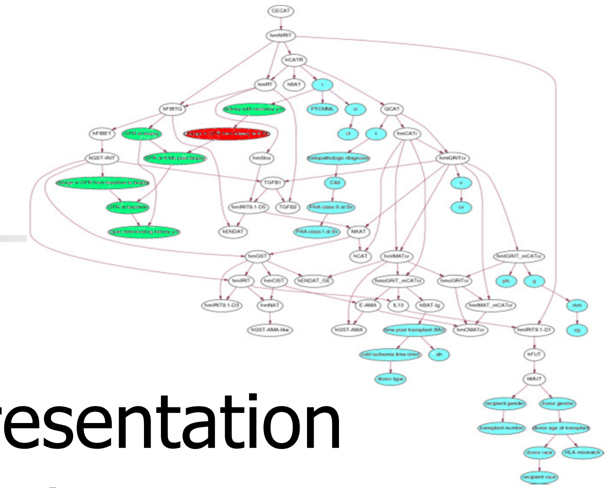
$I(B, AD, C)$ B independent of C, given A, D

Space of Topics

Learning…
• Parameter, Structure
• Data: Complete, Missing
• Framework: Frequentist, Bayesian



Summary



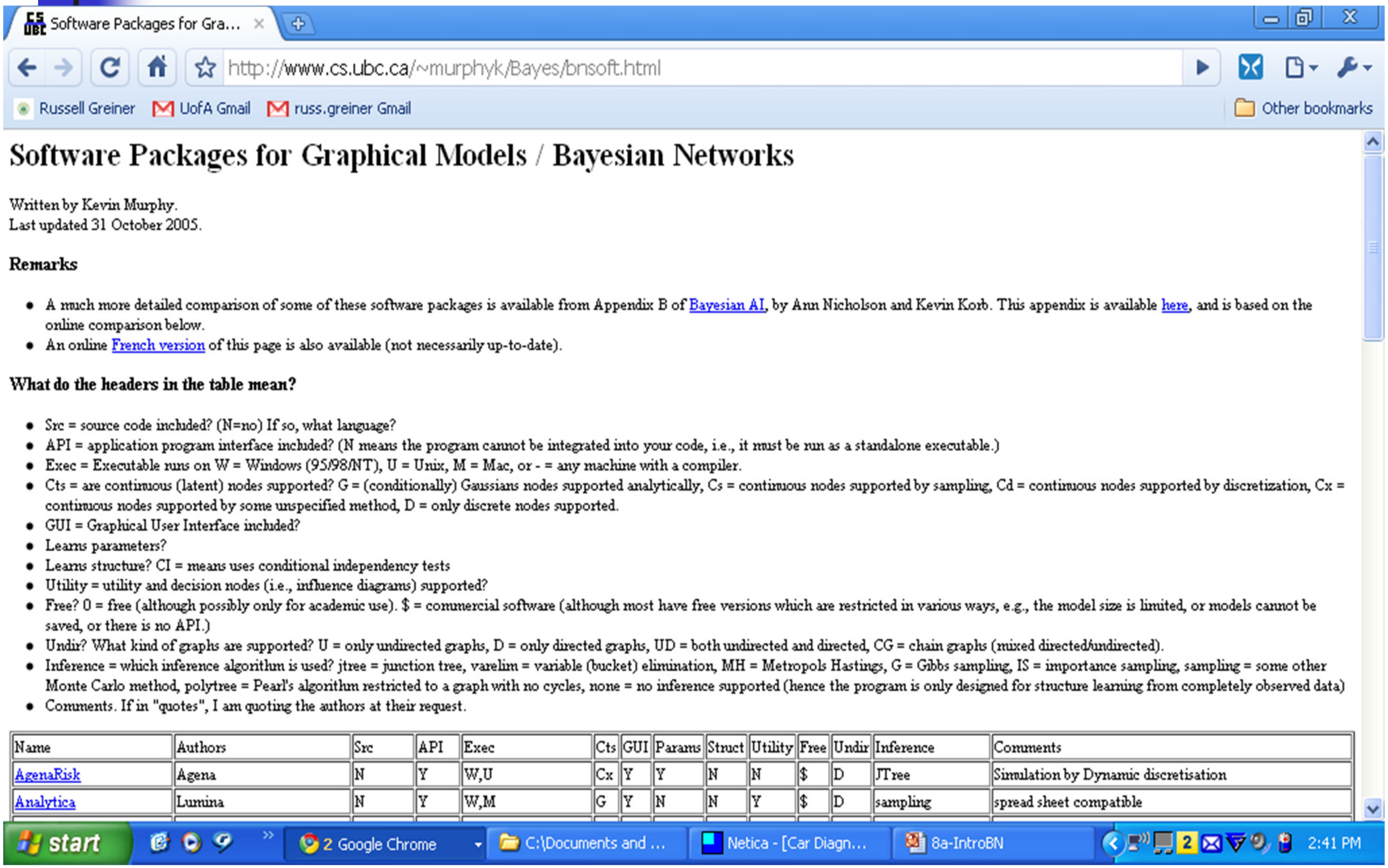
- Necessary to use Probabilistic Representation
  - … use *connections*… just *some* connections
  - Factored Distribution
  - ⇒ Belief Nets
- Belief Nets *are PROVEN TECHNOLOGY*
  - Lots of deployed applications
    - Medical Diagnosis, DSS for complex machines, …
  - Forecasting, Modeling, InfoRetrieval…
- Remaining Challenges:
  - Efficient inference
  - Effective learning

Resources

- General Webpage
  - http://webdocs.cs.ualberta.ca/~greiner/bn.html
- Texts:
  - Russell/Norvig
  - Koller/Friedman
- Software
  - Murphy's summary
  - Netica
- UofA Courses:
  - Cmput499/659 web-page
  - Cmput466 web-page
- Coursera course
  - https://www.coursera.org/course/pgm

http://www.cs.ubc.ca/~murphyk/Bayes/bnsoft.html

Software Systems



Software Packages for Graphical Models / Bayesian Networks

Written by Kevin Murphy.
Last updated 31 October 2005.

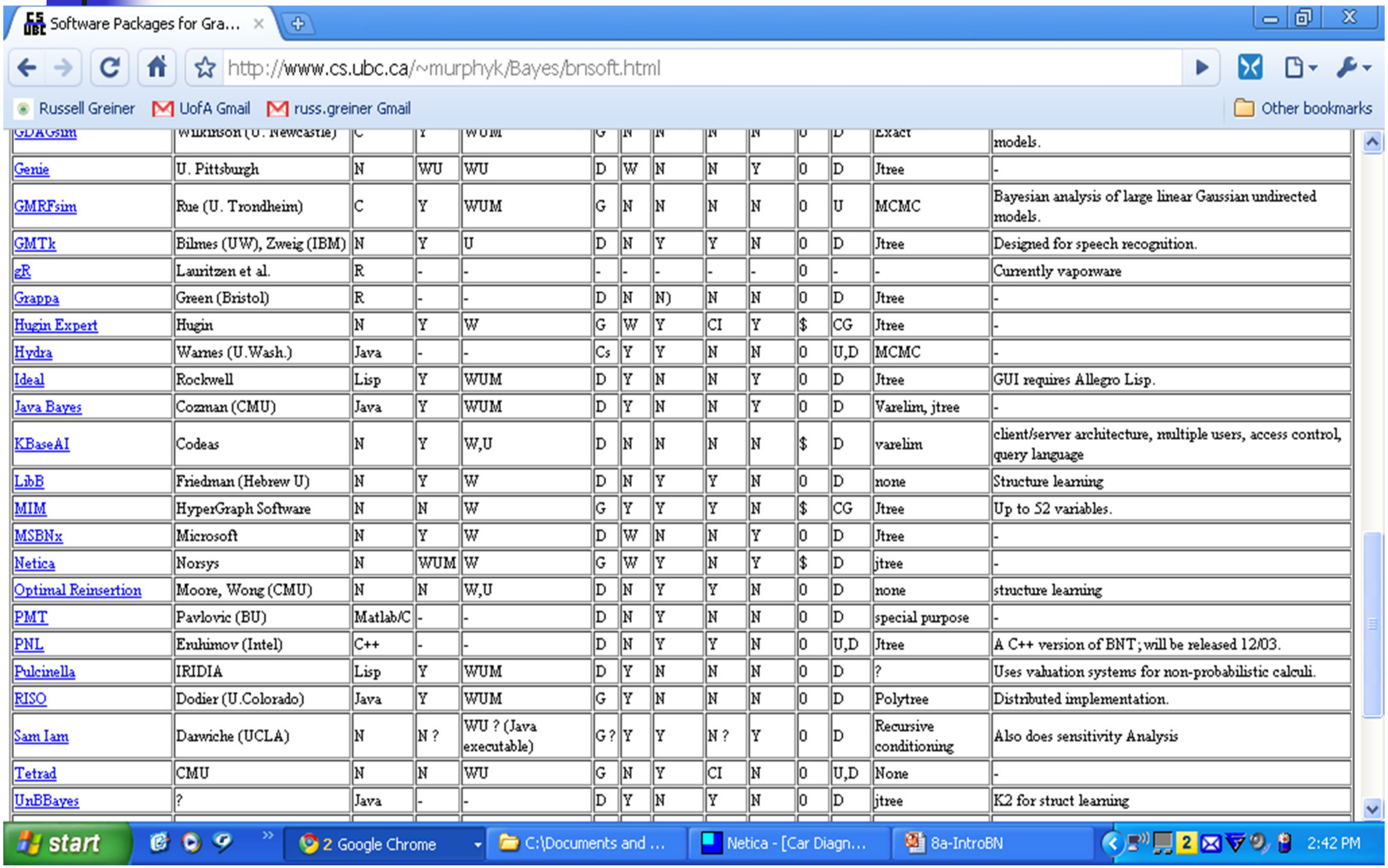
Remarks

- A much more detailed comparison of some of these software packages is available from Appendix B of Bayesian AI, by Ann Nicholson and Kevin Korb. This appendix is available here, and is based on the online comparison below.
- An online French version of this page is also available (not necessarily up-to-date).

What do the headers in the table mean?

- Src = source code included? (N=no) If so, what language?
- API = application program interface included? (N means the program cannot be integrated into your code, i.e., it must be run as a standalone executable.)
- Exec = Executable runs on W = Windows (95/98/NT), U = Unix, M = Mac, or - = any machine with a compiler.
- Cts = are continuous (latent) nodes supported? G = (conditionally) Gaussians nodes supported analytically, Cs = continuous nodes supported by sampling, Cd = continuous nodes supported by discretization, Cx = continuous nodes supported by some unspecified method, D = only discrete nodes supported.
- GUI = Graphical User Interface included?
- Learns parameters?
- Learns structure? CI = means uses conditional independency tests
- Utility = utility and decision nodes (i.e., influence diagrams) supported?
- Free? 0 = free (although possibly only for academic use). $ = commercial software (although most have free versions which are restricted in various ways, e.g., the model size is limited, or models cannot be saved, or there is no API.)
- Undir? What kind of graphs are supported? U = only undirected graphs, D = only directed graphs, UD = both undirected and directed, CG = chain graphs (mixed directed/undirected).
- Inference = which inference algorithm is used? jtree = junction tree, varelim = variable (bucket) elimination, MH = Metropols Hastings, G = Gibbs sampling, IS = importance sampling, sampling = some other Monte Carlo method, polytree = Pearl's algorithm restricted to a graph with no cycles, none = no inference supported (hence the program is only designed for structure learning from completely observed data)
- Comments. If in "quotes", I am quoting the authors at their request.

Name	Authors	Src	API	Exec	Cts	GUI	Params	Struct	Utility	Free	Undir	Inference	Comments
AgenaRisk	Agena	N	Y	W,U	Cx	Y	Y	N	N	$	D	JTree	Simulation by Dynamic discretisation
Analytica	Lumina	N	Y	W,M	G	Y	N	N	Y	$	D	sampling	spread sheet compatible



GDAGsim	Wilkinson (U. Newcastle)	C	Y	WUM	G	N	N	N	N	0	D	Exact	models.
Genie	U. Pittsburgh	N	WU	WU	D	W	N	N	Y	0	D	Jtree	-
GMRFsim	Rue (U. Trondheim)	C	Y	WUM	G	N	N	N	N	0	U	MCMC	Bayesian analysis of large linear Gaussian undirected models.
GMTk	Bilmes (UW), Zweig (IBM)	N	Y	U	D	N	Y	Y	N	0	D	Jtree	Designed for speech recognition.
gR	Lauritzen et al.	R	-	-	-	-	-	-	-	0	-	-	Currently vaporware
Grappa	Green (Bristol)	R	-	-	D	N	N)	N	N	0	D	Jtree	-
Hugin Expert	Hugin	N	Y	W	G	W	Y	CI	Y	$	CG	Jtree	-
Hydra	Warnes (U.Wash.)	Java	-	-	Cs	Y	Y	N	N	0	U,D	MCMC	-
Ideal	Rockwell	Lisp	Y	WUM	D	Y	N	N	Y	0	D	Jtree	GUI requires Allegro Lisp.
Java Bayes	Cozman (CMU)	Java	Y	WUM	D	Y	N	N	Y	0	D	Varelim, jtree	-
KBaseAI	Codeas	N	Y	W,U	D	N	N	N	N	$	D	varelim	client/server architecture, multiple users, access control, query language
LibB	Friedman (Hebrew U)	N	Y	W	D	N	Y	Y	N	0	D	none	Structure learning
MIM	HyperGraph Software	N	N	W	G	Y	Y	Y	N	$	CG	Jtree	Up to 52 variables.
MSBNx	Microsoft	N	Y	W	D	W	N	N	Y	0	D	Jtree	-
Netica	Norsys	N	WUM	W	G	W	Y	N	Y	$	D	jtree	-
Optimal Reinsertion	Moore, Wong (CMU)	N	N	W,U	D	N	Y	Y	N	0	D	none	structure learning
PMT	Pavlovic (BU)	Matlab/C	-	-	D	N	Y	N	N	0	D	special purpose	-
PNL	Eruhimov (Intel)	C++	-	-	D	N	Y	Y	N	0	U,D	Jtree	A C++ version of BNT; will be released 12/03.
Pulcinella	IRIDIA	Lisp	Y	WUM	D	Y	N	N	N	0	D	?	Uses valuation systems for non-probabilistic calculi.
RISO	Dodier (U.Colorado)	Java	Y	WUM	G	Y	N	N	N	0	D	Polytree	Distributed implementation.
Sam Iam	Darwiche (UCLA)	N	N ?	WU ? (Java executable)	G ?	Y	Y	N ?	Y	0	D	Recursive conditioning	Also does sensitivity Analysis
Tetrad	CMU	N	N	WU	G	N	Y	CI	N	0	U,D	None	-
UnBBayes	?	Java	-	-	D	Y	N	Y	N	0	D	jtree	K2 for struct learning