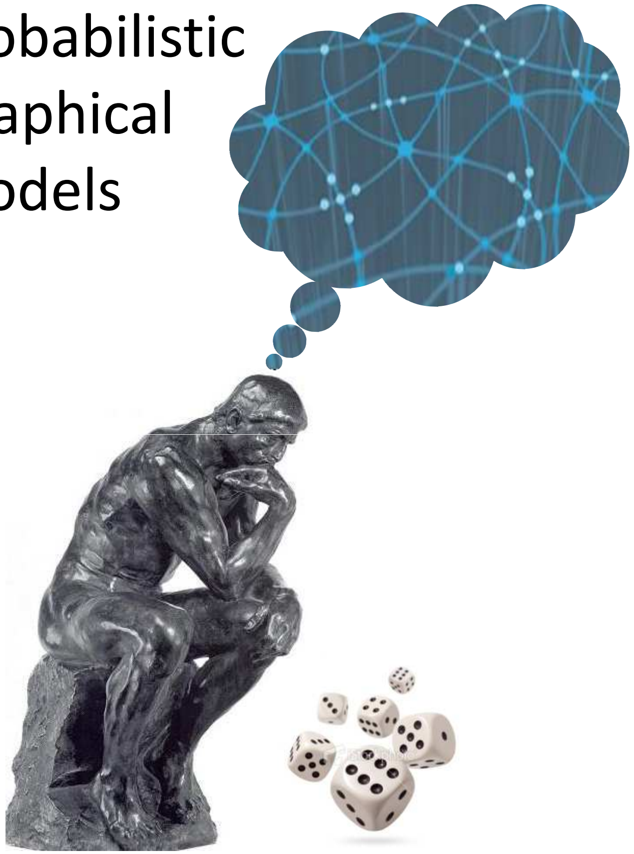
Probabilistic Graphical Models



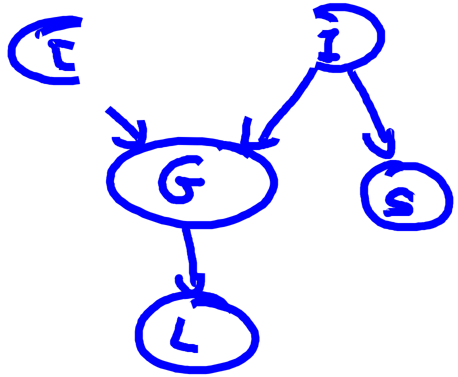
Representation

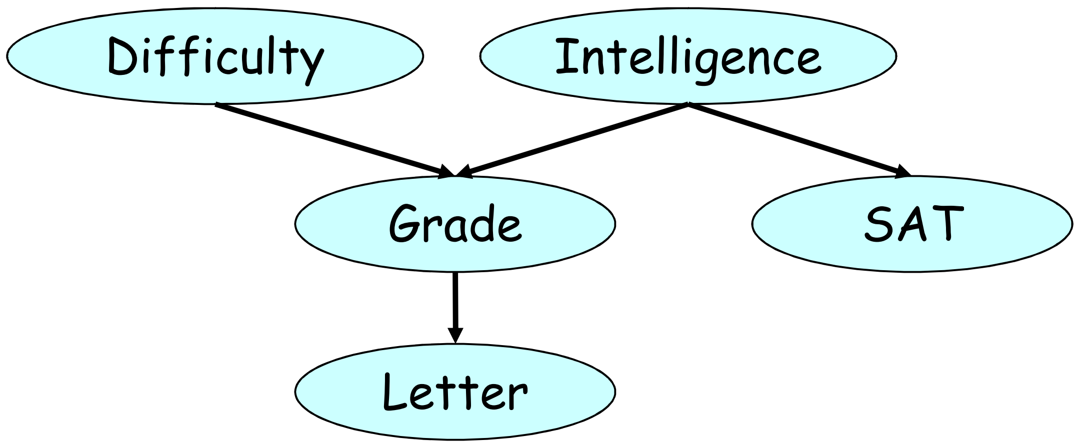
Bayesian Networks

Semantics & Factorization

- Grade
- Course Difficulty
- Student Intelligence
- Student SAT
- Reference Letter

P(G,D,I,S,L)




Difficulty
Intelligence
Grade
SAT
Letter

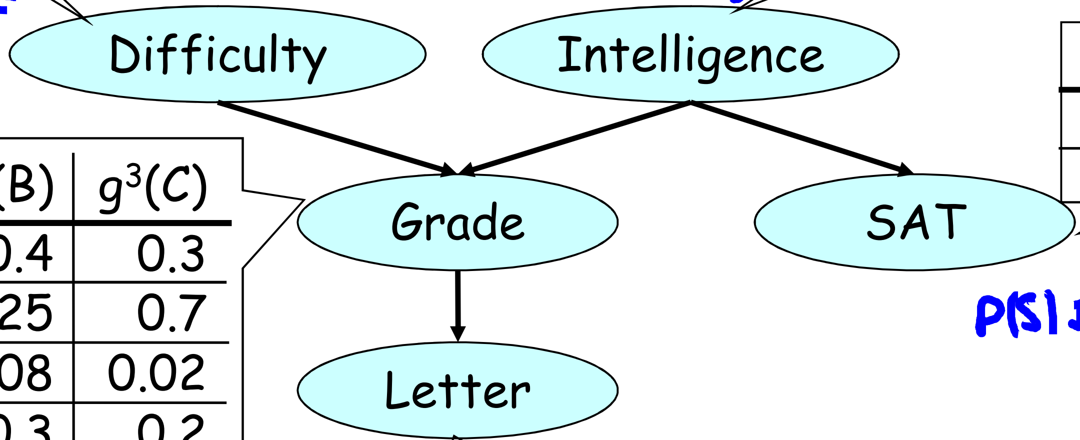
CPD = cond. prob. dist.

P(D)

d^0	d^1
0.6	0.4

P(I)

i^0	i^1
0.7	0.3



→ P(G | I, D)

	g^1(A)	g^2(B)	g^3(C)
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

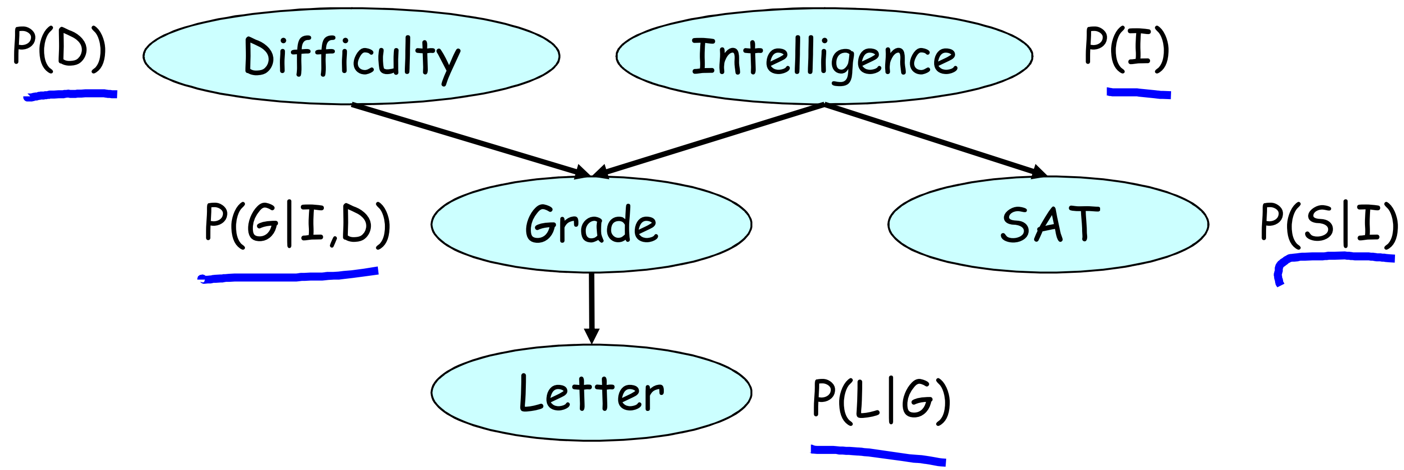
P(S | I)

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

P(L | G)

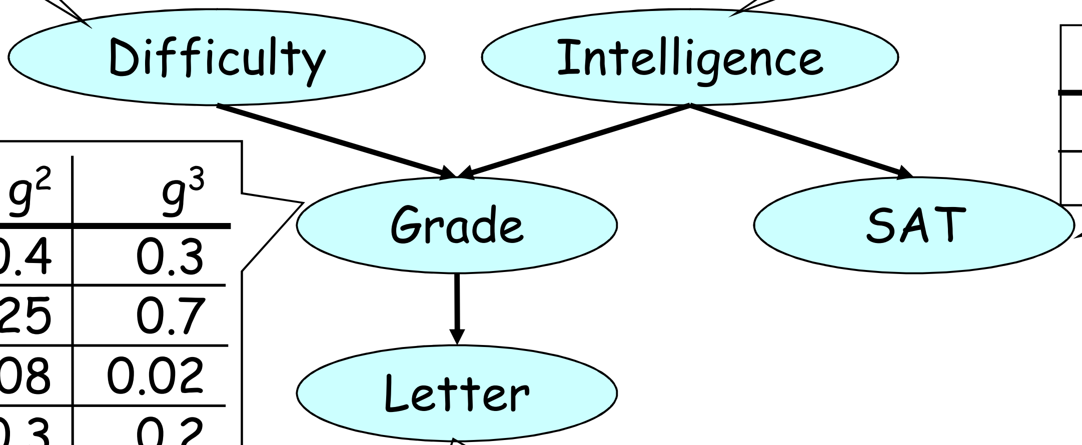
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Chain Rule for Bayesian Networks



$$\underbrace{P(D,I,G,S,L)} = P(D)\,P(I)\,P(G|I,D)\,P(S|I)\,P(L|G)$$

Distribution defined as a product of factors!



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

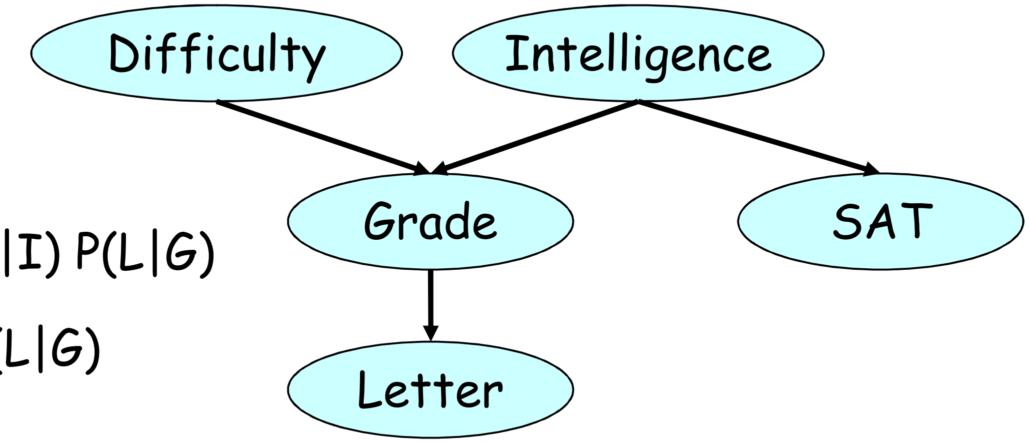
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$P(d^0, i^1, g^3, s^1, l^1) =$

$0.6 \times 0.3 \times 0.02 \times 0.01 \times 0.8$

Defining a joint distribution

- What is the joint distribution P(D,I,G,S,L)?



- ○ P(D) P(I) P(G|I) P(G|D) P(S|I) P(L|G)
- ○ P(D) P(I) P(G|I,D) P(S|I) P(L|G)
- ○ P(D) P(I) P(G) P(S) P(L)
- ○ P(D|G) P(I|D) P(S|I) P(G|L,I,D) P(L|G)

Bayesian Network

- A Bayesian network is:
  - A directed acyclic graph (DAG) G whose nodes represent the random variables $X_1,\ldots,X_n$
  - For each node X_i a CPD $P(X_i \mid Par_G(X_i))$
- The BN represents a joint distribution via the chain rule for Bayesian networks

$$\boxed{P(X_1,\ldots,X_n) = \Pi_i P(X_i \mid Par_G(X_i))}$$

BN Is a Legal Distribution: $P \geq 0$

P is a product of CPDs

CPDs are non-negative

BN Is a Legal Distribution: $\sum P = 1$

$$\sum_{D,I,G,S,L} P(D,I,G,S,L) = \sum_{D,I,G,S,L} \underbrace{P(D)\,P(I)\,P(G|I,D)\,P(S|I)\,P(L|G)}_{\text{chain rule}}$$

$$= \sum_{D,I,G,S} P(D)\,P(I)\,P(G|I,D)\,P(S|I)\,\overset{=1}{\cancel{\sum_{L} P(L|G)}}$$

$$= \sum_{D,I,G,S} P(D)\,P(I)\,P(G|I,D)\,P(S|I)$$

$$= \sum_{D,I,G} P(D)\,P(I)\,P(G|I,D)\,\overset{1}{\cancel{\sum_{S} P(S|I)}}$$

$$= \sum_{D,I} P(D)\,P(I)\,\underset{=1}{\cancel{\sum_{G} P(G|I,D)}}$$

What is the value of $\sum_L P(L \mid G)$

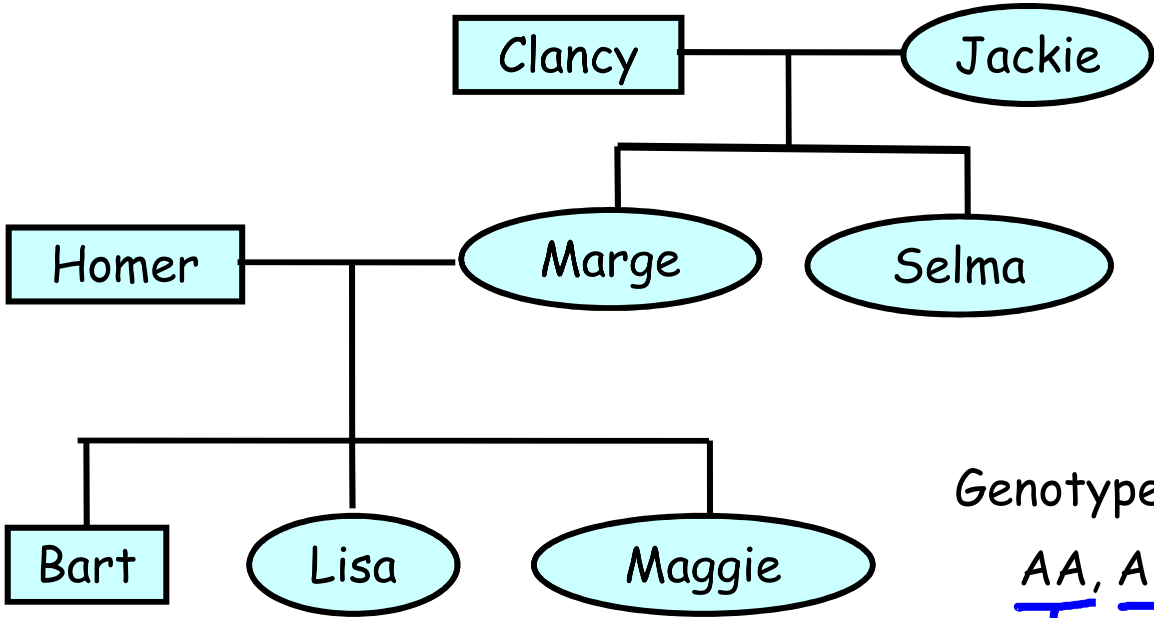
- 1
- P(L)
- P(G)
- None of the above

P Factorizes over G

- Let G be a graph over $X_1,\ldots,X_n$.
- P factorizes over G if

$$P(X_1,\ldots,X_n) = \prod_i P(X_i \mid Par_G(X_i))$$

Genetic Inheritance



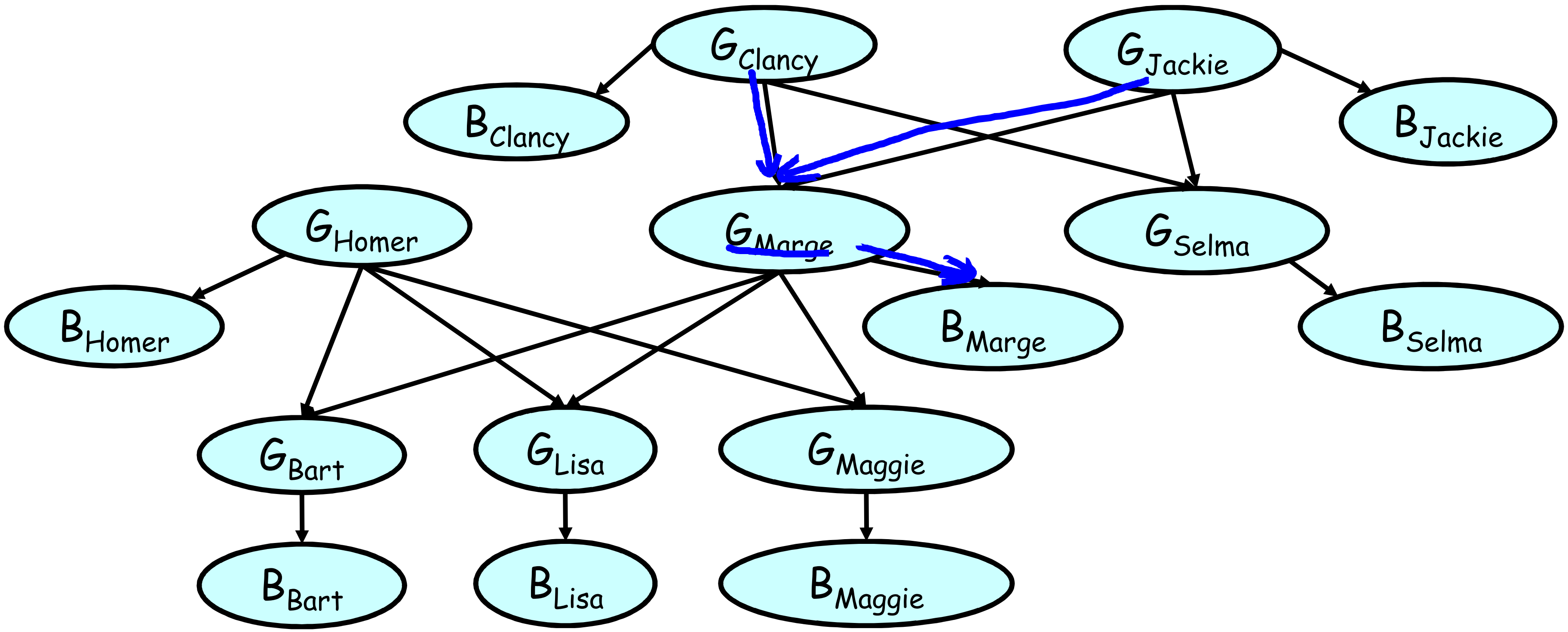
Genotype

AA, AB, AO, BO, BB, OO

Phenotype

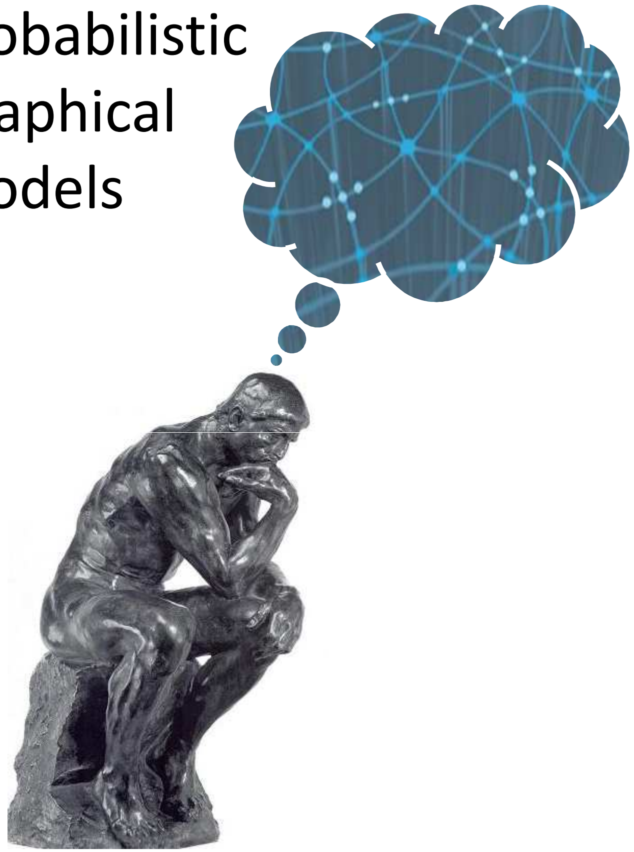
A, B, AB, O

BNs for Genetic Inheritance



Daphne Koller

Probabilistic
Graphical
Models



Representation

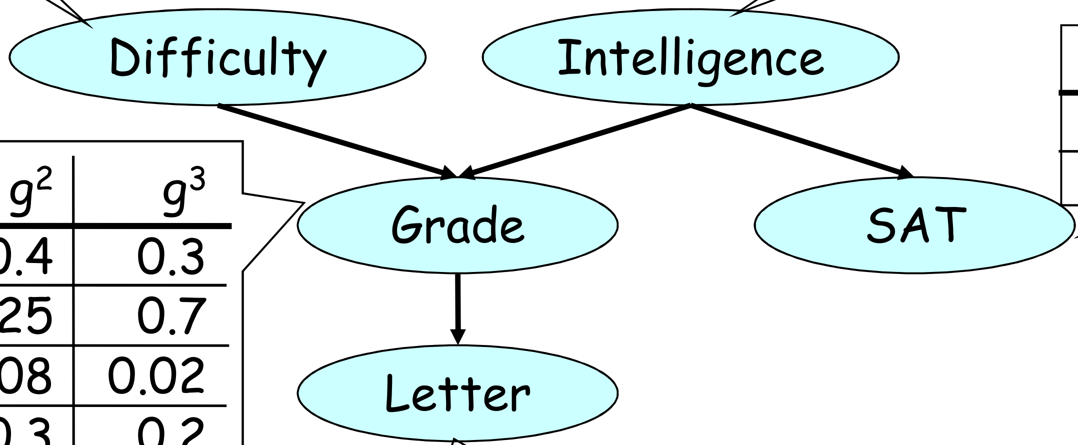
Bayesian Networks

Reasoning Patterns

The Student Network

d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3



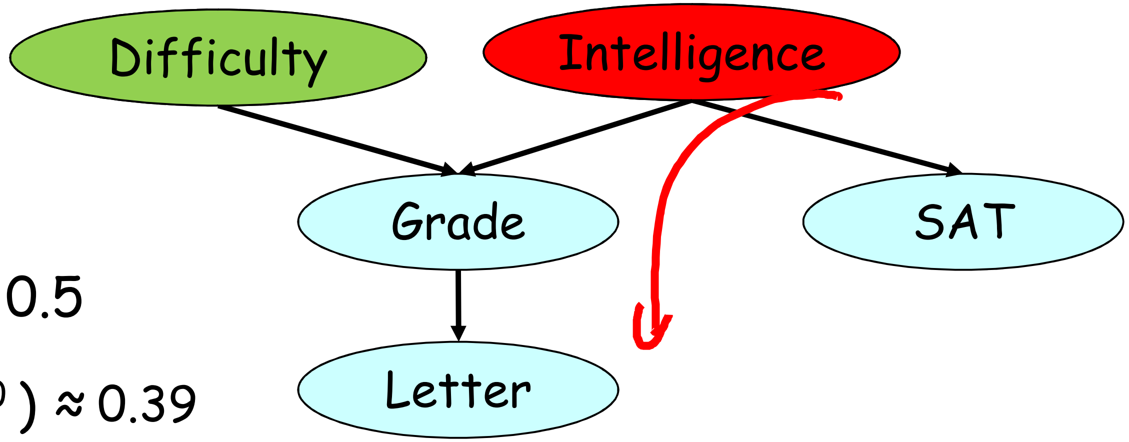
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Daphne Koller

Causal Reasoning



$P(l^1) \approx 0.5$

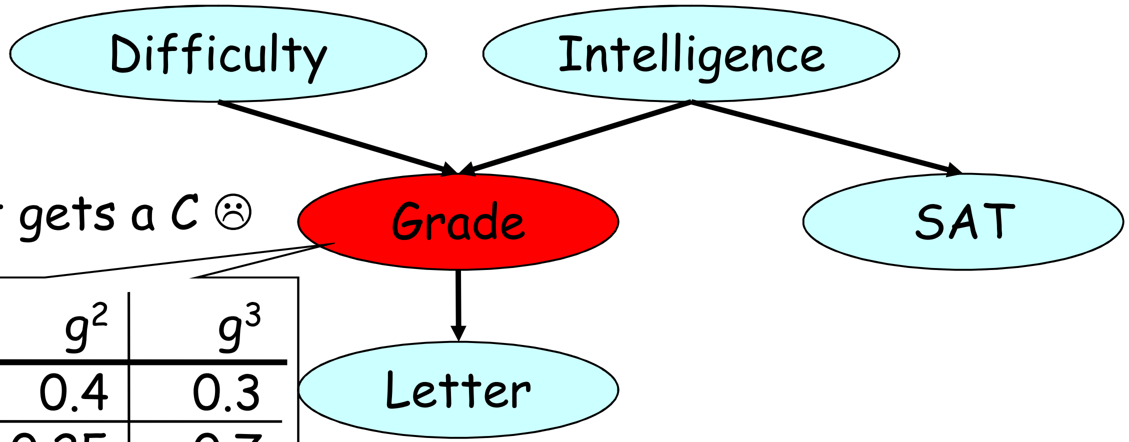
$P(l^1 \mid i^0) \approx 0.39$

$P(l^1 \mid i^0, d^0) \approx 0.51$

Evidential Reasoning

$P(d^1) = 0.4$ $P(i^1) = 0.3$

$P(d^1 \mid g^3) \approx 0.63$ $P(i^1 \mid g^3) \approx 0.08$



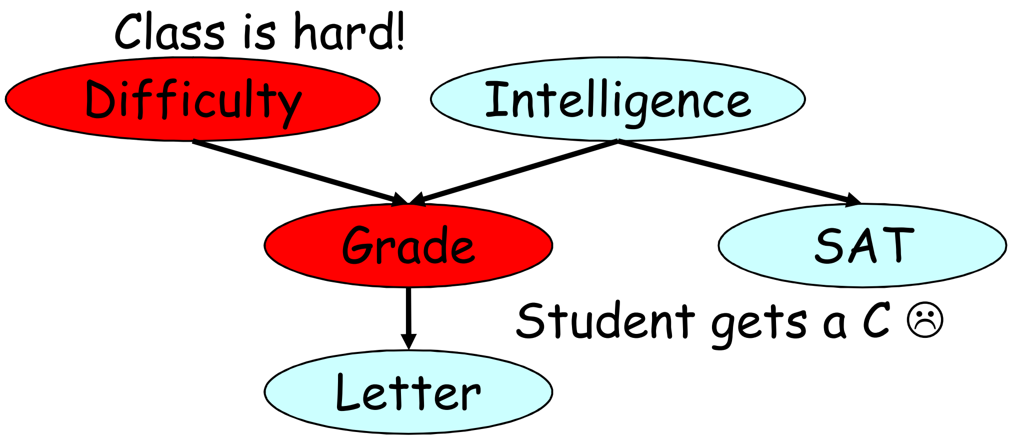
Student gets a C ☹

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

We find out that class is hard

- What happens to the posterior probability of high intelligence?

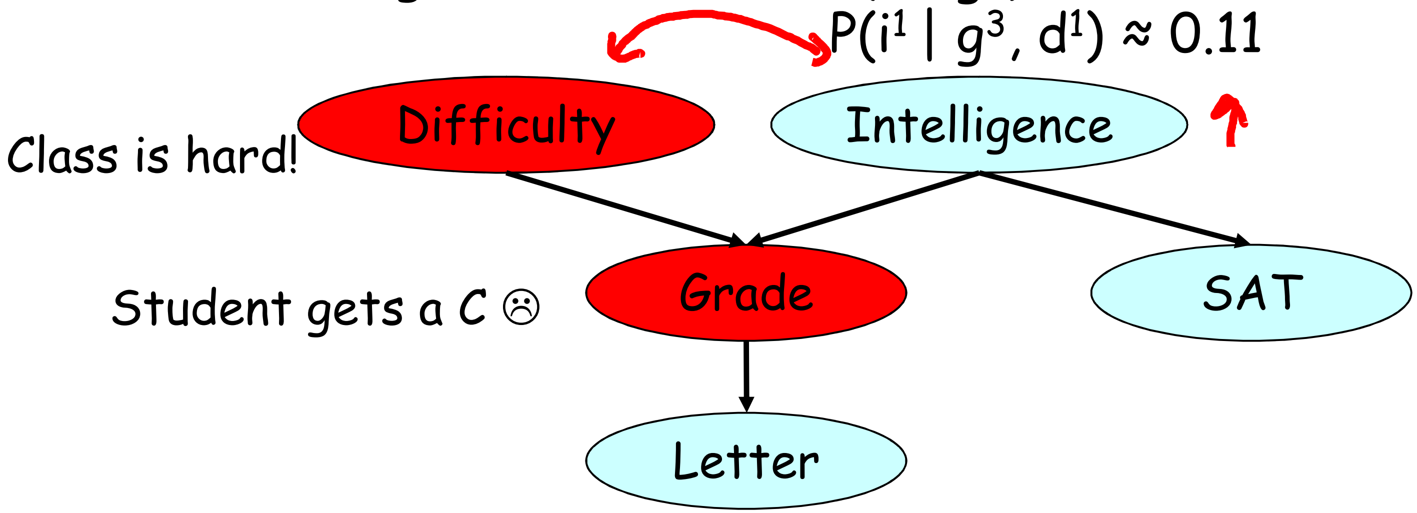
- Goes up
- Goes down
- Doesn't change
- We can't know



Intercausal Reasoning

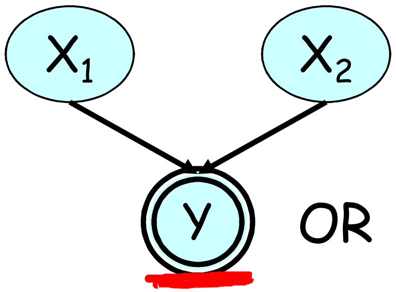
$P(d^1) = 0.4$
$P(d^1 \mid g^3) \approx 0.63$

$P(i^1) = 0.3$
$P(i^1 \mid g^3) \approx 0.08$
$P(i^1 \mid g^3, d^1) \approx 0.11$



Intercausal Reasoning Explained

explaining away



Y=1

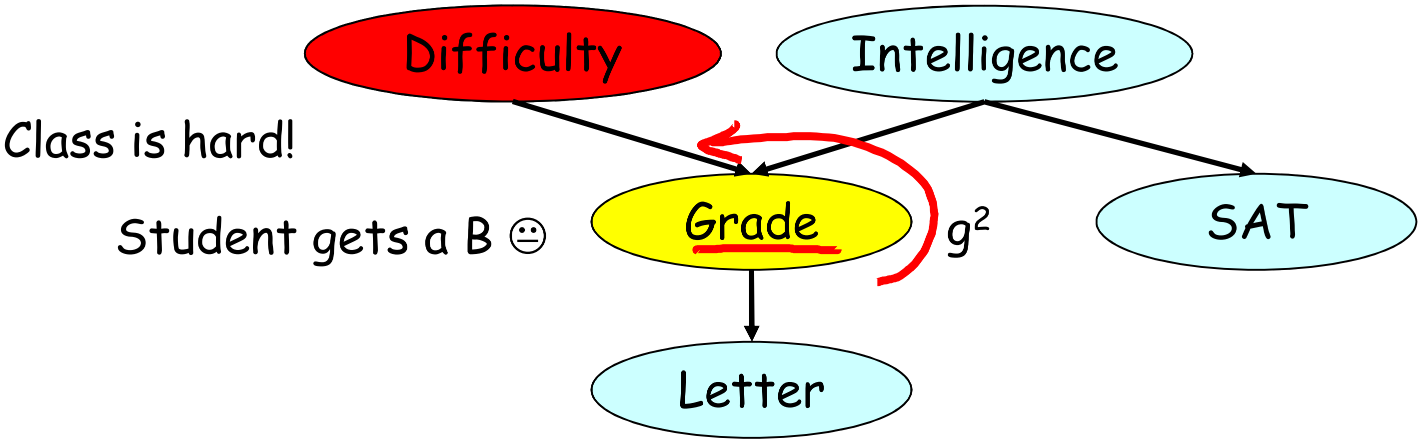
X_1	X_2	Y	**Prob**	
~~0~~	~~0~~	~~0~~	~~0.25~~	
~~0~~	~~1~~	~~1~~	~~0.25~~	0.33
1	0	1	0.25	0.33
1	1	1	0.25	0.33

[illegible] $P(X_1=1)=\frac{2}{3}$ $P(X_2=1)=\frac{2}{3}$

condition $X_1=1$ $P(X_2=1 \mid X_1=1)=0.5$

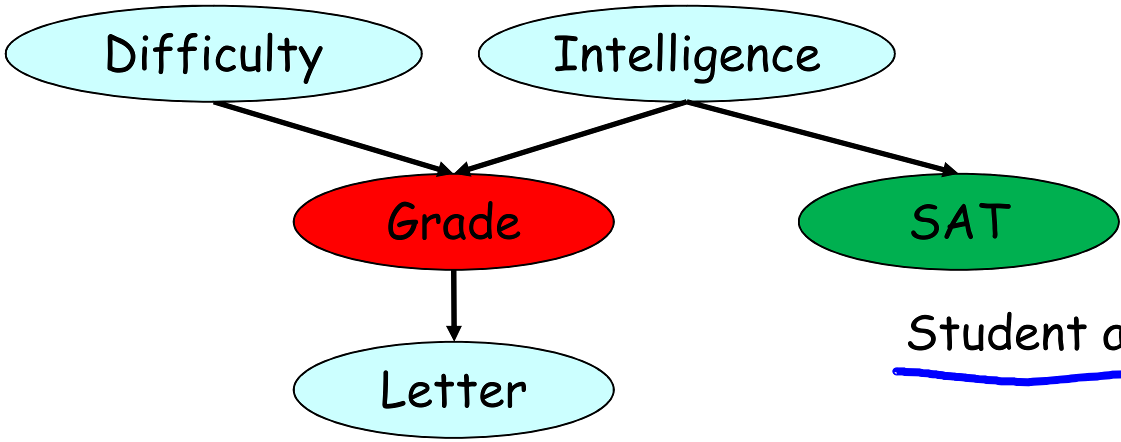
Intercausal Reasoning II

$P(i^1) = 0.3$
$P(i^1 \mid g^2) \approx 0.175$
$P(i^1 \mid g^2, d^1) \approx 0.34$



Student Aces the SAT

- What happens to the posterior probability that the class is hard?

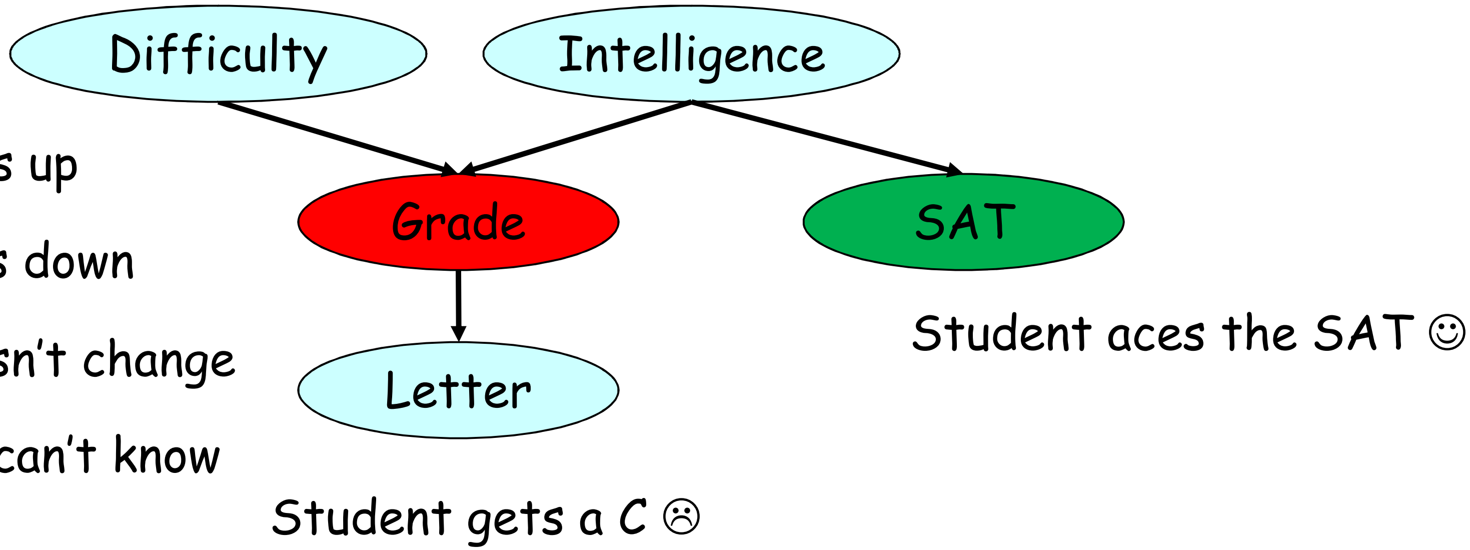


Student aces the SAT ☺

Student gets a C ☹

Student Aces the SAT

- What happens to the posterior probability that the class is hard?

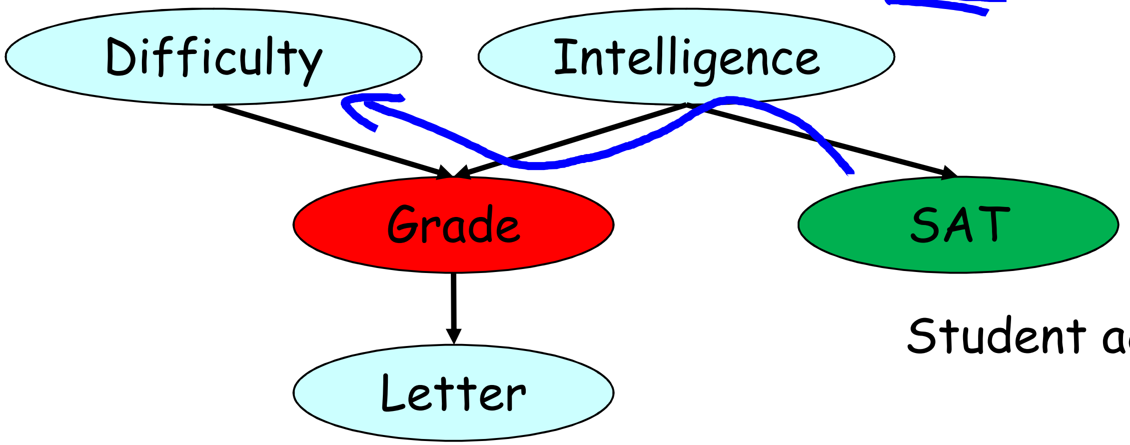


- Goes up
- Goes down
- Doesn't change
- We can't know

Student Aces the SAT

$P(d^1) = 0.4$
$P(d^1 \mid g^3) \approx 0.63$
$P(d^1 \mid g^3, s^1) \approx 0.76$

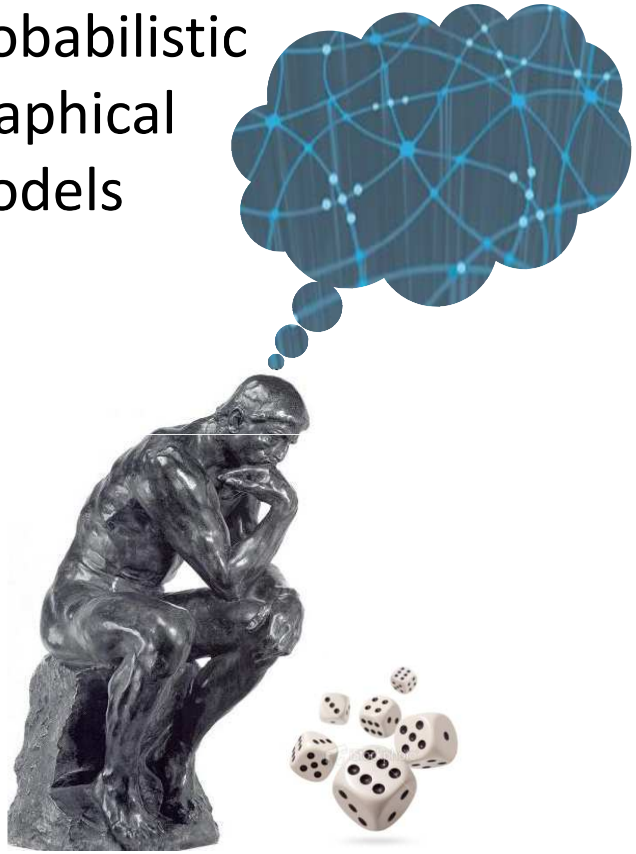
$P(i^1) = 0.3$
$P(i^1 \mid g^3) \approx 0.08$
$P(i^1 \mid g^3, s^1) \approx 0.58$



Student aces the SAT ☺

Student gets a C ☹

Probabilistic
Graphical
Models



Representation

Bayesian Networks

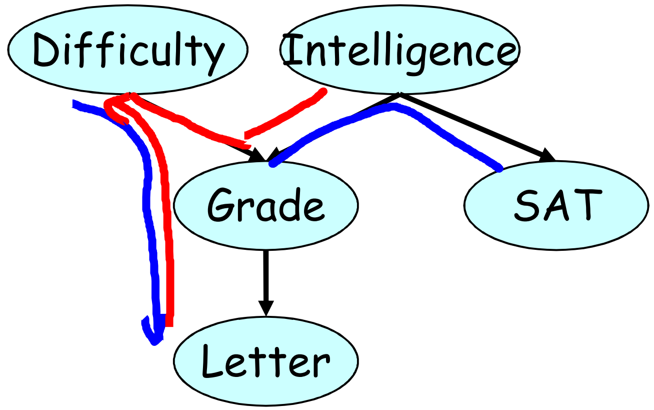
Flow of Probabilistic Influence

When can X influence Y?

condition on W changes beliefs about Y

- $X \rightarrow Y$ ✓
- $X \leftarrow Y$ ✓
- $X \rightarrow W \rightarrow Y$ ✓
- $X \leftarrow W \leftarrow Y$ ✓
- $X \leftarrow \underline{W} \rightarrow Y$ ✓
- $X \rightarrow W \leftarrow Y$ ✗

v-structure

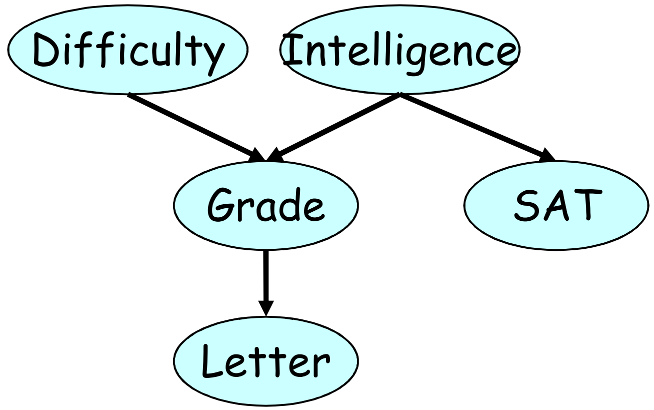


Active Trails

- A trail $X_1 — \dots — X_n$ is active if: it has no v-structures $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$

When can X influence Y Given evidence about **Z**

	$W \notin Z$	$W \in Z$
$X \rightarrow Y$		
$X \leftarrow Y$		
$X \rightarrow W \rightarrow Y$	✓	✗
$X \leftarrow W \leftarrow Y$	✓	✗
$X \leftarrow W \rightarrow Y$	✓	✗
$X \rightarrow W \leftarrow Y$	✗	✓



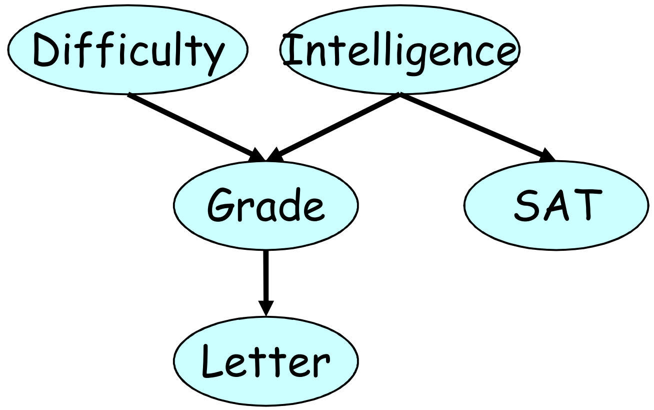
When can X influence Y given evidence about **Z**

- S — I — G — D allows influence to flow when:

I is observed ✗

I not observed, nothing else ✗

I not observed G is observed



Active Trails

- A trail $X_1 — \dots — X_n$ is active given **Z** if:
  - for any v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ we have that X_i or one of its descendants $\in Z$
  - no other X_i is in Z

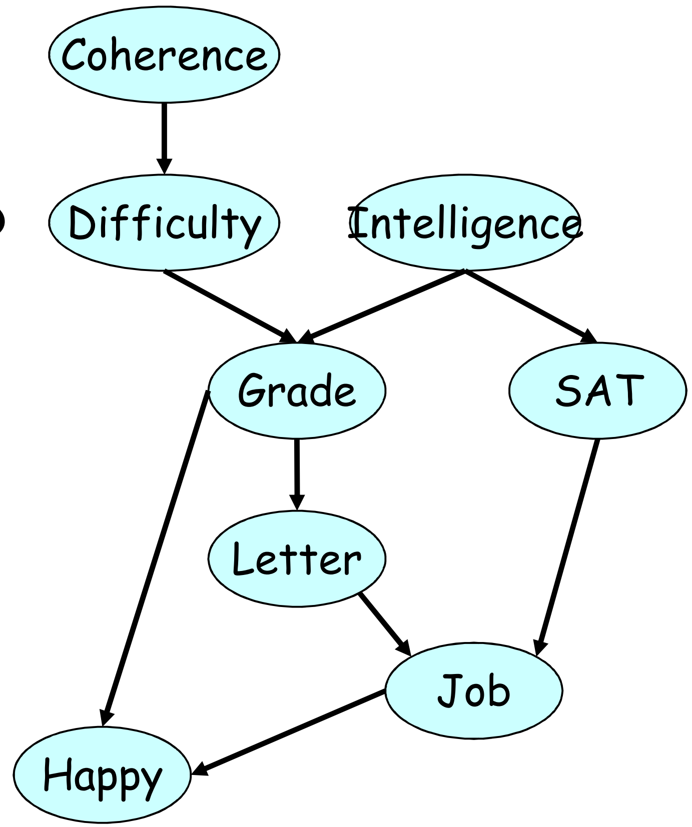
activate v-structures

not in v-structures

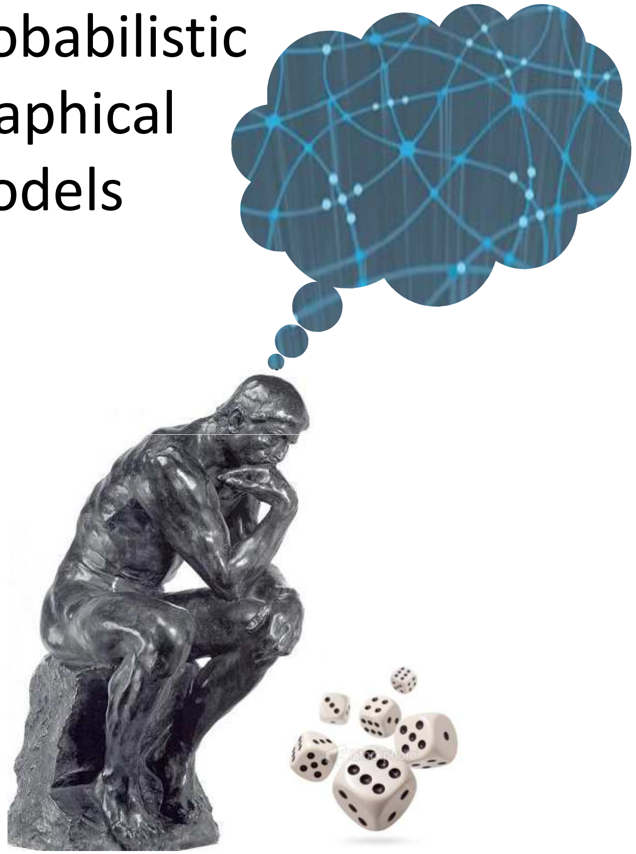
Active trails

- Which of the following are active trails if we observe G? (Mark all that apply.)

  - ○ C — D — G — I — S
  - ○ I — G — L — J — H
  - ○ I — S — J — H
  - ○ C — D — G — I — S — J — L



Probabilistic
Graphical
Models



Independencies

Preliminaries

Independence

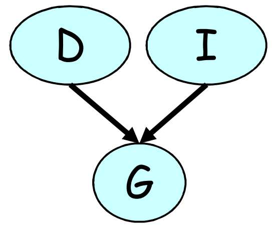
- For events α, β, $P \models \alpha \perp \beta$ if:
  - $P(\alpha \cap \beta) = P(\alpha) \cdot P(\beta)$ satisfies independence
  - $P(\alpha | \beta) = P(\alpha)$
  - $P(\beta | \alpha) = P(\beta)$
- For random variables X, Y, $P \models X \perp Y$ if:
  - $P(X, Y) = P(X)\, P(Y)$
  - $P(X | Y) = P(X)$
  - $P(Y | X) = P(Y)$

universal

$\forall x, y \quad P(x,y) = P(x) \cdot P(y)$

Daphne Koller

Independence



I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

$P(I,D)$ = ×

I	D	Prob
i^0	d^0	0.42
i^0	d^1	0.18
i^1	d^0	0.28
i^1	d^1	0.12

$P(I)$

I	Prob
i^0	0.6
i^1	0.4

$P(D)$

D	Prob
d^0	0.7
d^1	0.3

Daphne Koller

Conditional Independence

- For (sets of) random variables **X**,**Y**,**Z**

  satisfies indep conditioning

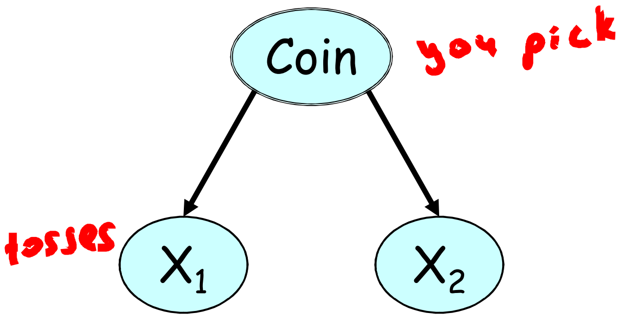
  $P \models (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if:

  given

  – $P(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})\, P(\mathbf{Y}|\mathbf{Z})$
  – $P(\mathbf{X}|\mathbf{Y},\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})$
  – $P(\mathbf{Y}|\mathbf{X},\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})$
  – $P(\mathbf{X},\mathbf{Y},\mathbf{Z}) \propto \phi_1(\mathbf{X},\cancel{\mathbf{Z}})\, \phi_2(\mathbf{Y},\mathbf{Z})$

Daphne Koller

Conditional Independence



$P \not\models X_1 \perp X_2$

$P \models (X_1 \perp X_2 \mid C)$

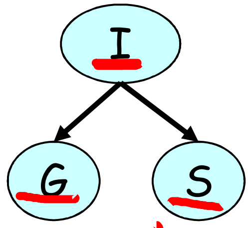
Conditional Independence

P(I, S, G)

I	S	G	Prob.
i^0	s^0	g^1	0.126
i^0	s^0	g^2	0.168
i^0	s^0	g^3	0.126
i^0	s^1	g^1	0.009
i^0	s^1	g^2	0.045
i^0	s^1	g^3	0.126
i^1	s^0	g^1	0.252
i^1	s^0	g^2	0.0224
i^1	s^0	g^3	0.0056
i^1	s^1	g^1	0.06
i^1	s^1	g^2	0.036
i^1	s^1	g^3	0.024

$P(S,G \mid \underline{i^0})$

S	G	Prob.
s^0	g^1	0.19
s^0	g^2	0.323
s^0	g^3	0.437
s^1	g^1	0.01
s^1	g^2	0.017
s^1	g^3	0.023



P(S | i^0)

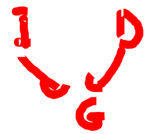
S	Prob
s^0	0.95
s^1	0.05

P(G | i^0)

G	Prob.
g^1	0.2
g^2	0.34
g^3	0.46

Daphne Koller

Conditioning can Lose Independences

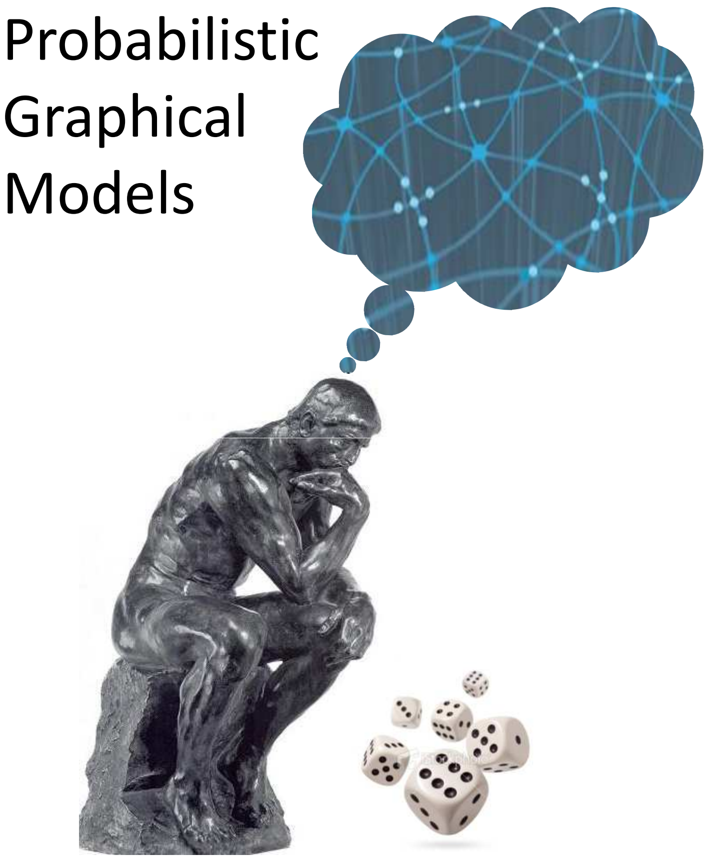


I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

$P(I,D \mid g^1)$

I	D	Prob.
i^0	d^0	0.282
i^0	d^1	0.02
i^1	d^0	0.564
i^1	d^1	0.134

Daphne Koller



Representation

Independencies

Bayesian Networks

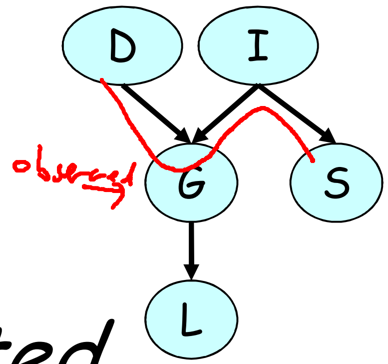
Independence & Factorization

$P(X,Y) = P(X)\,P(Y)$ X,Y independent

$P(\mathbf{X},\mathbf{Y},\mathbf{Z}) \propto \phi_1(\mathbf{X},\mathbf{Z})\,\phi_2(\mathbf{Y},\mathbf{Z})$ $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$

- Factorization of a distribution P implies independencies that hold in P
- If P factorizes over G, can we read these independencies from the structure of G?

Flow of influence & d-separation



Definition: **X** and **Y** are *d-separated* in G given **Z** if there is no active trail in G between **X** and **Y** given **Z**

Notation: d-sep$_G$(**X**, **Y** | **Z**)

Daphne Koller

d-separation

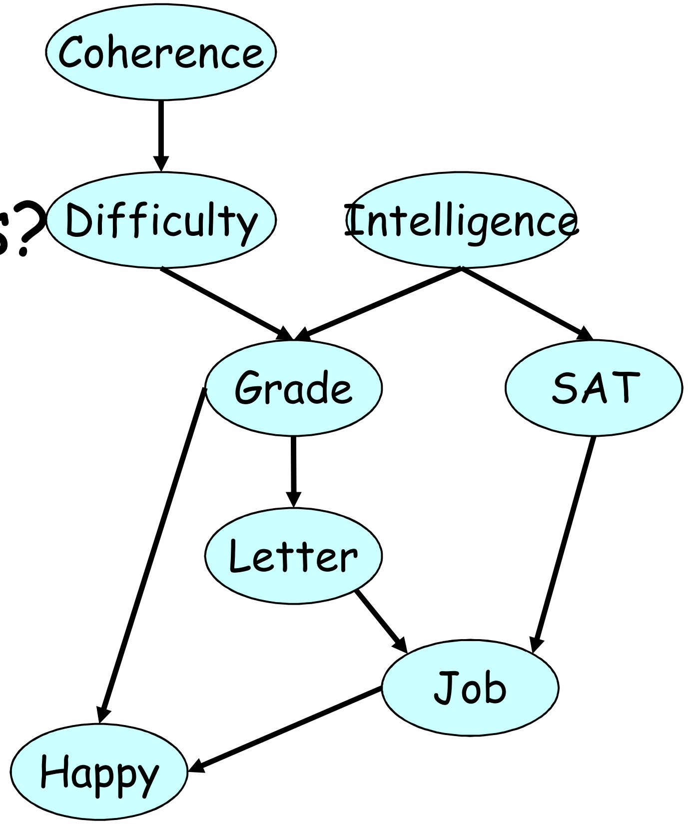
- Which of the following are true d-separation statements? (Mark all that apply.)

  d-sep(D,I | L)

  d-sep(D,J | L)

  d-sep(D,J | L,I)

  d-sep(D,J | L,H,I)



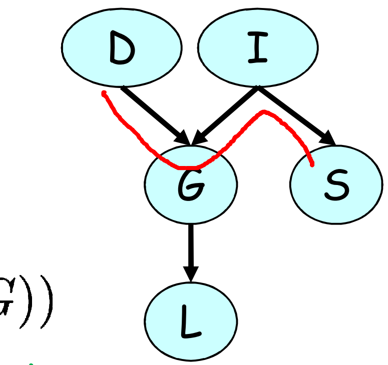
Factorization ⇨ Independence: BNs

Theorem: If P factorizes over G, and d-sep$_G$(**X**, **Y** | **Z**) then P satisfies (**X** ⊥ **Y** | **Z**)

chain rule

$$P(D,I,G,S,L) = P(D)P(I)P(G \mid D,I)P(S \mid I)P(L \mid G)$$

$P \models D \perp S$



$$P(D,S) = \sum_{G,L,I} P(D)P(I)P(G \mid D,I)P(S \mid I)P(L \mid G)$$

$\sum_L$

$$= \sum_I P(D)P(I)P(S \mid I) \underbrace{\sum_G (\underbrace{P(G \mid D,I)}_{1} \underbrace{\sum_L P(L \mid G)}_{1})}_{1}$$

$\phi_2(S)$

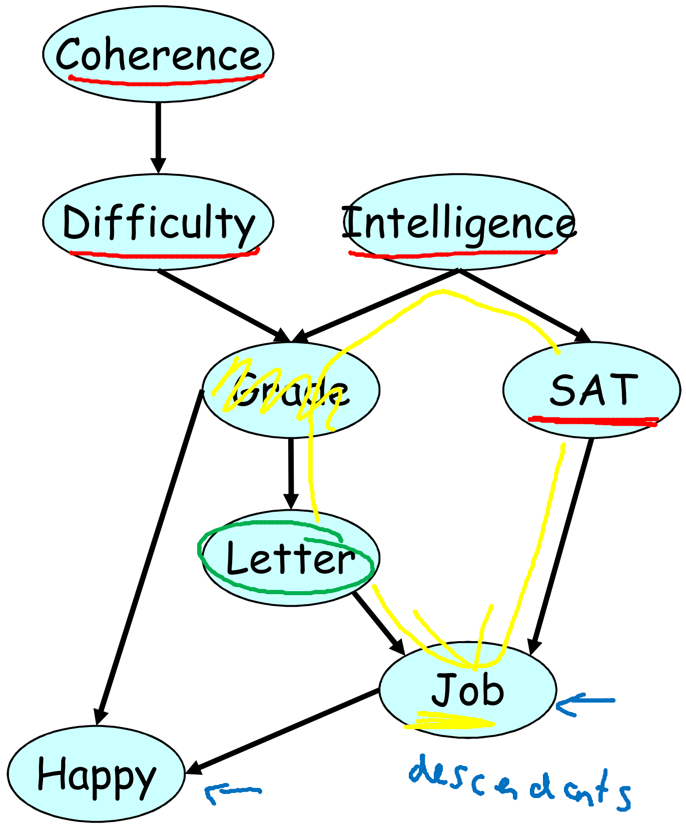
$$= \underbrace{P(D)}_{\phi_1(D)} \underbrace{(\sum_I P(I)P(S \mid I))}_{\phi_2(S)}$$

Daphne Koller

Any node is d-separated from its non-descendants given its parents

Grade

If P factorizes over G, then in P, any variable is independent of its non-descendants given its parents



Daphne Koller

I-maps

- d-separation in $G \Rightarrow P$ satisfies corresponding independence statement

$$I(G) = \{(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \; : \; \text{d-sep}_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})\}$$

- Definition: If P satisfies I(G), we say that G is an I-map (independency map) of P

I-maps

$P_1 \models D \perp I$

I	D	Prob
i^0	d^0	0.42
i^0	d^1	0.18
i^1	d^0	0.28
i^1	d^1	0.12

$P_2 \not\models D \perp I$

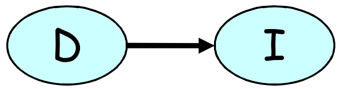
I	D	Prob.
i^0	d^0	0.282
i^0	d^1	0.02
i^1	d^0	0.564
i^1	d^1	0.134

I-map

G_1

$I(G_1) = \{D \perp I\}$

G_2



$I(G_2) = \emptyset$

Daphne Koller

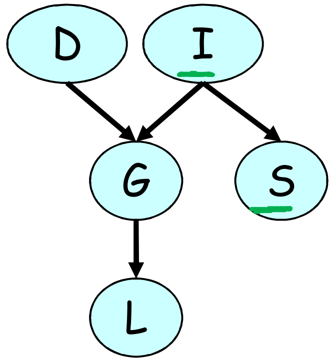
Factorization ⇨ Independence: BNs

Theorem: If P factorizes over G, then G is an I-map for P

Can read from G independencies in P
regardless of parameters

Independence ⇨ Factorization

Theorem: If G is an I-map for P, then P factorizes over G



$I \perp D$

P(I,D) chain rule for probabilities

$$P(D,I,G,S,L) = P(D)P(I \mid \cancel{D})P(G \mid D,I)P(S \mid \cancel{D},I,\cancel{G})P(L \mid \cancel{D},\cancel{I},G,\cancel{S})$$

$$P(D,I,G,S,L) = P(D)P(I)P(G \mid D,I)P(S \mid I)P(L \mid G)$$

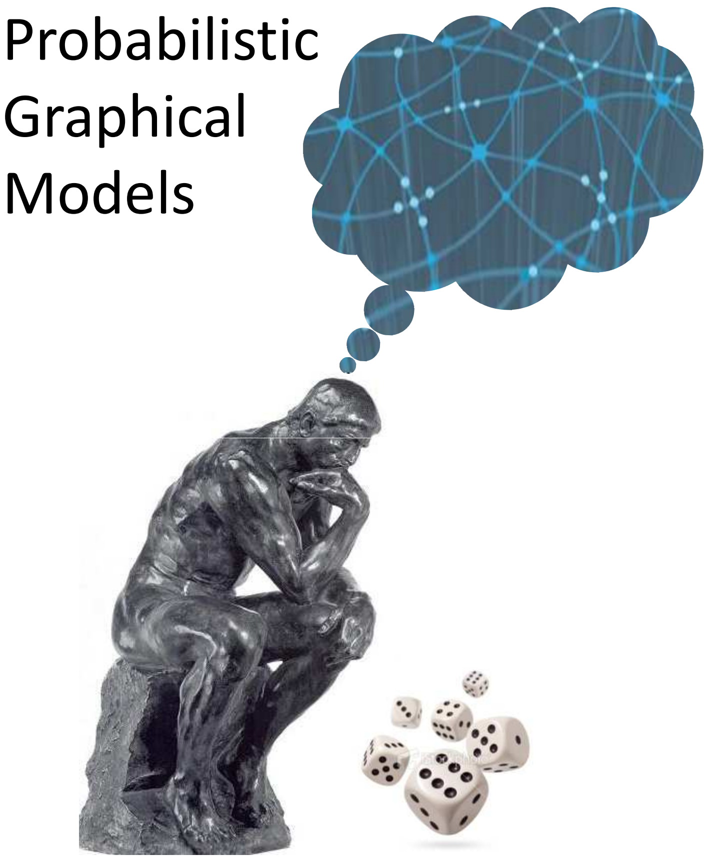
Daphne Koller

Summary

Two equivalent views of graph structure:

- Factorization: G allows P to be represented
- I-map: Independencies encoded by G hold in P

If P factorizes over a graph G, we can read from the graph independencies that must hold in P (an independency map)

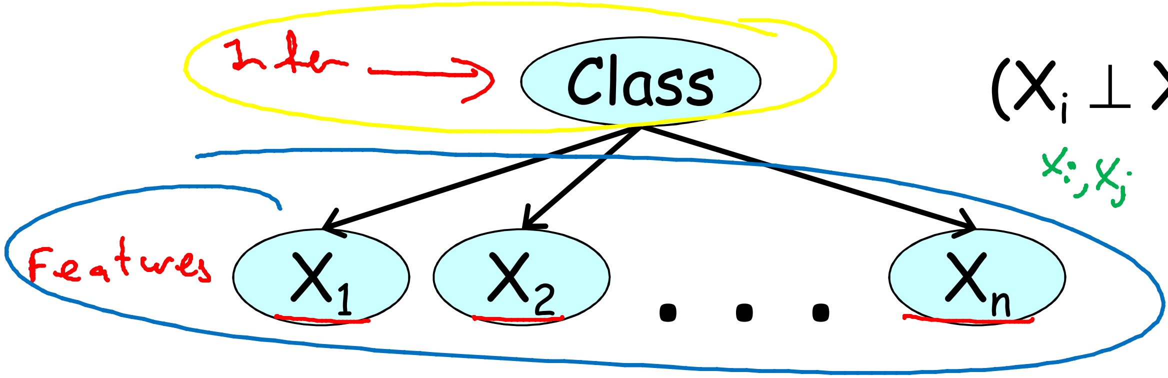


Representation

Bayesian Networks

Naïve Bayes

Naïve Bayes Model



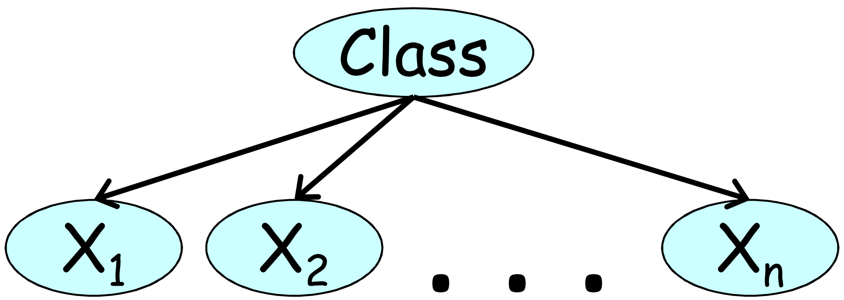
$(X_i \perp X_j \mid C)$ for all X_i, X_j

x_i, x_j conditionally independent given C

$$P(C, X_1, \ldots, X_n) = P(C) \prod_{i=1}^{n} P(X_i \mid C)$$

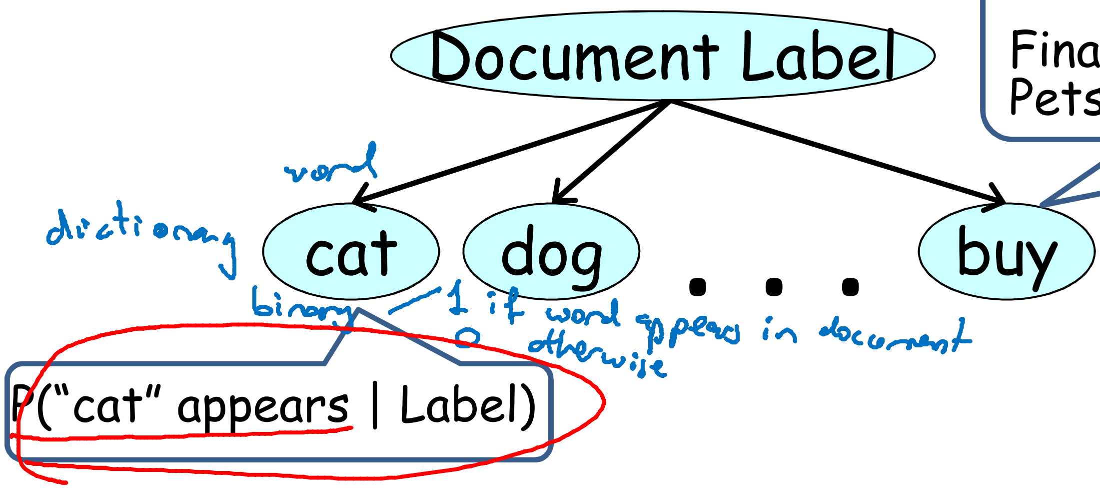
Daphne Koller

Naïve Bayes Classifier



$$\frac{P(C=c^1 \mid x_1,\ldots,x_n)}{P(C=c^2 \mid x_1,\ldots,x_n)} = \underbrace{\frac{P(C=c^1)}{P(C=c^2)}}\ \underbrace{\prod_{i=1}^{n} \frac{P(x_i \mid C=c^1)}{P(x_i \mid C=c^2)}}_{\text{odds ratios}}$$

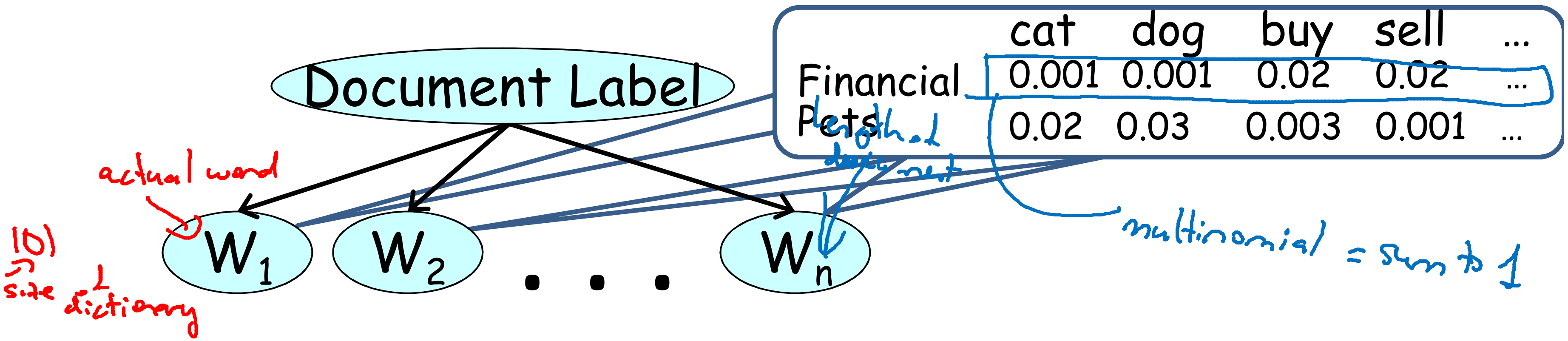
Bernoulli Naïve Bayes for Text



	cat	dog	buy	sell	...
Financial	0.001	0.001	0.2	0.3	...
Pets	0.3	0.4	0.02	0.0001	...

$$\frac{P(C = c^1 \mid x_1, \ldots, x_n)}{P(C = c^2 \mid x_1, \ldots, x_n)} = \frac{P(C = c^1)}{P(C = c^2)} \prod_{i=1}^{n} \frac{P(x_i \mid C = c^1)}{P(x_i \mid C = c^2)}$$

Multinomial Naïve Bayes for Text



	cat	dog	buy	sell	...
Financial	0.001	0.001	0.02	0.02	...
Pets	0.02	0.03	0.003	0.001	...

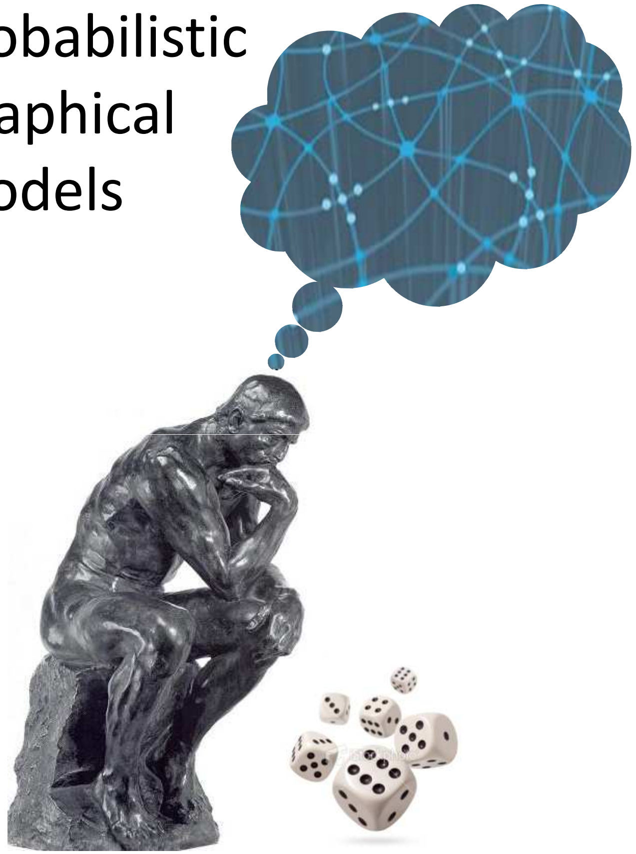
$$\frac{P(C=c^1 \mid x_1,\ldots,x_n)}{P(C=c^2 \mid x_1,\ldots,x_n)} = \frac{P(C=c^1)}{P(C=c^2)} \prod_{i=1}^{n} \frac{P(x_i \mid C=c^1)}{P(x_i \mid C=c^2)}$$

Daphne Koller

Summary

- Simple approach for classification
  - Computationally efficient
  - Easy to construct
- Surprisingly effective in domains with many weakly relevant features
- Strong independence assumptions reduce performance when many features are strongly correlated

Probabilistic
Graphical
Models



Representation

Bayesian Networks

Application: Diagnosis

Medical Diagnosis: Pathfinder (1992)

- Help pathologist diagnose lymph node pathologies (60 different diseases)
- Pathfinder I: Rule-based system
- Pathfinder II used naïve Bayes and got superior performance

Heckerman et al.

Medical Diagnosis: Pathfinder (1992)

- Pathfinder III: Naïve Bayes with better knowledge engineering
- No incorrect zero probabilities
- Better calibration of conditional probabilities
  - P(finding | disease$_1$) to P(finding | disease$_2$)
  - Not P(finding$_1$ | disease) to P(finding$_2$ | disease)

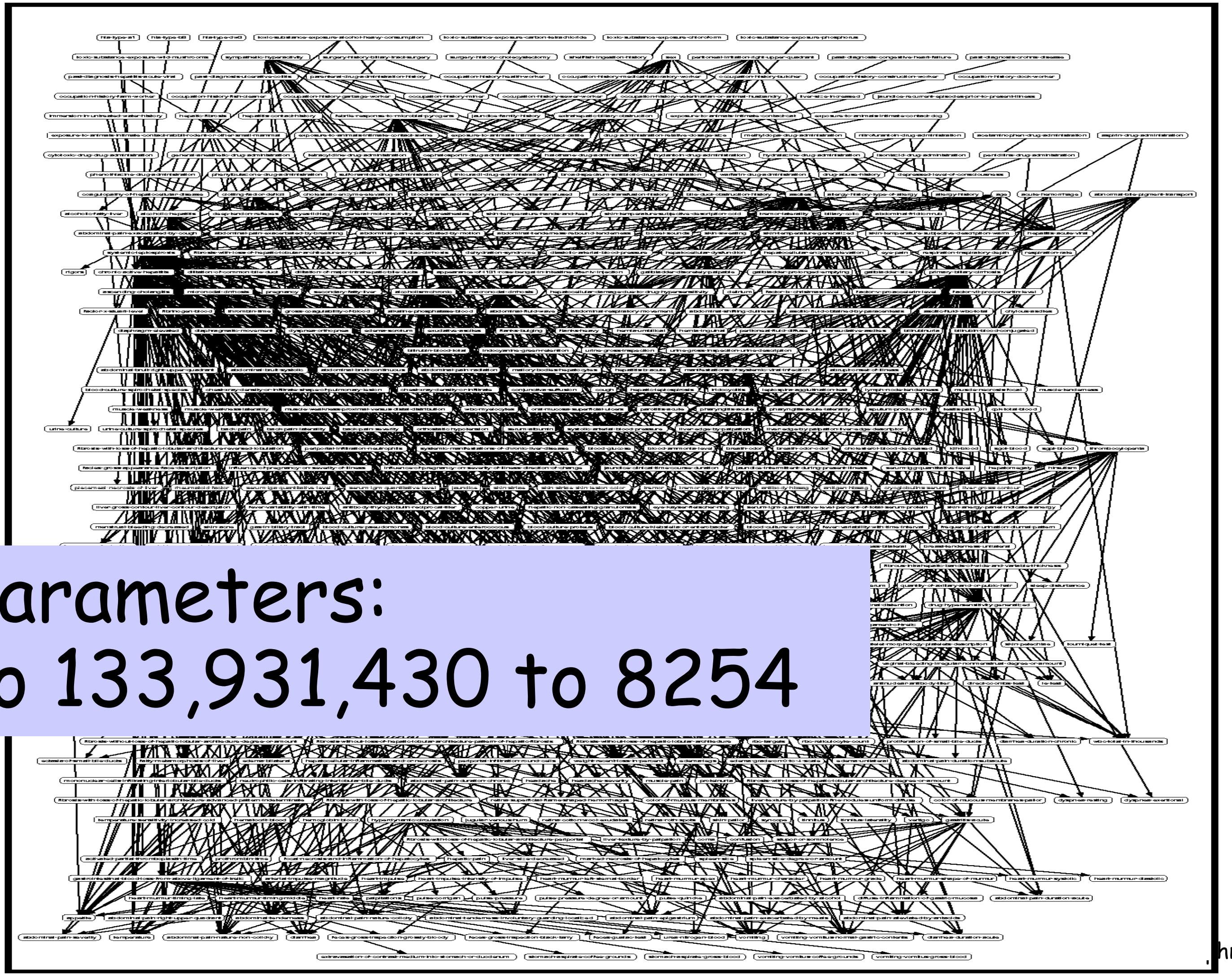
Heckerman et al.

Medical Diagnosis: Pathfinder (1992)

- Pathfinder IV: Full Bayesian network
  - Removed incorrect independencies
  - Additional parents led to more accurate estimation of probabilities
- BN model agreed with expert panel in 50/53 cases, vs 47/53 for naïve Bayes model
- Accuracy as high as expert that designed the model

Heckerman et al.

CPCS

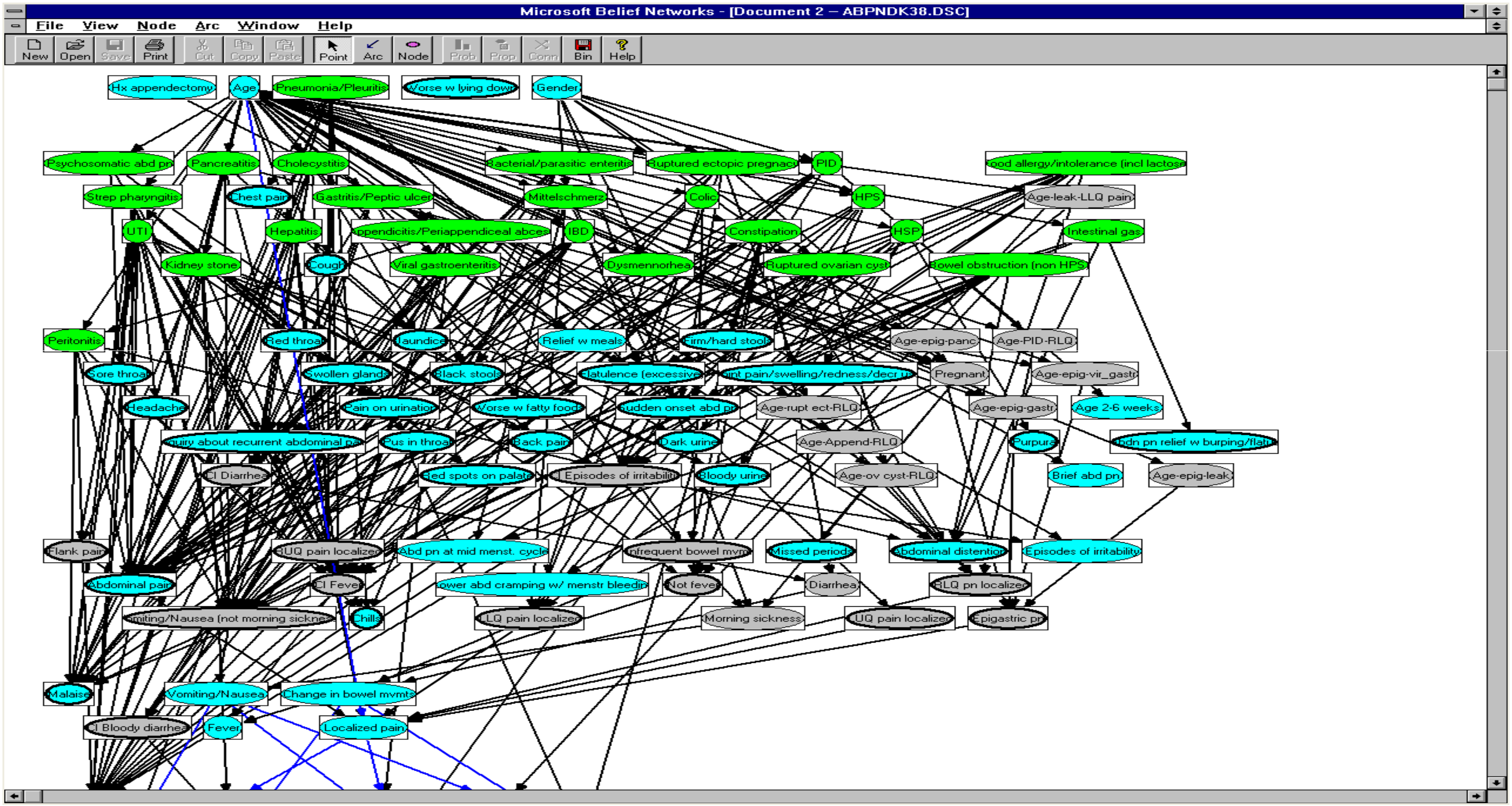


# of parameters:
2^{1000} to 133,931,430 to 8254

Pradhan et al.

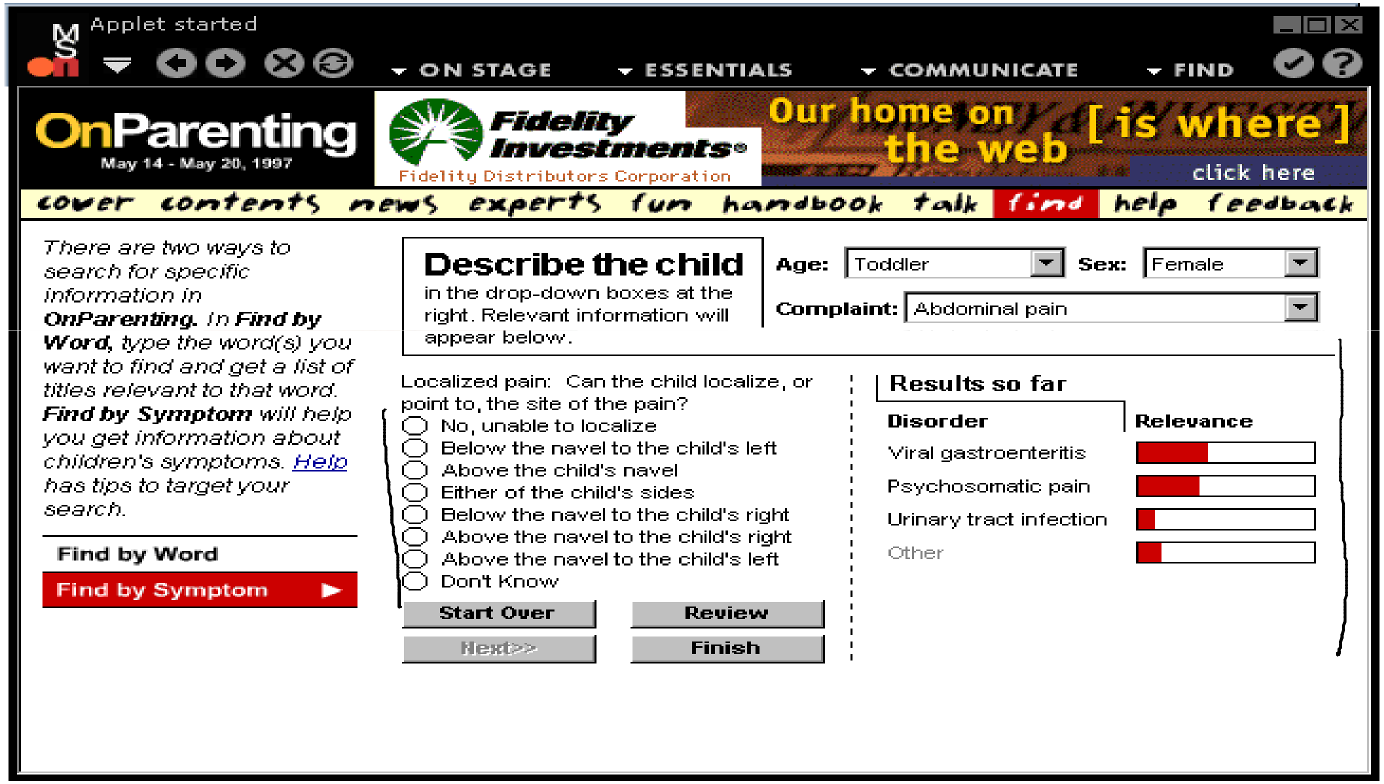
hne Koller

Medical Diagnosis (Microsoft)



Daphne Koller

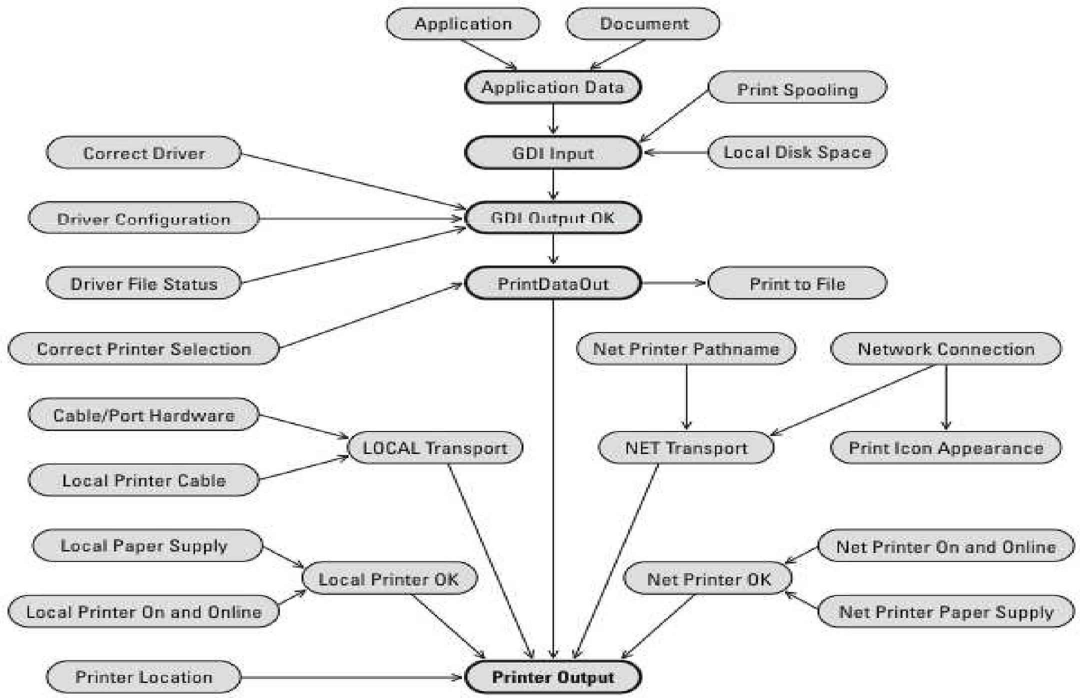
Medical Diagnosis (Microsoft)



Daphne Koller

Fault Diagnosis

- Microsoft troubleshooters



Daphne Koller

Fault Diagnosis

- Many examples:
  - Microsoft troubleshooters
  - Car repair
- Benefits:
  - Flexible user interface
  - Easy to design and maintain