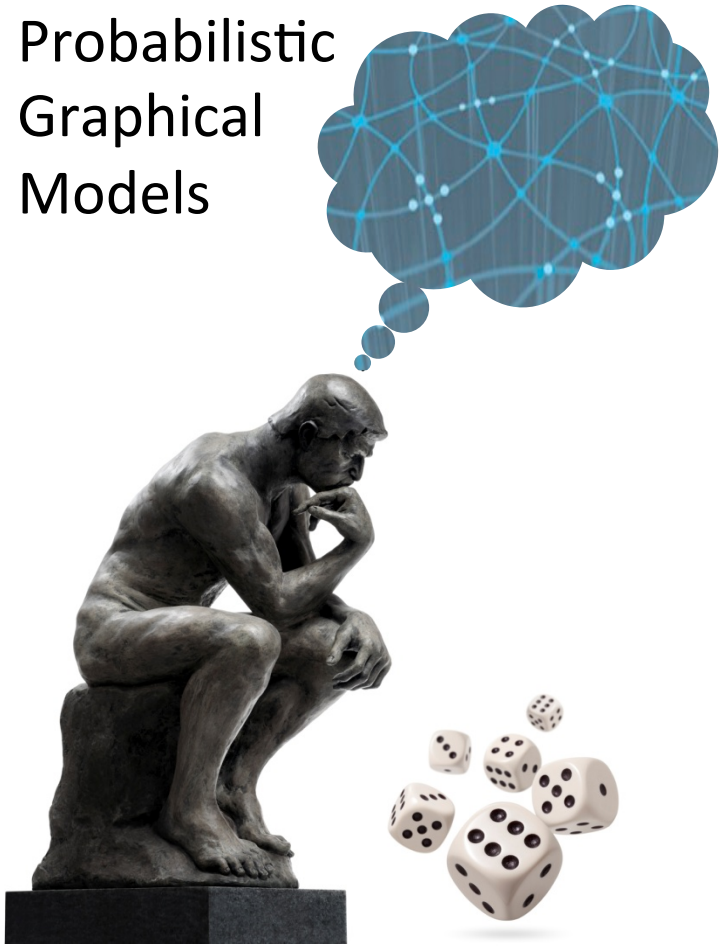
Probabilistic Graphical Models



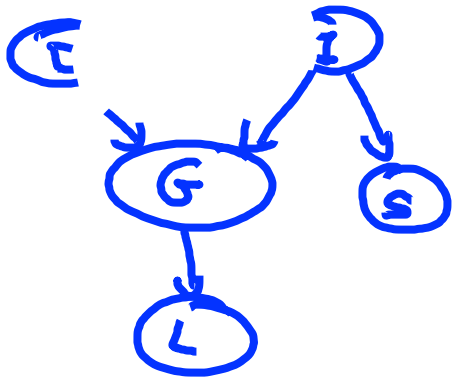
Representation

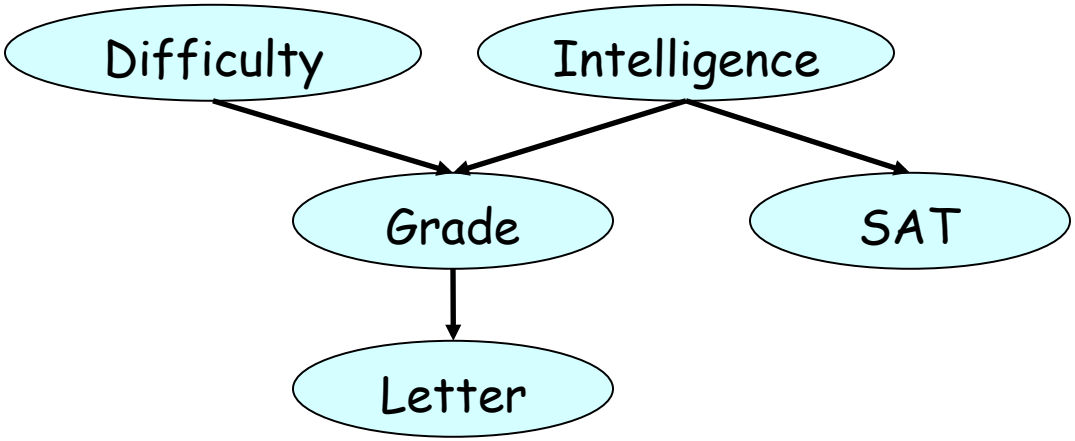
Bayesian Networks

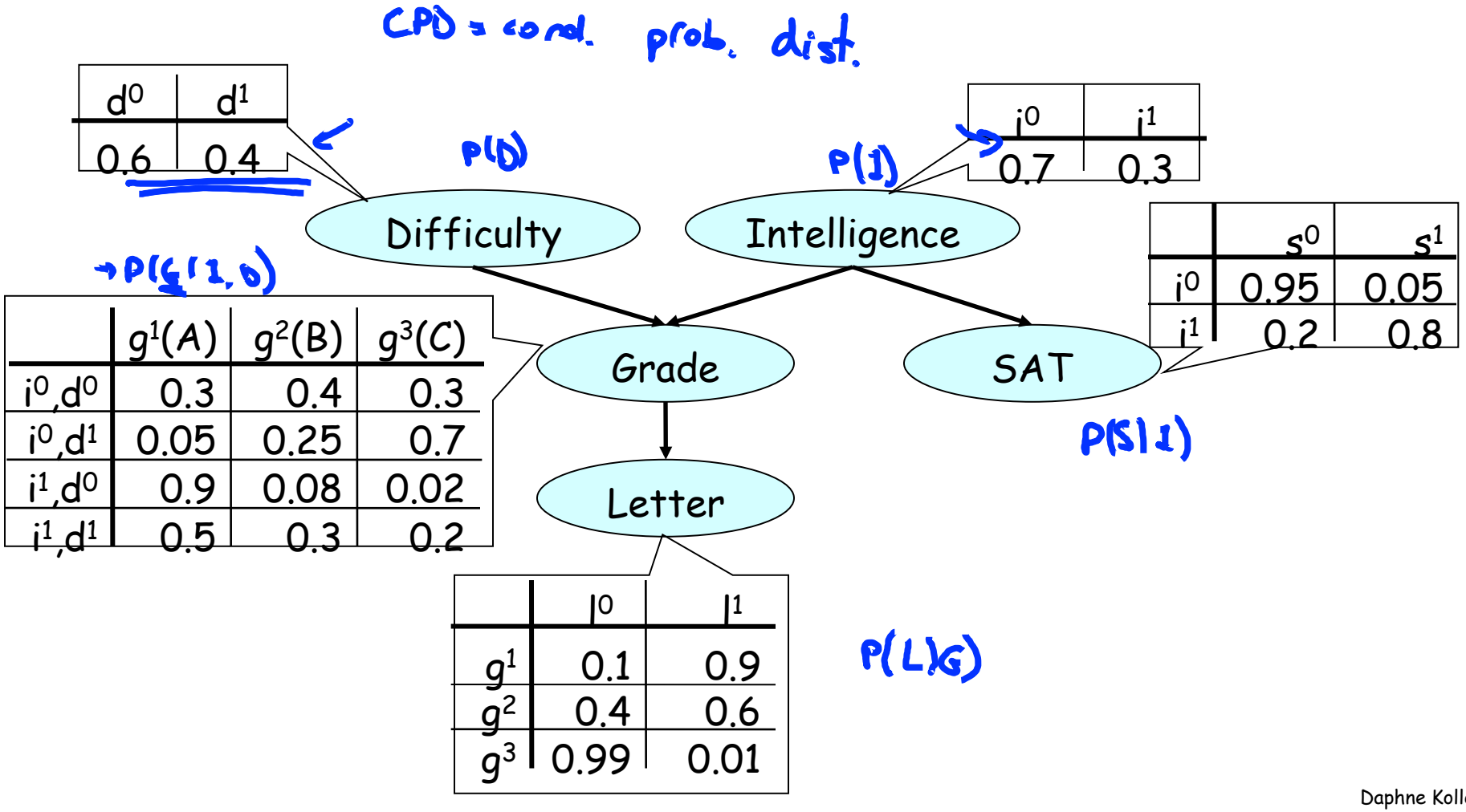
Semantics & Factorization

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

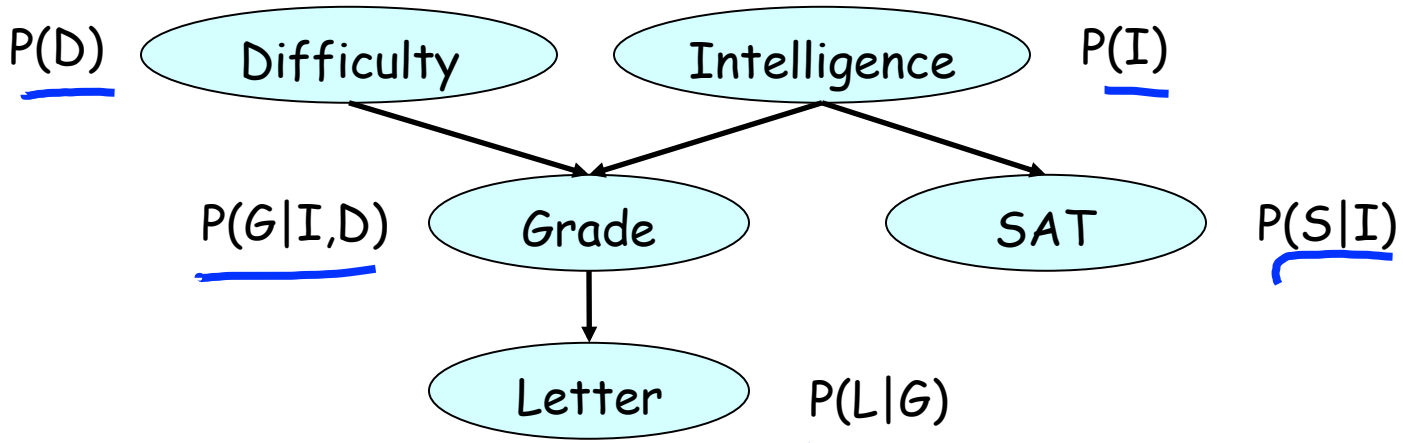
P(G,D,I,S,L)




Difficulty
Intelligence
Grade
SAT
Letter

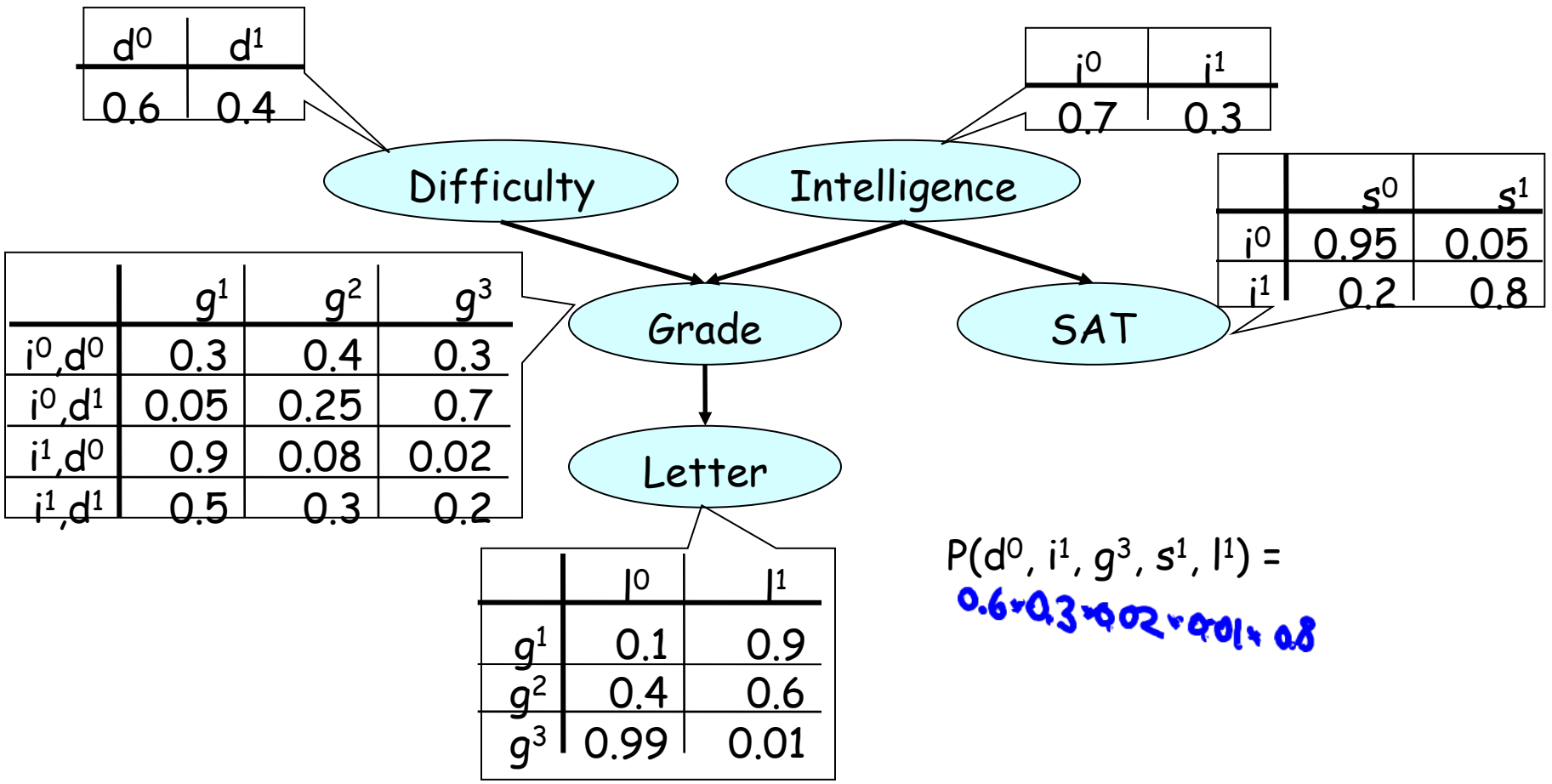

CPD = cond. prob. dist.
d0 d1
0.6 0.4
P(D)
i0 i1
0.7 0.3
P(I)
Difficulty
Intelligence
Grade
SAT
Letter
P(G|I,D)
g1(A) g2(B) g3(C)
i0,d0 0.3 0.4 0.3
i0,d1 0.05 0.25 0.7
i1,d0 0.9 0.08 0.02
i1,d1 0.5 0.3 0.2
s0 s1
i0 0.95 0.05
i1 0.2 0.8
P(S|I)
l0 l1
g1 0.1 0.9
g2 0.4 0.6
g3 0.99 0.01
P(L|G)

Chain Rule for Bayesian Networks



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

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

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

$0.6 \cdot 0.3 \cdot 0.02 \cdot 0.01 \cdot 0.8$

Daphne Koller

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

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

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

P is a product of CPDs

CPDs are non-negative

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

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

chain rule

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

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

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

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

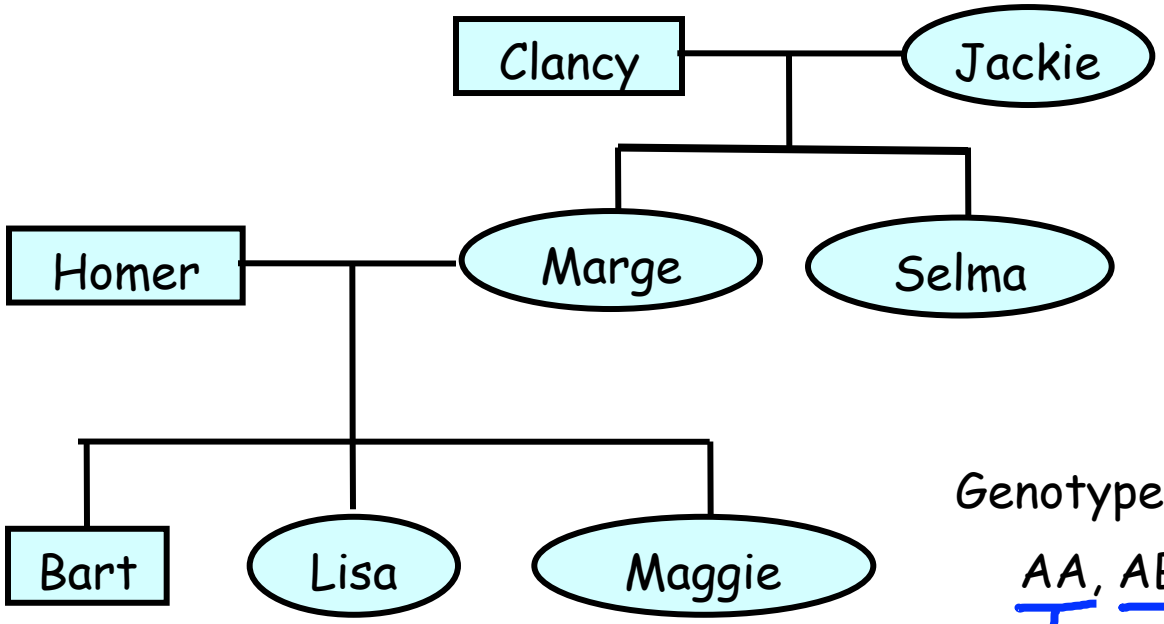
Daphne Koller

P Factorizes over G

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

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

Genetic Inheritance



Genotype

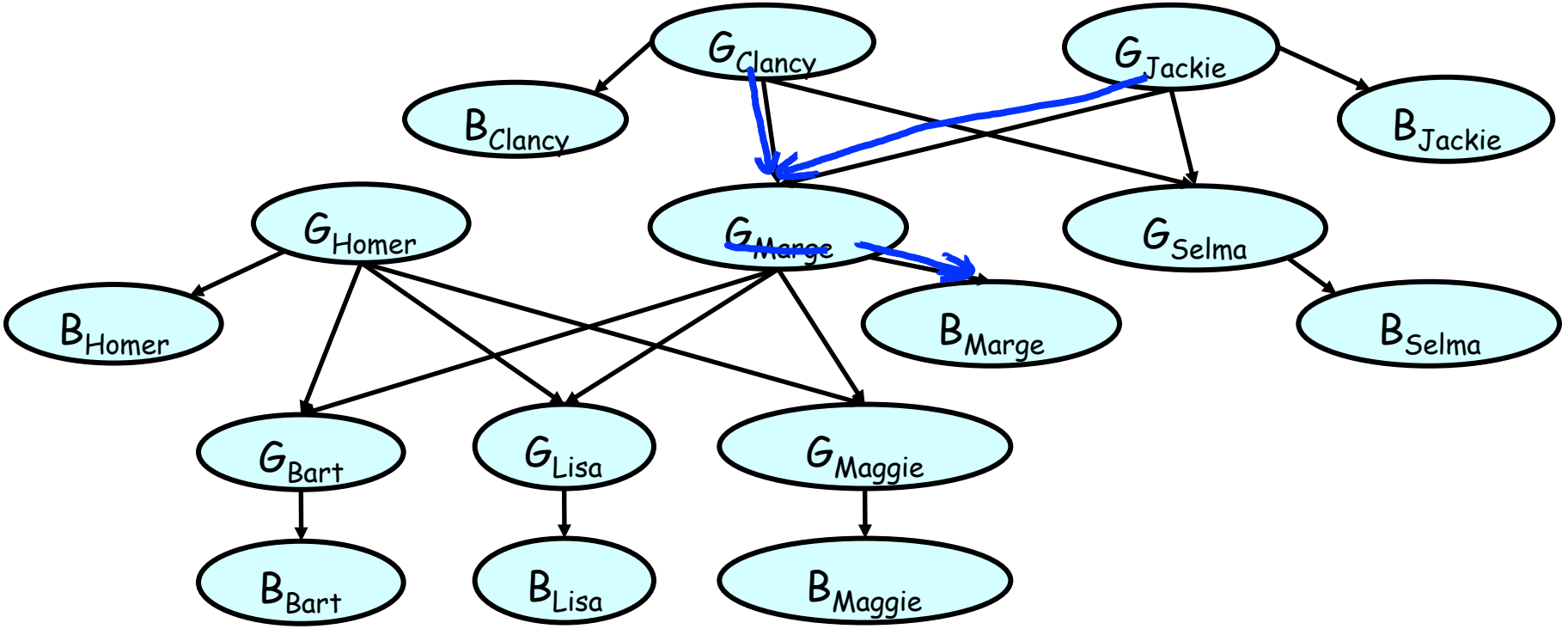
AA, AB, AO, BO, BB, OO

Phenotype

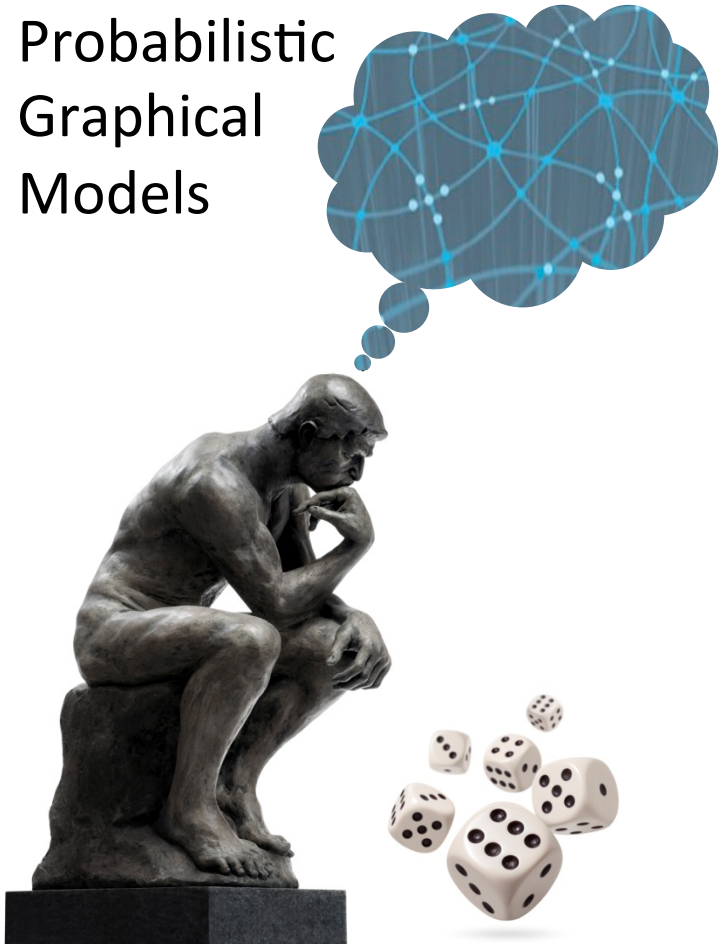
A, B, AB, O

Daphne Koller

BNs for Genetic Inheritance



Probabilistic Graphical Models

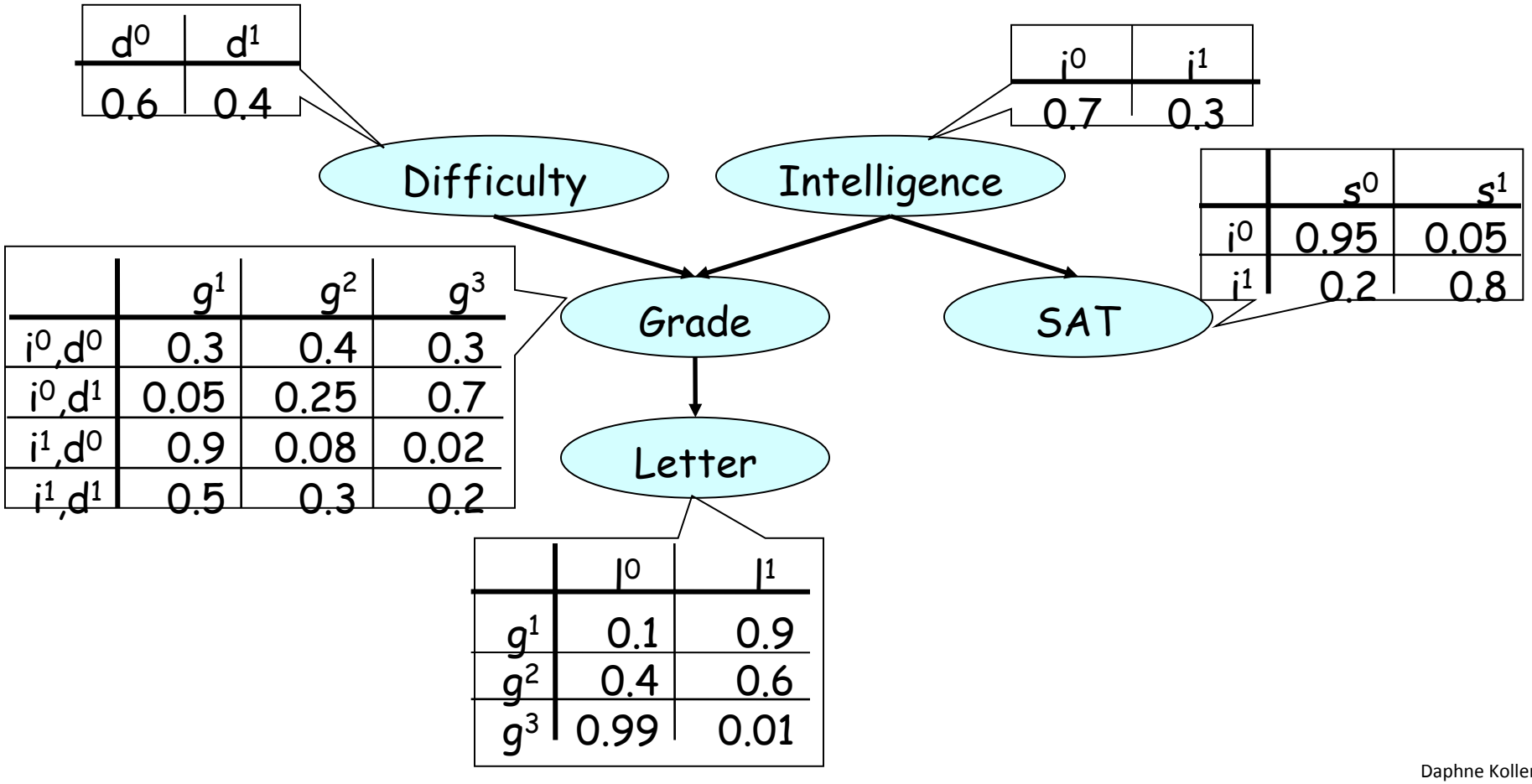


Representation

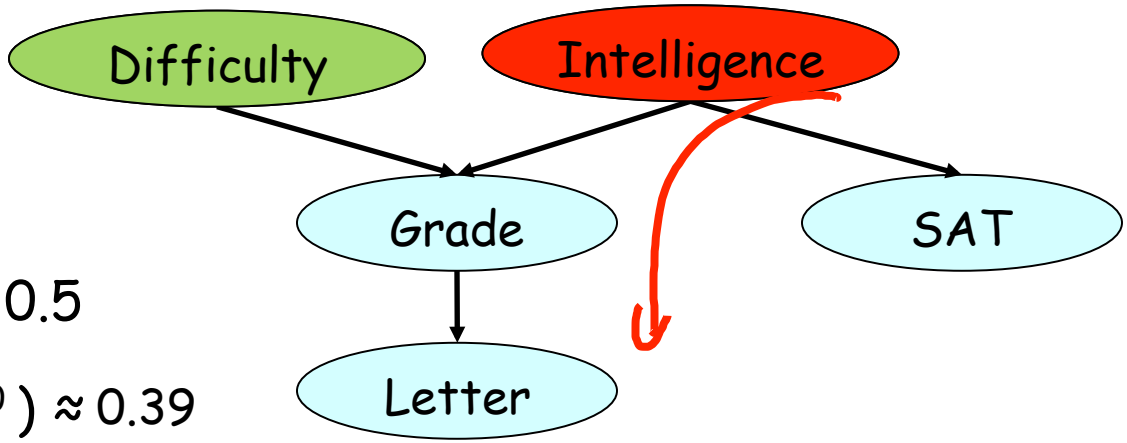
Bayesian Networks

Reasoning Patterns

The Student Network



Causal Reasoning



$P(l^1) \approx 0.5$

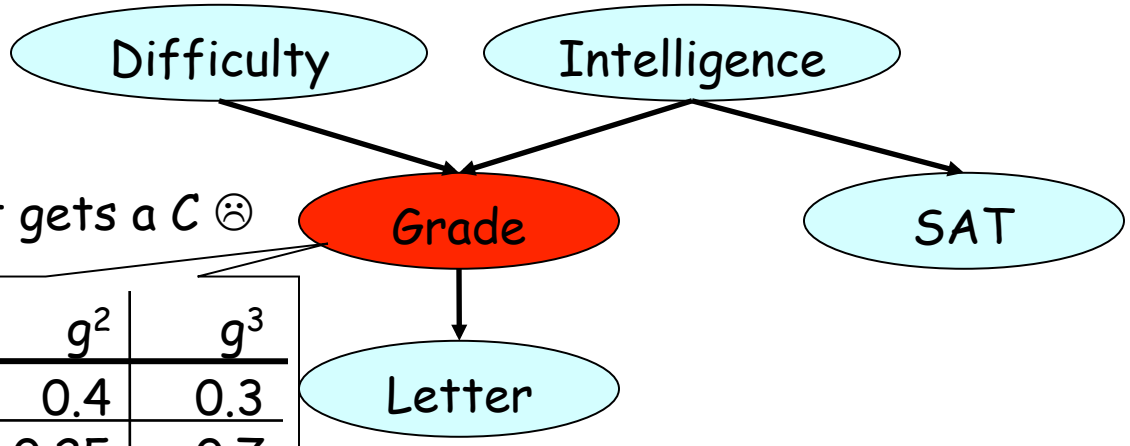
$P(l^1 \mid i^0) \approx 0.39$

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

Evidential Reasoning

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

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



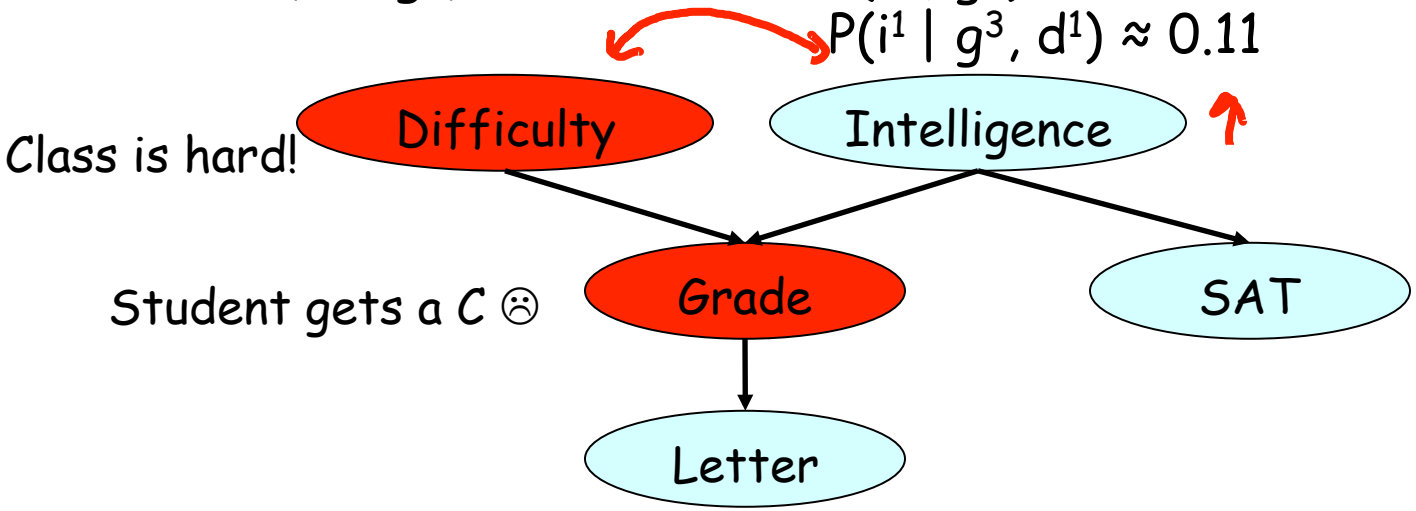
Student gets a C ☹

	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

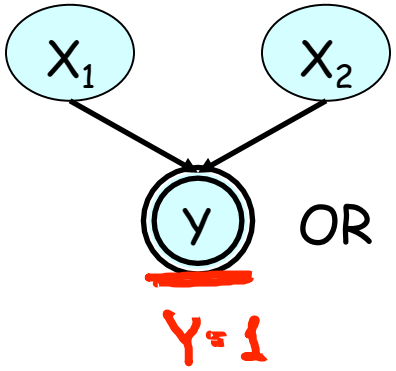
Intercausal Reasoning

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

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



Intercausal Reasoning Explained



explaining away

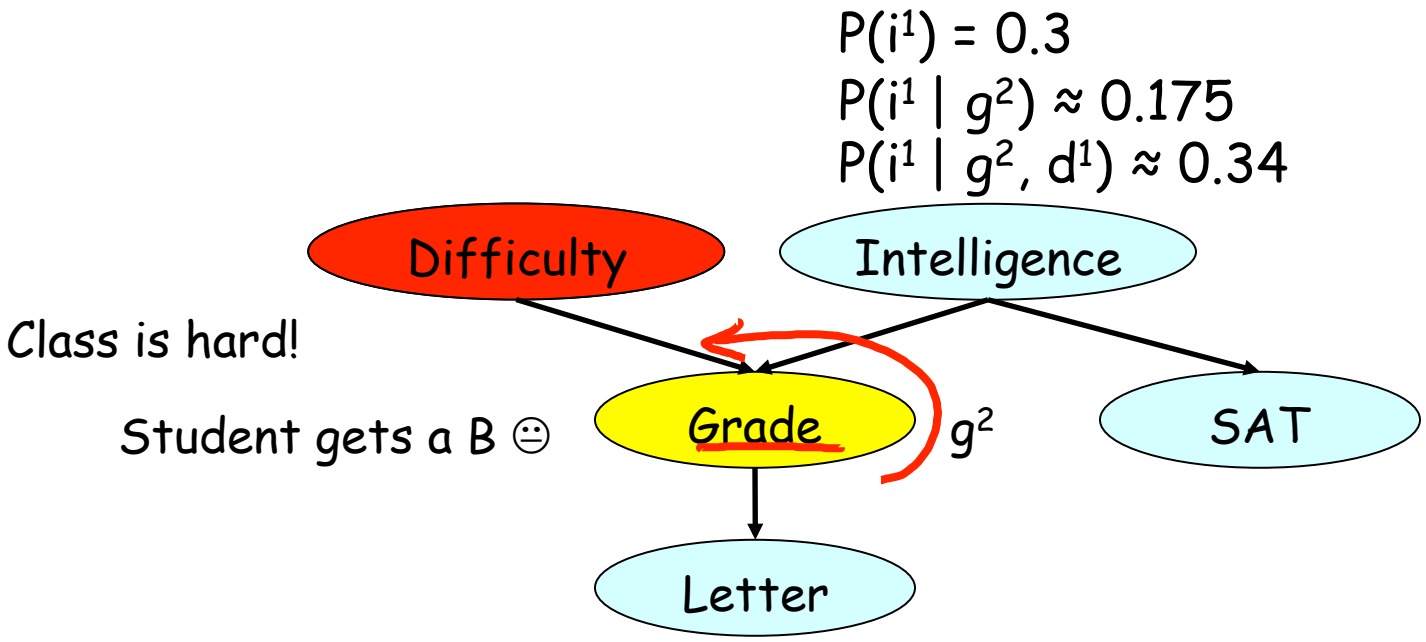
X_1	X_2	Y	Prob
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	1	0.25

0.33
0.33
0.33

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

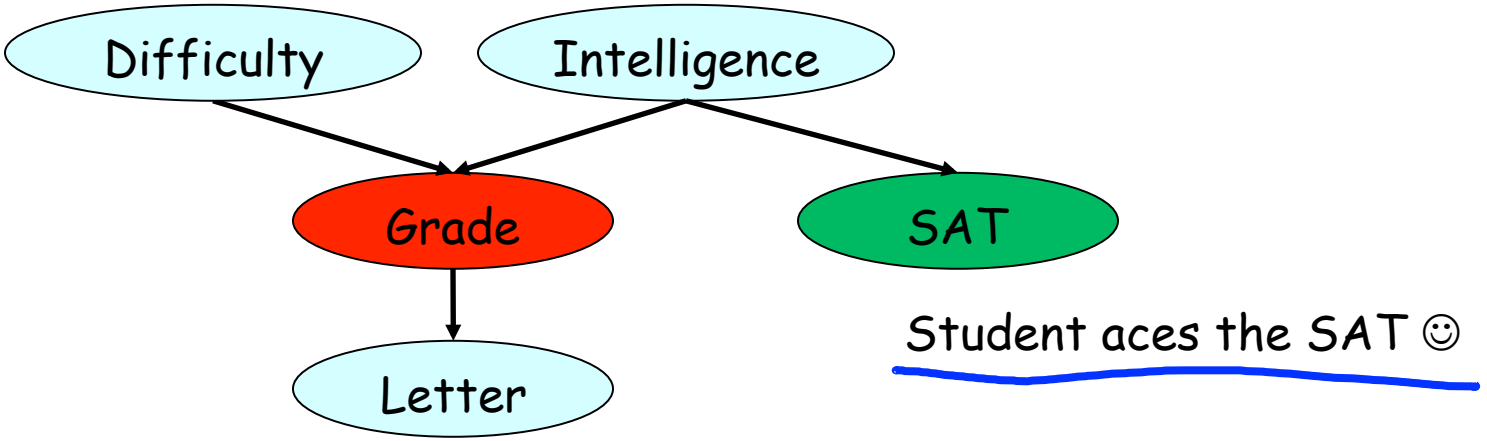
conditional $P(X_2=1 \mid X_1=1) = 0.5$

Intercausal Reasoning II



Student Aces the SAT

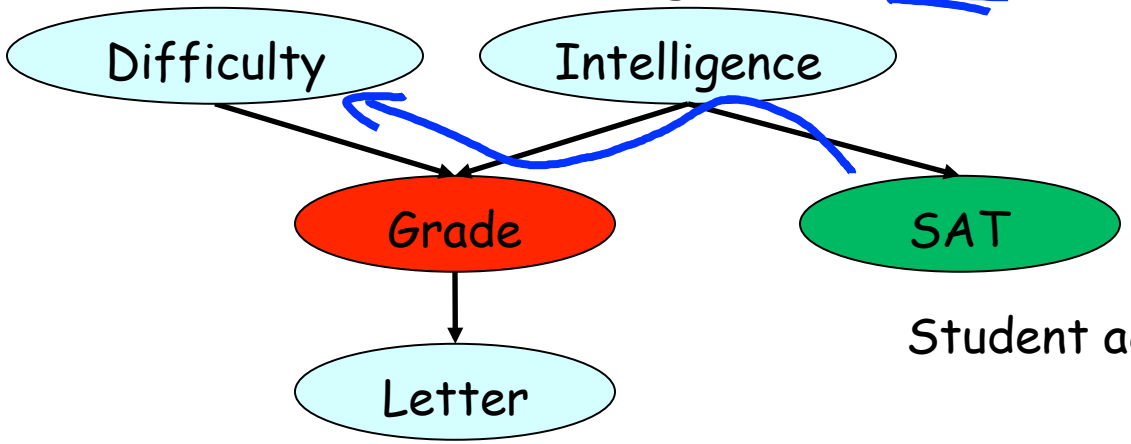
- What happens to the posterior probability that the class is hard?



Student Aces the SAT

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

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

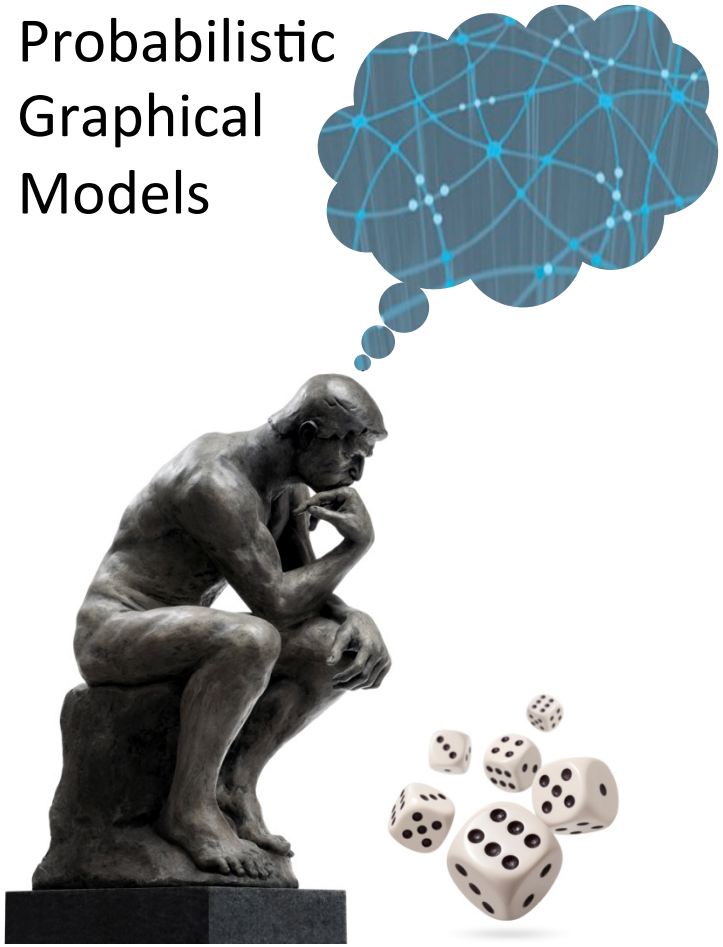


Student aces the SAT ☺

Student gets a C ☹

Daphne Koller

Probabilistic Graphical Models



Representation

Bayesian Networks

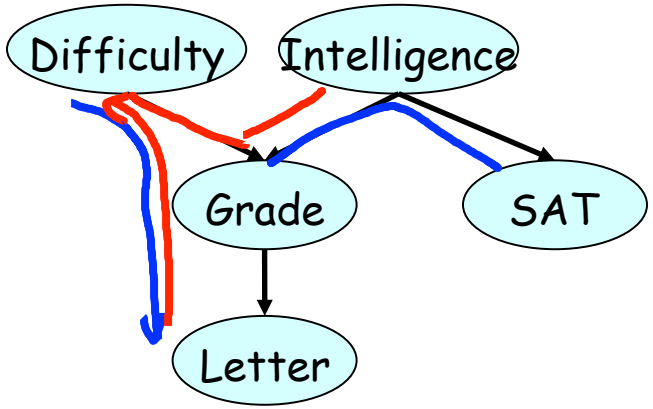
Flow of Probabilistic Influence

When can X influence Y?

condition on W changes beliefs about Y

- X → Y ✓
- X ← Y ✓
- X → W → Y ✓
- X ← W ← Y ✓
- X ← W → Y ✓
- X → W ← Y ✗

v-structure



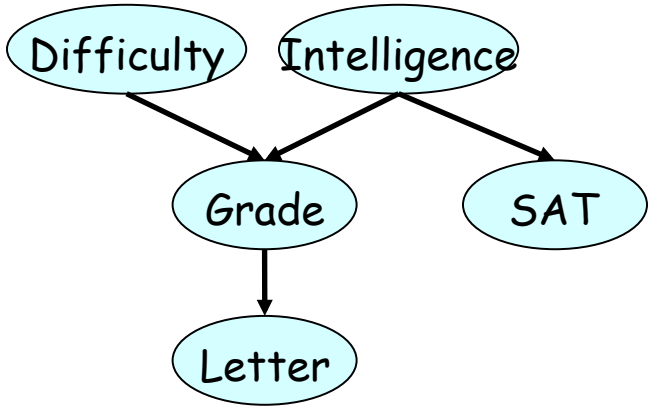
Active Trails

- A trail $X_1 - \ldots - X_n$ 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

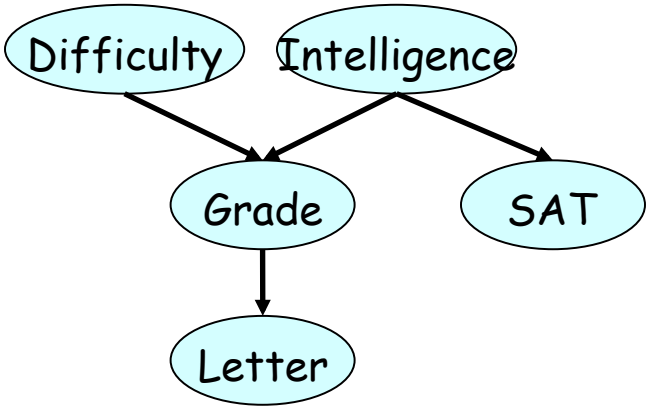
- $X \rightarrow Y$
- $X \leftarrow Y$
- $X \rightarrow W \rightarrow Y$
- $X \leftarrow W \leftarrow Y$
- $X \leftarrow W \rightarrow Y$
- $X \rightarrow W \leftarrow Y$

$W \notin Z$	$W \in Z$
✓	✗
✓	✗
✓	✗
✗	✓



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

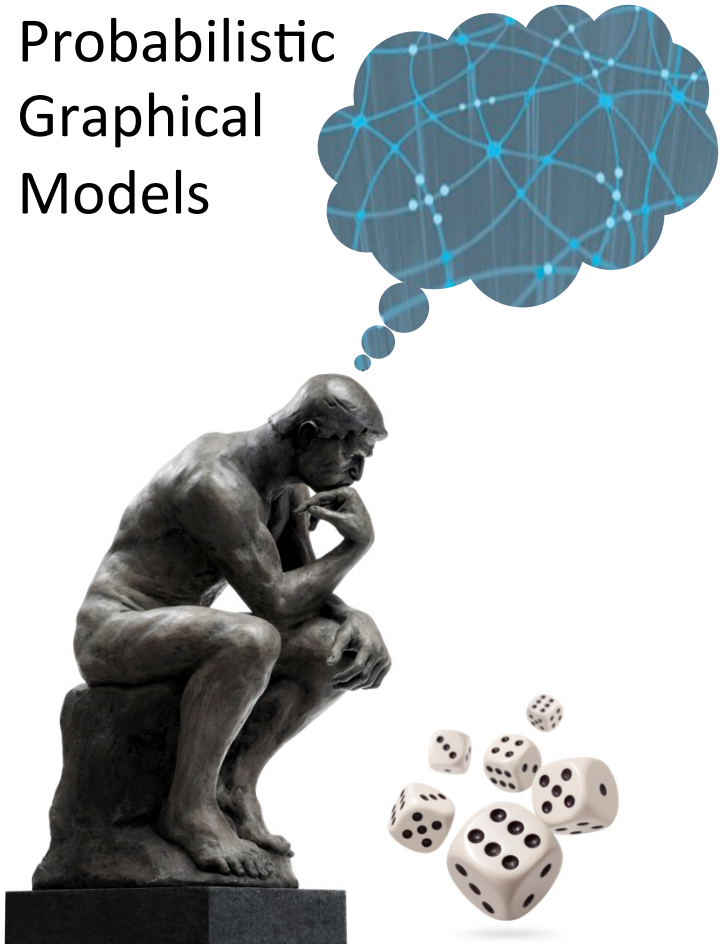
Active Trails

- A trail $X_1 - \ldots - X_n$ is active given $\mathbf{Z}$ 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

Probabilistic Graphical Models



Representation

Independencies

Preliminaries

Independence

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

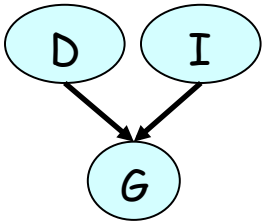
Independence

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

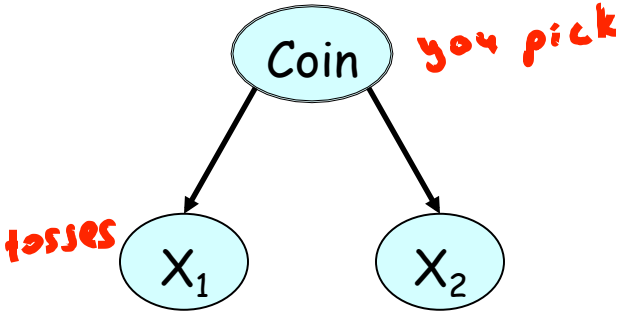
P(D)

D	Prob
d^0	0.7
d^1	0.3

Conditional Independence

- For (sets of) random variables X, Y, Z
  $P \models (X \perp Y \mid Z)$ if:
  - $P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$
  - $P(X \mid Y, Z) = P(X \mid Z)$
  - $P(Y \mid X, Z) = P(X \mid Z)$
  - $P(X, Y, Z) \propto \phi_1(X, Z)\, \phi_2(Y, Z)$

Conditional Independence



$P \not\models (X_1 \perp X_2)$

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

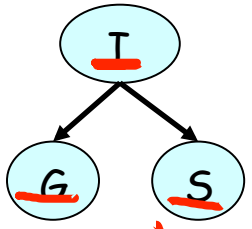
Conditional Independence

P(I, S, G)

I	S	G	Prob.
i^0	s^0	g^1	0.114
i^0	s^0	g^2	0.1938
i^0	s^0	g^3	0.2622
i^0	s^1	g^1	0.006
i^0	s^1	g^2	0.0102
i^0	s^1	g^3	0.0138
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	0.108
i^1	s^1	g^2	0.0096
i^1	s^1	g^3	0.0024

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

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)

S	Prob
s^0	0.95
s^1	0.05

P(G | i)

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

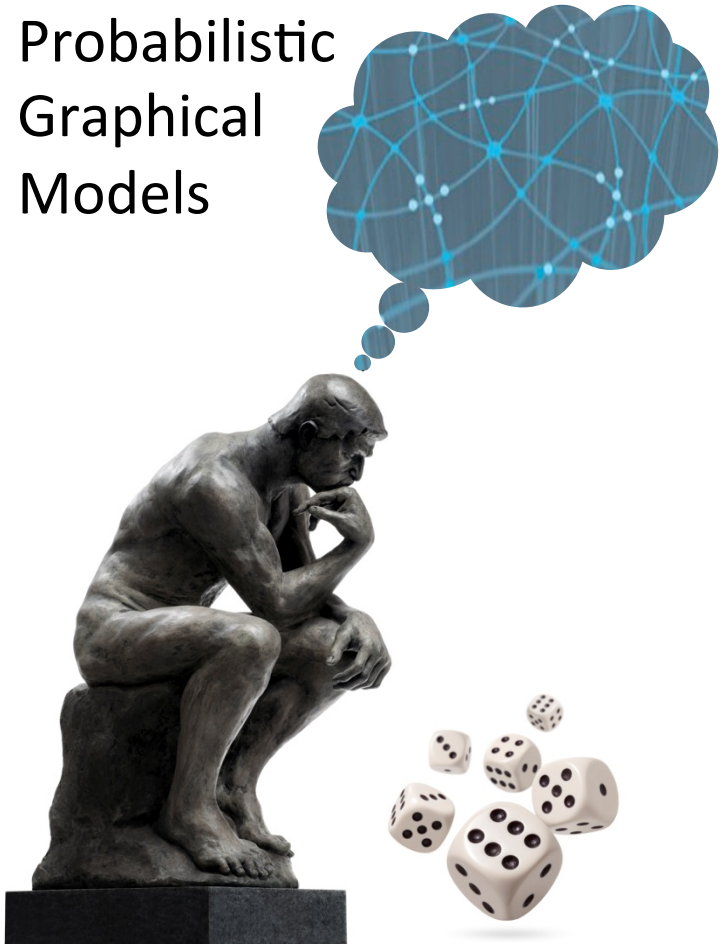
Conditioning can Lose Independences

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

Probabilistic Graphical Models



Representation

Independencies

Bayesian Networks

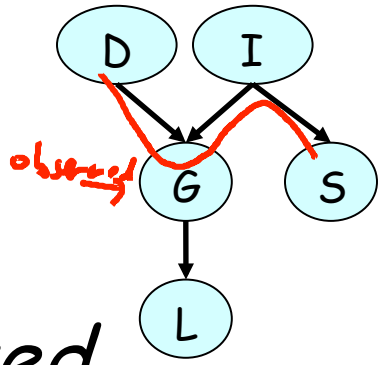
Independence & Factorization

$P(X,Y) = P(X)\,P(Y)$ — 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**)

Factorization ⇨ Independence: BNs

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

chain rule

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

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

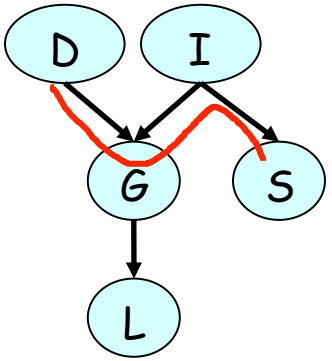
$$= \sum_{I} P(D)P(I)P(S \mid I) \sum_{G} (P(G \mid D,I) \sum_{L} P(L \mid G))$$

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

$\phi_2(S)$

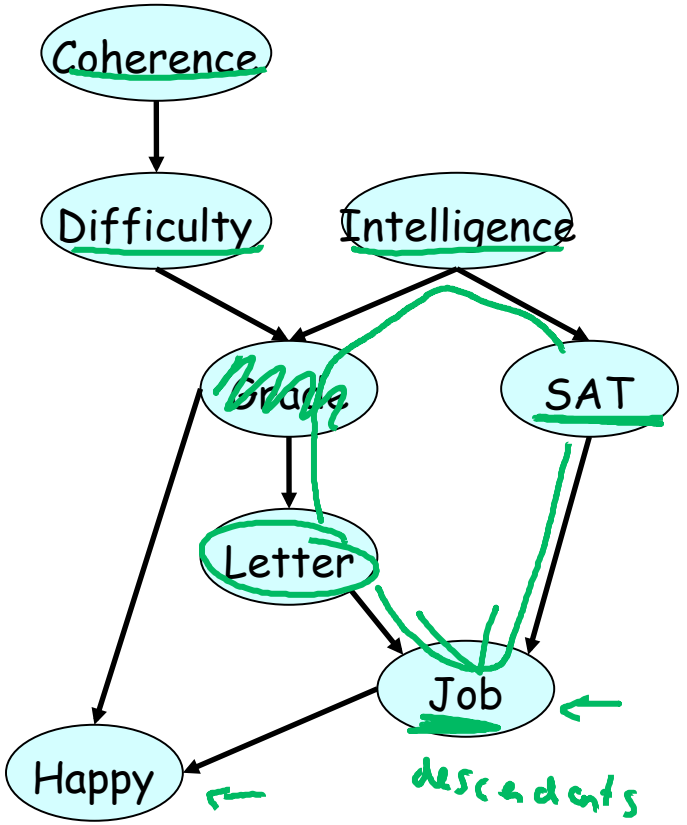
$\phi_1(D)$

P ⊨ D ⊥ S



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

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 ⇨ 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

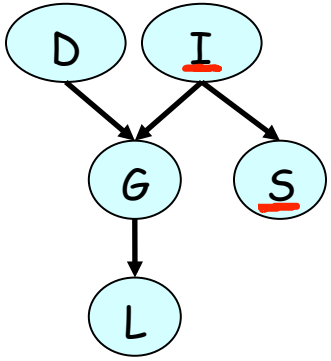
Factorization ⇨ Independence: BNs

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



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

Daphne Koller

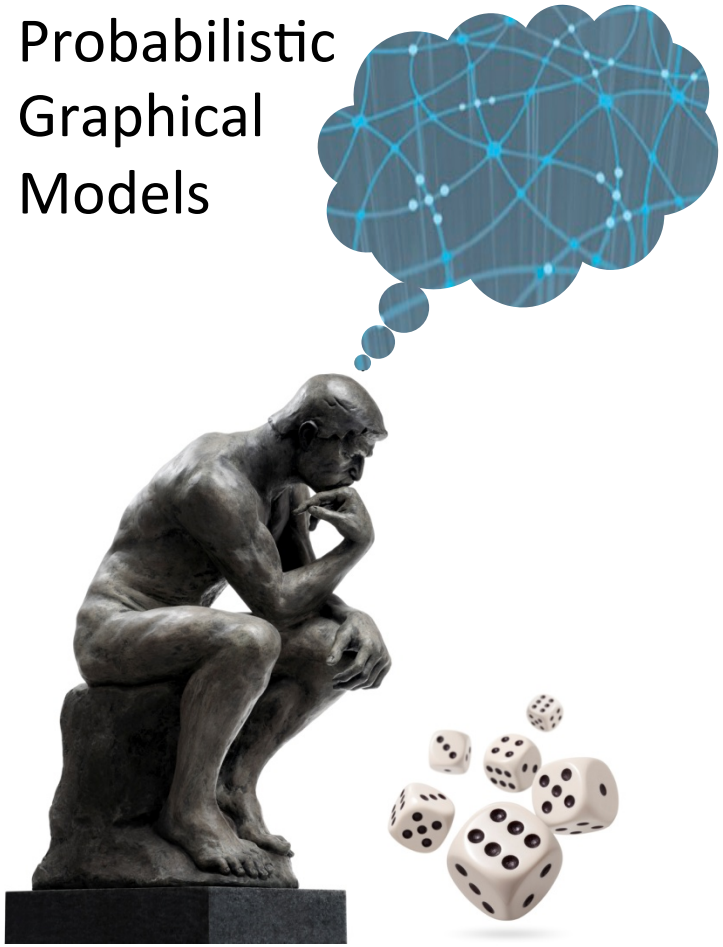
Summary

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)

Probabilistic Graphical Models

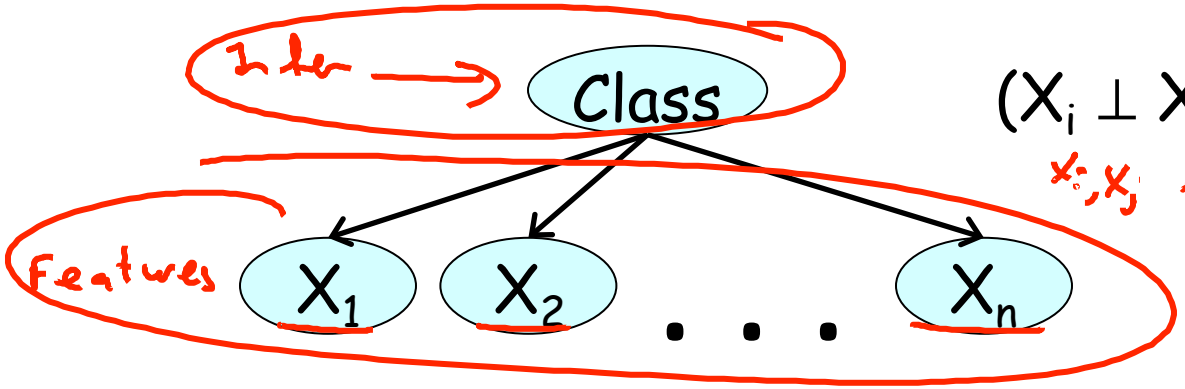


Representation

Bayesian Networks

Naïve Bayes

Naïve Bayes Model

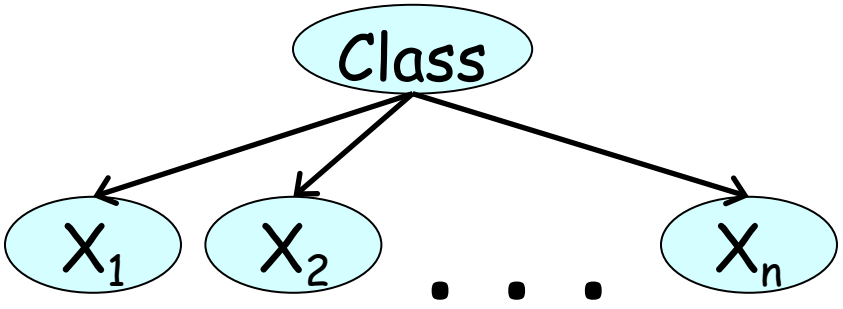


$(X_i \perp X_j \mid C)$ for all X_i, X_j

X_i, X_j conditionally independent given C

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

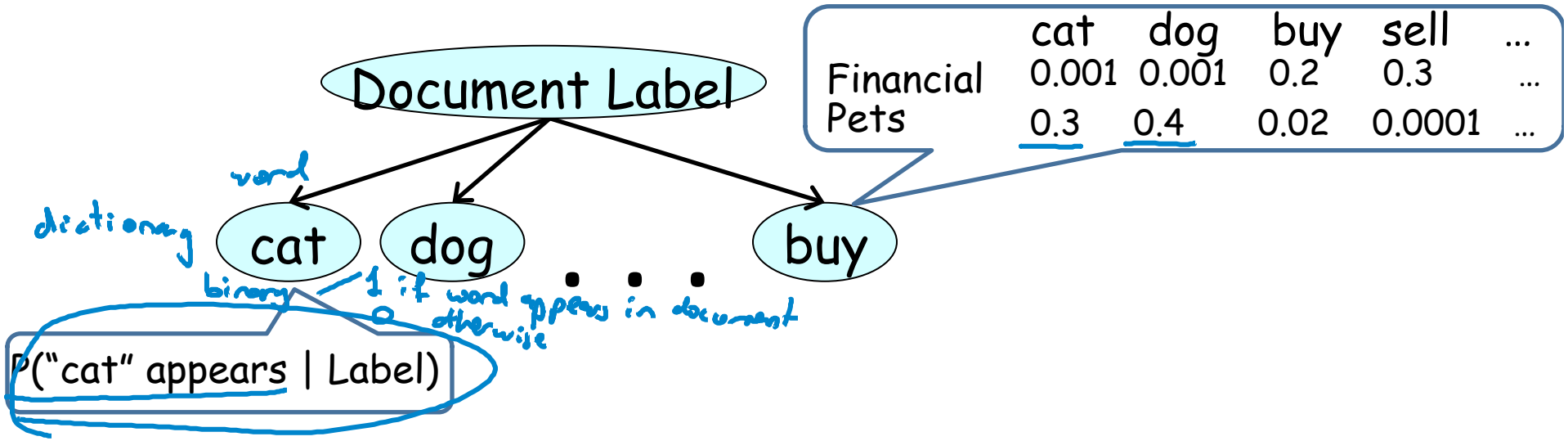
Naïve Bayes Classifier



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

odds ratios

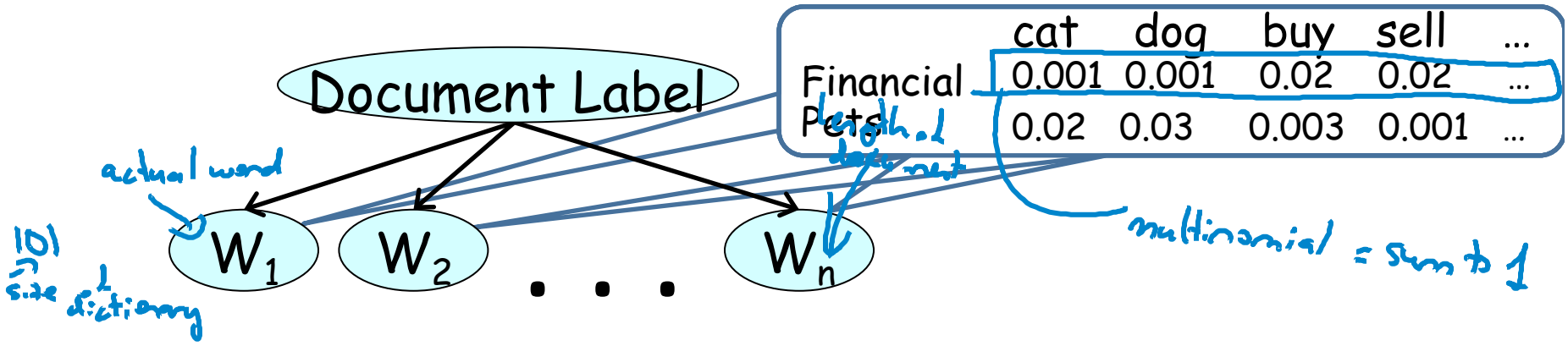
Bernoulli Naïve Bayes for Text



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

Daphne Koller

Multinomial Naïve Bayes for Text



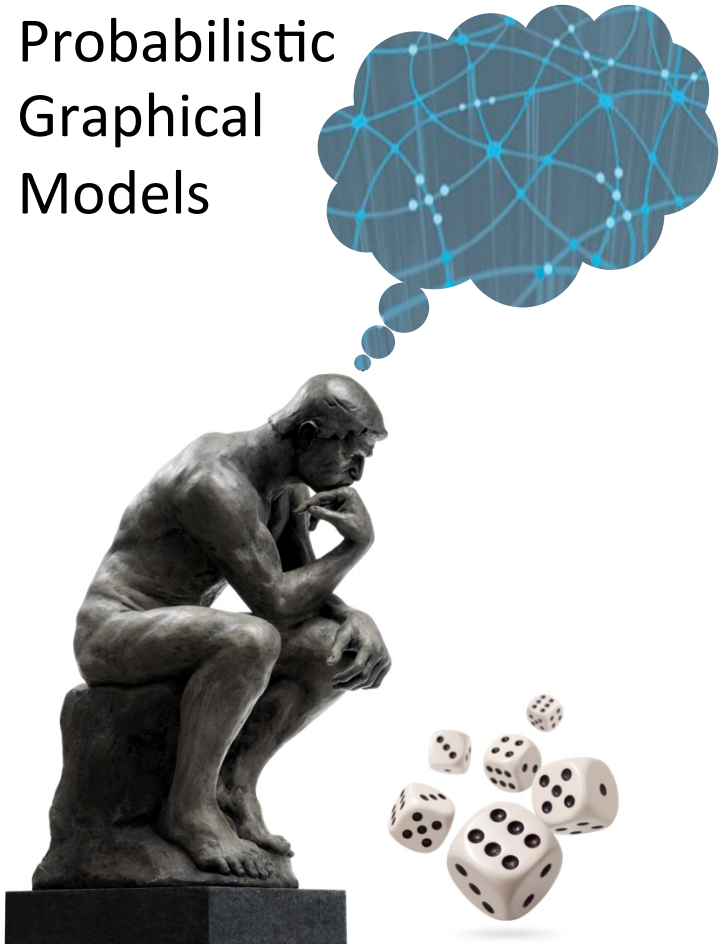
$$\frac{P(C = c^1 \mid x_1, \ldots, x_n)}{P(C = c^2 \mid x_1, \ldots, x_n)} = \frac{P(C = c^1)}{P(C = c^2)} \prod_{i=1}^{n} \frac{P(x_i \mid C = c^1)}{P(x_i \mid C = c^2)}$$

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)

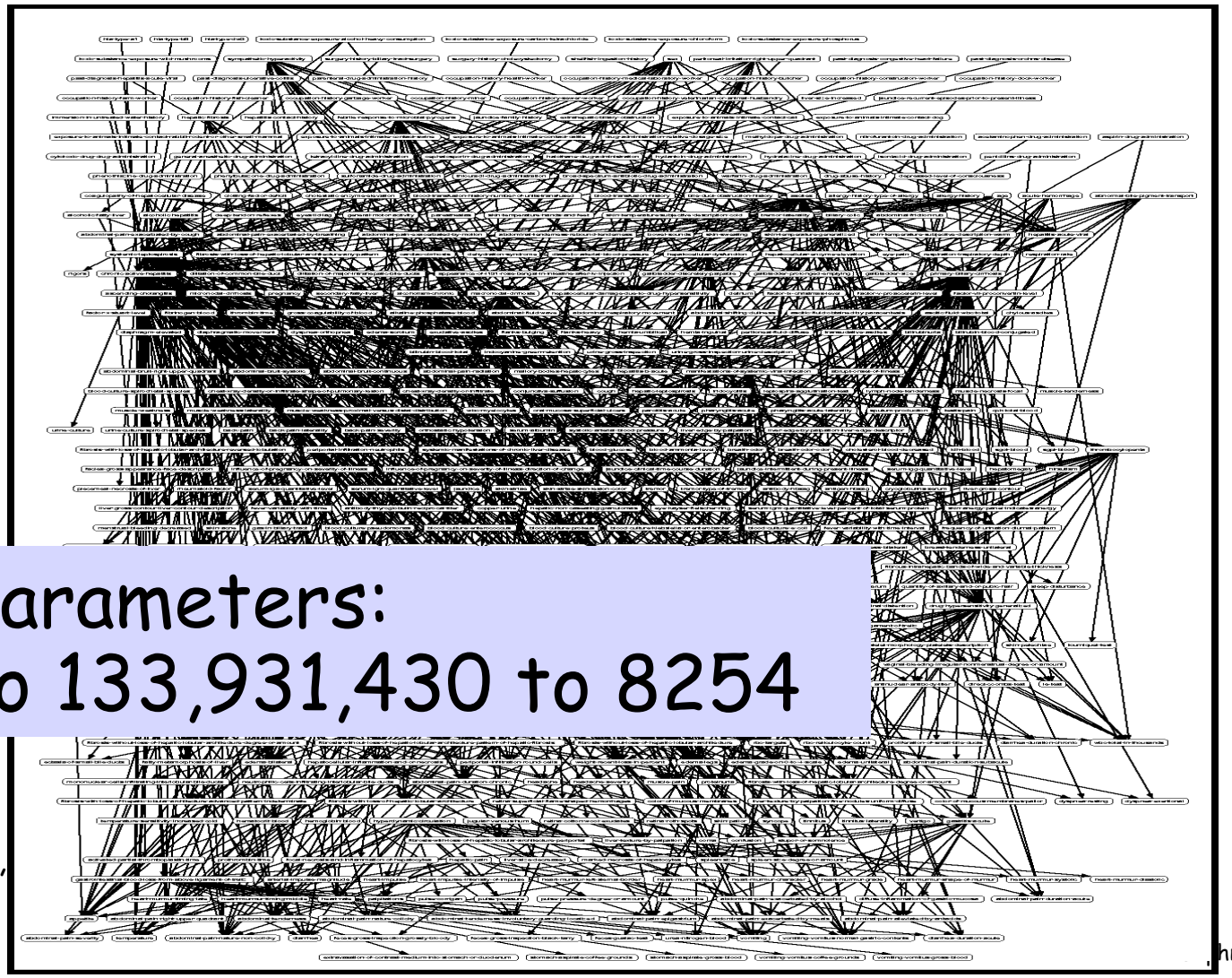
Heckerman et al.

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.

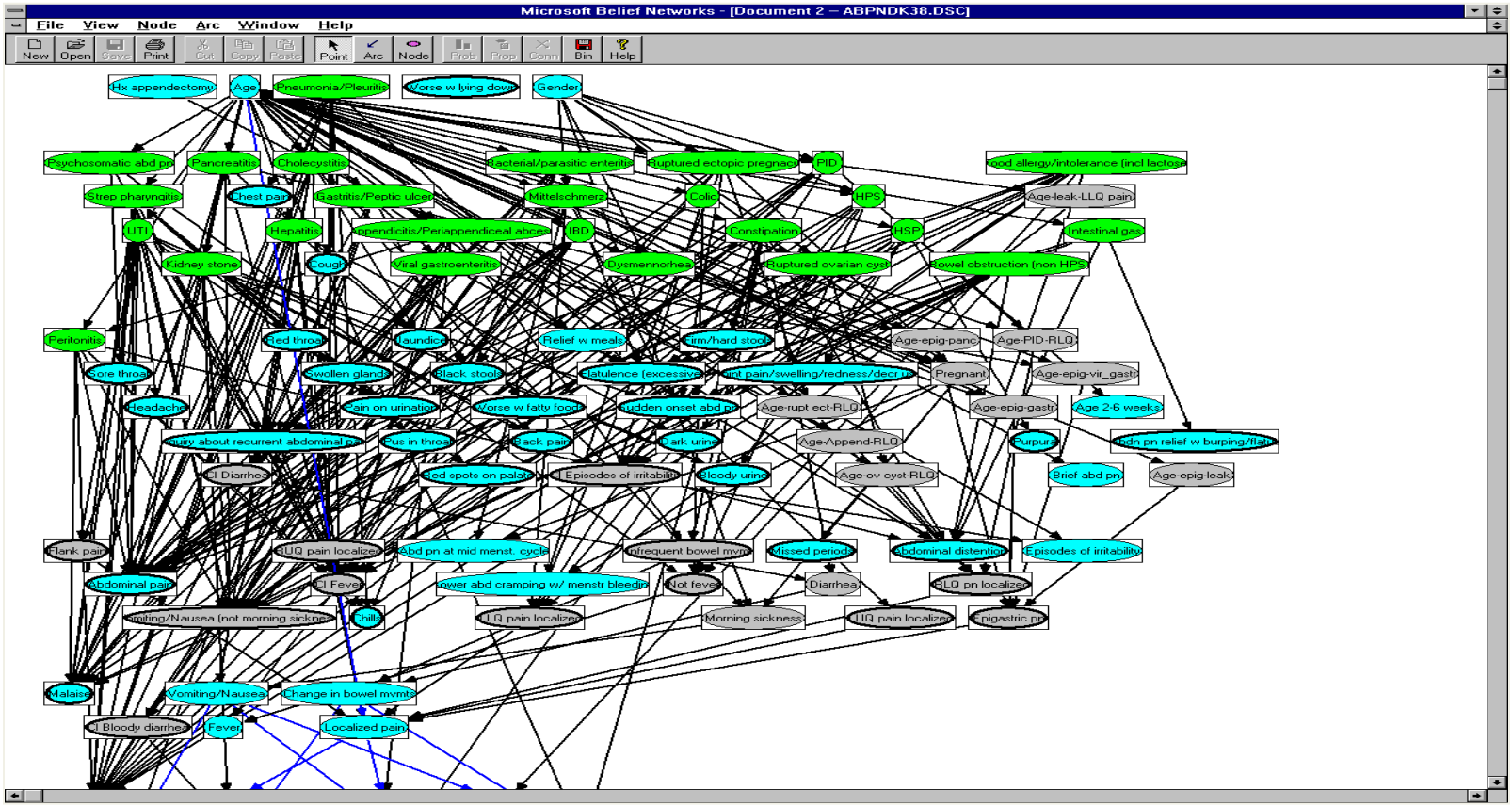
CPCS



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

M. Pradhan , G. Provan ,
B. Middleton , M.Henrion,
UAI 94

Medical Diagnosis (Microsoft)



Thanks to: Eric Horvitz, Microsoft Research

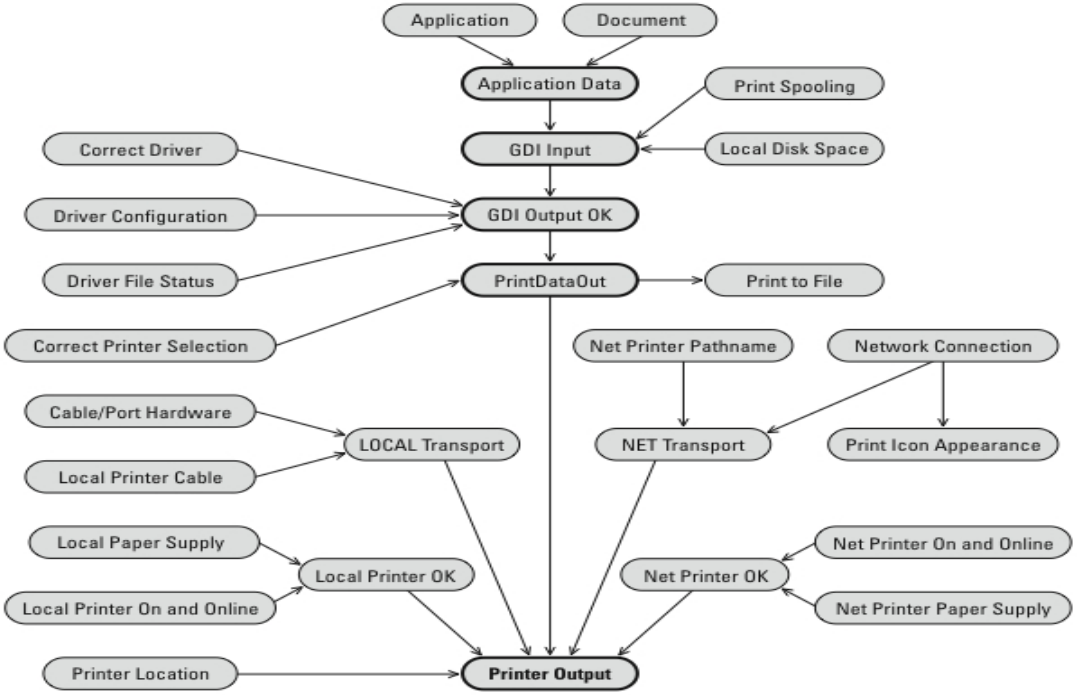
Daphne Koller

Medical Diagnosis (Microsoft)

Thanks to: Eric Horvitz, Microsoft Research

Fault Diagnosis

- Microsoft troubleshooters



Daphne Koller

Fault Diagnosis

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