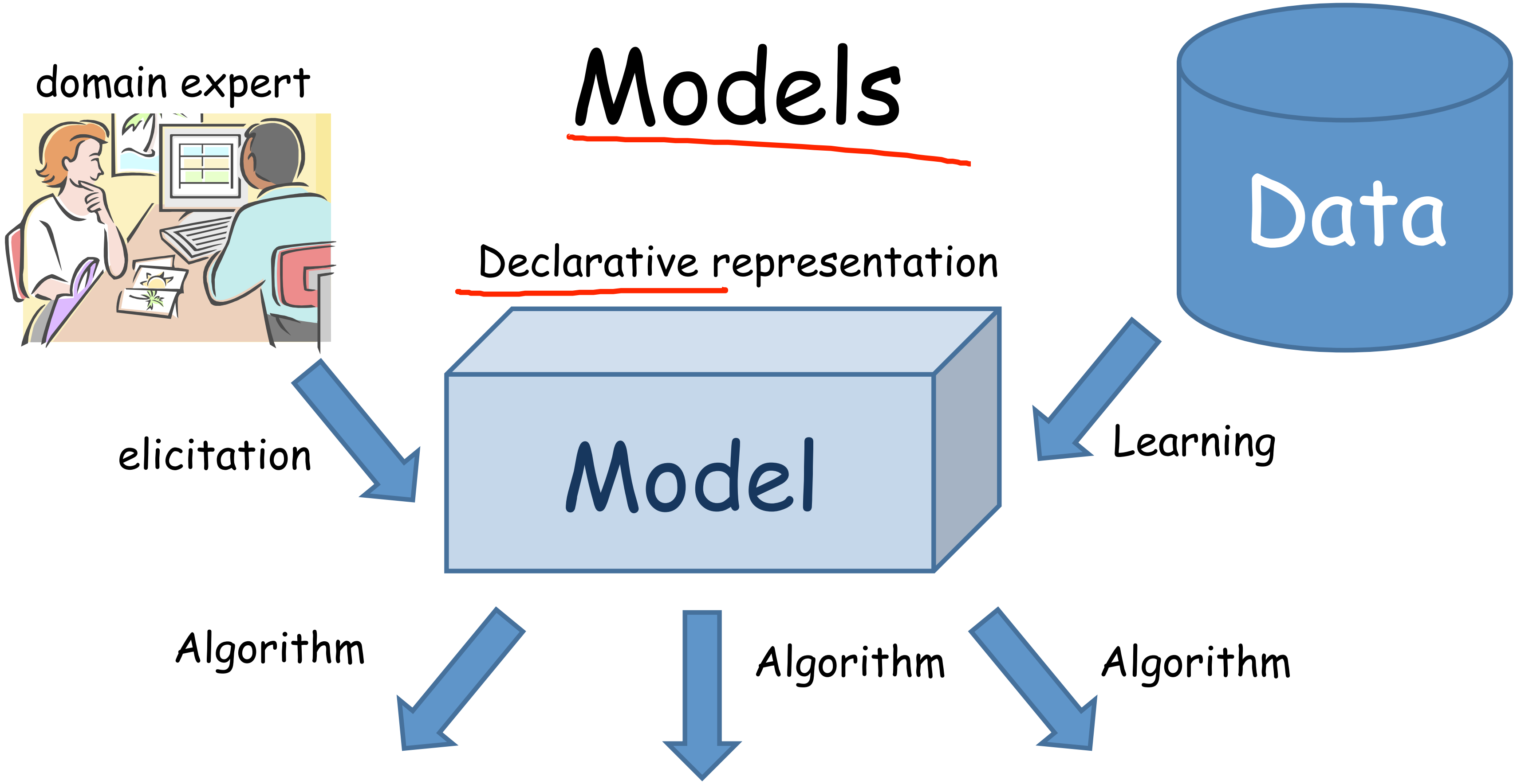
Models



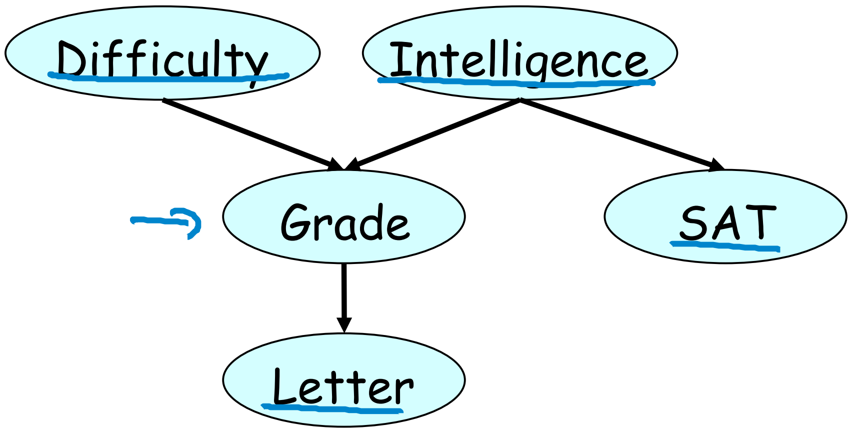
domain expert
Declarative representation
Data
elicitation
Model
Learning
Algorithm
Algorithm
Algorithm

$X_1 \ldots X_n$ — nodes

Graphical Models

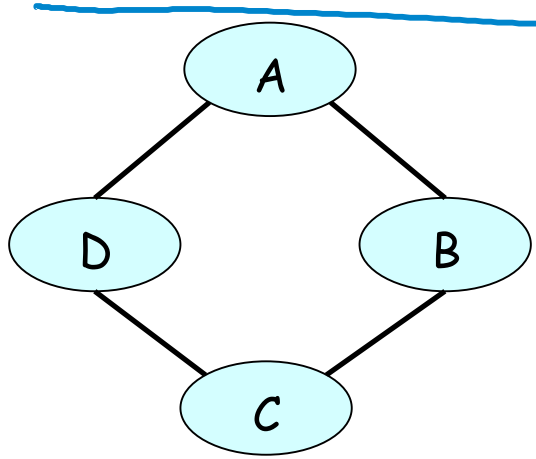
directed graph

Bayesian networks



undirected graph

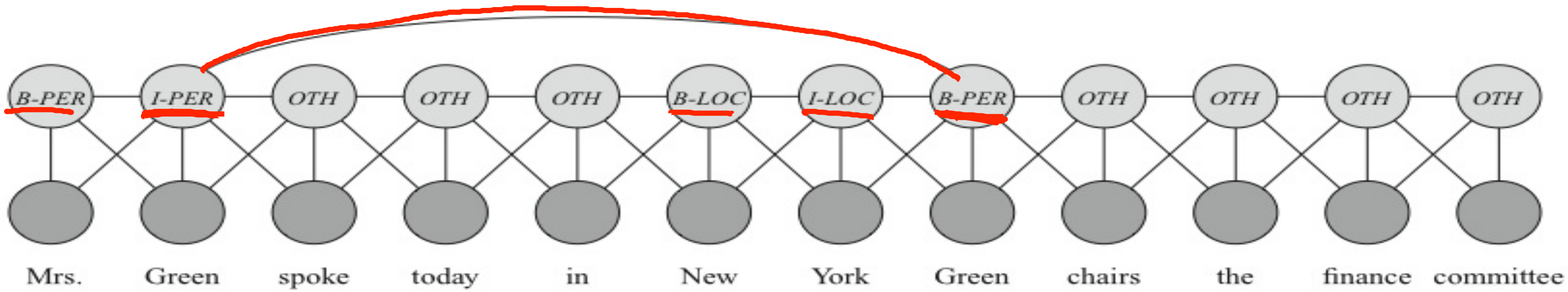
Markov networks



Textual Information Extraction

Mrs. Green spoke today in New York. Green chairs the finance committee.

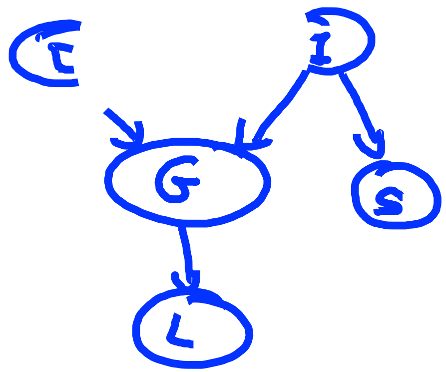
Person Location Person organization

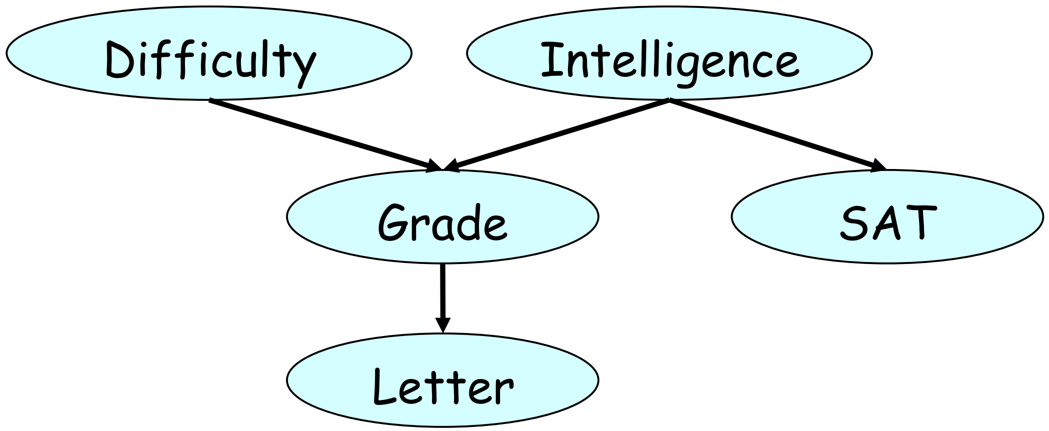


Daphne Koller

- **G**rade
- Course **D**ifficulty
- Student **I**ntelligence
- Student **S**AT
- Reference **L**etter

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




Difficulty
Intelligence
Grade
SAT
Letter

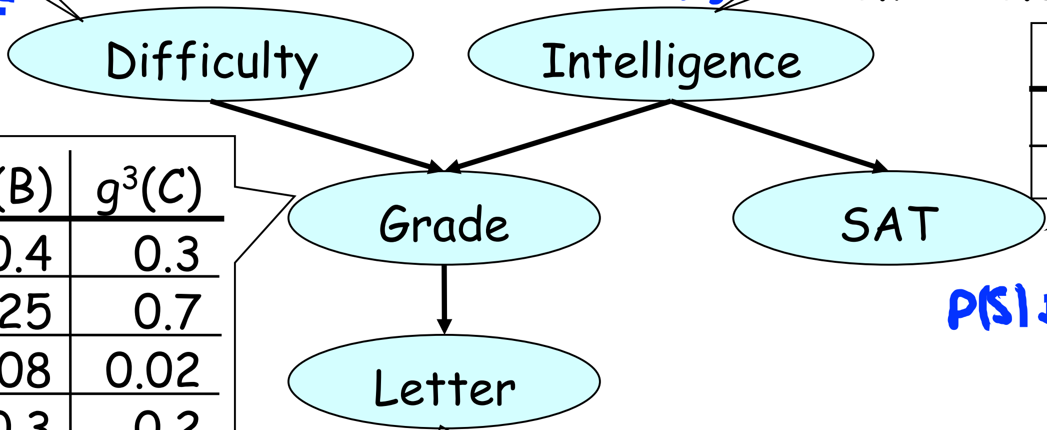
CPD = cond. prob. dist.

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)

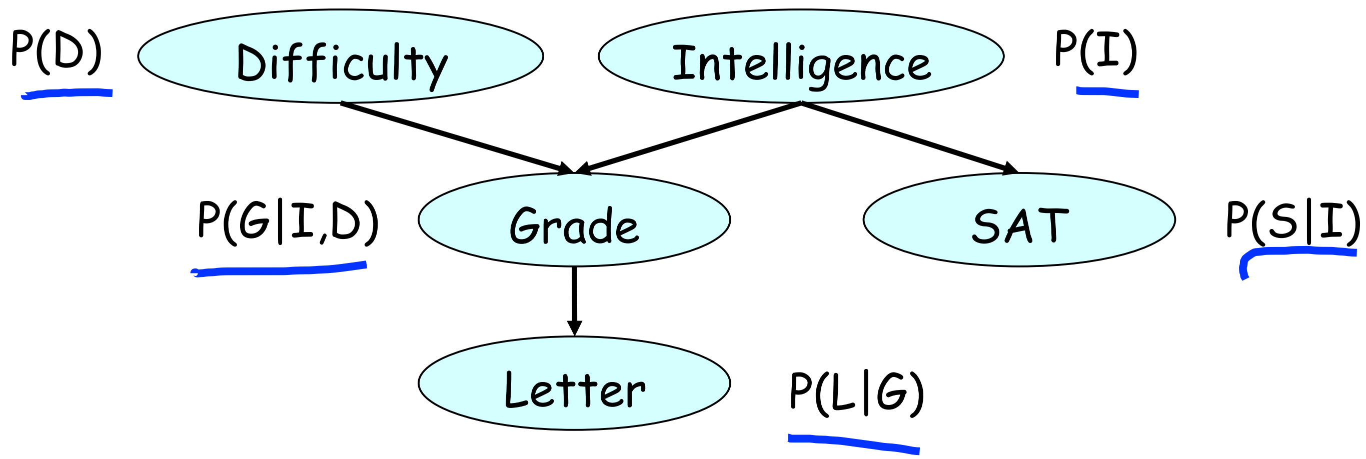
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

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

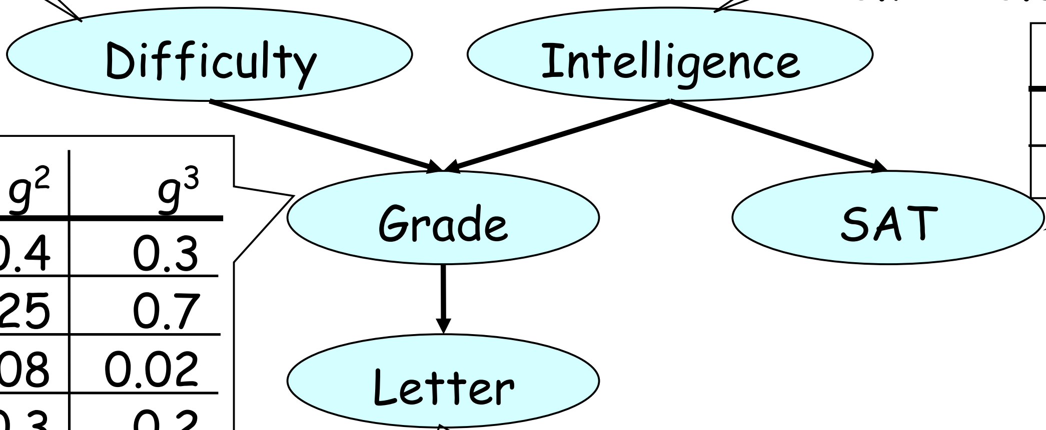
Daphne Koller

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

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

$0.6 * 0.3 * 0.02 * 0.01 * 0.8$

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

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

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

P is a product of CPDs

CPDs are non-negative

Daphne Koller

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

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

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

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

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

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

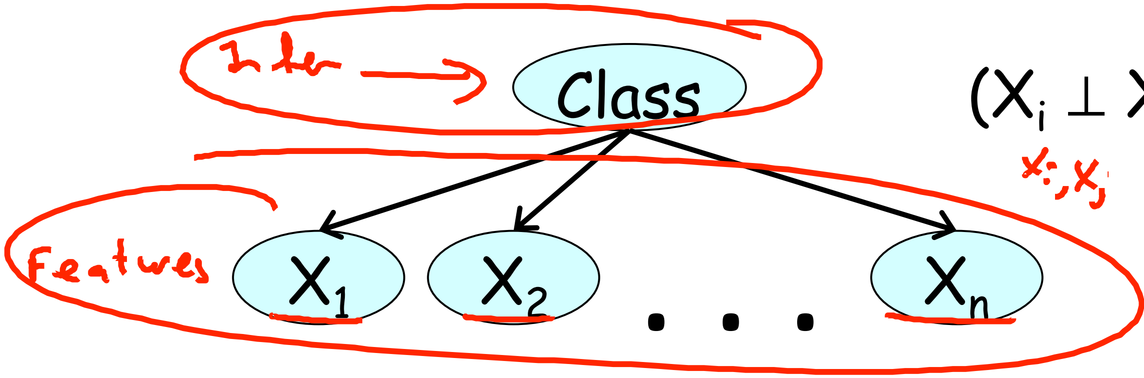
Daphne Koller

P Factorizes over G

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

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

Naïve Bayes Model



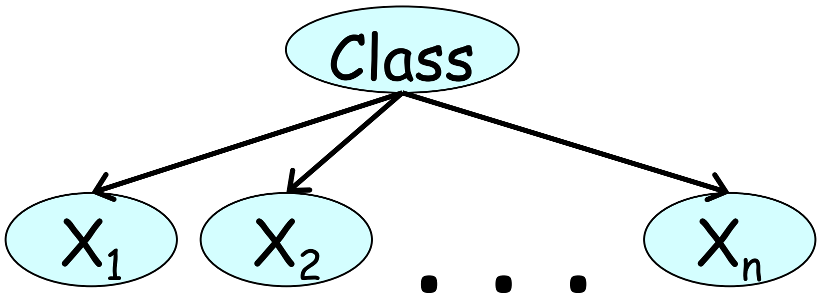
$(X_i \perp X_j \mid C)$ for all X_i, X_j

x_i, x_j conditionally independent given C

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

Daphne Koller

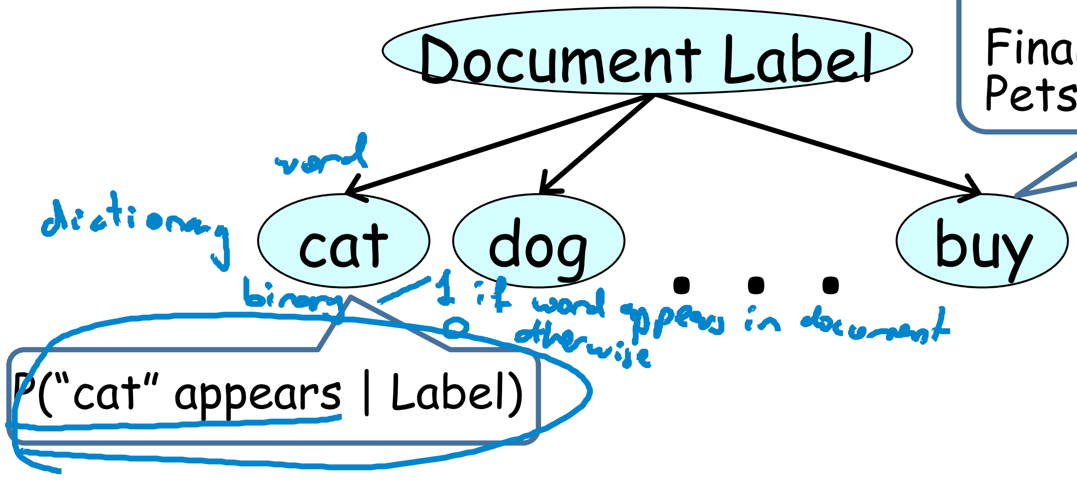
Naïve Bayes Classifier



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

Bernoulli Naïve Bayes for Text

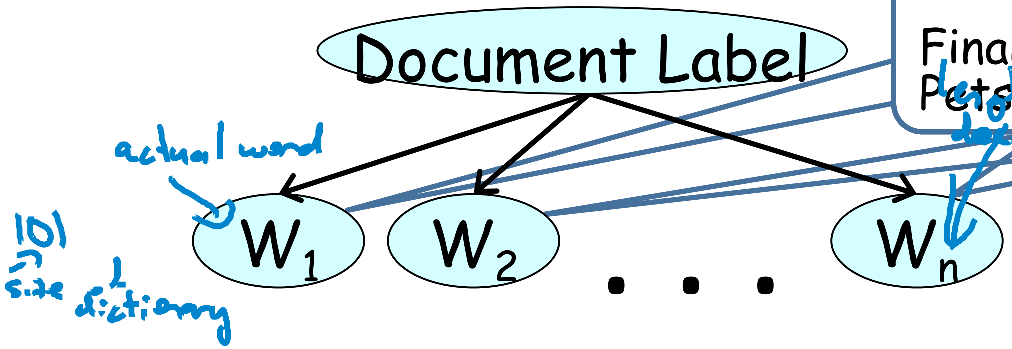
	cat	dog	buy	sell	...
Financial	0.001	0.001	0.2	0.3	...
Pets	0.3	0.4	0.02	0.0001	...



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

Daphne Koller

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	...

multinomial = sum to 1

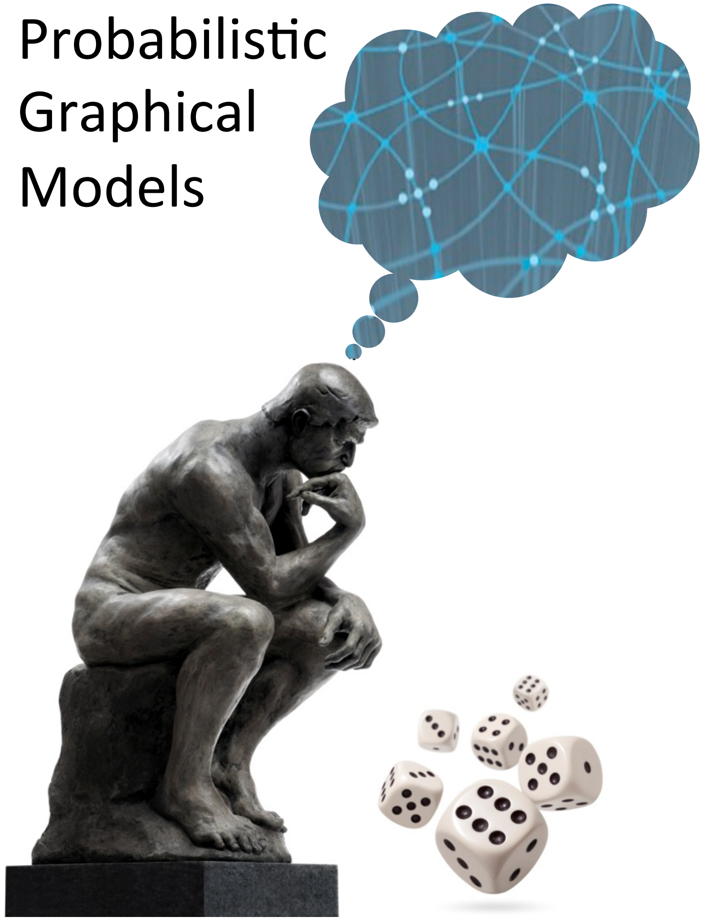
$$\frac{P(C=c^1 \mid x_1,\ldots,x_n)}{P(C=c^2 \mid x_1,\ldots,x_n)} = \frac{P(C=c^1)}{P(C=c^2)} \prod_{i=1}^{n} \frac{P(x_i \mid C=c^1)}{P(x_i \mid C=c^2)}$$

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.