Graphical Models

Lecture 3:

Local Conditional Probability Distributions

Andrew McCallum
mccallum@cs.umass.edu

Thanks to Noah Smith and Carlos Guestrin for some slide materials.

Conditional Probability Distributions

- Proper CPD:

$$\sum_{x \in \mathrm{Val}(X)} P(X = x \mid \mathbf{Parents}(X)) \quad = \quad 1$$

(for continuous case, integral)

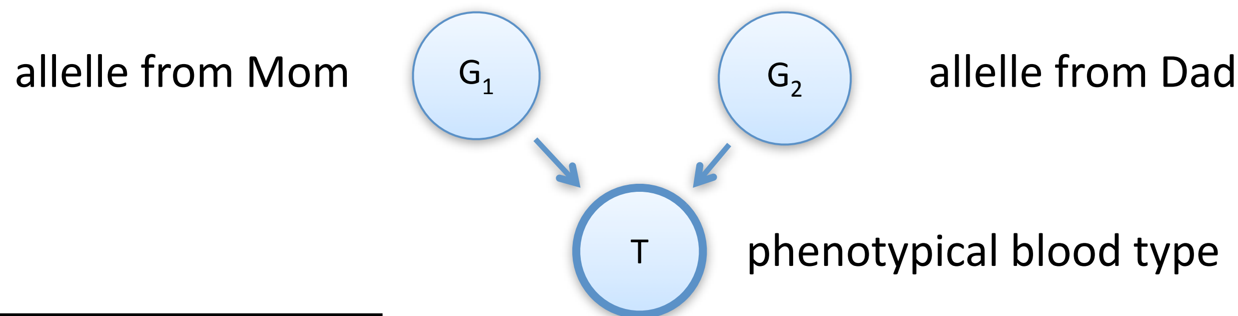
- Everything breaks down into distributions over one variable given some others.

Conditional Probability Distributions

- So far, we've seen **table** representations.
  - How many parameters?
  - Where will they come from?
  - Alternatives?
- If we know more about the form of a CPD, we may be able to infer more properties of P.
  - Independence assertions
  - Efficient inference (later)

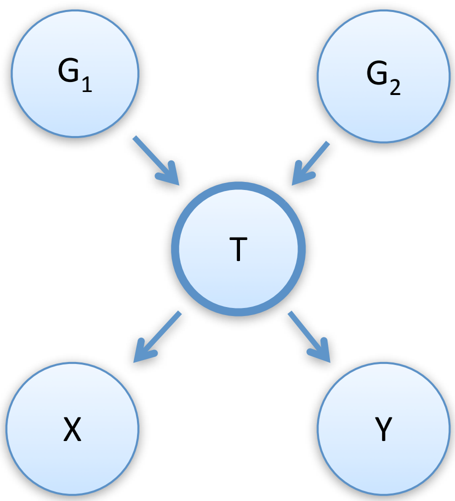
Deterministic CPDs

- Sometimes a variable's value is a deterministic function of its parents' values.



$P(T \mid G_1, G_2)$	G_i = a and G_{2-i} = b	G_i = a and $G_{2-i} \in \{a, o\}$	G_i = b and $G_{2-i} \in \{b, o\}$	$G_1 = G_2$ = o
ab	1	0	0	0
a	0	1	0	0
b	0	0	1	0
o	0	0	0	1

Deterministic CPDs Affect I(G)



- $X \perp Y \mid T$
  (local Markov assumption)
- $X \perp Y \mid \{G_1, G_2\}$
- Can we derive such independence assertions?

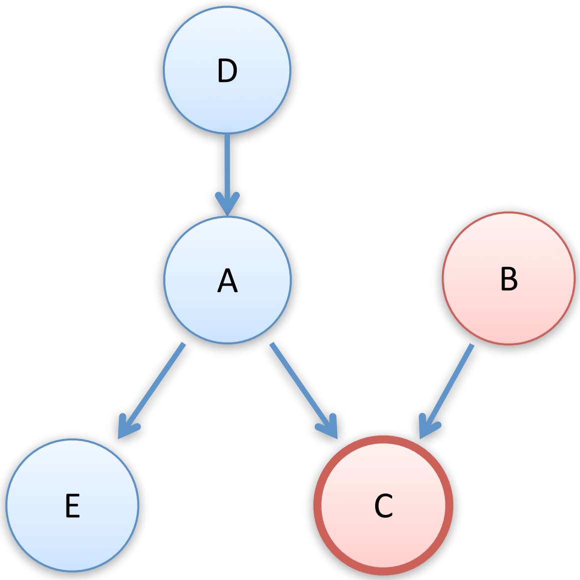
Review: D-Separation

- Three sets of nodes:
  X, **Y**, and observed nodes **Z**
- **X** and **Y** are **d-separated** given **Z** if there is no active trail from any $X \in \mathbf{X}$ to any $Y \in \mathbf{Y}$ given **Z**.

D-Separation with Deterministic CPDs

- Query: $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$?
- Deterministic variables: **D**
- Algorithm:
  - Let **Z'** = **Z**
  - While there is an $X_i \in \mathbf{D}$ such that **Parents**(X_i) $\subseteq$ **Z'**, add X_i to **Z'**
  - Calculate d-separation between **X** and **Y** given **Z'**
- This is sound and complete.
  - But if we know *still more* about the deterministic functions involved, we may be able to go farther!

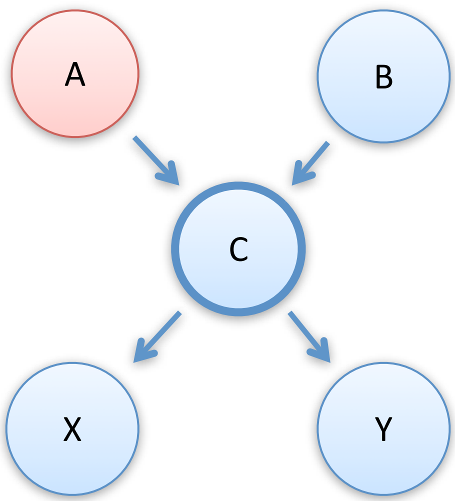
Another Example



- Suppose C = A XOR B
- In the case of XOR,
  B and C fully determine A

- D ⊥ E | {B, C}

Context-Specific Independence



- Suppose C = A OR B
- We are given A = 1.
  - This implies that C = 1.
- P(X | B, A=1) = P(X | A=1)
  - Independence!
- This does not work if A = 0.

Previously we defined X ⊥ Y | Z to represent the assumption P(X|Y, Z) = P(X|Z) for all values of X, Y, Z.
Deterministic functions can imply a type of independence that holds only for particular values of some variables.

Context-Specific Independence

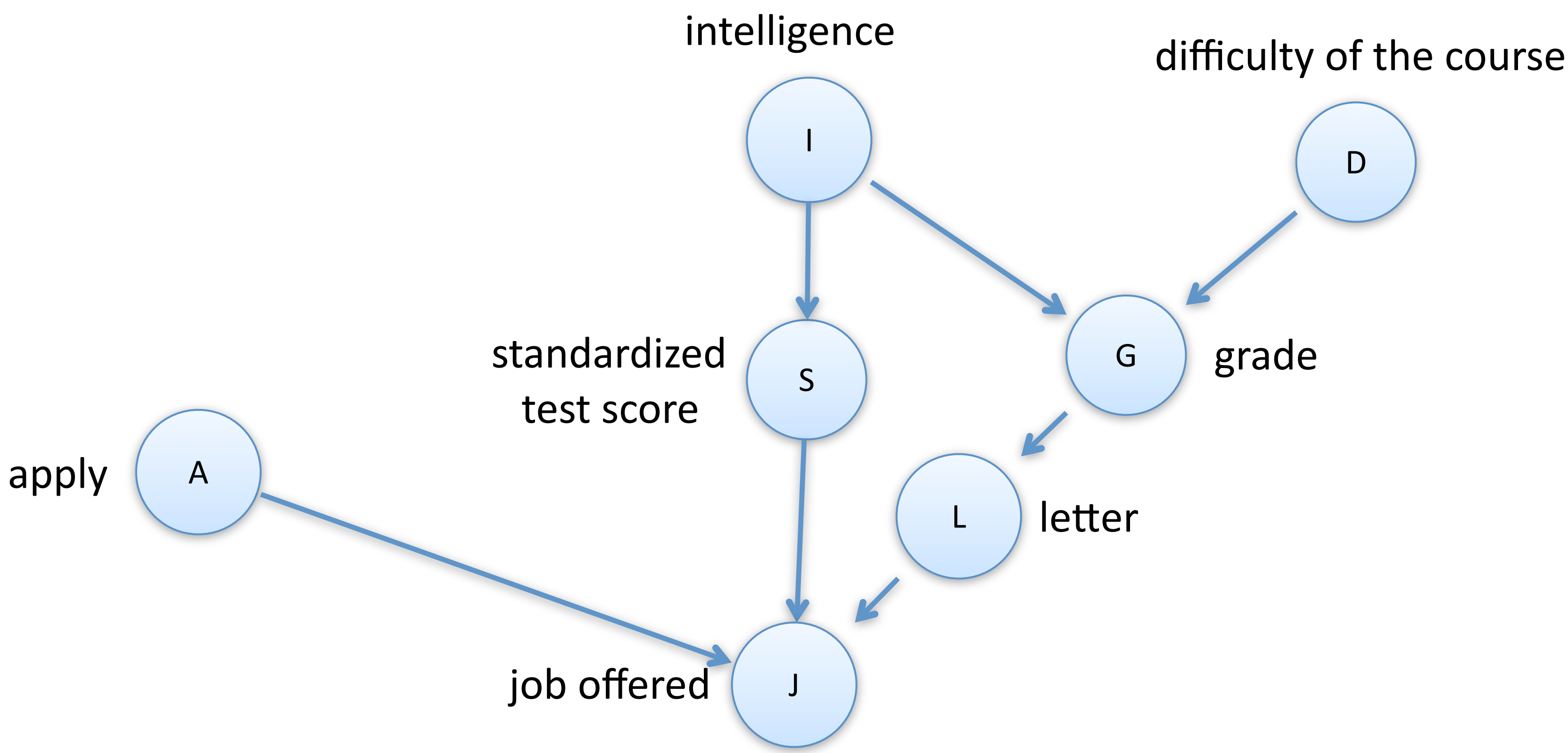
Definition

- Let **X**, **Y**, and **Z** be disjoint sets of variables; **C** is a set of variables that could overlap.
- Let **c** $\in$ Val(**C**).
- **X** and **Y** are conditionally independent given **Z** and the context **c**:
  - $\mathbf{X} \perp_{\mathbf{c}} \mathbf{Y} \mid \mathbf{Z}$
  - $P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{C} = \mathbf{c}) = P(\mathbf{X} \mid \mathbf{Z}, \mathbf{C} = \mathbf{c})$
    whenever $P(\mathbf{Y}, \mathbf{Z}, \mathbf{C} = \mathbf{c}) > 0$

Tree CPDs

- Use a tree to represent P(X | **Parents**(X))
  - Each leaf is a distribution over X
  - Each path defines a context

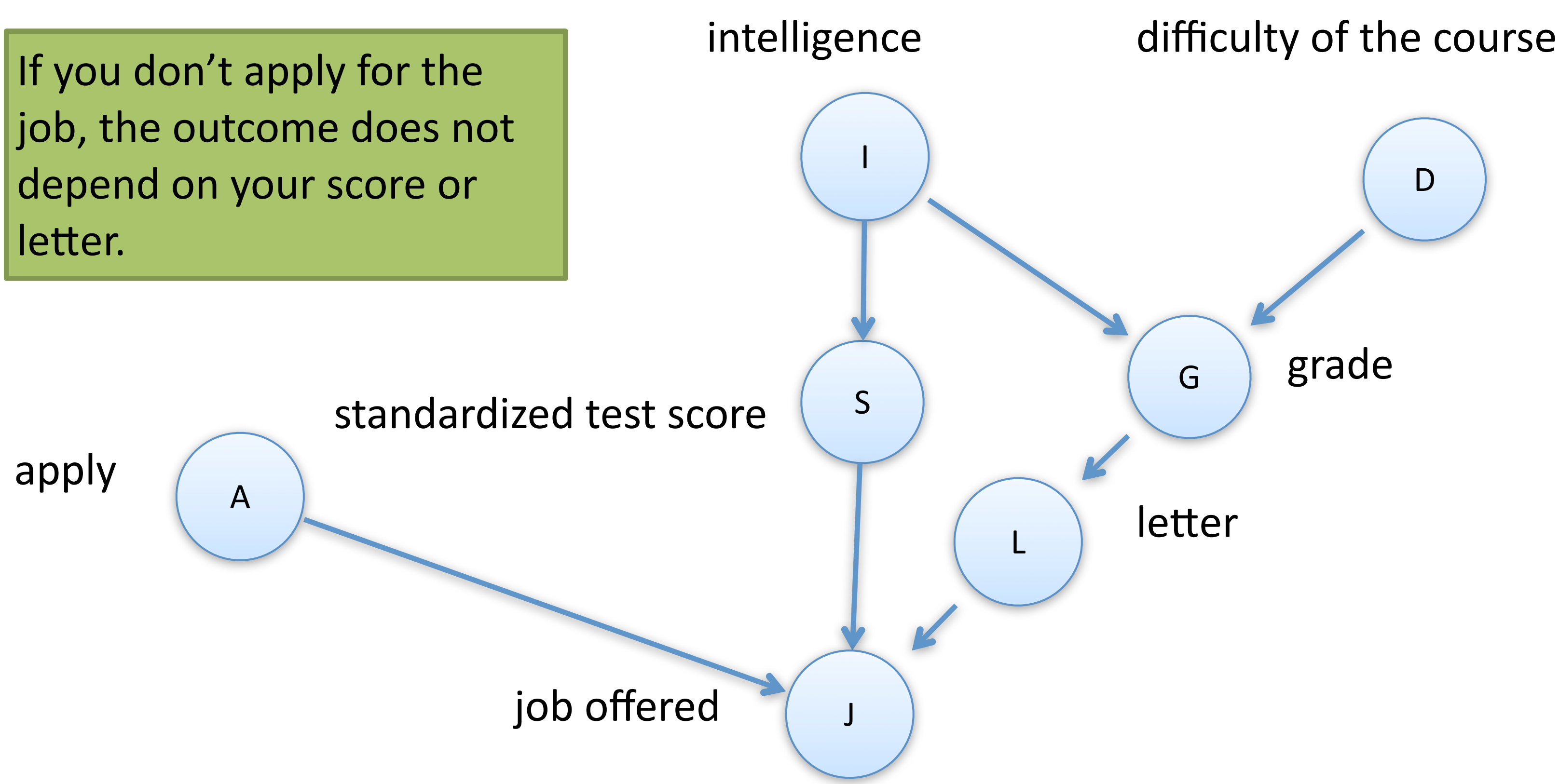
Student Example



P(J | A, S, L)	000	001	010	011	100	101	110	111
yes								
no								

Student Example

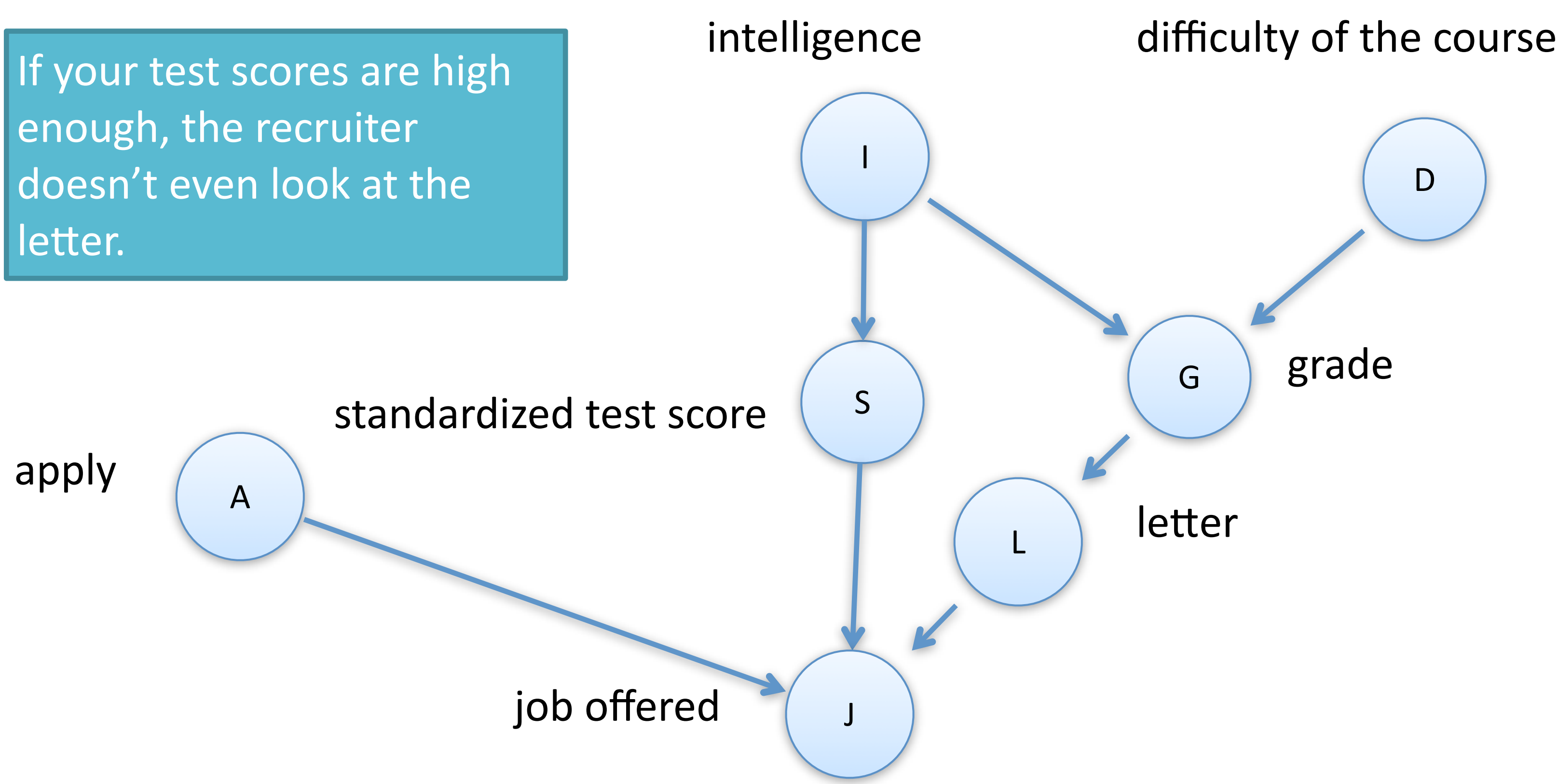
If you don't apply for the job, the outcome does not depend on your score or letter.



P(J | A, S, L)	0??	100	101	110	111
yes					
no					

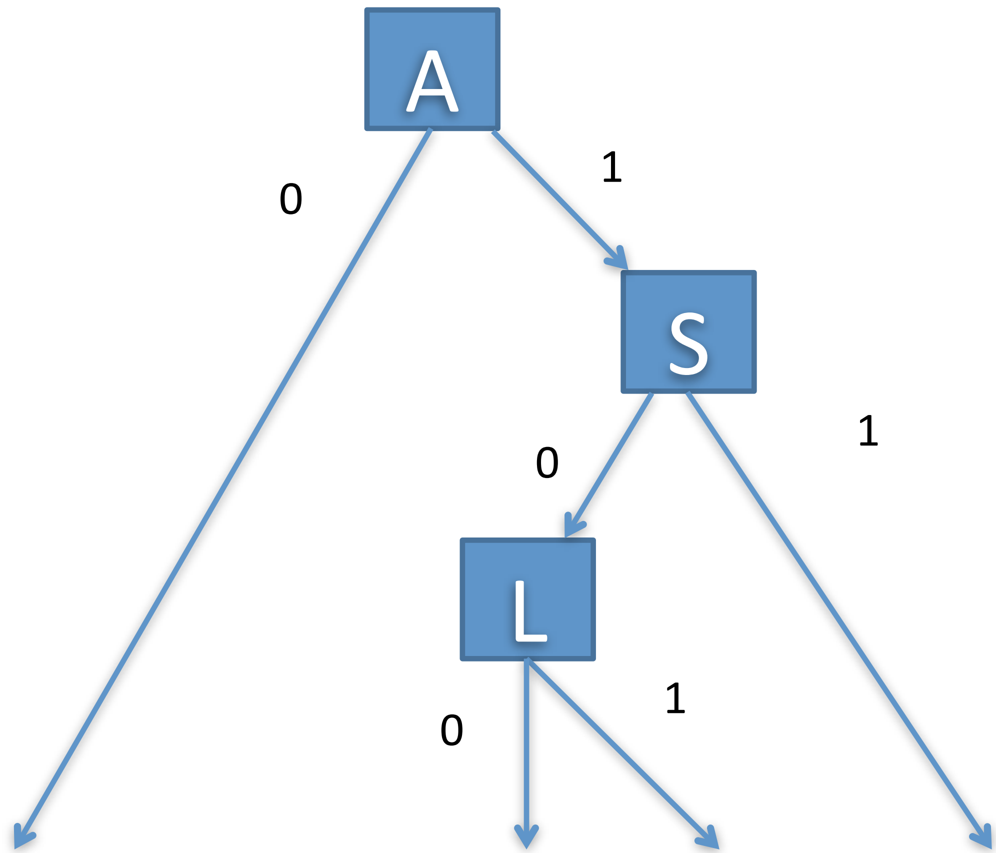
Student Example

If your test scores are high enough, the recruiter doesn't even look at the letter.



P(J | A, S, L)	0??	100	101	11*
yes				
no				

Student Example



P(J | A, S, L)	0??	100	101	11*
yes				
no				

Other Structured CPDs

- Rules
  - CPD trees can always be represented compactly as rules, but the converse does not hold
- Decision diagrams
- Any kind of partition structure of Val(X) $\times$ Val(**Parents**(X))

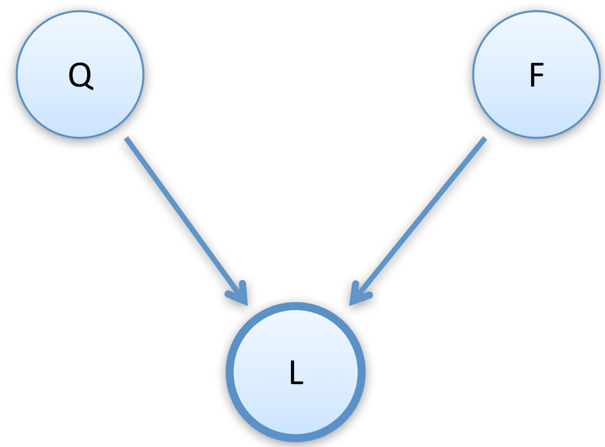
- Context-specific CPDs can make some edges spurious (given the context)!

Independent Causes

- A different kind of CPD structure
- Consider random variable Y and its parents, **Parents**(Y) = **X**
- Two examples
  - Noisy OR
  - Generalized linear models

Another Professor (Perfect World)

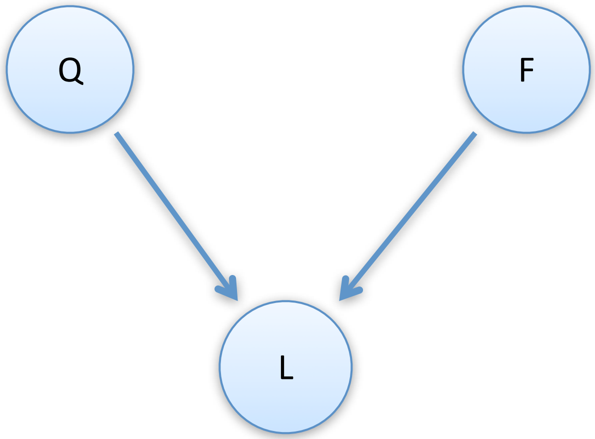
- Letter quality depends on whether you participated by asking good questions (Q), and on whether you wrote a good final paper (F)



P(L | Q, F)	Q = 1 and F = 1	Q = 1 and F = 0	Q = 0 and F = 1	Q = 0 and F = 0
high	1	1	1	0
low	0	0	0	1

Another Professor (Real World)

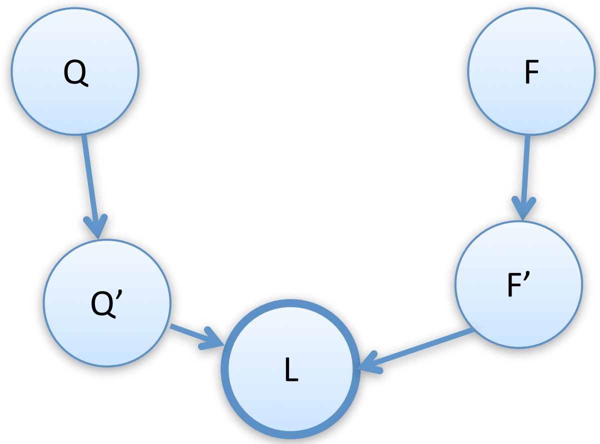
- Letter quality depends on whether you participated by asking good questions (Q), and on whether you wrote a good final paper (F)



P(L | Q, F)	Q = 1 and F = 1	Q = 1 and F = 0	Q = 0 and F = 1	Q = 0 and F = 0
high	0.98	0.8	0.9	0
low	0.02	0.2	0.1	1

Another Professor (Real World)

P(Q’ | Q)	Q = 1	Q = 0
high	0.8	0
low	0.2	1

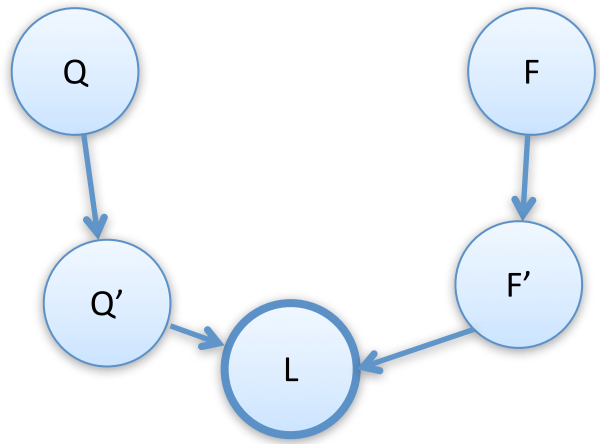


$L = Q' \vee F'$

P(F’ | F)	F = 1	F = 0
high	0.9	0
low	0.1	1

Another Professor

P(Q' | Q)	Q = 1	Q = 0
high	0.8	0
low	0.2	1



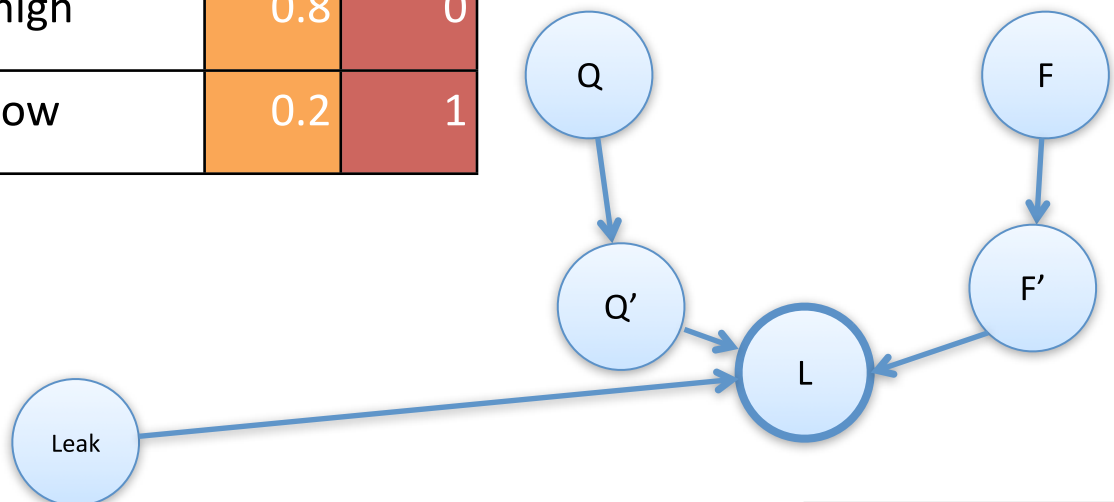
noise parameter, λ_i

$L = Q' \vee F'$

P(F' | F)	F = 1	F = 0
high	0.9	0
low	0.1	1

Another Professor

P(Q' | Q)	Q = 1	Q = 0
high	0.8	0
low	0.2	1



leak probability, λ_0

1	ε
0	$1 - \varepsilon$

$L = Q' \vee F' \vee Leak$

P(F' | F)	F = 1	F = 0
high	0.9	0
low	0.1	1

Noisy OR Model

λ_i is the noise parameter for X_i.

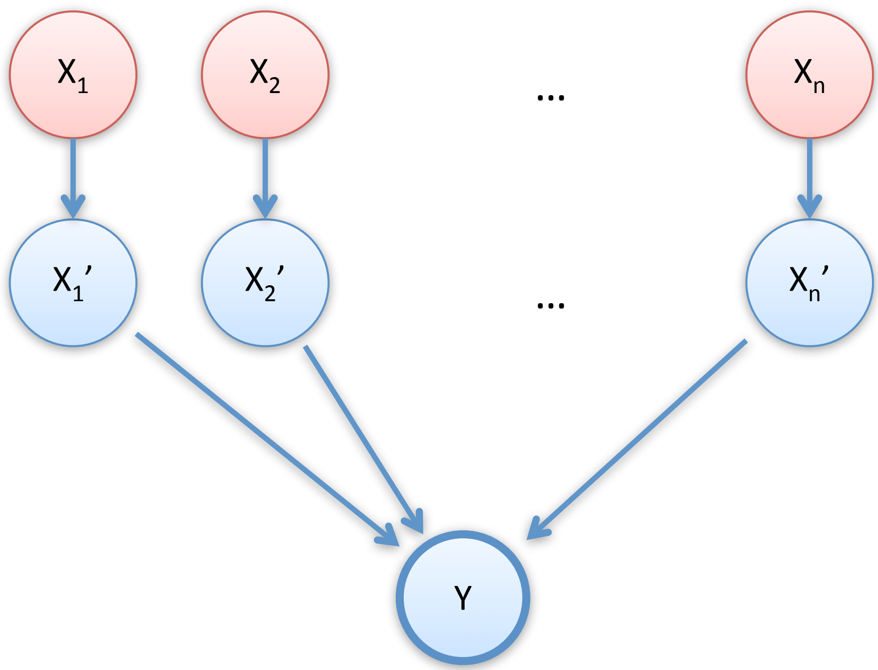
$$P(Y=0 \mid \boldsymbol{X}) = (1-\lambda_0) \prod_{i:X_i=1} (1-\lambda_i)$$

$$P(Y=1 \mid \boldsymbol{X}) = 1 - \left[(1-\lambda_0) \prod_{i:X_i=1} (1-\lambda_i)\right]$$

Or equivalently written (where $x^1 = 1$ and $x^0 = 0$)

$$P(y^0|x_1, ..., x_k) = (1-\lambda_0)\prod_{i=1}^{k}(1-\lambda_i)^{x_i}$$

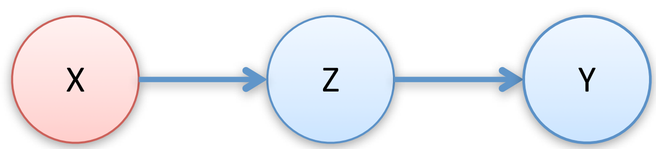
Noisy OR as a Conditional Bayesian Network



(What is a Conditional Bayesian Network?)

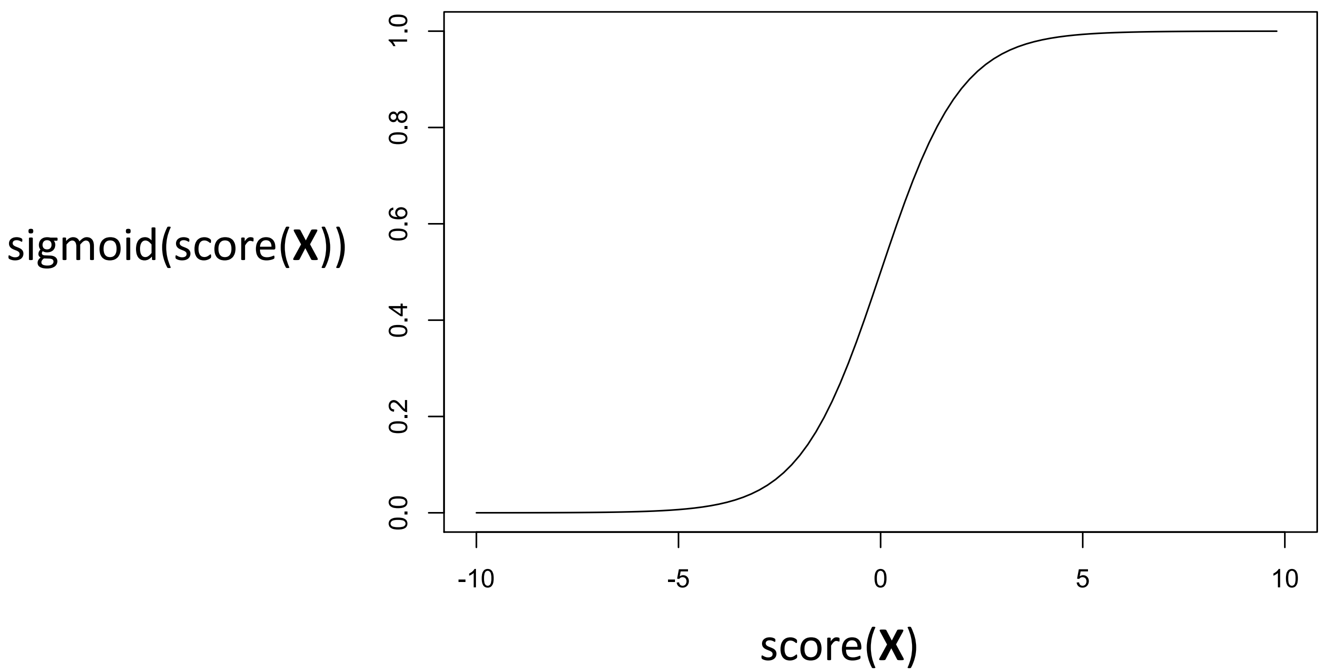
- Conditional Bayesian Network is a BN with three types of variables:
  - Inputs, always observed, no parents: **X**
  - Outputs: **Y**
  - Encapsulated: **Z**

$$P(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{X}) = \prod_{W \in \boldsymbol{Y} \cup \boldsymbol{Z}} P(W \mid \mathbf{Parents}(W))$$



Independent Causes

- Many "additive" effects combine to score **X**
- P(Y = 1) is defined as a function of **X**



$$\text{sigmoid}(z) = \frac{e^z}{1 + e^z}$$

Generalized Linear Model

- Score is defined as a *linear* function of **X**:

Z = f(X) is a random variable!

$$f(\boldsymbol{X}) = \underbrace{w_0 + \sum_i w_i X_i}_{Z}$$

- Probability distribution over binary value Y is commonly* defined by:

$$P(Y=1) = \text{sigmoid}(f(\boldsymbol{X}))$$

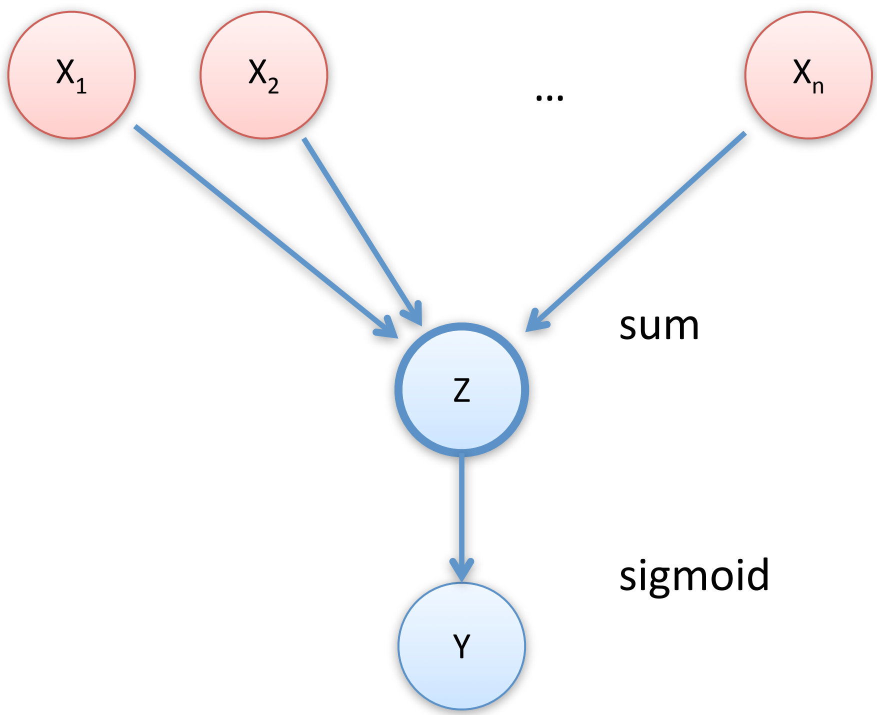
Logistic regression
Maximum entropy classifier

$$\text{sigmoid}(z) = \frac{e^z}{1+e^z}$$

aka "logit" function

* not the only choice

Logistic CPD Model as a Conditional Bayesian Network



Compare: Naïve Bayes model

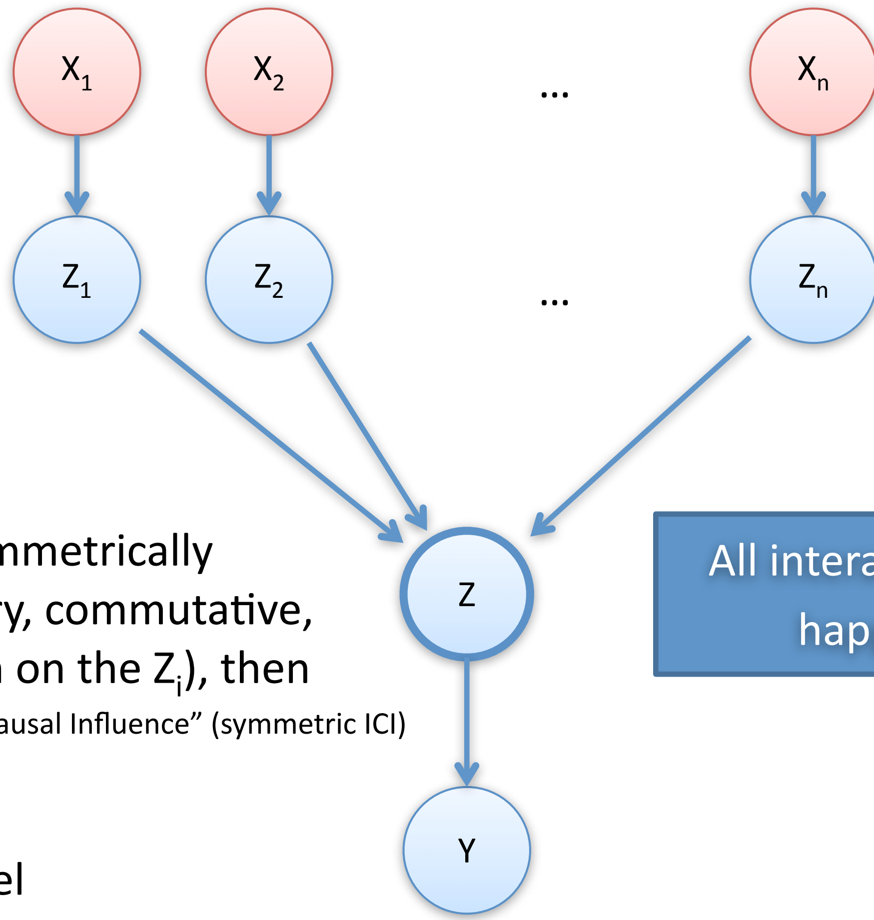
Logistic Models

- Weight w_i can be positive or negative.
- The **X** and Y do not need to be binary.
  - Very useful:
    multinomial logistic to allow many values for Y;
    indicator variables to allow many values for **X**
- "Multinomial logit"

CPDs with Causal Independence

Captures Noisy-OR and Generalized linear models

individual noise models



All interaction among X_i happens here!

deterministic and symmetrically decomposable (binary, commutative, associative operation on the Z_i), then

“Symmetric Independence of Causal Influence” (symmetric ICI)

Noisy OR:

$X \rightarrow Z_i$ = simple noisy model

Z_i s $\rightarrow$ Z = OR

$Z \rightarrow Y$ = copy

GLM:

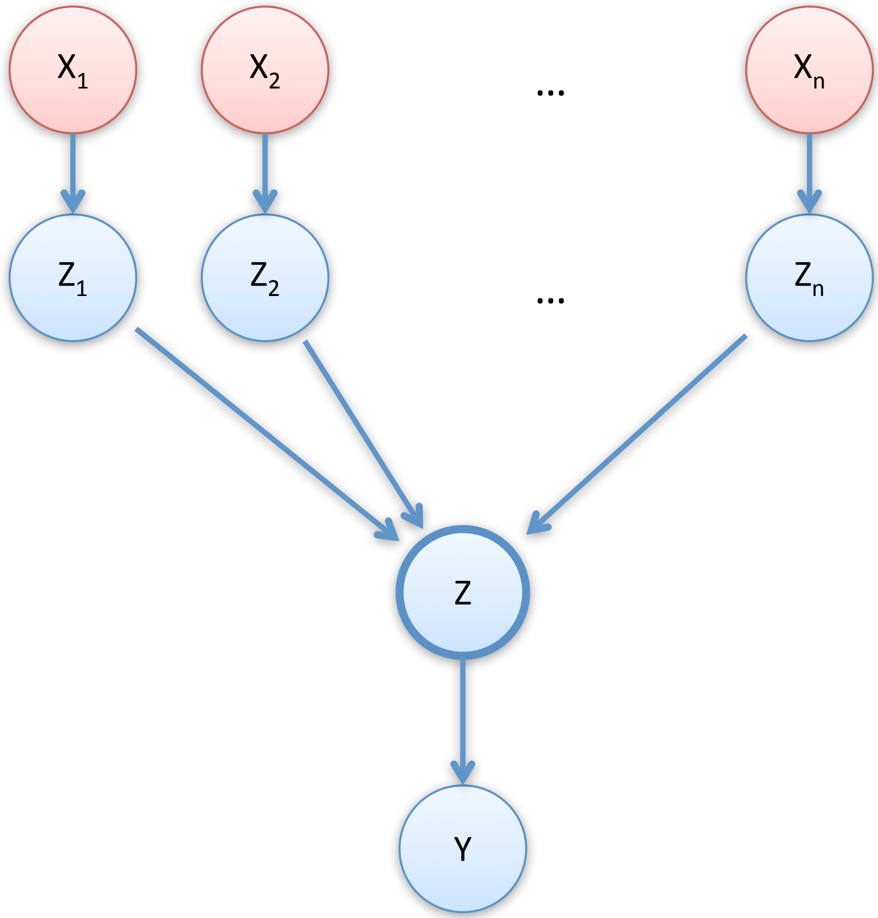
$X \rightarrow Z_i = w_i X_i$

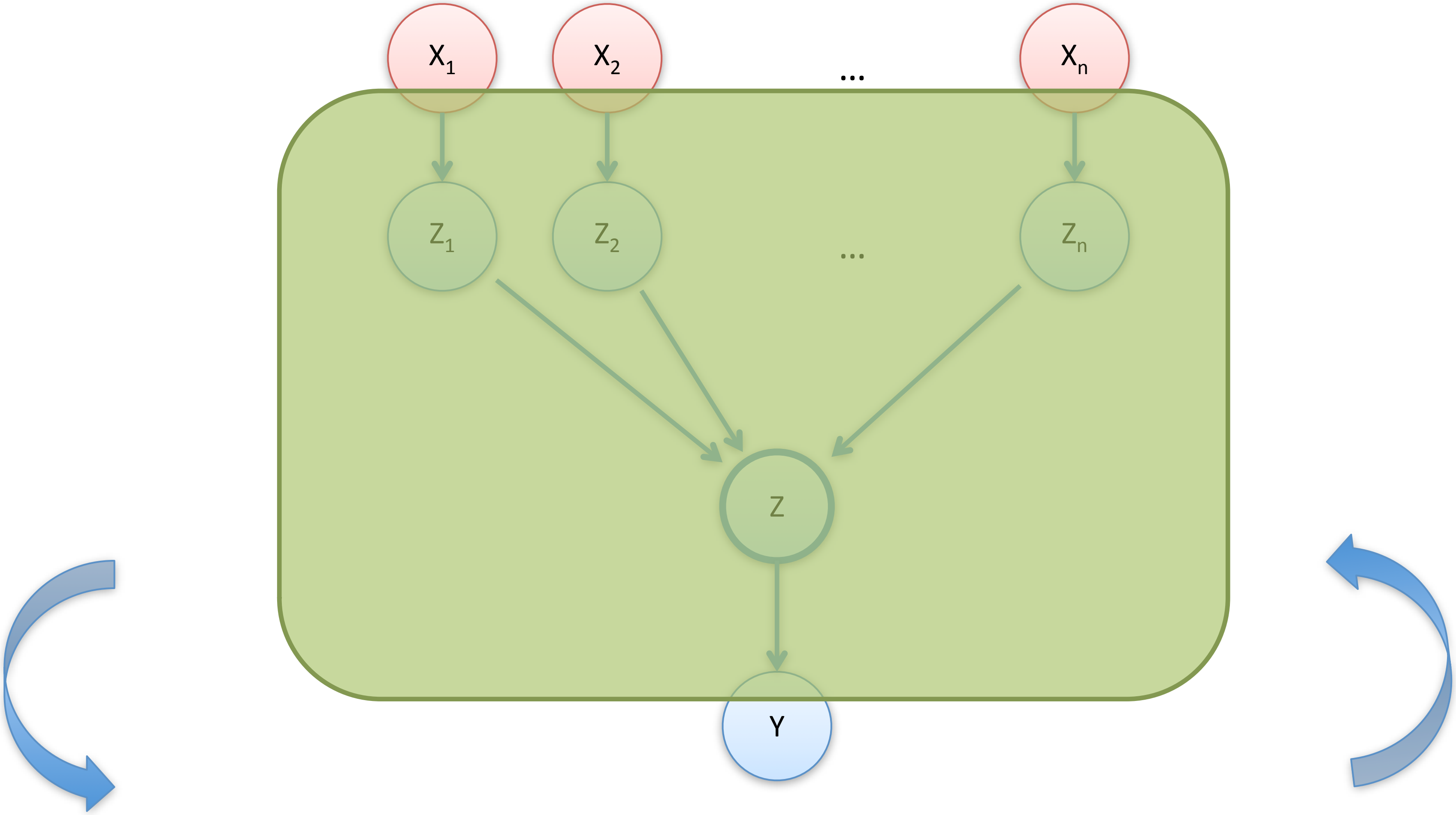
Z_i s $\rightarrow$ Z = sum

$Z \rightarrow Y$ = sigmoid

Something Interesting Happened!

- We are now representing conditional probability distributions as conditional Bayesian Networks!


X1
X2
...
Xn
Z1
Z2
...
Zn
Z
Y



P(Y | **X**)	000	001	010	011	100	101	110	111
1								
0								

Continuous Random Variables

First case: continuous child and parents

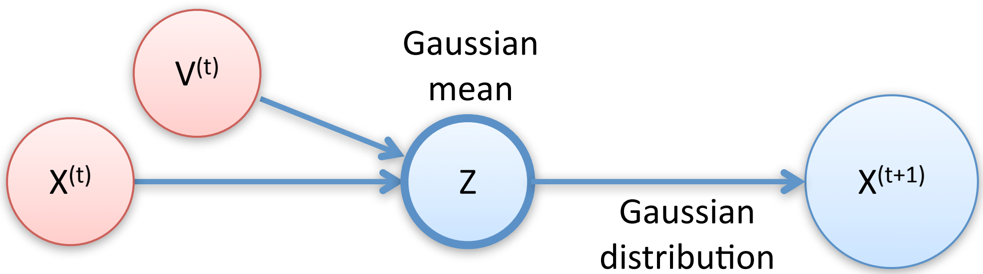
- Gaussian distribution is a CPD:
  $P(Y \mid \text{Mean} = \mu, \text{Variance} = \sigma^2) = \text{Normal}(\mu, \sigma^2)$
- **Linear Gaussian** CPD: Normal(linear(**x**), σ^2)
  - Y is a linear function of the variables **X**, with Gaussian noise that has variance σ^2

Example

- $X^{(t)}$ is a vehicle's position at time t
- $V^{(t)}$ is its velocity at time t

$$X^{(t+1)} \approx X^{(t)} + V^{(t)}$$

- Allow for some randomness in the motion:



Linear Gaussian Model

$$p(Y|x_1, ...x_k) = \mathcal{N}(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k; \sigma^2)$$

Continuous Random Variables

Second case: continuous child with discrete *and* continuous parents ($\mathbf{X}_{\text{disc}}$ and $\mathbf{X}_{\text{cont}}$)

- "Conditional linear Gaussian":

$$P(Y \mid \boldsymbol{X}_{\text{disc}} = \boldsymbol{x}_{\text{disc}}, \boldsymbol{X}_{\text{cont}} = \boldsymbol{x}_{\text{cont}}) = \mathcal{N}\left(w_{\boldsymbol{x}_{\text{disc}},0} + \sum_j w_{\boldsymbol{x}_{\text{disc}},j}\, x_{\text{cont},j},\, \sigma^2_{\boldsymbol{x}_{\text{disc}}}\right)$$

  – different weights for each $\mathbf{x}_{\text{disc}}$

- Induces a **Gaussian mixture** for Y

Continuous Random Variables

Third case:

discrete child Y with continuous parent X

- Threshold model
  - Makes P(Y | X) discontinuous in X's value
- Multinomial logit (see slide 27 "Generalized Linear Model")

CPDs can be Bayesian Networks!

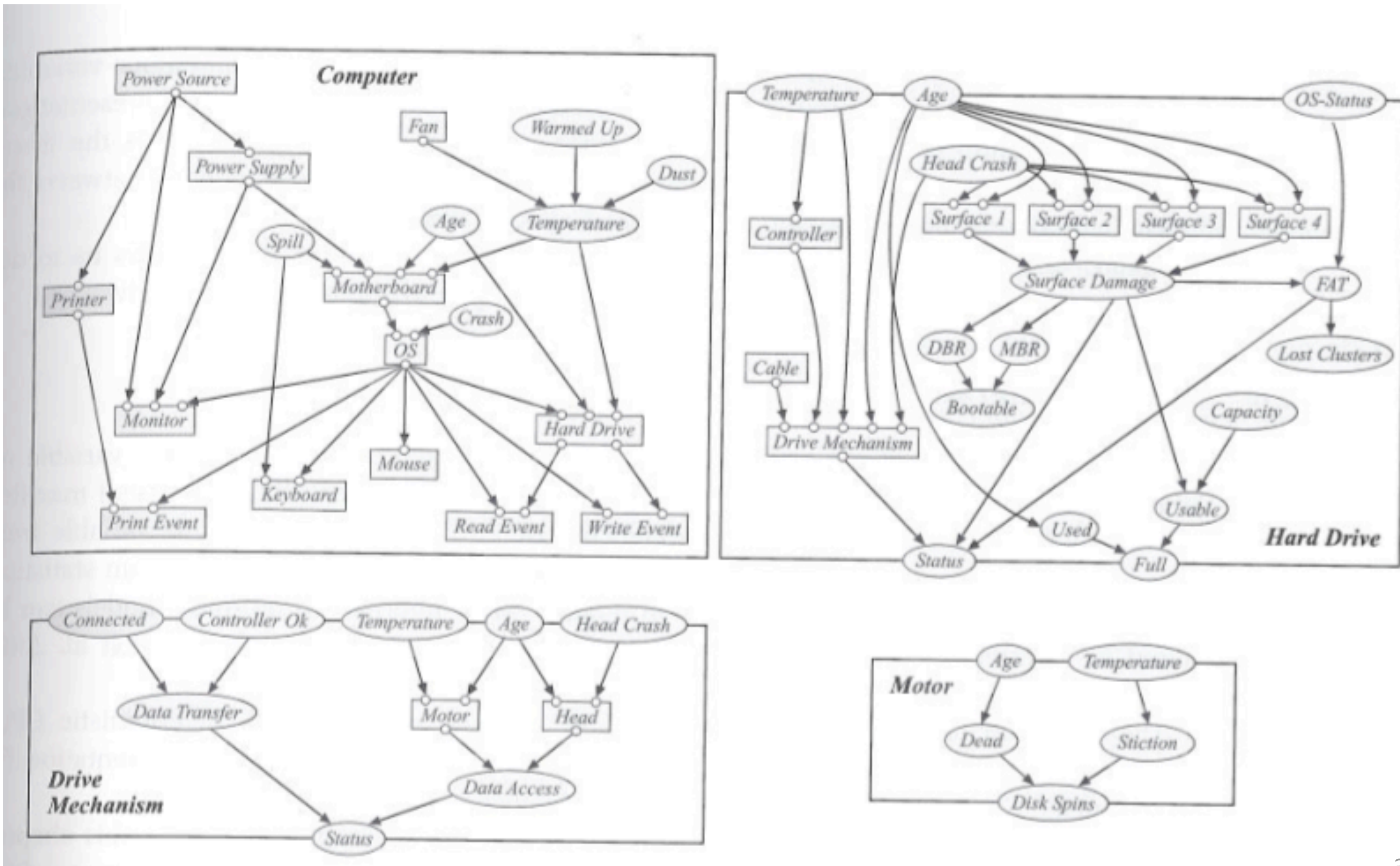
- *"Conditional Bayesian Network"* is a BN with three types of variables:
  - Inputs, always observed, no parents: **X**
  - Outputs: **Y**
  - Encapsulated: **Z**

$$P(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{X}) = \prod_{W \in \boldsymbol{Y} \cup \boldsymbol{Z}} P(W \mid \mathbf{Parents}(W))$$

Conditional Probability Distribution

- A CPD is an "*Encapsulated CPD*" if it can be represented by a *Conditional Bayesian Network*.
  - Construct a complex BN, with components composed of other BN subcomponents! ...object-oriented style!

Encapsulated CPDs (K&F Figure 5.15)



Where We Are

- Structure in CPDs
- Effects on independence assertions
- Examples:
  - Determinism
  - Context-specificity
  - Independent causes
  - Continuous distributions
  - CPDs represented by Conditional BNs