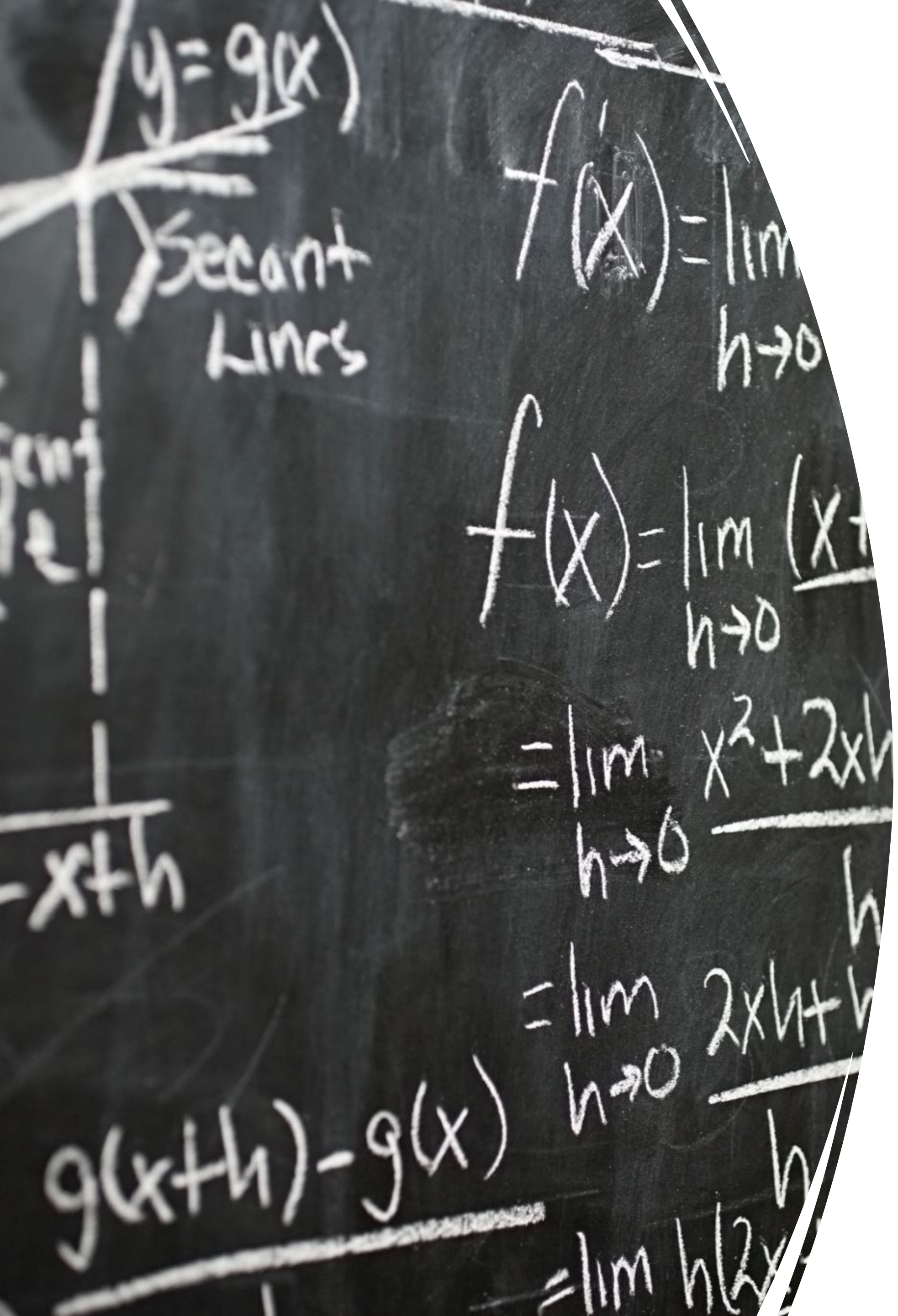


Lecture 21 -- Probabilistic Graphical Models

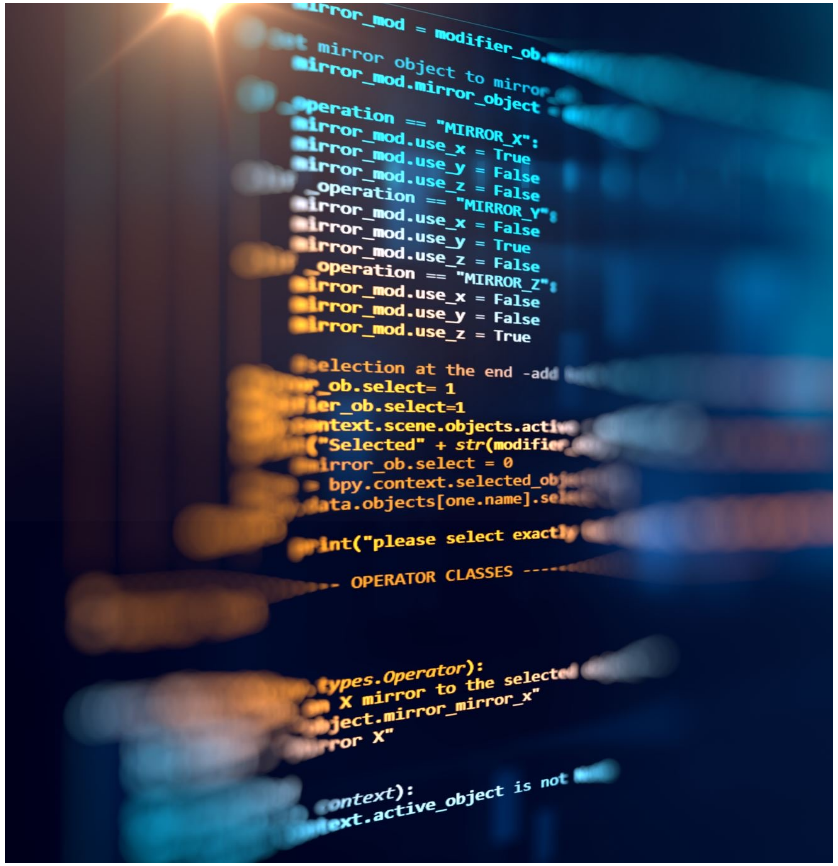
Prof Xiaowei Huang

https://cgi.csc.liv.ac.uk/~xiaowei/

(Attendance Code: **453995**)

Up to now,

- Traditional Machine Learning Algorithms
- Adversarial Attack and Defence
- Deep learning



Up to now, the contents are heavily statistical.

We are certainly pondering:

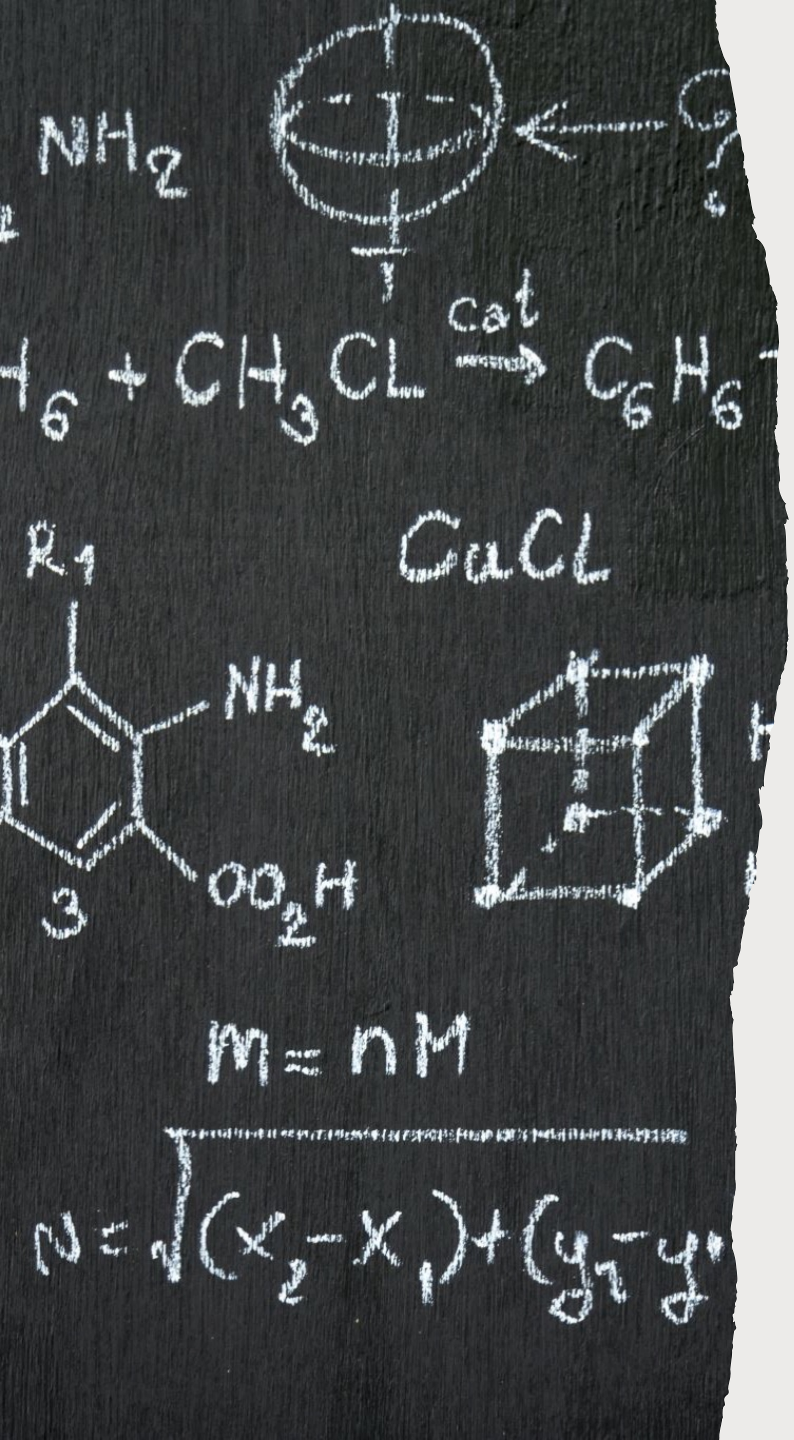
(1) How can we embed our expert knowledge?
(2) How can we summarise human-understandable information from statistics?

(1)

Complex distribution

Knowledge as Graphs

(2)

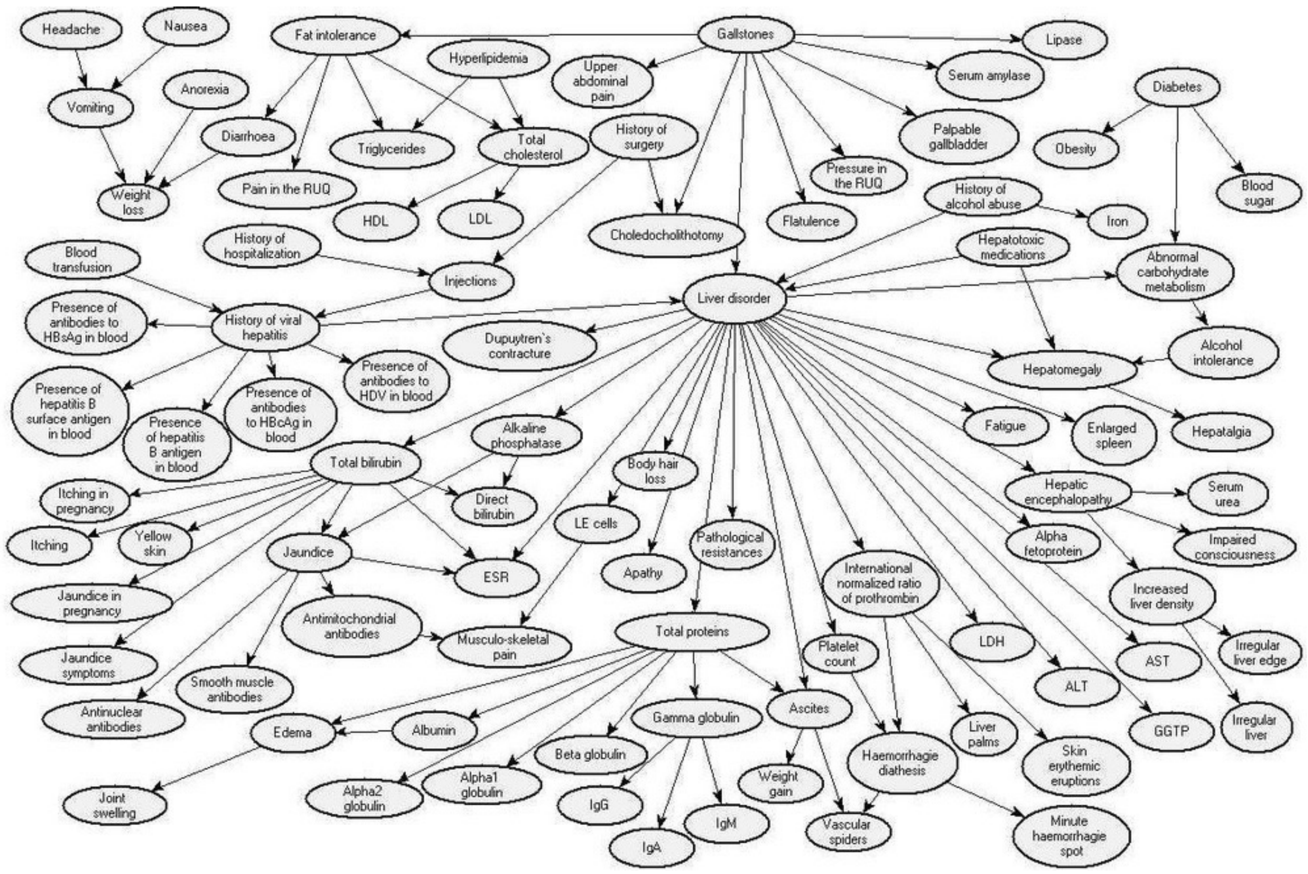


Topics

- From Naïve Bayes to Bayesian Network
- Example Bayes Networks
- Example Simple Probability Query
- Summarisation on what is Graphical Model
- Abstraction of Neural Networks into Graphical Model
- Further Example: Alarm Network

Will take about more intriguing probabilistic reasoning in later lectures

What are Graphical Models?



Model $\mathcal{M}$

Data:

$$\mathcal{D} \equiv \{X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)}\}_{i=1}^{N}$$

Top 10 Real-world Bayesian Network Applications – Know the importance!

- https://data-flair.training/blogs/bayesian-network-applications/
  - Gene Regulatory Network
  - Medicine
  - Biomonitoring
  - Document Classification
  - Information Retrieval
  - Semantic Search
  - Image Processing
  - Spam Filter
  - Turbo Code
  - System Biology

Gene Regulatory Network

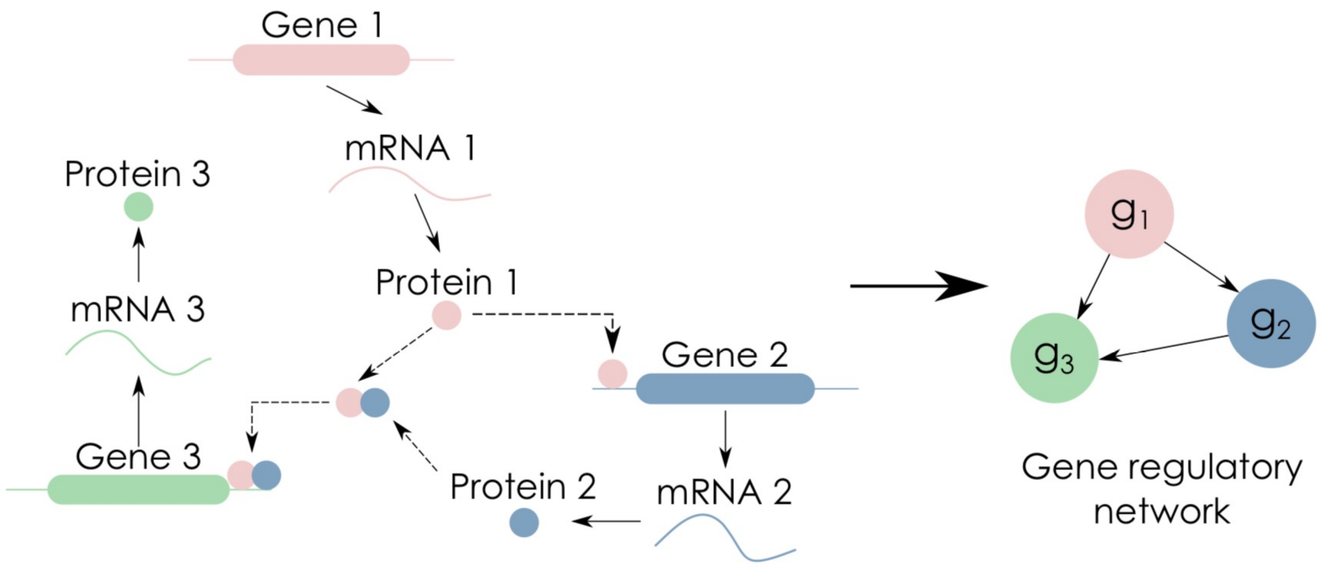
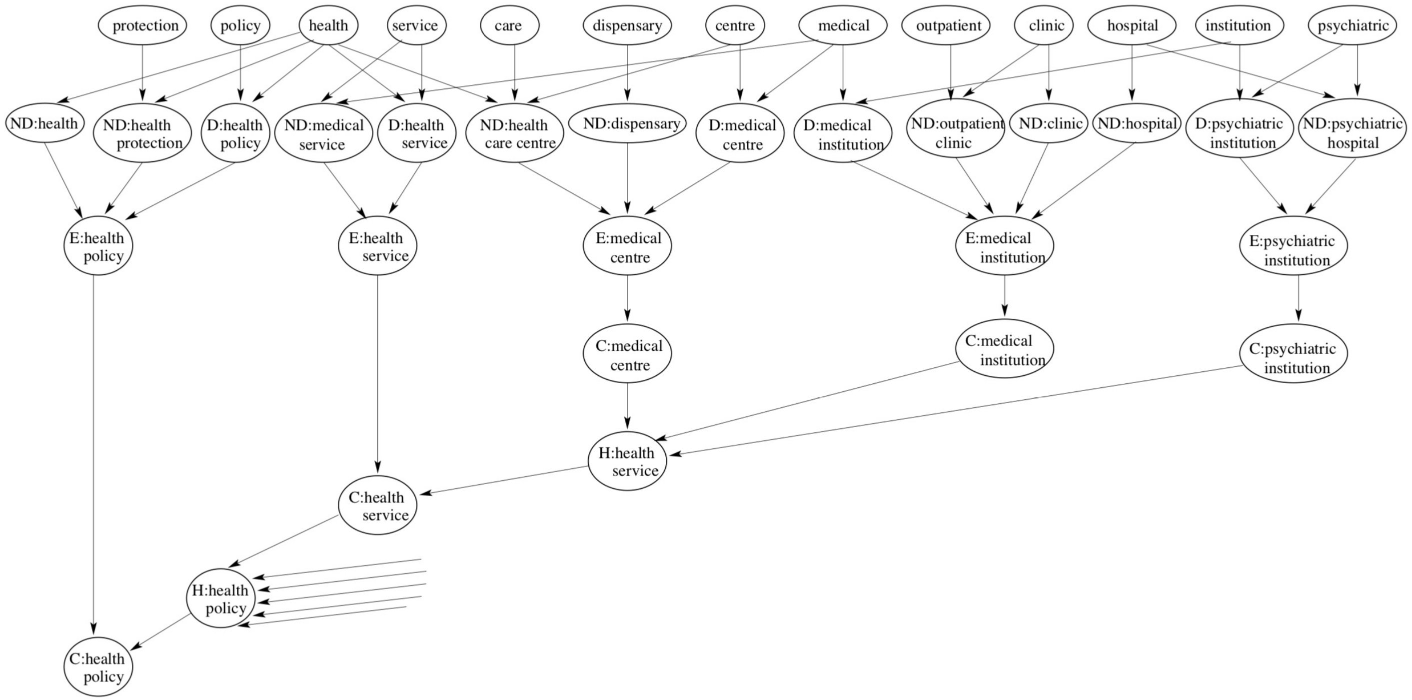
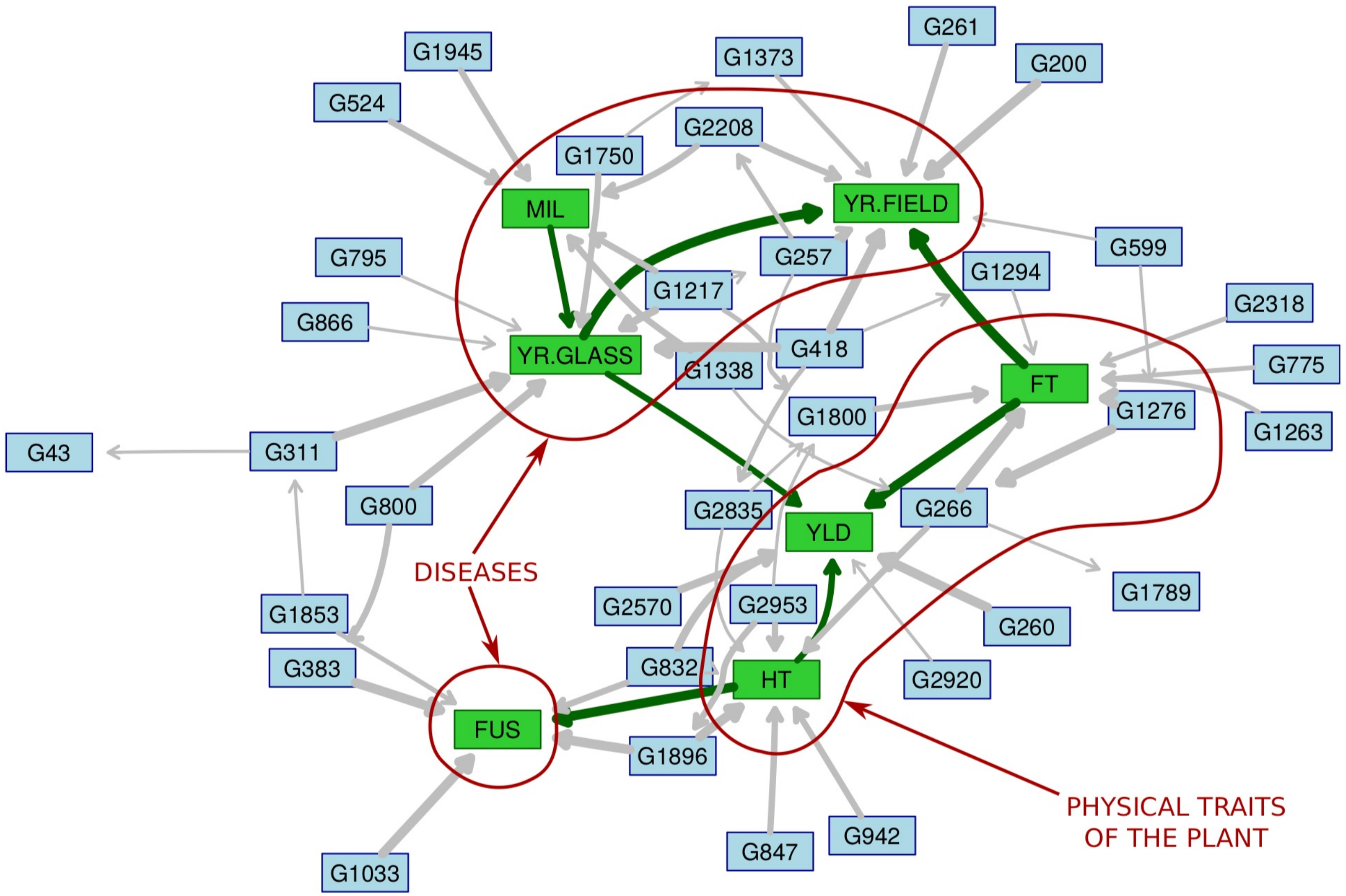


Fig. 1 A cartoon schematic of a gene regulatory network. A complex biophysical model describes the interaction between three genes, involving both direct regulation (gene 2 by gene 1) and combinatorial regulation via complex formation (gene 3 by genes 1 and 2). The abstracted structure of the system is given in the (directed) network on the right.

Document Classification



WHEAT: a Bayesian Network (44 nodes, 66 arcs)



Fundamental Questions in ML

- Representation
  - How to capture/model uncertainties in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?

Now we focus on this in the last component

- Inference
  - How do I answer questions/queries according to my model and/or based on given data? e.g.: $P(X_i \mid \mathbf{D})$
- Learning
  - Which model is "right" for the data: e.g.: $\mathcal{M} = \arg\max_{\mathcal{M} \in \mathrm{M}} F(\mathbf{D}; \mathcal{M})$

We have discussed this quite a lot by now

From Naïve Bayes to Bayesian Network

Recap of Basic Prob. Concepts

- What is the joint probability distribution on multiple variables?

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

  - How many state configuration in total?
  - Are they all needed to be represented?
  - **Do we get any scientific insight?**

Recall: naïve Bayes

Parameters for Joint Distribution

- Each X_i represents outcome of tossing coin *i*
  - Assume coin tosses are marginally independent
  - i.e., $X_i \perp X_j$ therefore

$$P(X_1, X_2, ..., X_n) = P(X_1)P(X_2)...P(X_n)$$

Recall: assumption for naïve Bayes, except the latter is based on conditional independence

- If we use standard parameterization of the joint distribution, the independence structure is obscured and required 2^n parameters
- However, we can use a more natural set of parameters: *n* parameters

$\theta_1, ..., \theta_n$

Parameterization

- Example: Company is trying to hire recent graduates
- Goal is to hire intelligent employees
  - No way to test intelligence directly
  - But have access to Student's score
    - Which is informative but not fully indicative
- Two random variables
  - Intelligence: $Val(I) = \{i^1, i^0\}$, high and low
  - Score: $Val(S) = \{s^1, s^0\}$, high and low
- Joint distribution has 4 entries
  - Need three parameters

I	S	P(I,S)
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.06
i^1	s^1	0.24

Joint distribution

Alternative Representation: Conditional Parameterization

$$P(I, S) = P(I)P(S|I)$$

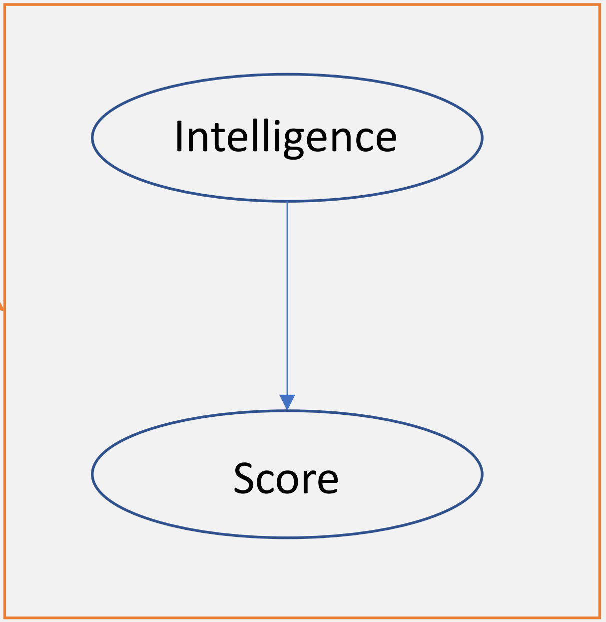
- Representation more compatible with causality
  - Intelligence influenced by Genetics, upbringing
  - Score influenced by Intelligence

- Note: BNs are not required to follow causality but they often do
- Need to specify $P(I)$ and $P(S|I)$

i^0	i^1
0.7	0.3

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

- Three binomial distributions (3 parameters) needed
  - One marginal, two conditionals $P(S|I = i^0)$ $P(S|I = i^1)$



Bayesian Networks (from two nodes to more nodes)

- One more variable G with $Val(G) = \{g^1, g^2, g^3\}$ to represents grades A, B, C, in addition to I and S

- If we have the following conditional independence, i.e., $P \models (S \perp G \mid I)$

- That is, Score and Grade are independent given Intelligence, i.e., knowing Intelligence, Score gives no information about class grade

Use of Conditional Independence

- Assertions
  - From probabilistic reasoning $P(I, S, G) = P(I)P(S, G \mid I)$
  - From assumption $P \models (S \perp G \mid I)$
- Combining, we have

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

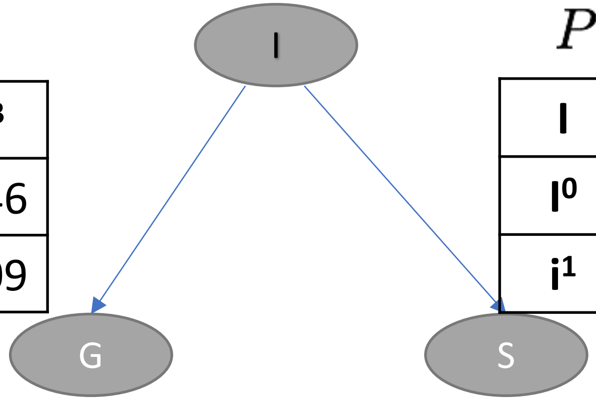
$$P(I, S, G) = P(I)P(S \mid I)P(G \mid I)$$

$P(I)$

i^0	i^1
0.7	0.3

$P(G|I)$

I	g^1	g^2	g^3
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09



$P(S|I)$

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

Two binomials, one 3-value multinomials:
7 params.
More compact than joint distribution

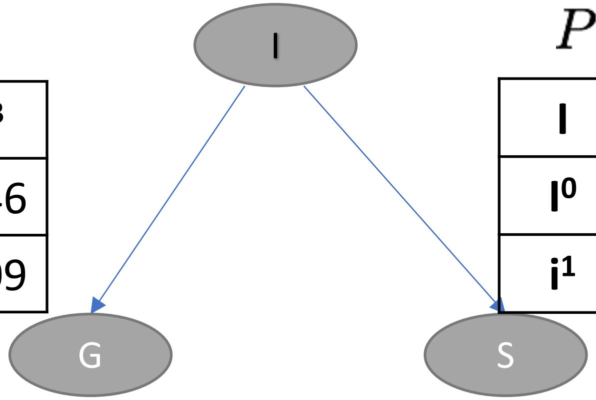
Bayesian Networks (from two nodes to more nodes)

$P(I)$

i^0	i^1
0.7	0.3

$P(G|I)$

I	g^1	g^2	g^3
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09



$P(S|I)$

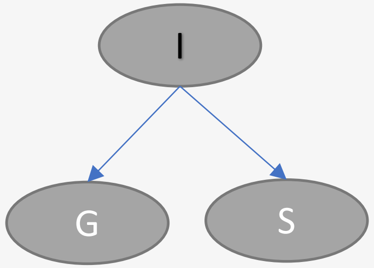
I	s^0	s^1
I^0	0.95	0.05
i^1	0.2	0.8

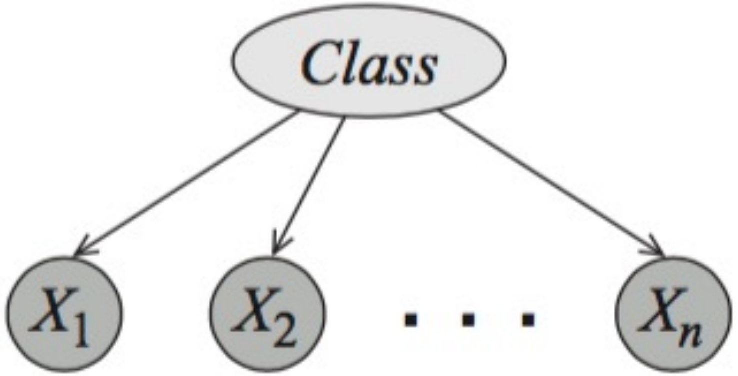
$$P(i^1, s^1, g^2) = P(i^1)P(s^1 \mid i^1)P(g^2 \mid i^1)$$
$$= 0.3 * 0.8 * 0.17 = 0.0408$$

BN for General Naive Bayes Model

- Encoded using a very small number of parameters
- Linear in the number of variables

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





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

Bayesian Networks: Conditional Parameterization and Conditional Independences

Conditional probability tables

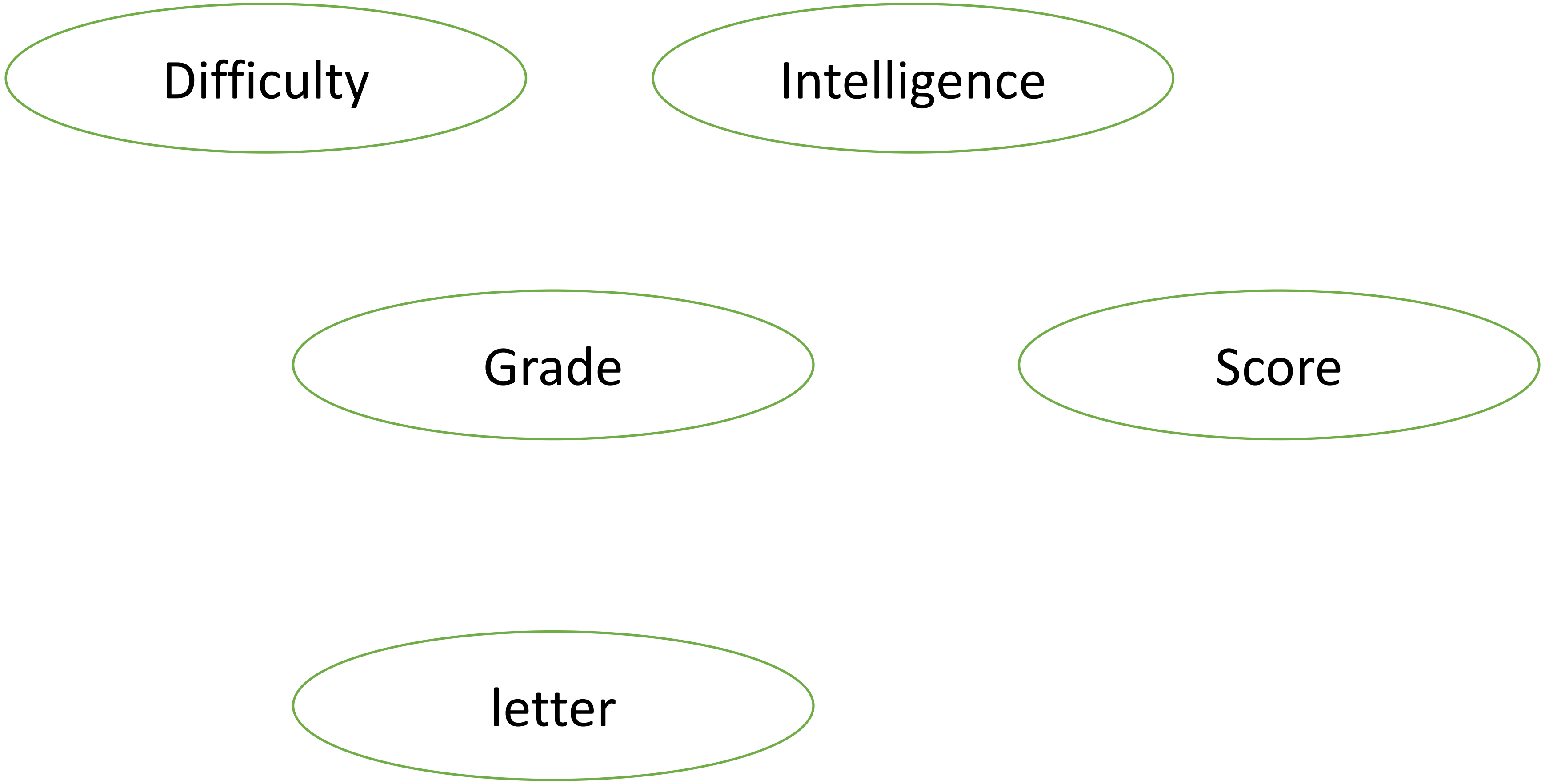
- Conditional Parameterization is combined with Conditional Independence assumptions to produce very compact representations of high dimensional probability distributions

Application of Naive Bayes Model

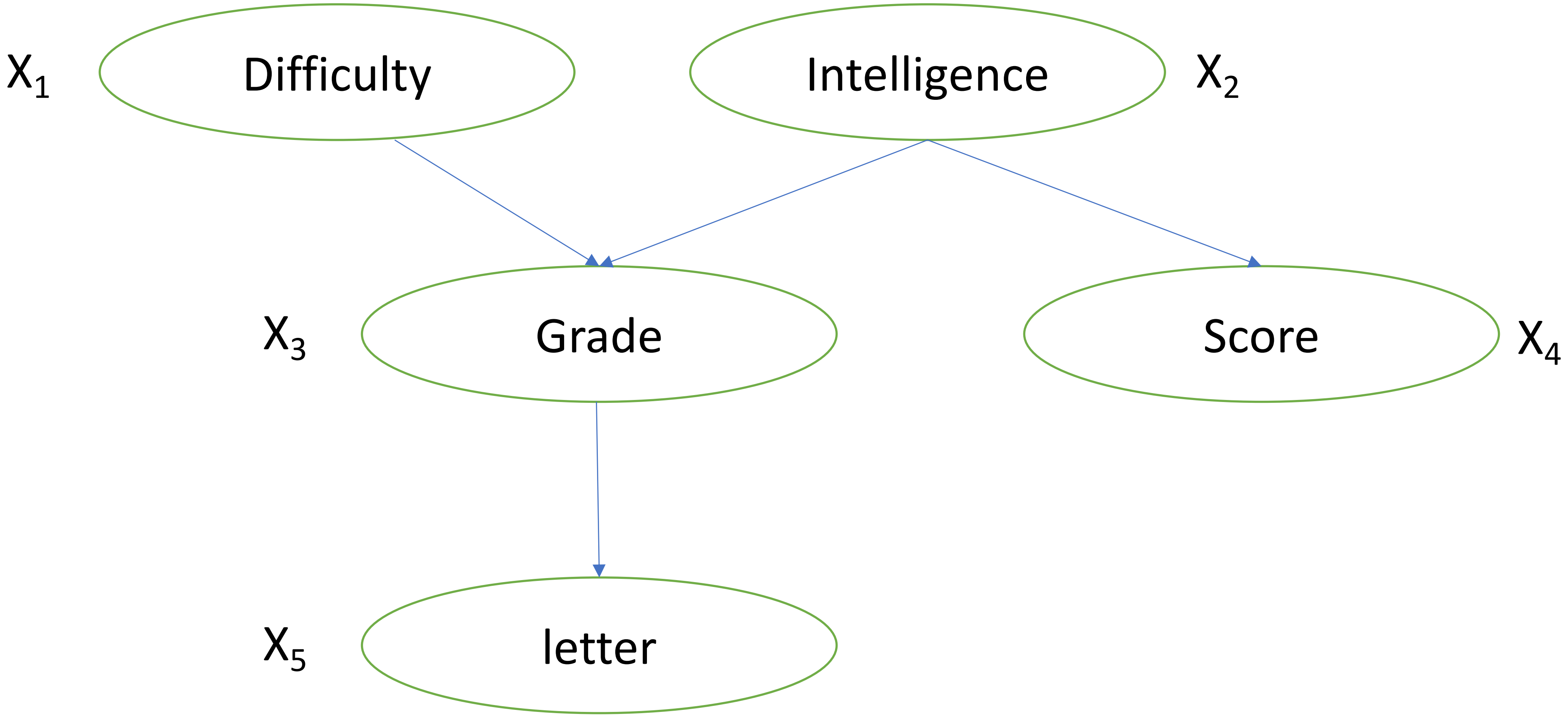
- Medical Diagnosis
  – Pathfinder expert system for lymph node disease (Heckerman et.al., 1992)
- Full BN agreed with human expert 50/53 cases
- Naive Bayes agreed 47/53 cases

Example Bayes Networks

Student Bayesian Network

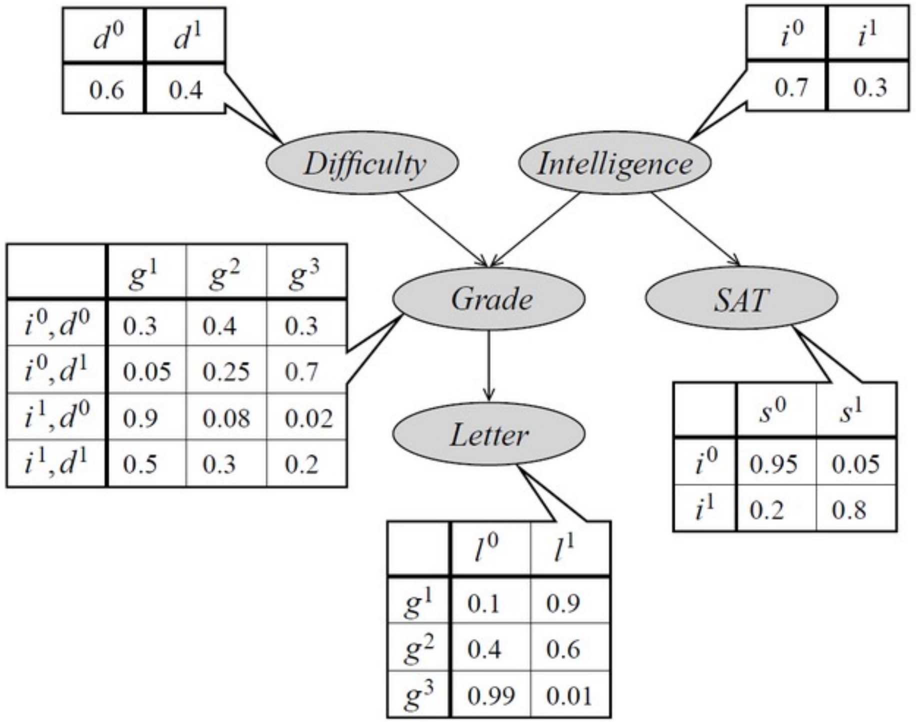


Student Bayesian Network

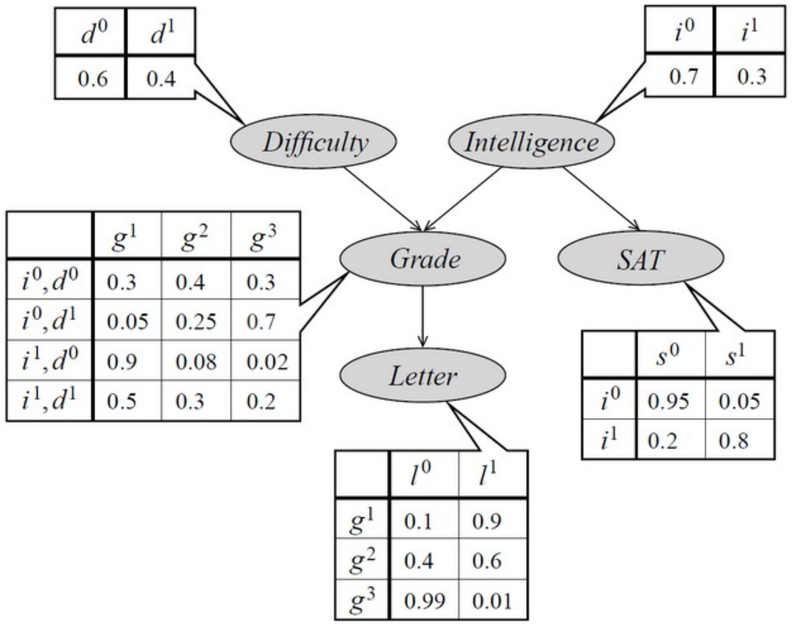


Student Bayesian Network

- Represents joint probability distribution over multiple variables
- BNs represent them in terms of graphs and conditional probability distributions (CPDs)
- Resulting in great savings in number of parameters needed



Joint distribution from Student BN



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

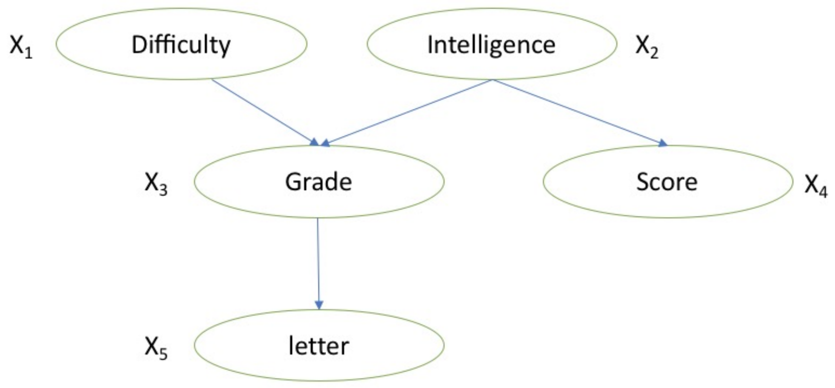
pa: parent nodes

- CPDs: $P(X_i \mid pa(X_i))$
- Joint Distribution:

$$P(X) = P(X_1, X_2, ..., X_n)$$
$$P(X) = \prod_{i=1}^{n} P(X_i \mid pa(X_i))$$
$$P(D, I, G, S, L) = P(D)P(I)P(G \mid D, I)P(S \mid I)P(L \mid G)$$

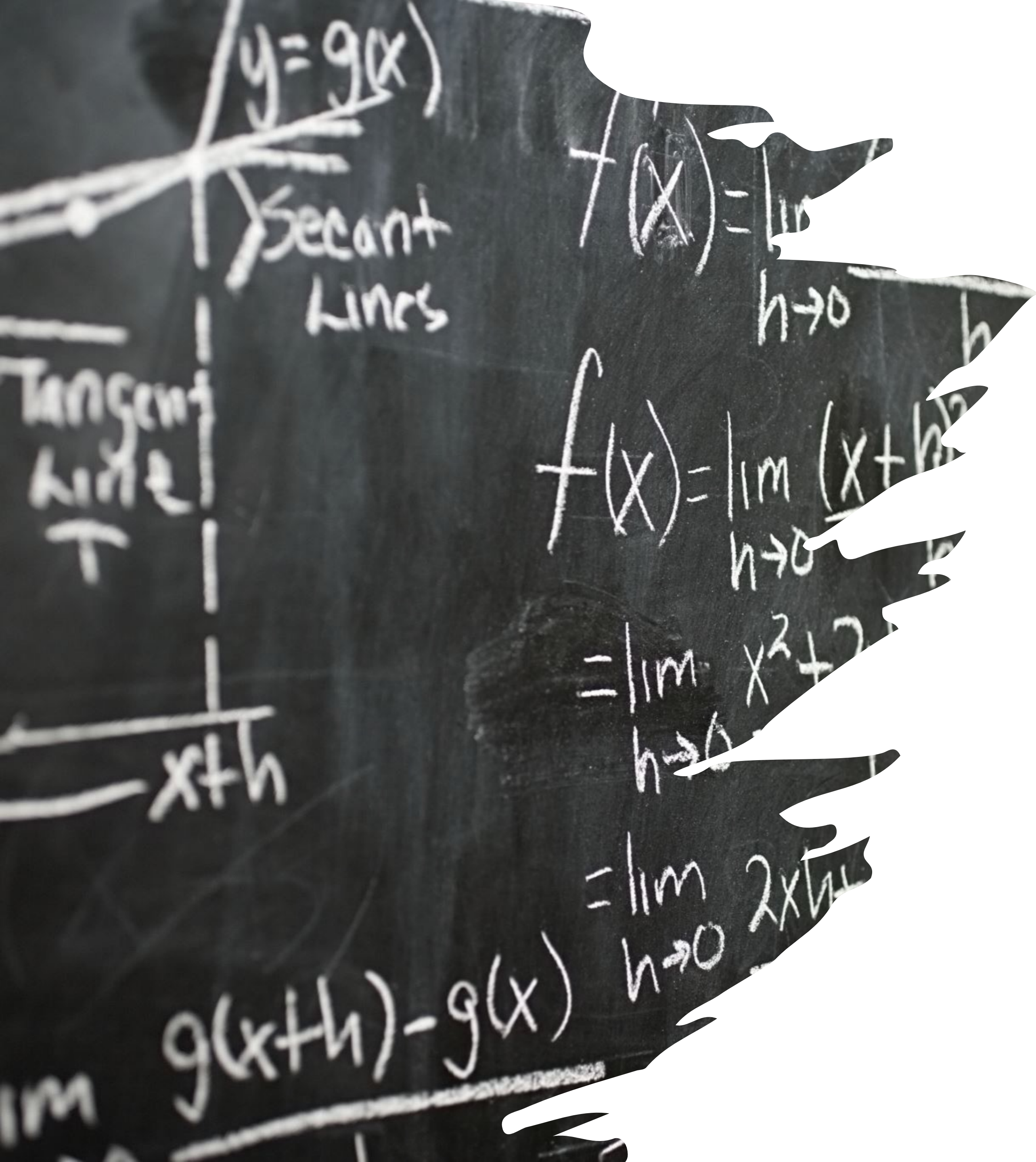
Student Bayesian Network

- If Xs are conditionally independent (as described by a PGM), the joint distribution can be factored into a product of simpler terms, e.g.,



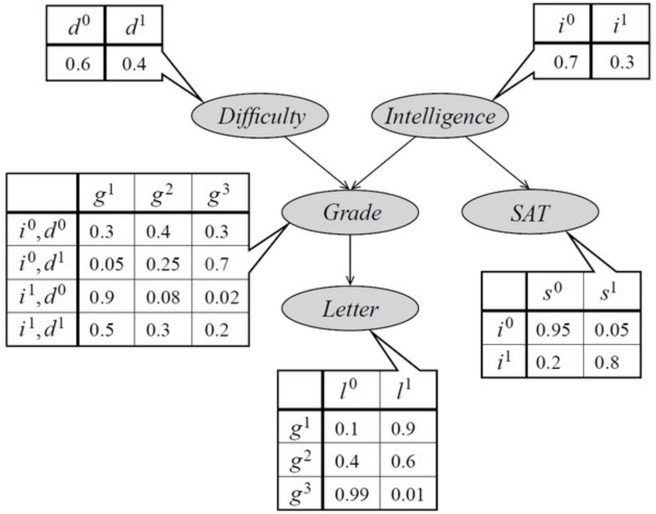
$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_2)P(X_3 \mid X_1, X_2)P(X_4 \mid X_2)P(X_5 \mid X_3)$$

- What's the benefit of using a PGM:
    - Incorporation of domain knowledge and causal (logical) structures
    - 1+1+7+3+3=14, a reduction from 2^5-1 = 31



Example Probability Query

Example of Probability Query



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

$$P(Y = y_i \mid E = e) = \frac{P(Y = y_i, E = e)}{P(E = e)}$$

Posterior Marginal   Probability of Evidence

Posterior Marginal Estimation: $P(I = i^1 \mid L = l^0, S = s^1) = ?$

Probability of Evidence: $P(L = l^0, S = s^1) = ?$

- Here we are asking for a specific probability rather than a full distribution

Computing the Probability of Evidence

- Probability Distribution of Evidence

$$P(L,S) = \sum_{D,I,G} P(D,I,G,L,S)$$ Sum Rule of Probability

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

- Probability of Evidence

$$P(L = l^0, s = s^1) = \sum_{D,I,G} P(D)P(I)P(G \mid D,I)P(L = l^0 \mid G)P(S = s^1 \mid I)$$

- More Generally $P(E = e) = \sum_{X \setminus E} \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \mid_{E=e}$

Computing the Posterior Marginal (Method 1)

$$P(I = i^1 | L = l^0, S = s^1) = \frac{P(I = i^1, L = l^0, S = s^1)}{P(L = l^0, S = s^1)}$$

Now we know how to compute $P(L = l^0, S = s^1)$

Can you do the other one? $P(I = i^1, L = l^0, S = s^1)$

Alternatively, Rational Statistical Inference

The Bayes Theorem:

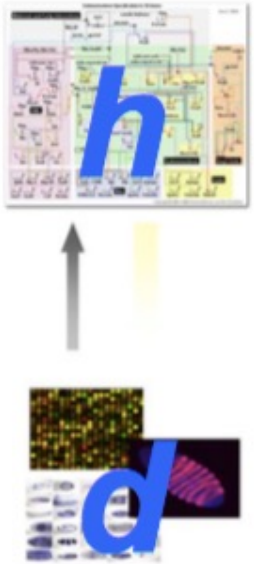
Likelihood

Prior probability

Posterior probability

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$

Sum over space of hypotheses

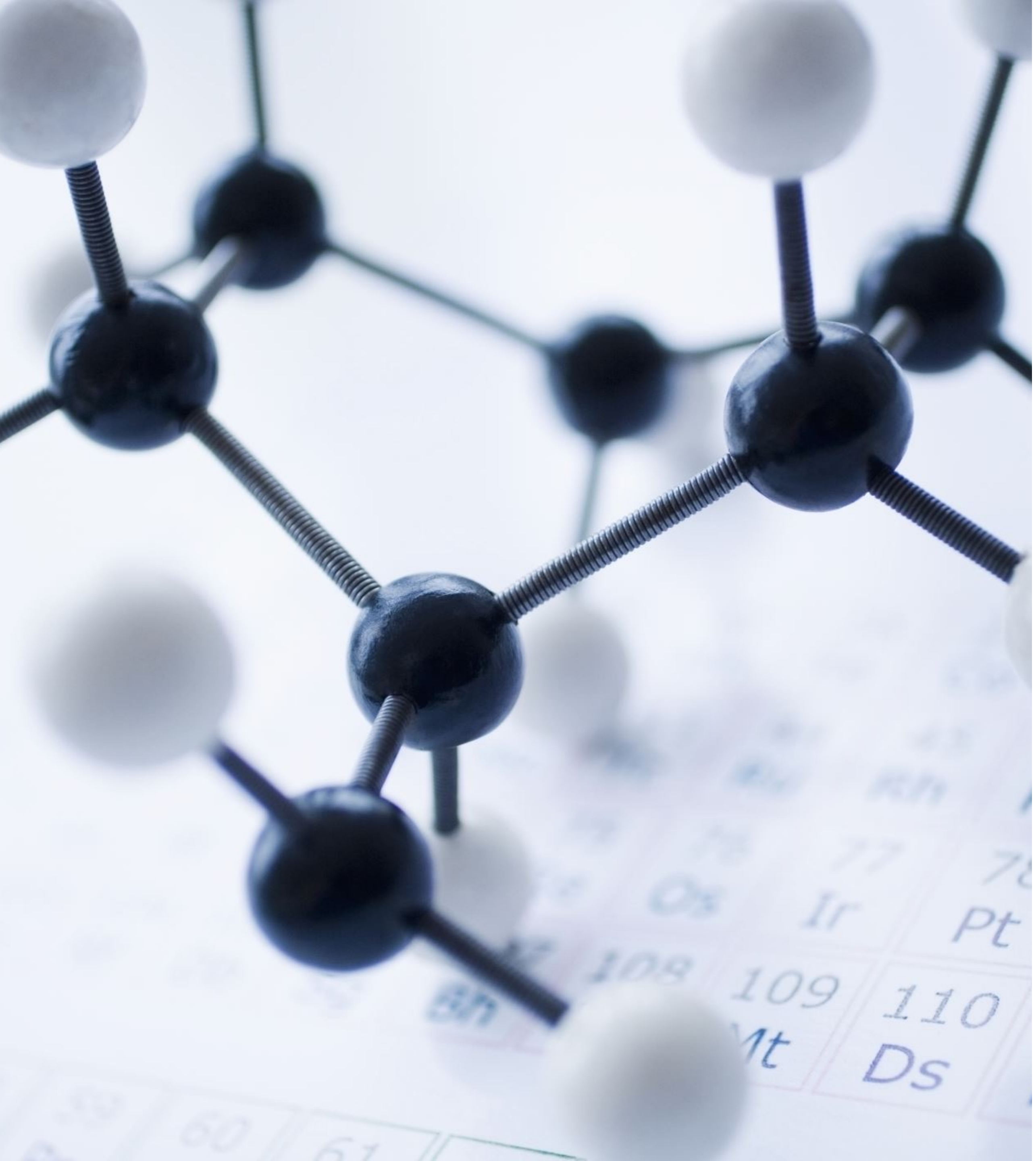


Rational Statistical Inference (Method 2)

$$P(I = i^1 | L = l^0, S = s^1) = \frac{P(L = l^0, S = s^1 | I = i^1) P(I = i^1)}{\sum_{i \in \{i^0, i^1\}} P(L = l^0, S = s^1 | I = i) P(I = i)}$$

If we know that $P \models L \perp S | I$

$$P(I = i^1 | L = l^0, S = s^1) = \frac{P(L = l^0 | I = i^1) P(S = s^1 | I = i^1) P(I = i^1)}{\sum_{i \in \{i^0, i^1\}} P(L = l^0 | I = i) P(S = s^1 | I = i) P(I = i)}$$



What is a Graphical Model?

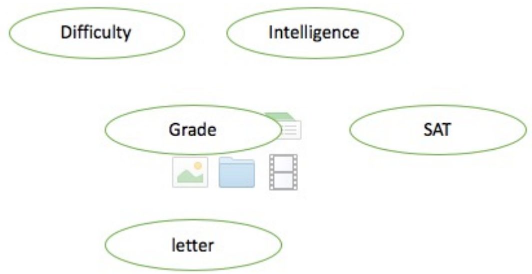
So What is a Graphical Model?

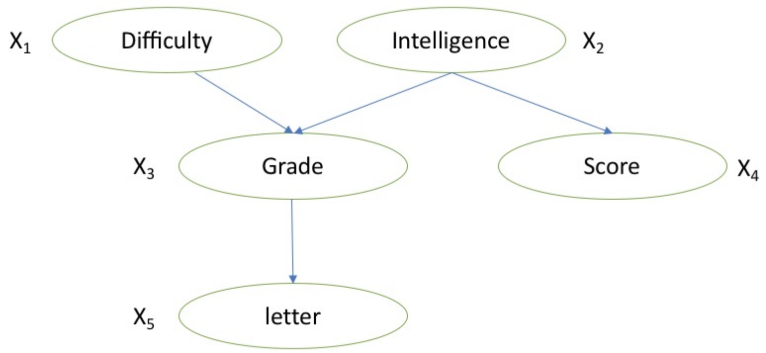
- In a nutshell,

GM = **Multivariate Statistics + Structure**

What is a Graphical Model?

- The informal blurb:
  - It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***





- A more formal description:
  - It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables

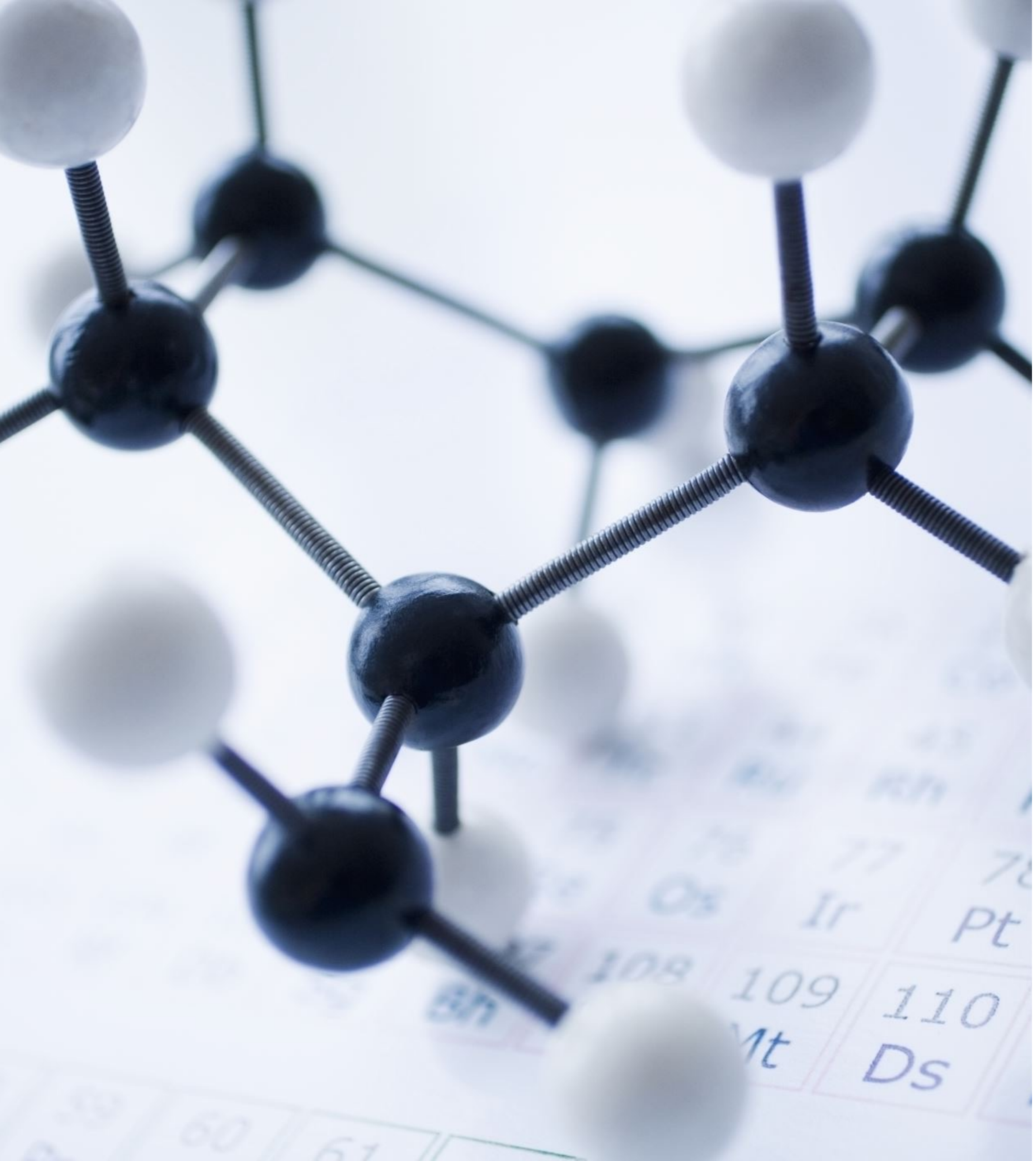
Two types of GMs

- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

- Undirected edges simply give correlations between variables (**Markov Random Field** or **Undirected Graphical model**):

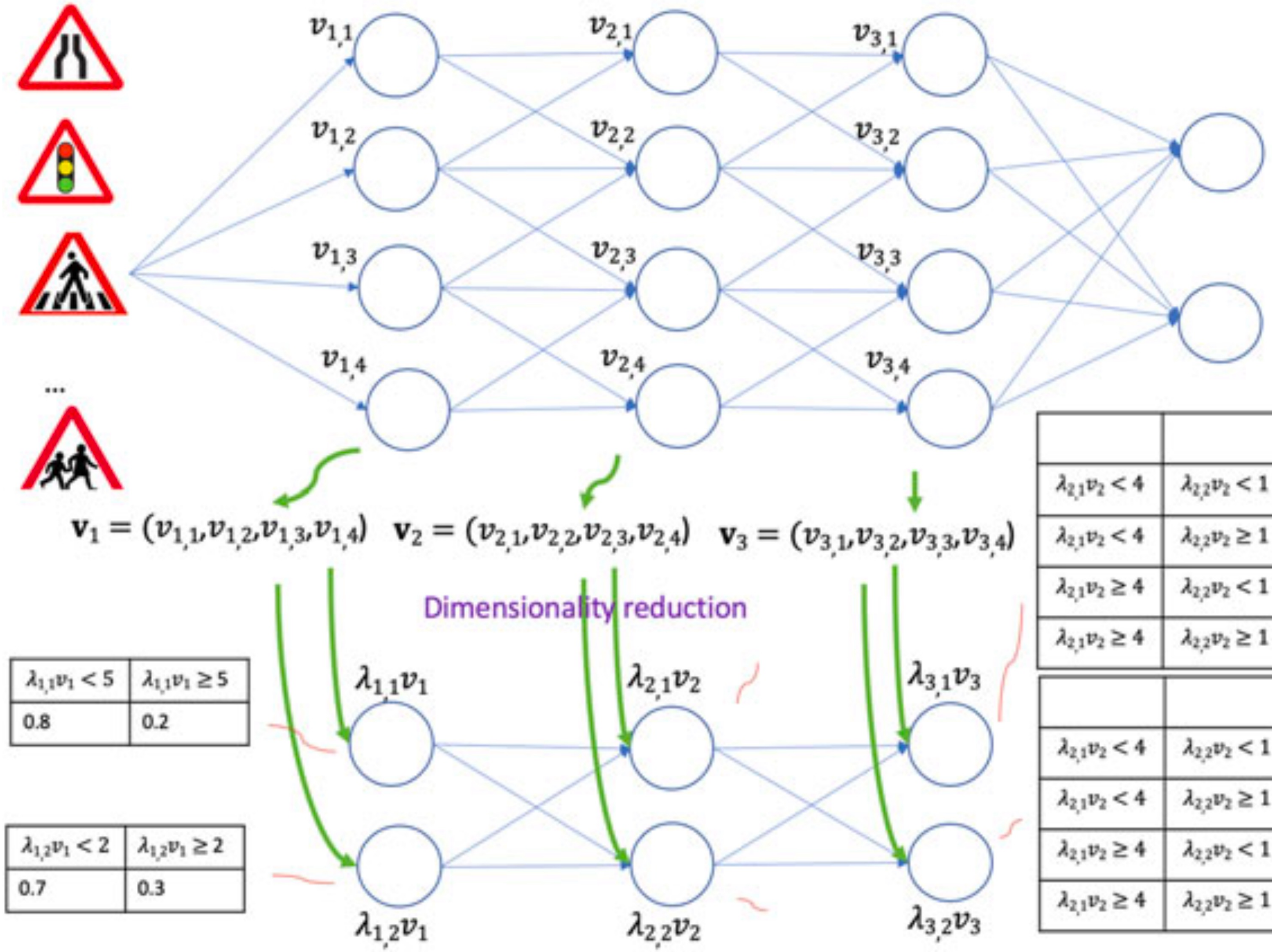
Bayesian Network vs. Bayesian Neural Network

- Bayesian network is the probabilistic graphical model we discuss here.
- Bayesian neural network is a neural network with Bayesian assumption on its weights.



Abstraction of Neural Network as Probabilistic Graphical Model

Abstraction of Neural Network as Probabilistic Graphical Model



$\lambda_{1,1}v_1 < 5$	$\lambda_{1,1}v_1 \geq 5$
0.8	0.2

$\lambda_{1,2}v_1 < 2$	$\lambda_{1,2}v_1 \geq 2$
0.7	0.3

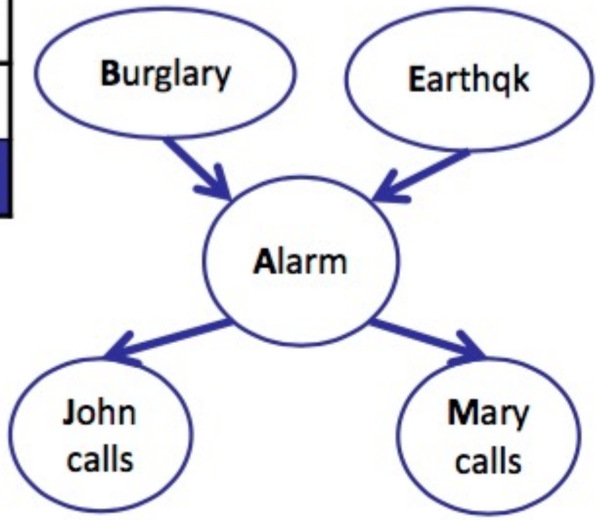
		$\lambda_{3,1}v_3 < 3$	$\lambda_{3,1}v_3 \geq 3$
$\lambda_{2,1}v_2 < 4$	$\lambda_{2,2}v_2 < 1$	1.0	0
$\lambda_{2,1}v_2 < 4$	$\lambda_{2,2}v_2 \geq 1$	0.8	0.2
$\lambda_{2,1}v_2 \geq 4$	$\lambda_{2,2}v_2 < 1$	0.5	0.5
$\lambda_{2,1}v_2 \geq 4$	$\lambda_{2,2}v_2 \geq 1$	0.6	0.4

		$\lambda_{3,2}v_3 < 3$	$\lambda_{3,2}v_3 \geq 3$
$\lambda_{2,1}v_2 < 4$	$\lambda_{2,2}v_2 < 1$	0.4	0.6
$\lambda_{2,1}v_2 < 4$	$\lambda_{2,2}v_2 \geq 1$	0.9	0.1
$\lambda_{2,1}v_2 \geq 4$	$\lambda_{2,2}v_2 < 1$	0	1.0
$\lambda_{2,1}v_2 \geq 4$	$\lambda_{2,2}v_2 \geq 1$	0.3	0.7

Further Example: Alarm Network

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

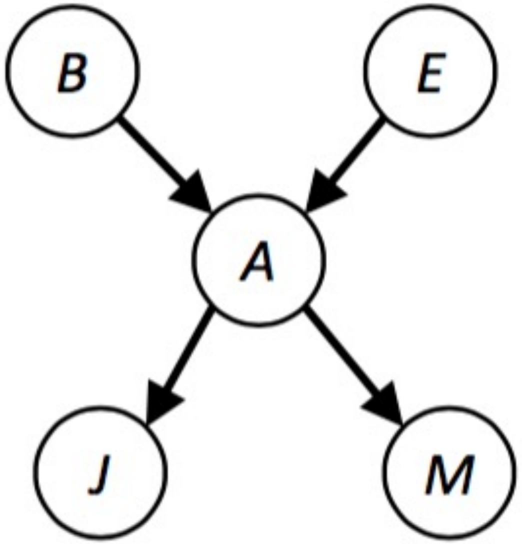
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999

E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$$P(+b, -e, +a, -j, +m) =$$
$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$
$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$