Lecture 12: Parameter Estimation in Probabilistic Models

GENOME 560, Spring 2017

Su-In Lee, CSE & GS (suinlee@uw.edu)

Survey results

- Homework assignment
  - Weekly homework
- Longer vs. shorter
- R-session
  - More examples

Review of Last Lecture

- What did we learn in Tuesday's class?

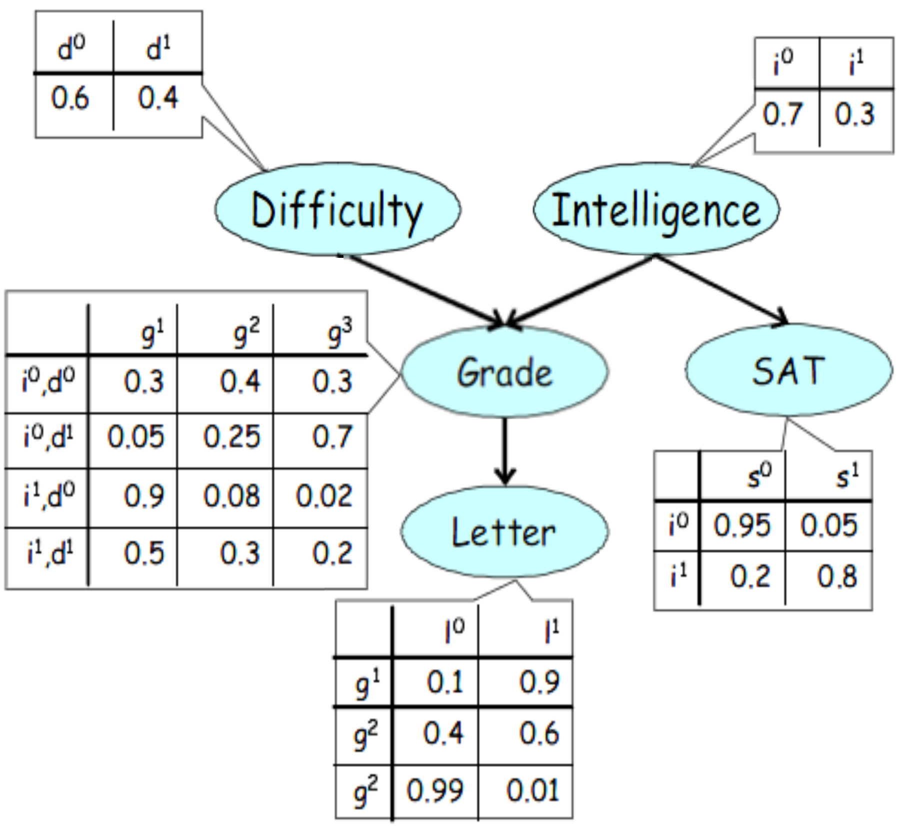
Review of Last Lecture

- Conditional probability distribution
- Bayesian networks

Outline

- Conditional distribution and Bayesian networks
- Special cases of Bayesian networks
- Model Selection
- Basic concepts of parameter estimation
    - Maximum likelihood estimation (MLE)

The *Student* Example



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

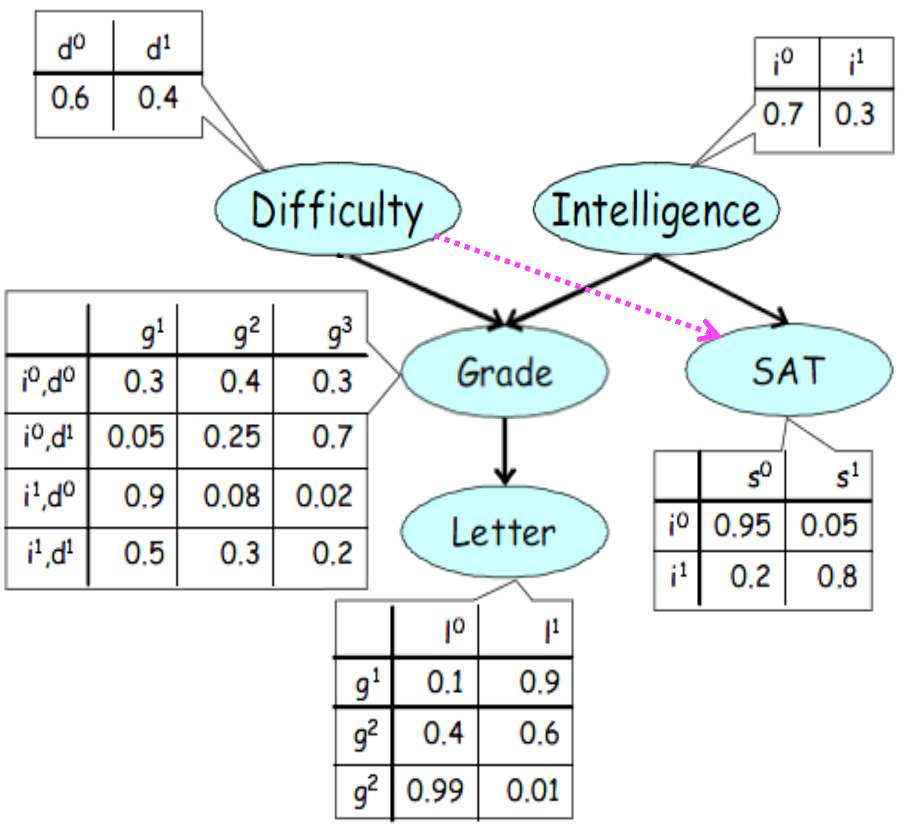
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

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

- Random variables
  - Course difficulty (D) = $\{d^0, d^1\}$
    Probability distribution, P(D)
  - Intelligence (I) = $\{i^0, i^1\}$
    Probability distribution, P(I)
  - SAT (S) = $\{s^0, s^1\}$
    Conditional probability distribution, P(S|I)
  - Grade (G) = $\{g^1, g^2, g^3\}$
    Conditional probability distribution, P(G|D,I)
  - Quality of Letter (L) = $\{l^0, l^1\}$
    Conditional probability distribution, P(L|G)

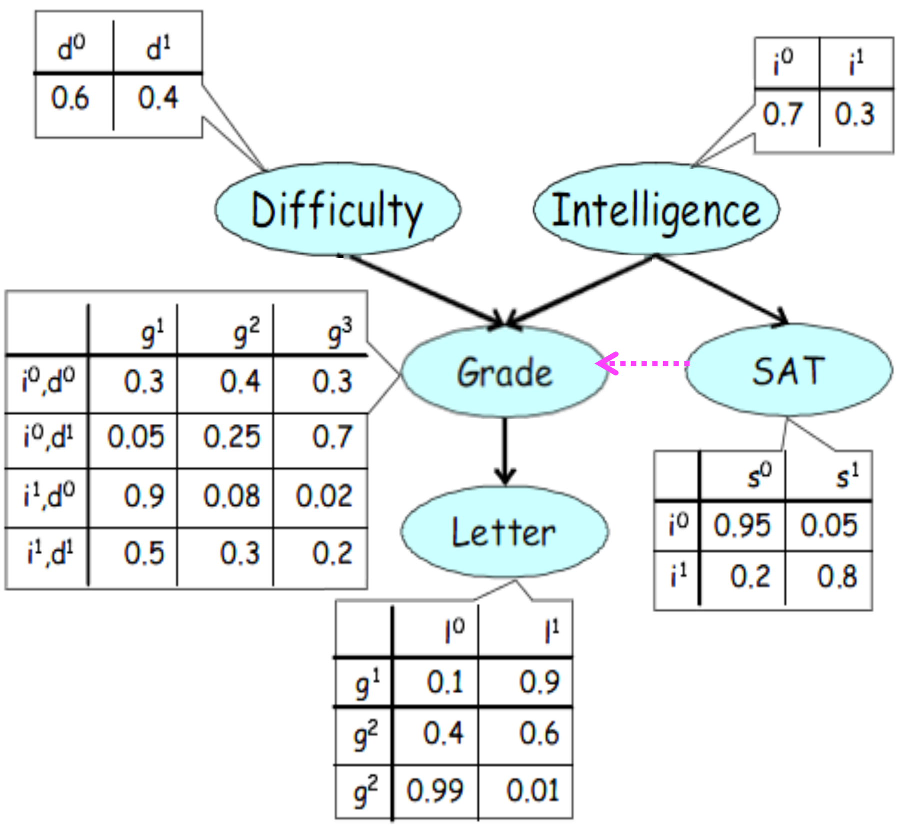
from Koller & Friedman (page 48 in K&F)

The *Student* Example



- Random variables
  - Course difficulty (D) = $\{d^0, d^1\}$
    Probability distribution, P(D)
  - Intelligence (I) = $\{i^0, i^1\}$
    Probability distribution, P(I)
  - SAT (S) = $\{s^0, s^1\}$
    Conditional probability distribution, P(S|I)
    P(S|I,D) ?
  - Grade (G) = $\{g^1, g^2, g^3\}$
    Conditional probability distribution, P(G|D,I)
  - Quality of Letter (L) = $\{l^0, l^1\}$
    Conditional probability distribution, P(L|G)

The *Student* Example



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

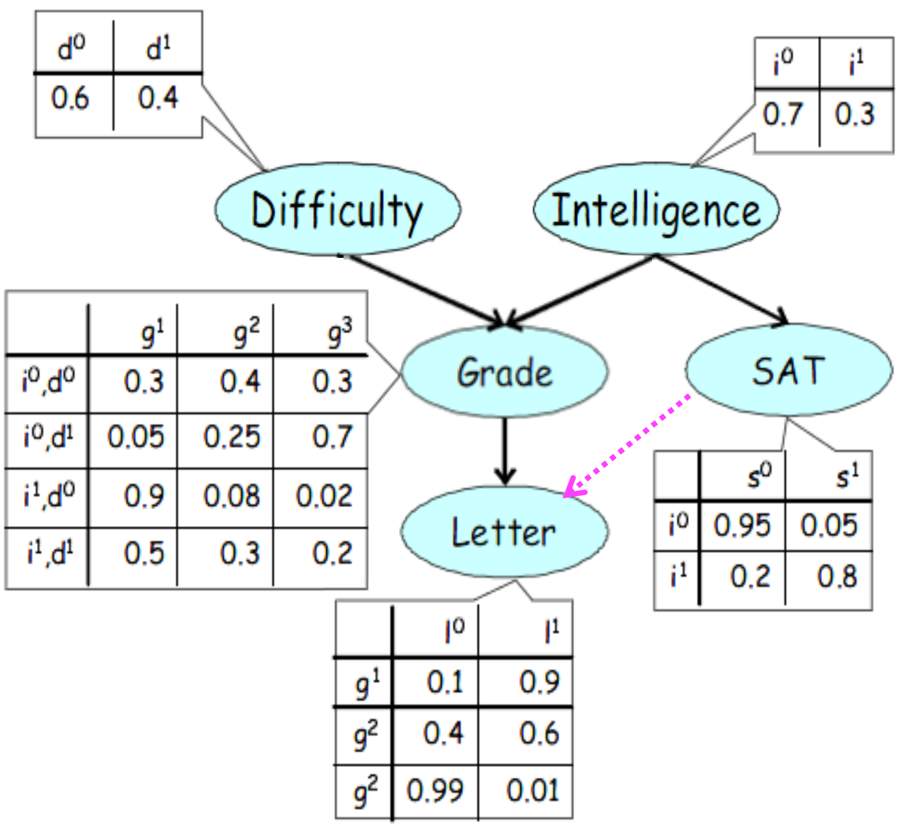
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	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^2	0.99	0.01

- Random variables
  - Course difficulty (D) = $\{d^0, d^1\}$
    Probability distribution, P(D)
  - Intelligence (I) = $\{i^0, i^1\}$
    Probability distribution, P(I)
  - SAT (S) = $\{s^0, s^1\}$
    Conditional probability distribution, P(S|I)
  - Grade (G) = $\{g^1, g^2, g^3\}$
    Conditional probability distribution, P(G|D,I)
    P(G|D,I,S) ?
  - Quality of Letter (L) = $\{l^0, l^1\}$
  Conditional probability distribution, P(L|G)

The *Student* Example



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^2	0.99	0.01

- Random variables
  - Course difficulty (D) = $\{d^0, d^1\}$
    Probability distribution, P(D)
  - Intelligence (I) = $\{i^0, i^1\}$
    Probability distribution, P(I)
  - SAT (S) = $\{s^0, s^1\}$
    Conditional probability distribution, P(S|I)
  - Grade (G) = $\{g^1, g^2, g^3\}$
    Conditional probability distribution, P(G|D,I)
  - Quality of Letter (L) = $\{l^0, l^1\}$
    Conditional probability distribution, P(L|G)
    P(L|G, S) ?

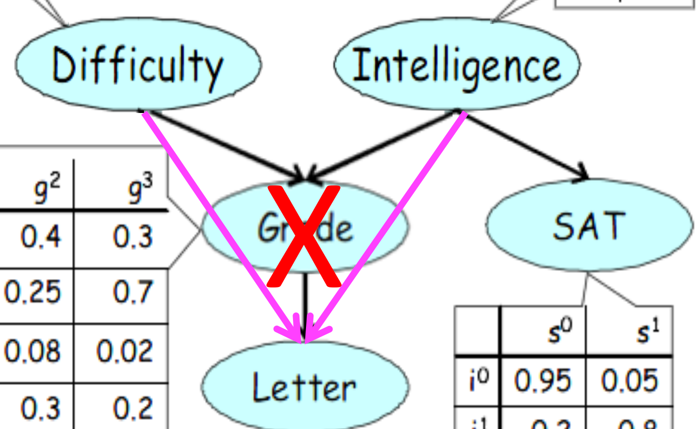
from Koller & Friedman

What if the instructor lost the grade book?

It's like we don't have a measurement of the protein level of some key gene. We always have incomplete data in some aspect.

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

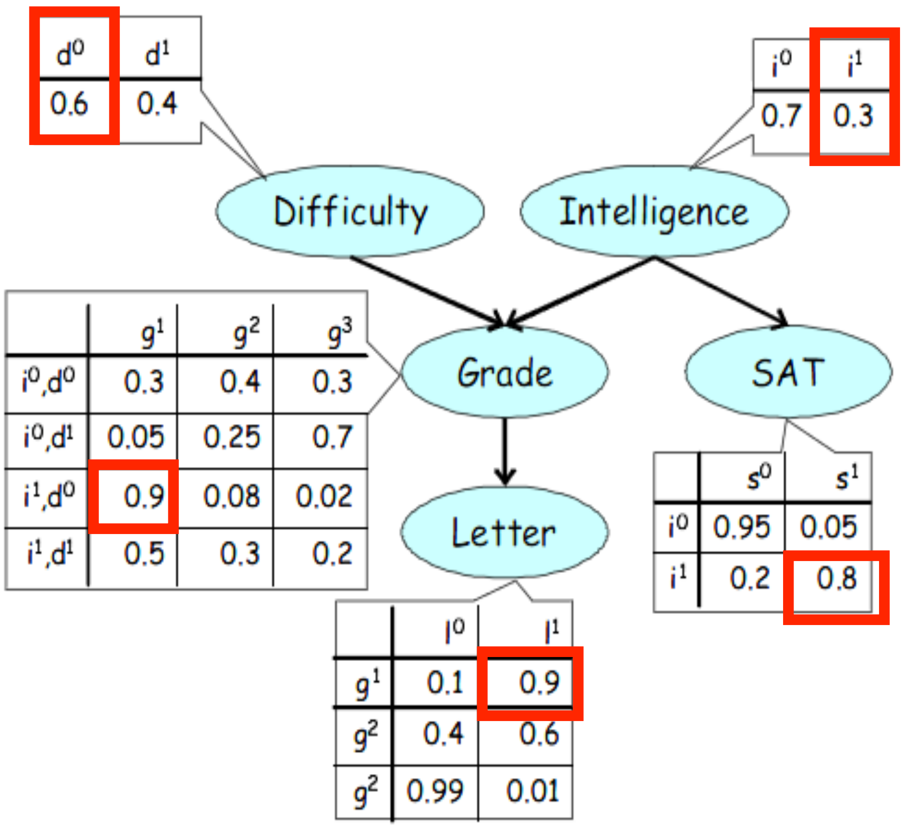
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

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

- Random variables
  - Course difficulty (D) = $\{d^0, d^1\}$
    Probability distribution, P(D)
  - Intelligence (I) = $\{i^0, i^1\}$
    Probability distribution, P(I)
  - SAT (S) = $\{s^0, s^1\}$
    Conditional probability distribution, P(S|I)
  - Grade (G) = $\{g^1, g^2, g^3\}$
    Conditional probability distribution, P(G|D,I)
  - Quality of Letter (L) = $\{l^0, l^1\}$
    Conditional probability distribution, P(L|G)
    P(L|D, I)

The *Student* Example

- What is the probability of observing {D=easy, I=intelligent, G=good, L=strong, S=high} ?

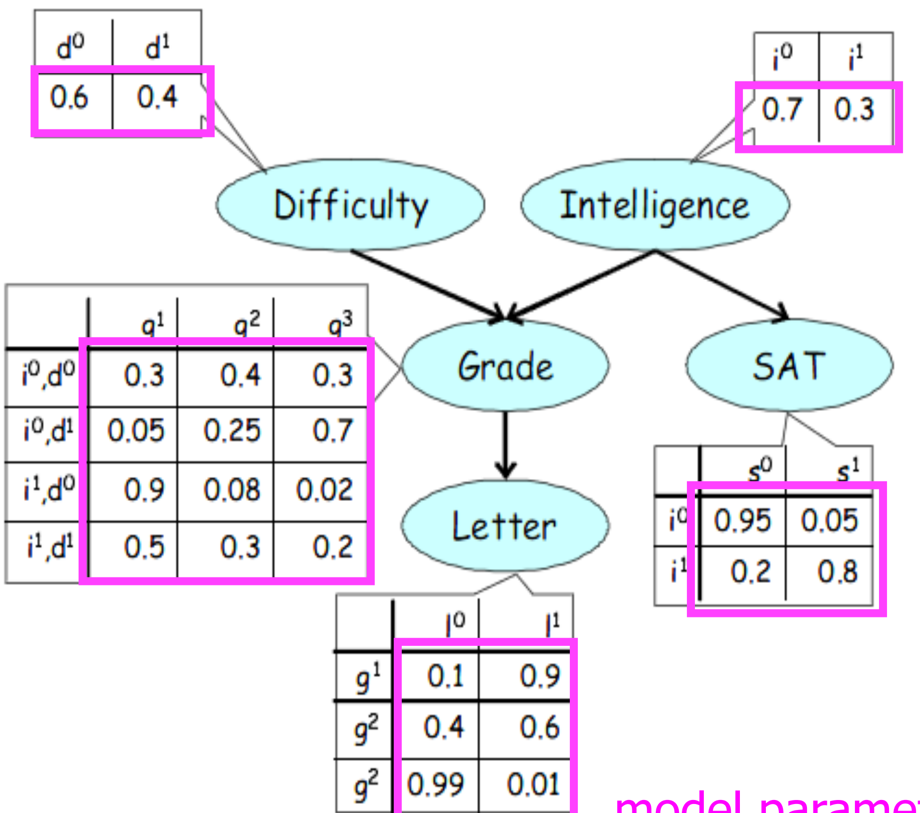


- P(D,I,G,L,S)
  = P(D) P(I) P(G|D,I) P(S|I) P(L|G)

- P(D=easy, I=intelligent, G=good, L=strong, S=high)
  = P(D=easy) P(I=intelligent) P(G=good | D=easy, I=intelligent) P(S=strong | I=intelligent) P(L=strong | G=good)
  = 0.6 x 0.3 x 0.9 x 0.9 x 0.8
  = 0.1166

Conditional probability tables (CPTs)

- What is the probability of observing {D=easy, I=intelligent, G=good, L=strong, S=high} ?



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^2	0.99	0.01

model parameters
(can be "learned" from data!)

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

- P(D=easy, I=intelligent, G=good, L=strong, S=high)

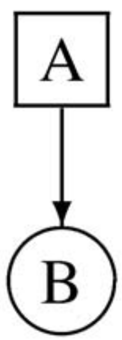
= P(D=easy) P(I=intelligent)
P(G=good | D=easy, I=intelligent)
P(S=strong | I=intelligent)
P(L=strong | G=good)

= 0.6 x 0.3 x 0.9 x 0.9 x 0.8

= 0.1166

How about continuous variables?

- Squares – discrete nodes
- Circles – continuous nodes

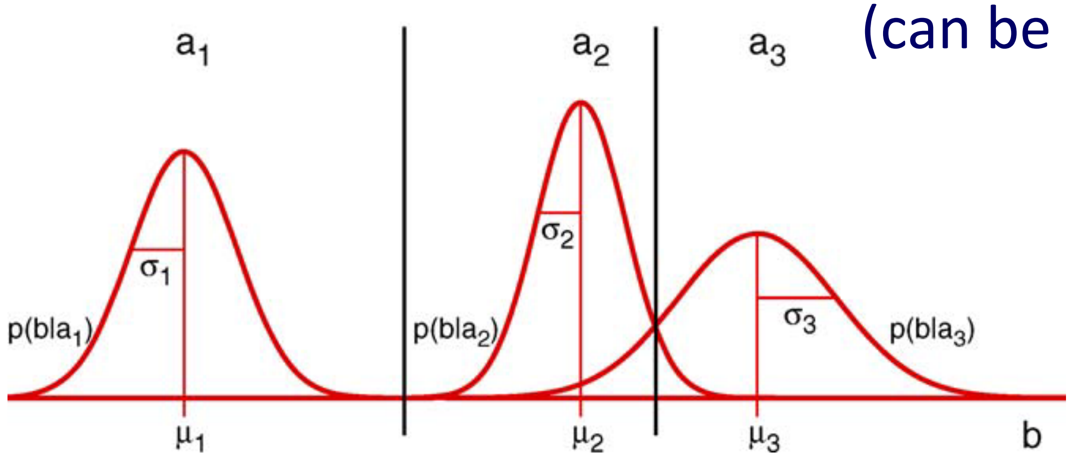


$\theta_B = p(B|A)$

	$p(B=b)$
$A=a_1$	μ_1, σ_1^2
$A=a_2$	μ_2, σ_2^2
$A=a_3$	μ_3, σ_3^2

- A: a variable with *k* =3 states
- B: a continuous node

Model parameters
(can be learned from data)



Joint probability distribution

- The JPD is expressed in terms of a product of CPDs, describing each variable in terms of its parents, i.e., those variables it depends upon.

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \mathbf{pa}(x_i), \theta_i)$$

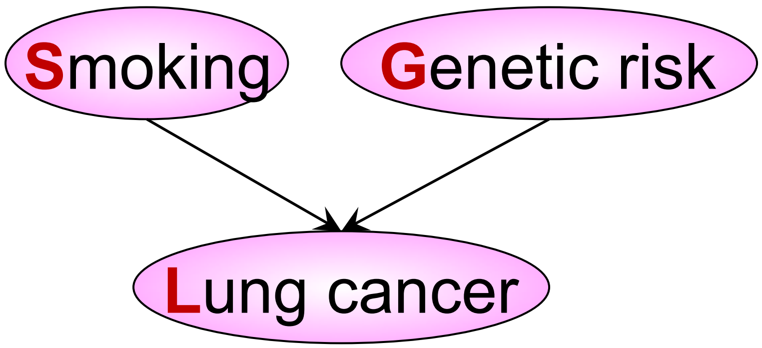
- where $\mathbf{x} = \{x_1, \ldots, x_n\}$ are the variables (nodes in the BN) and $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ denotes the model parameters, where is the set of parameters describing the distribution for the *i* th variable x_i and θ_i denotes the parents of x_i.

Outline

- Conditional distribution and Bayesian networks
- Special cases of Bayesian networks
- Model Selection
- Basic concepts of parameter estimation
  - Maximum likelihood estimation (MLE)

Regression Model

- The *Lung cancer* example



- G: genetic risk, Val(G) = $\{g^1, g^0\}$
- S: smoking, Val(D) = $\{s^1, s^0\}$
- L: lung cancer, Val(L) = $\{l^1, l^0\}$

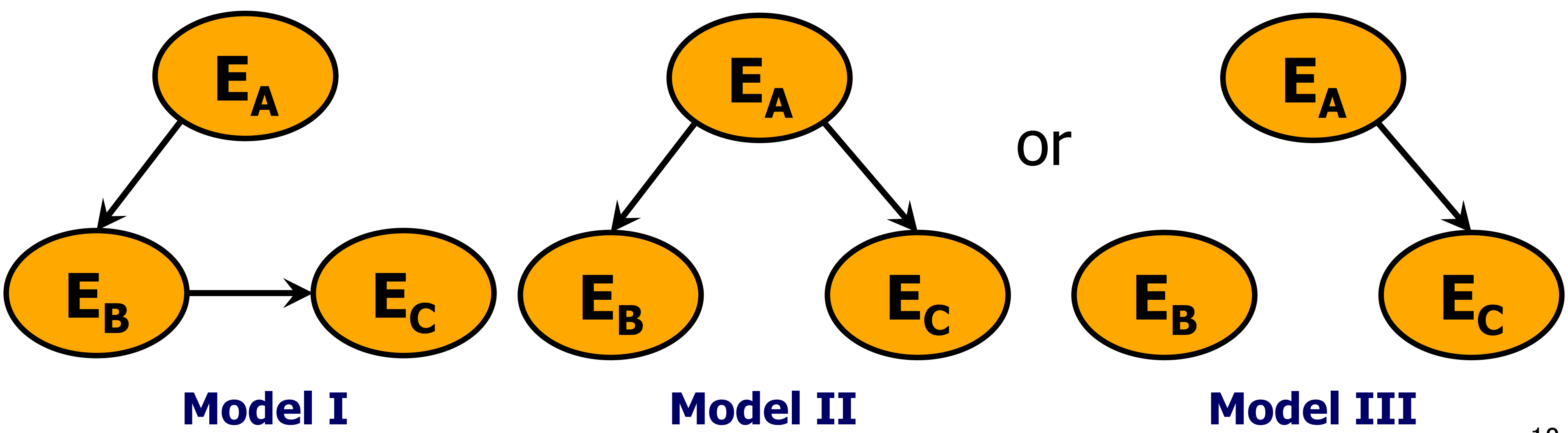
LET'S GO BACK TO THE MODEL SELECTION PROBLEM.

Model selection problem

- Which model do we think is the most likely?
- Given data **D**, let's solve argmax_x P (Model x is true | **D**)

$$P(\text{Model x is true} \mid \mathbf{D}) = \frac{P(\mathbf{D} \mid \text{Model x is true})P(\text{Model x is true})}{P(\mathbf{D})}$$

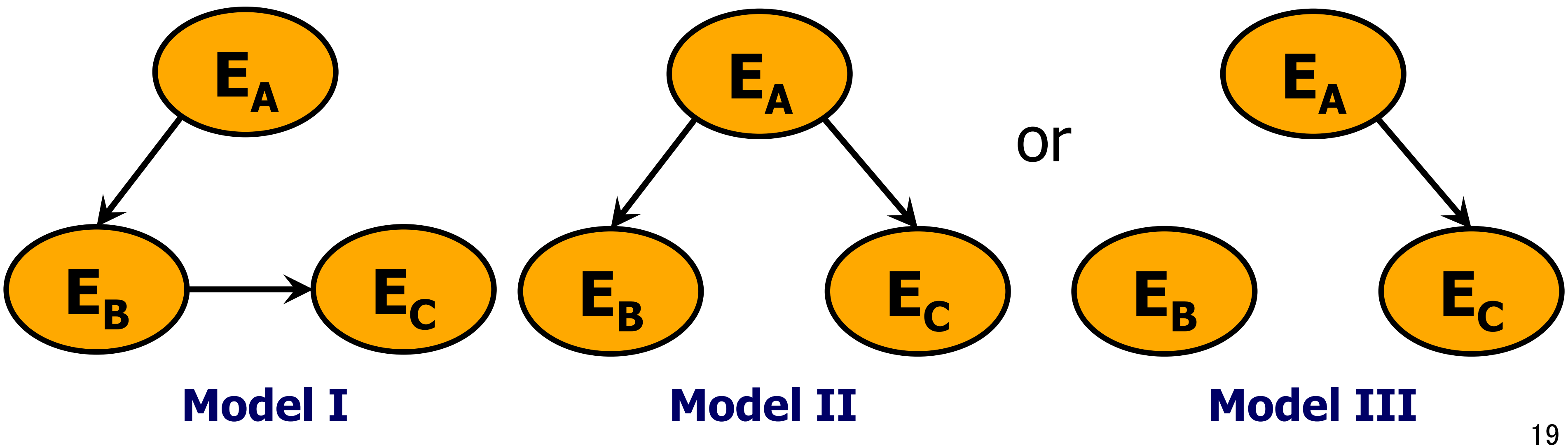
Doesn't depend on x



Model selection problem

- Which model do we think is the most likely?
- Given data **D**, let's solve argmax$_x$ P (Model x is true | **D**)

$$P(\text{Model x is true} \mid \mathbf{D}) \propto P(\mathbf{D} \mid \text{Model x is true}) P(\text{Model x is true})$$

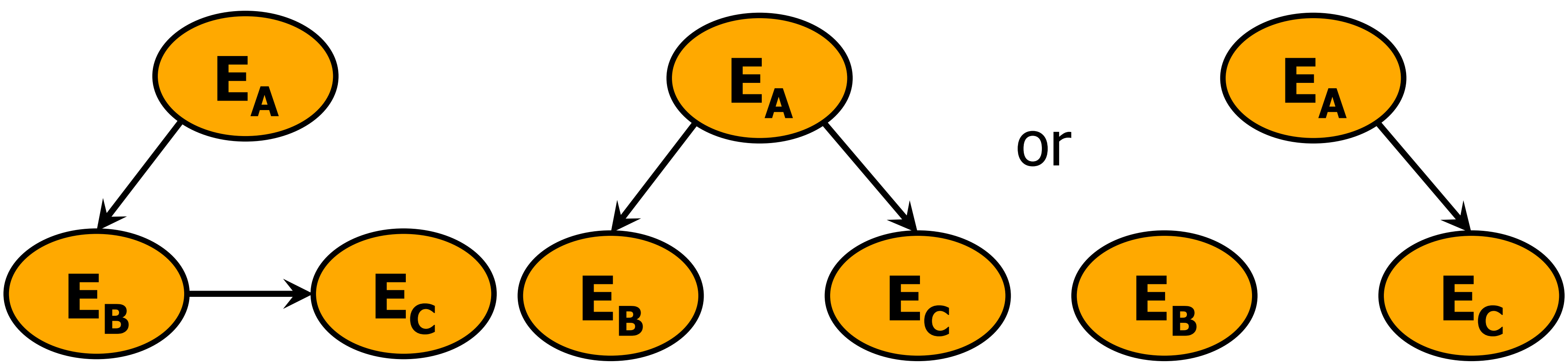


Model selection problem

- Which model do we think is the most likely?
- Given data **D**, let's solve argmax$_x$ P (Model x is true | **D**)

$$P(\text{Model I is true} \mid \mathbf{D}) \propto P(\mathbf{D} \mid \text{Model I is true})P(\text{Model I is true})$$
$$P(\text{Model II is true} \mid \mathbf{D}) \propto P(\mathbf{D} \mid \text{Model II is true})P(\text{Model II is true})$$
$$P(\text{Model III is true} \mid \mathbf{D}) \propto P(\mathbf{D} \mid \text{Model III is true})P(\text{Model III is true})$$

compare



Model I **Model II** **Model III**

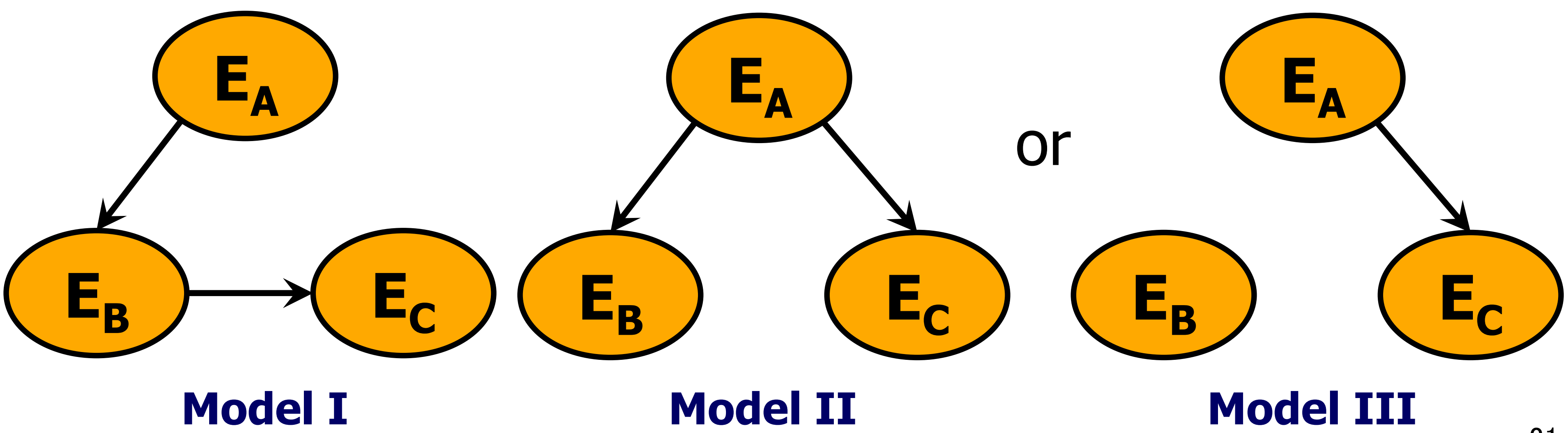
Model selection problem

- Which model do we think is the most likely?
- Given data **D**, let's solve argmax$_x$ P (Model x is true | **D**)

$$P(\mathbf{D} \mid \text{Model I is true})P(\text{Model I is true})$$

$$P(\mathbf{D} \mid \text{Model II is true})P(\text{Model II is true})$$

$$P(\mathbf{D} \mid \text{Model III is true})P(\text{Model III is true})$$



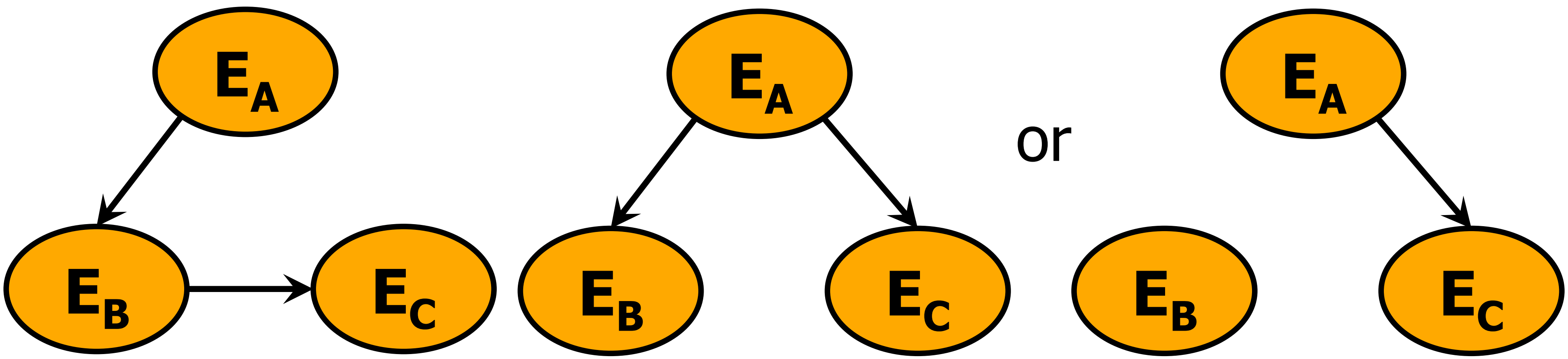
Model selection problem

- Which model do we think is the most likely?
- Given data **D**, let's solve argmax_x P (Model x is true | **D**)

$$P(\mathbf{D} \mid \text{Model I is true})$$

$$P(\mathbf{D} \mid \text{Model II is true})$$

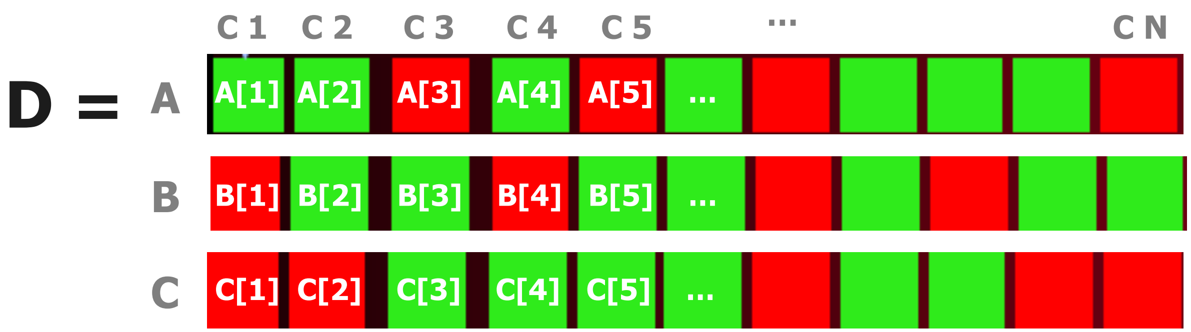
$$P(\mathbf{D} \mid \text{Model III is true})$$

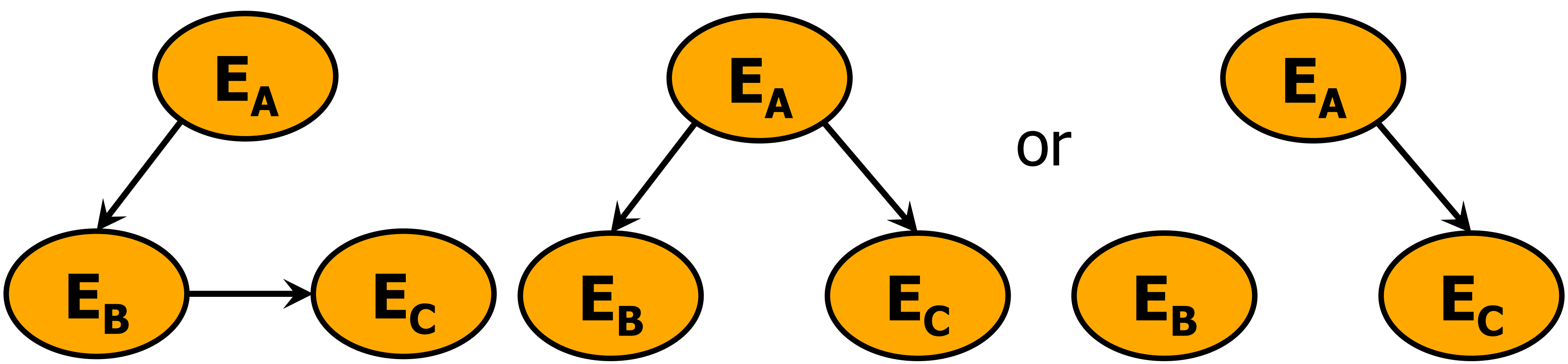


$P(E_A)P(E_B|E_A)P(E_C|E_B)$ $P(E_A)P(E_B|E_A)P(E_C|E_A)$ $P(E_A)P(E_B)P(E_C|E_A)$

Model selection problem

- Which model do we think is the most likely?

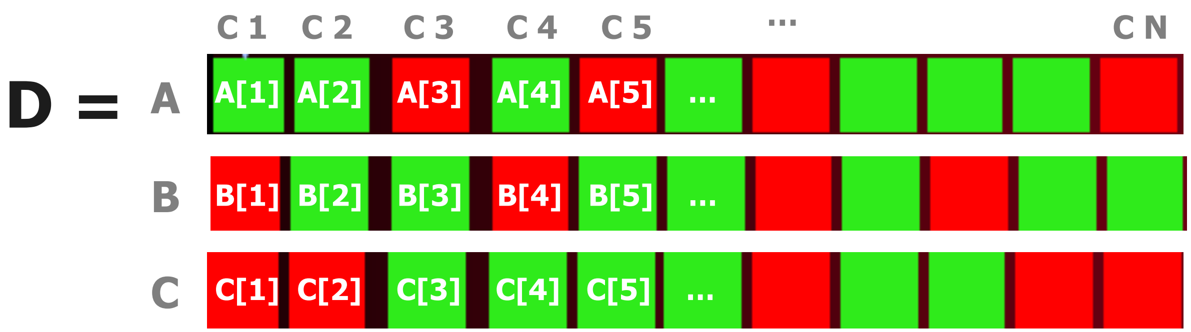


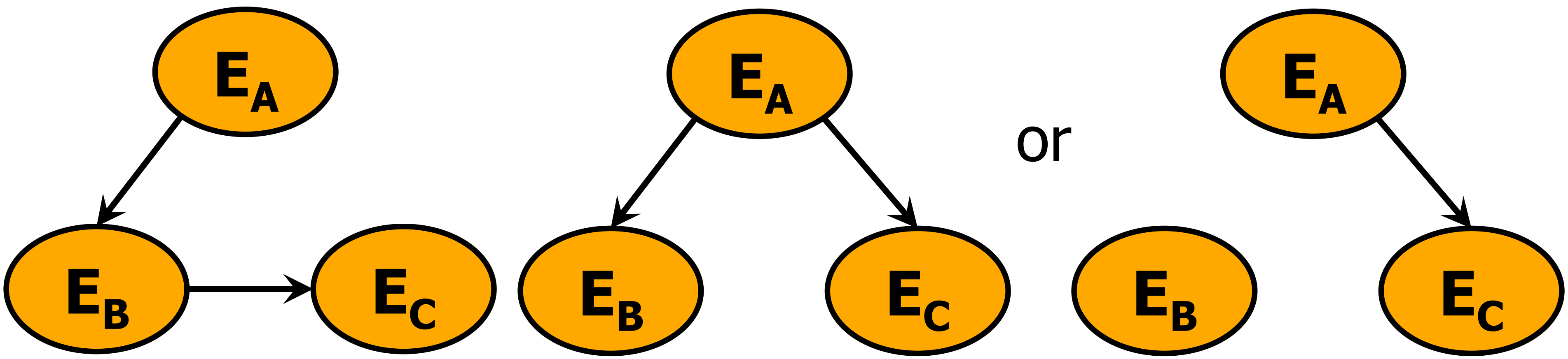


$P(E_A)P(E_B|E_A)P(E_C|E_B)$ $P(E_A)P(E_B|E_A)P(E_C|E_A)$ $P(E_A)P(E_B)P(E_C|E_A)$

Model selection problem

- Which model do we think is the most likely?





$P(E_A)P(E_B|E_A)P(E_C|E_B)$ $P(E_A)P(E_B|E_A)P(E_C|E_A)$ $P(E_A)P(E_B)P(E_C|E_A)$

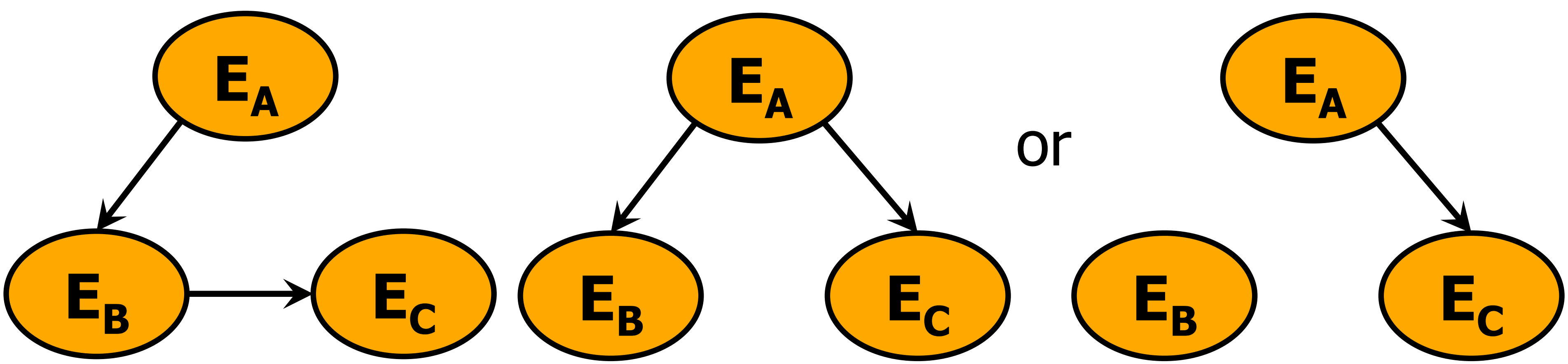
Model selection problem

- Which model do we think is the most likely?
- Given data **D**, let's solve argmax_x P (Model x is true | **D**)

$$P(\mathbf{D} \mid \text{Model I is true}) = \prod_i P(E_A=A[i])\, P(E_B=B[i] \mid E_A=A[i])\, P(E_C=C[i] \mid E_B=B[i])$$

$$P(\mathbf{D} \mid \text{Model II is true}) = \prod_i P(E_A=A[i])\, P(E_B=B[i] \mid E_A=A[i])\, P(E_C=C[i] \mid E_A=A[i])$$

$$P(\mathbf{D} \mid \text{Model III is true}) = \prod_i P(E_A=A[i])\, P(E_B=B[i])\, P(E_C=C[i] \mid E_A=A[i])$$



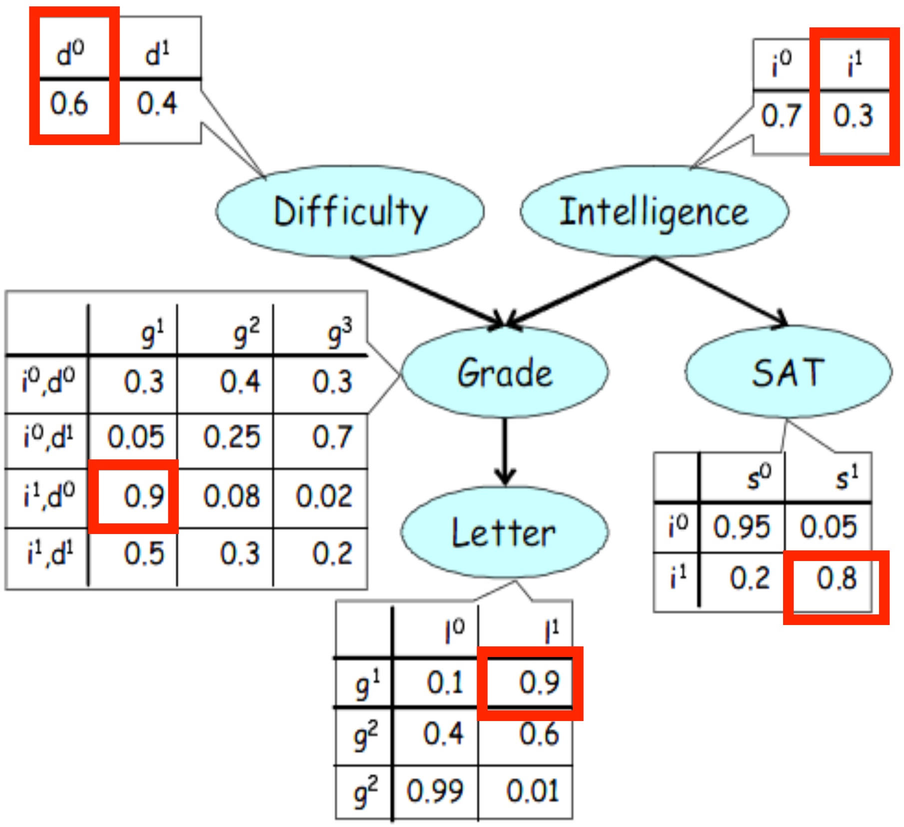
$P(E_A)P(E_B|E_A)P(E_C|E_B)$ $P(E_A)P(E_B|E_A)P(E_C|E_A)$ $P(E_A)P(E_B)P(E_C|E_A)$

Outline

- Conditional distribution and Bayesian networks
- Special cases of Bayesian networks
- Model Selection
- Basic concepts of parameter estimation
  - Maximum likelihood estimation (MLE)

Review: Joint Probability Distribution

- What is the probability of observing
  {D=easy, I=intelligent, G=good, L=strong, S=high} ?



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

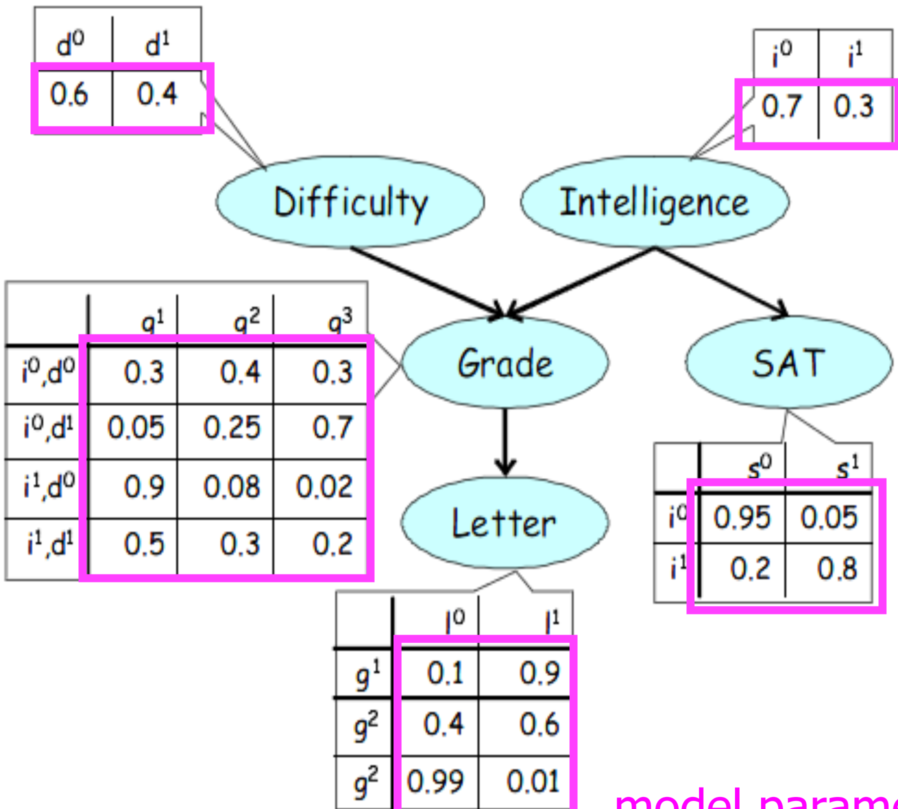
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

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

- P(D=easy, I=intelligent, G=good, L=strong, S=high)
  = P(D=easy) P(I=intelligent) P(G=good | D=easy, I=intelligent) P(S=strong | I=intelligent) P(L=strong | G=good)
  = 0.6 x 0.3 x 0.9 x 0.9 x 0.8
  = 0.1166

Parameters in Bayesian Networks

- What is the probability of observing {D=easy, I=intelligent, G=good, L=strong, S=high} ?



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^2	0.99	0.01

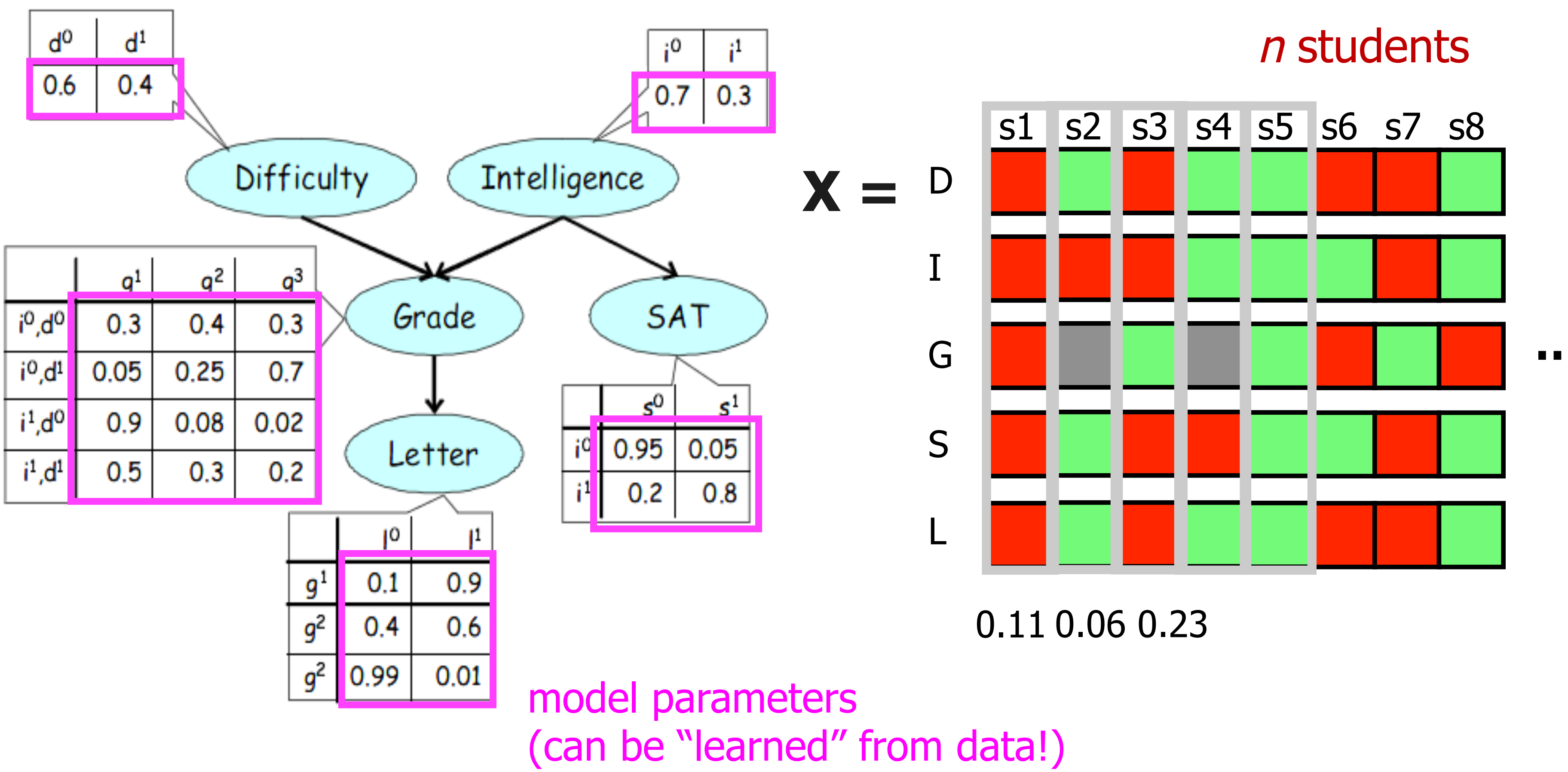
model parameters
(can be "learned" from data!)

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

- P(D=easy, I=intelligent, G=good, L=strong, S=high)
  = P(D=easy) P(I=intelligent)
  P(G=good | D=easy, I=intelligent)
  P(S=strong | I=intelligent)
  P(L=strong | G=good)
  = 0.6 x 0.3 x 0.9 x 0.9 x 0.8
  = 0.1166

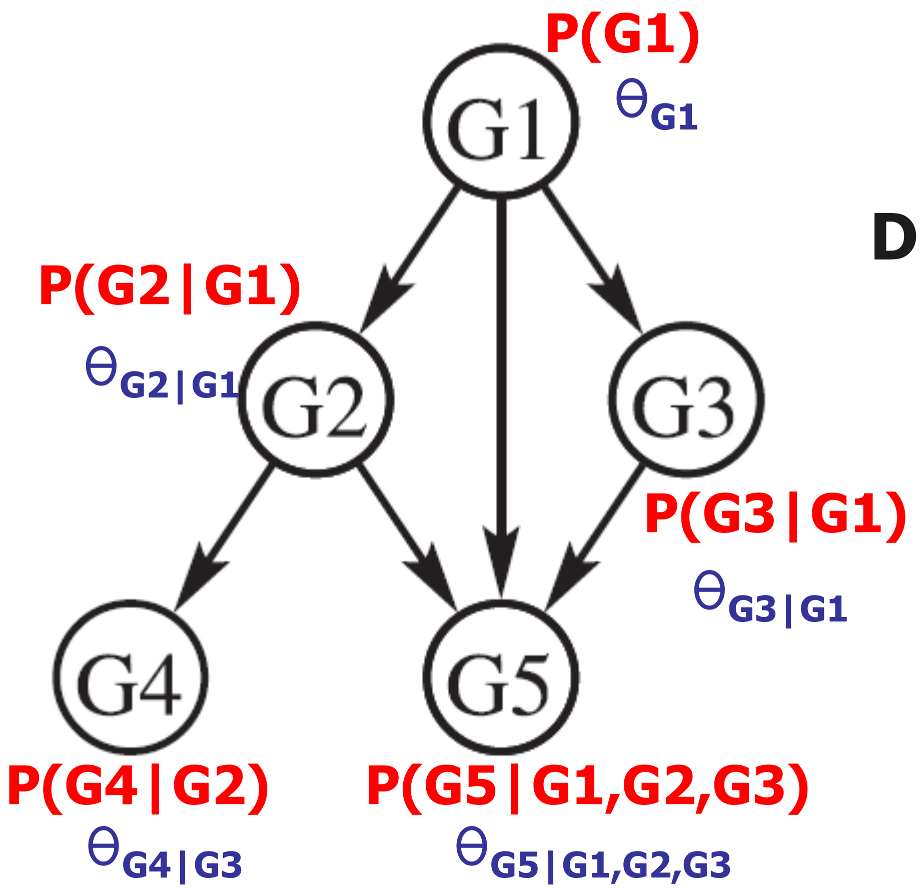
Data Likelihood

- What is the probability of observing multiple students with certain values on the five variables ?



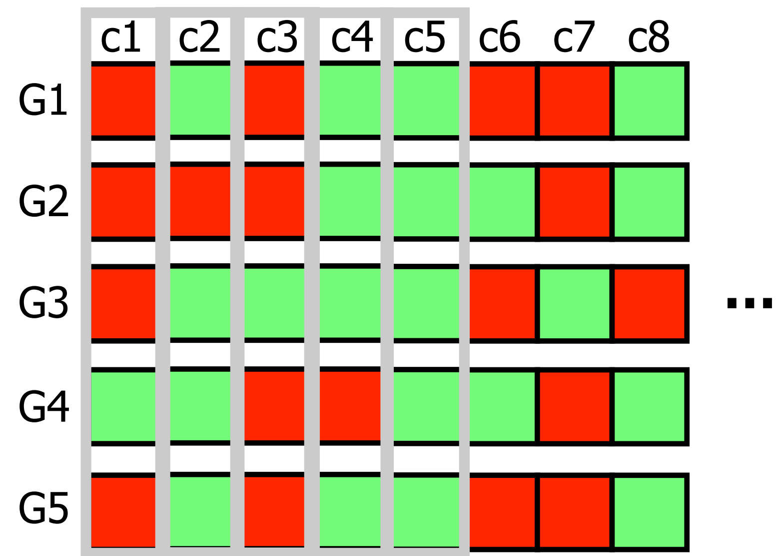
Data Likelihood of the 5-gene network

- Learn the parameters based on **D**



n instances

D =



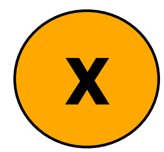
Outline

- Conditional distribution and Bayesian networks
- Special cases of Bayesian networks
- Model Selection
- Basic concepts of parameter estimation
  - Maximum likelihood estimation (MLE)

LET’S CONSIDER THE SIMPLEST EXAMPLE.

Properties of Good Parameter Estimates

- What are characteristics of good estimators?
- How well they *explain* the world?
- Say that you flip a coin
  - Let's say that a random variable *X* represents the outcome
  - *p* = probability of getting Head



- If you flip a coin many times, maybe we can figure out.
  - Realization of the random variable
  - Observation data D = {HHTHHTHTHTHTHTH ...}

samples (or instances)

Introduction to Likelihood

- **Before** an experiment is performed, the outcome is unknown
- Probability function allows us to predict the probability of any outcome based on **known** parameters:

  P(Data | θ)

- For example, say that we know that probability of getting a Head in a coin toss is p = 0.6
  - Then, we can calculate the probability P(Data| θ) for ANY data

$$D_1 = \{HTHHHTHHHT\} \qquad P(D \mid \theta) = p^7(1-p)^3$$
$$D_2 = \{HTH\} \qquad P(D \mid \theta) = p^2(1-p)$$
$$D_3 = \{TTTH\} \qquad P(D \mid \theta) = p^3(1-p)$$

:

  - If p were a different value, the above probabilities would have been different…

Introduction to Likelihood

- **After** an experiment is performed, the outcome is known.
- Now we talk about the likelihood that a parameter would generate the observed data:

$$L(\theta : D) = P(\text{Data} \mid \theta)$$

- Estimation proceeds by finding the value of θ that makes the observed data most *likely.*
  - *Maximum Likelihood Estimate (MLE)* $\hat{\theta}$
- We need to find what is a parameter and what the observed data are.

Motivating Example

- Suppose that there is a disease (let's say halitosis) which is partly genetically determined.
- The genotype aa has a 40% chance of getting the disease, and the other two possible genotypes, AA and Aa, each has a 10% chance of getting the disease.
- Suppose we observe 1000 individuals and find that the 182 of them have the disease.
- Based on the observation, *we want to estimate the frequency of the A allele*.
- What are the data? What is the parameter?

The Coin Example

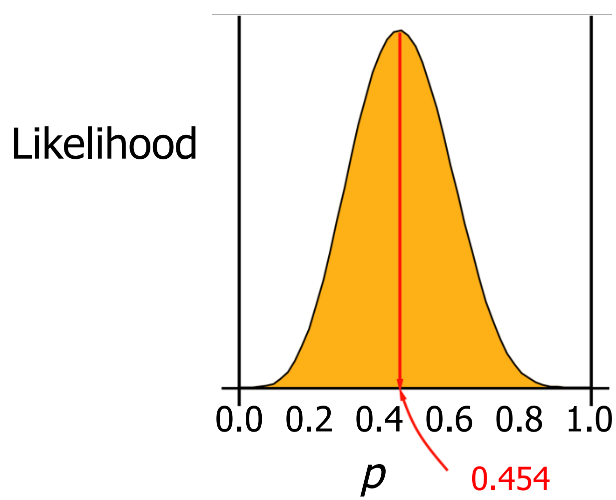
- Let's toss a coin *n* times with probability *p* of heads

- Probability of outcome D = {HHTHTTTTHTTH} is

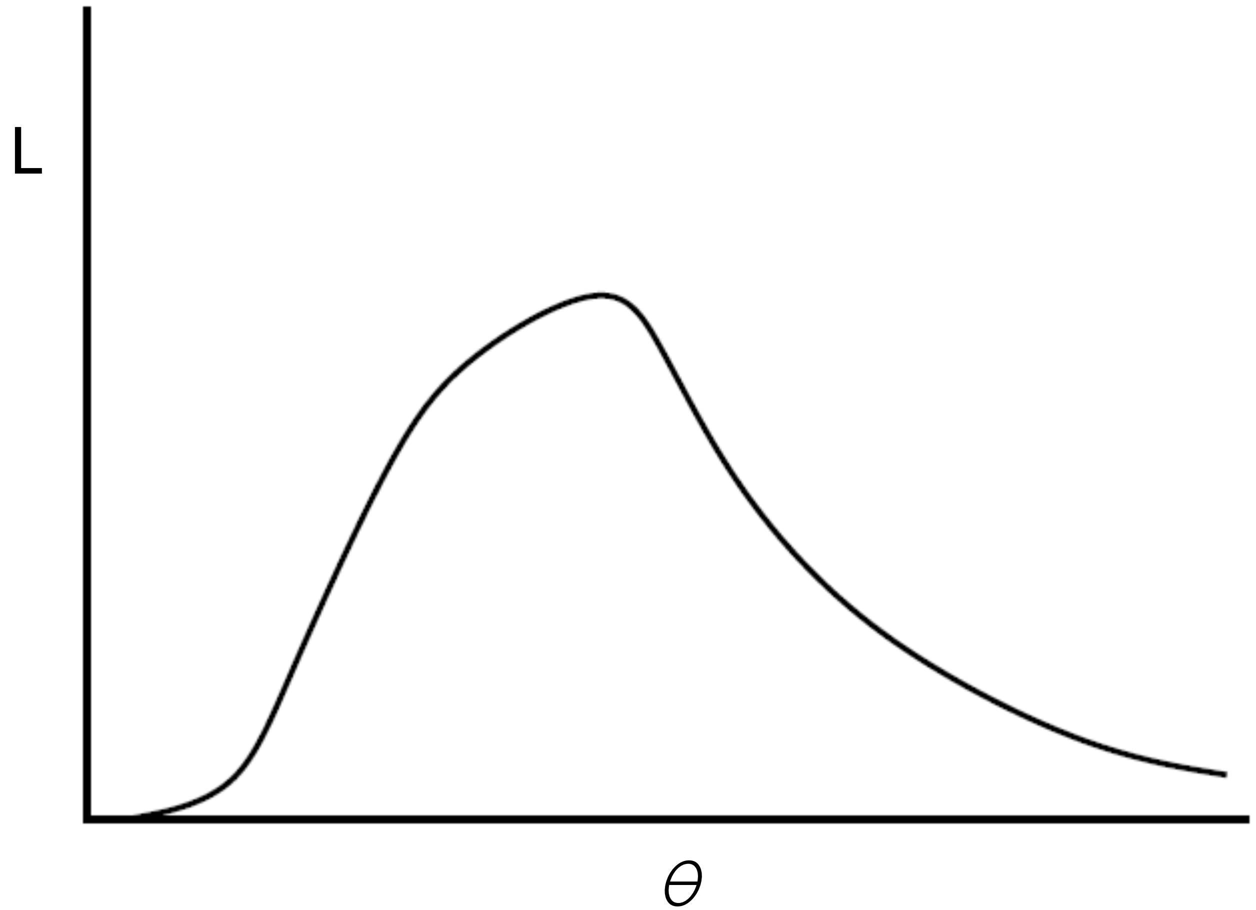
$$pp(1-p)p(1-p)(1-p)(1-p)(1-p)p(1-p)(1-p)p$$

- The likelihood is then $L = P(D \mid p) = p^5(1-p)^6$

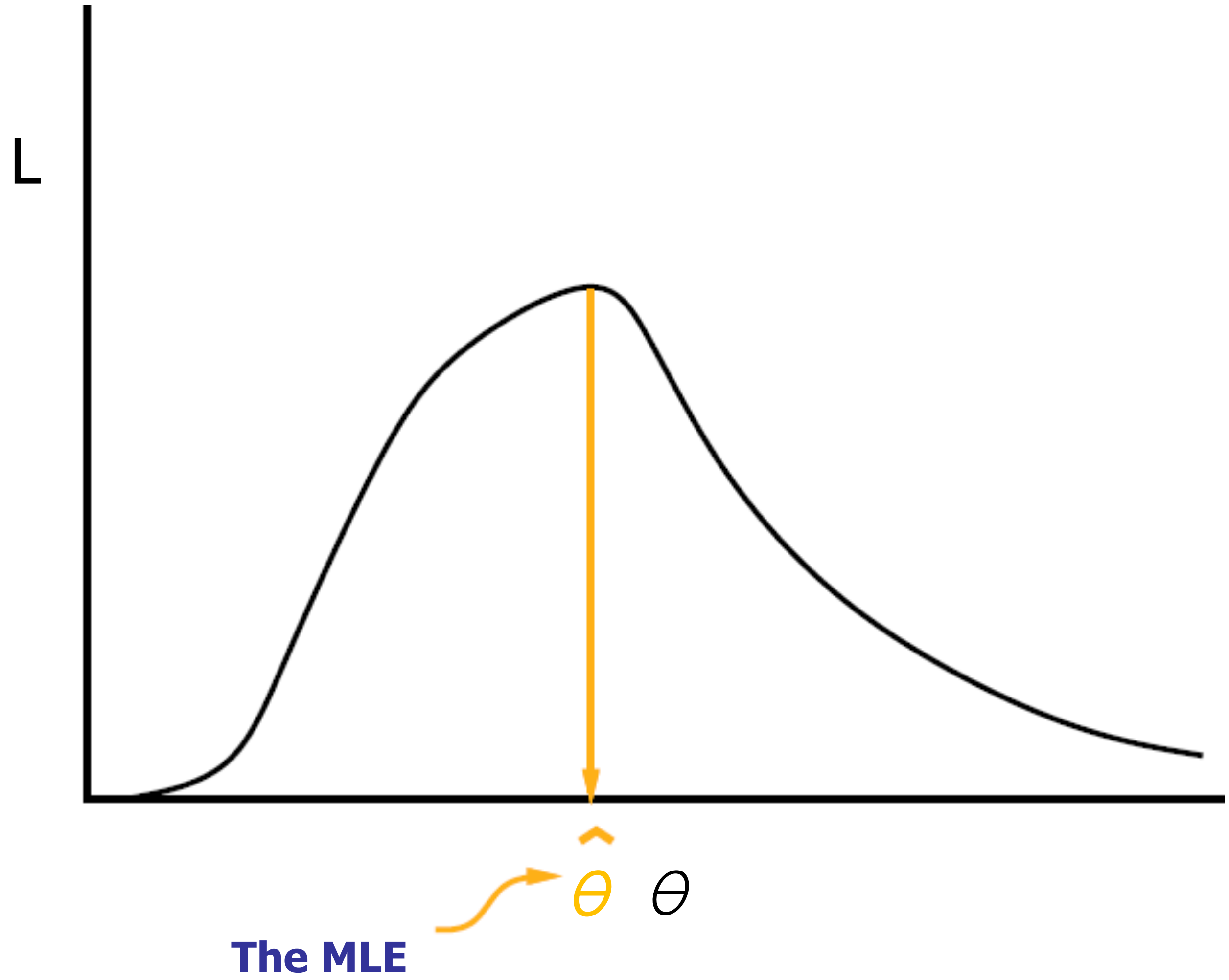
- Plotting *L* against *p* to find its maximum



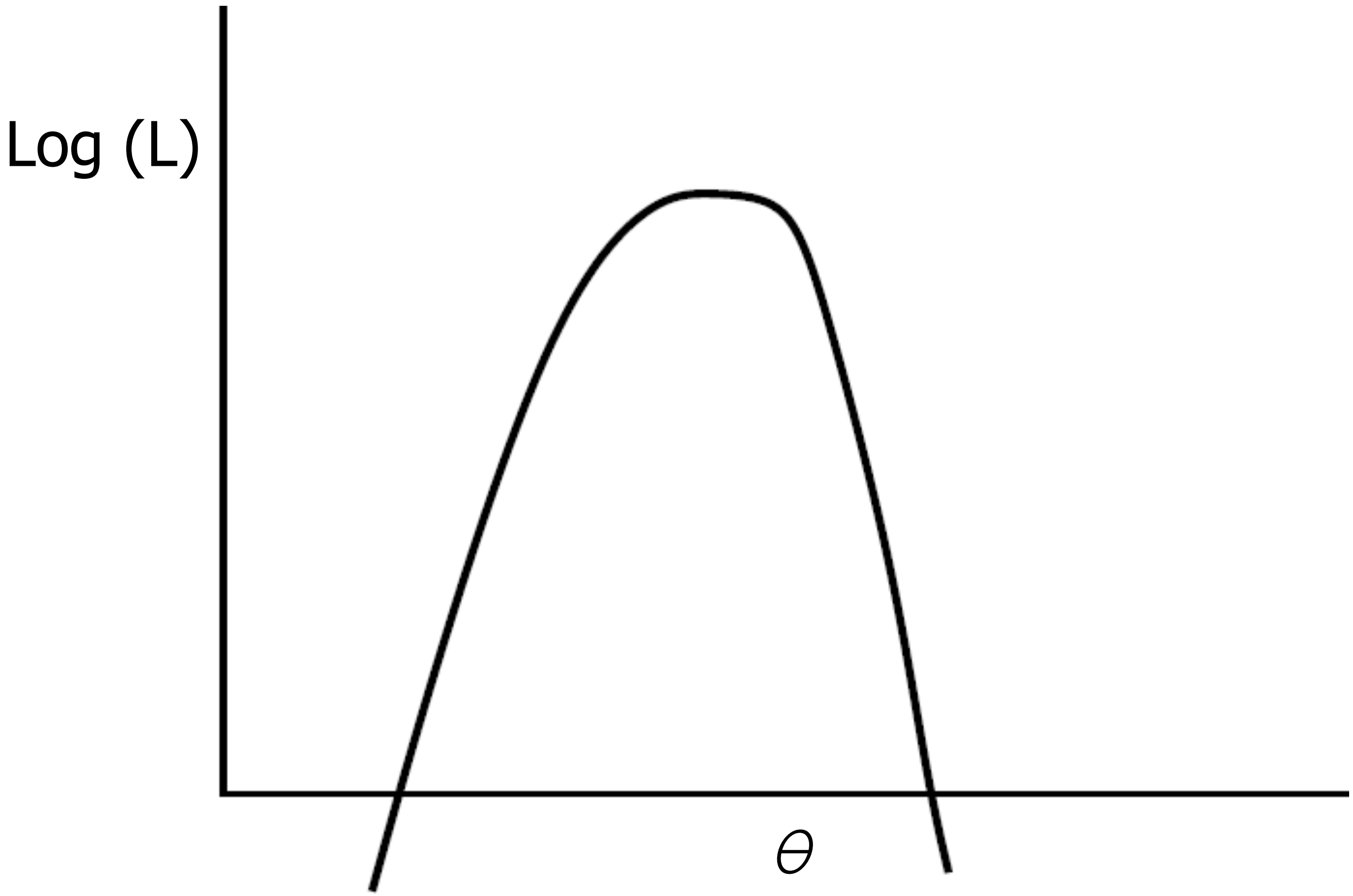
A Likelihood Curve



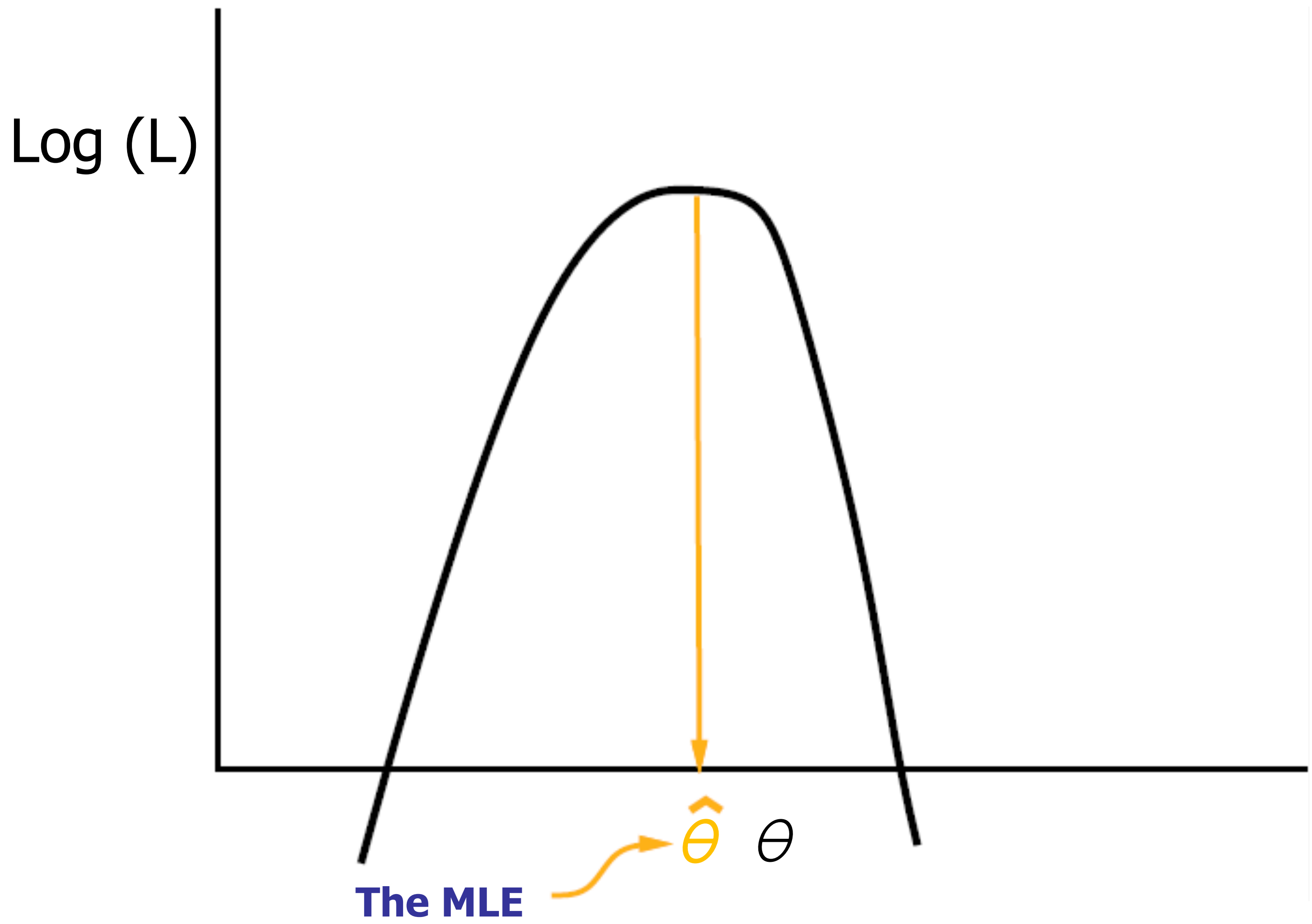
Its Maximum Likelihood Estimate



Better to Plot log (L) than L



Better to Plot log (L) than L



Differentiating to Find the Maximum

- Differentiate the expression for log (L) with respect to p

$$\log L = \log\left[p^5(1-p)^6\right] = 5\log p + 6\log(1-p)$$

- Equate the derivative to 0

$$\frac{\partial \log L}{\partial p} = \left(\frac{5}{p} - \frac{6}{1-p}\right) = 0$$

$$5 - 11p = 0 \quad \Rightarrow \quad \hat{p} = \frac{5}{11}$$

  - The value of p that is at the peak can be found to be $p = 5/11$

Formal Statement of MLE

- Let x[1], x[2], …, x[M] be a sequence of *M* observed values
  - e.g. x[m] = H or x[m] = T in coin tossing

- **Joint probability:**

$$P(D \mid \theta) = P(X = x[1])P(X = x[2]) \cdots P(X = x[M])$$

$$= \prod_{m=1}^{M} P(X = x[m])$$

- **Likelihood** is then:

$$L(\theta : D) = \prod_{m=1}^{M} P(X = x[m])$$

$$\log L(\theta : D) = \sum_{m=1}^{M} \log P(X = x[m])$$