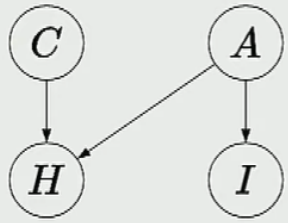
15. Bayesian Network 3 - Maximum Likelihood

Review: Bayesian network



$$\mathbb{P}(C=c, A=a, H=h, I=i) = p(c)p(a)p(h \mid c, a)p(i \mid a)$$

Definition: Bayesian network

Let $X = (X_1, \ldots, X_n)$ be random variables.

A **Bayesian network** is a directed acyclic graph (DAG) that specifies a joint distribution over X as a product of local conditional distributions, one for each node:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i \mid x_{\text{Parents}(i)})$$

Stanford

Review: probabilistic inference

Bayesian network:

$$\mathbb{P}(X = x) = \prod_{i=1}^{n} p(x_i \mid x_{\text{Parents}(i)})$$

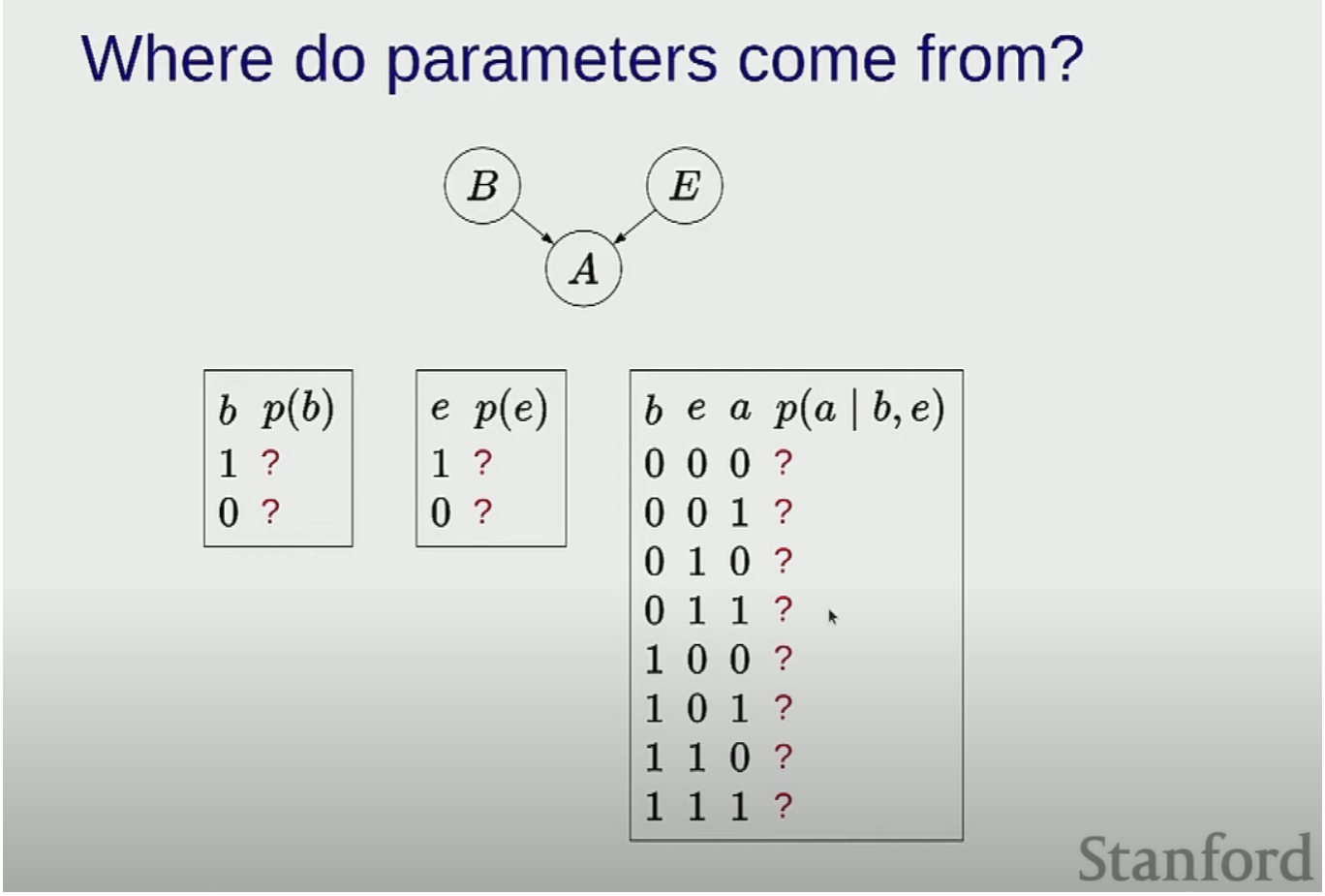
Probabilistic inference:

$$\mathbb{P}(Q \mid E = e)$$

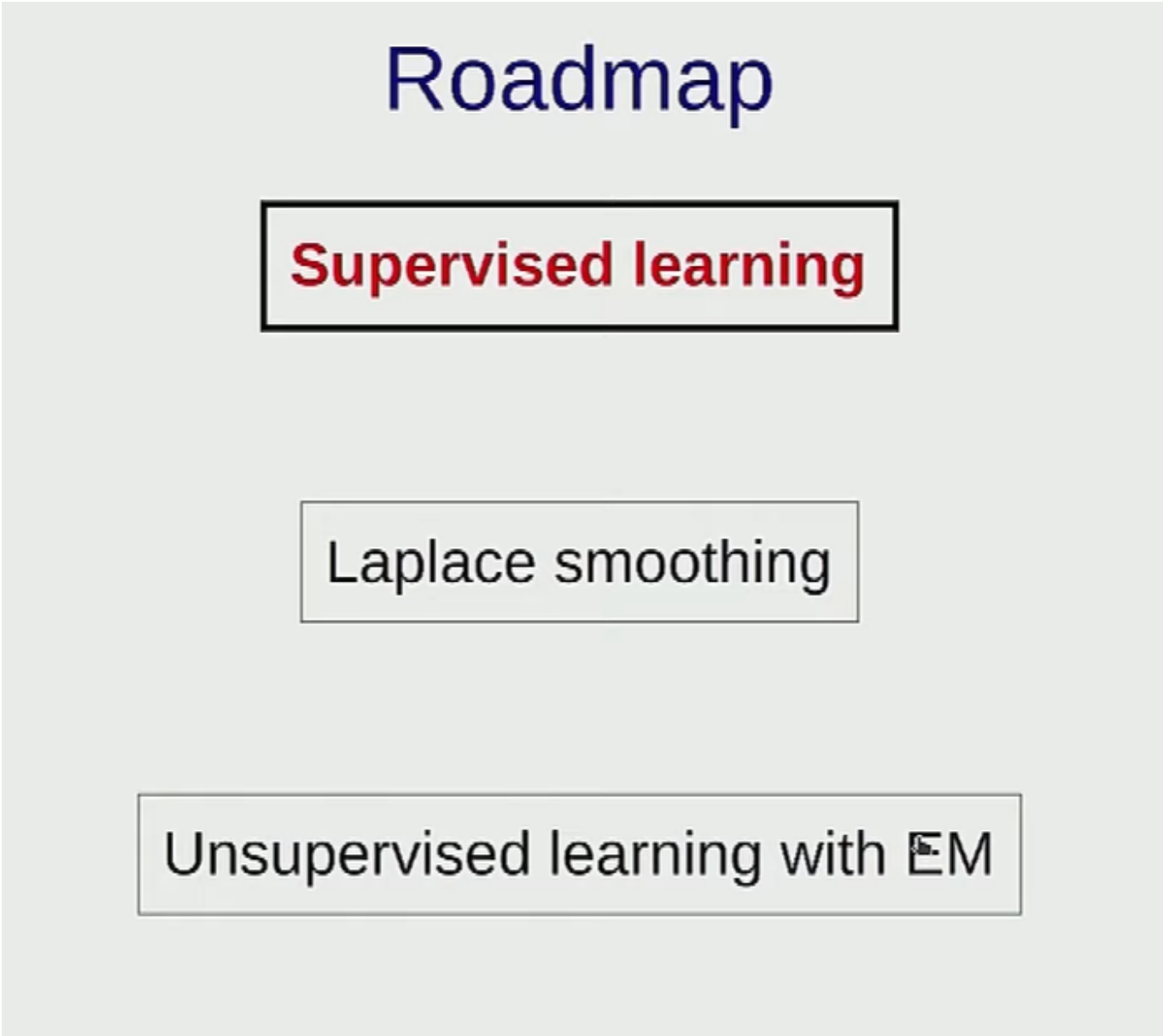
Algorithms:

- Forward-backward: HMMs, exact
- Particle filtering: HMMs, approximate
- Gibbs sampling: general, approximate

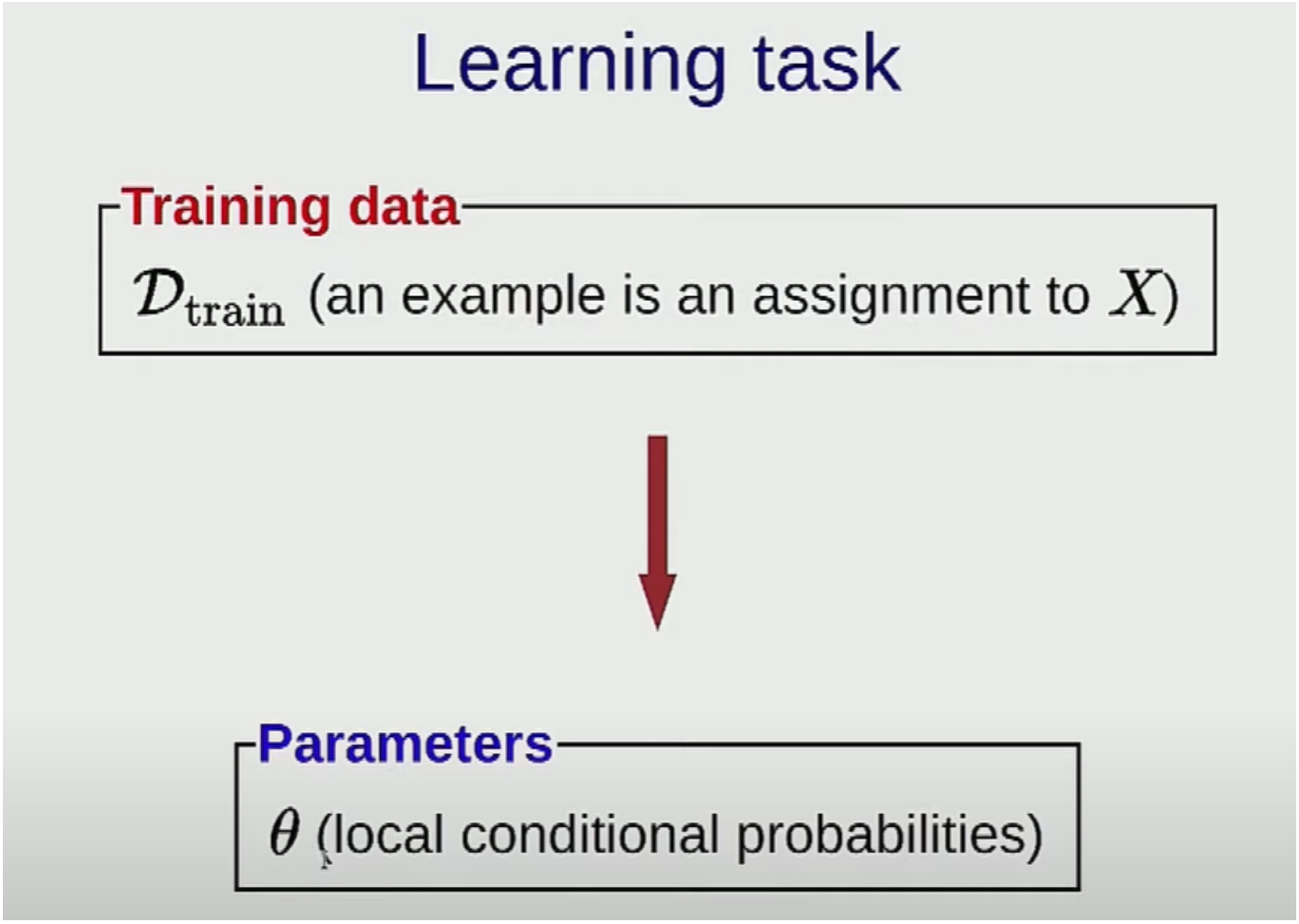
We are gonna talk about learning

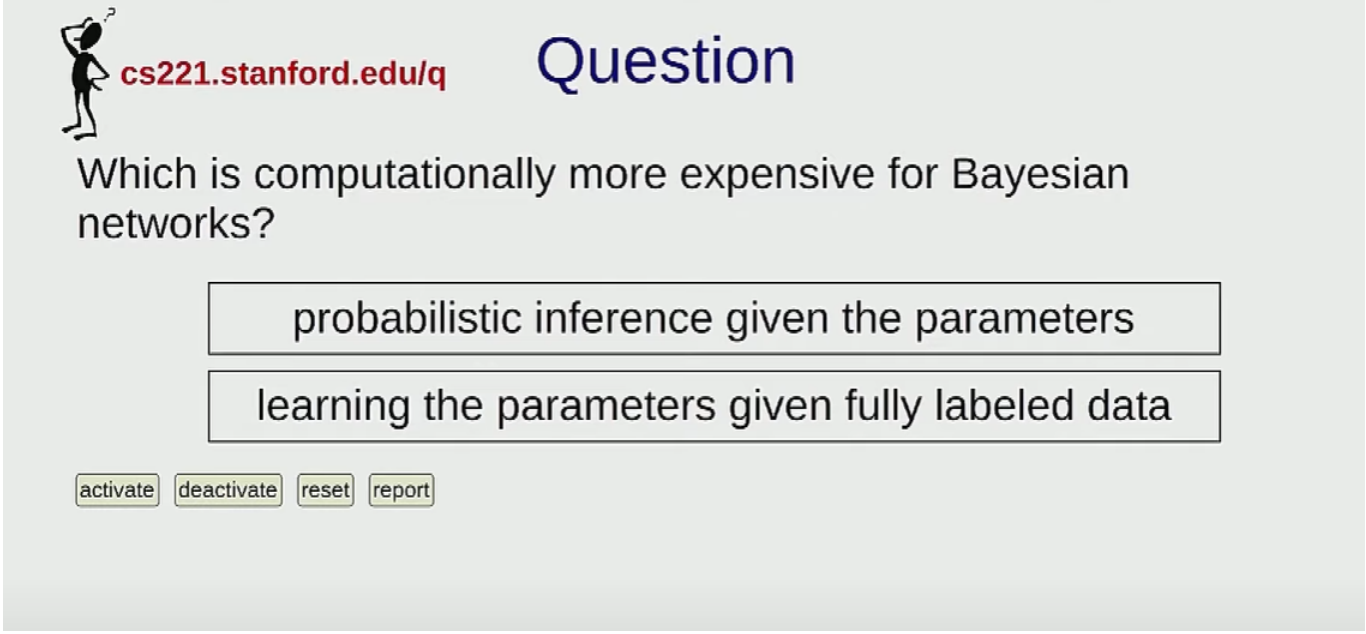


So far we have just assumed that someone hands you these local conditional distribution which has numbers filled out In real world we have to figure them out from data



Supervised learning





It turns out that probabilistic inference is more expensive So in fully supervised learning it should be easier

Example: one variable

Setup:

- One variable R representing the rating of a movie $\{1, 2, 3, 4, 5\}$

$$\mathbb{P}(R = r) = p(r)$$

Parameters:

$$\theta = (p(1), p(2), p(3), p(4), p(5))$$

Training data:

$$\mathcal{D}_{\text{train}} = \{1, 3, 4, 4, 4, 4, 4, 5, 5, 5\}$$

Fully observed

Example: one variable

Learning:

$$\mathcal{D}_{\text{train}} \quad \Rightarrow \quad \theta$$

Intuition: $p(r) \propto$ number of occurrences of r in $\mathcal{D}_{\text{train}}$

Example:

$$\mathcal{D}_{\text{train}} = \{1, 3, 4, 4, 4, 4, 4, 5, 5, 5\}$$

θ:

r	count(r)
1	1
2	0
3	1
4	5
5	3

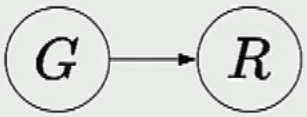
Stanford

count and divide, then normalize

Example: two variables

Variables:

- Genre $G \in \{\text{drama}, \text{comedy}\}$
- Rating $R \in \{1, 2, 3, 4, 5\}$



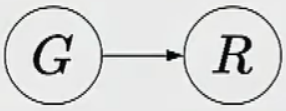
$$\mathbb{P}(G = g, R = r) = p_G(g)p_R(r \mid g)$$

$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4), (\text{d}, 4), (\text{d}, 5), (\text{c}, 1), (\text{c}, 5)\}$$

Parameters: $\theta = (p_G, p_R)$

We first access Pg and Pr separetely

Example: two variables



$$\mathbb{P}(G = g, R = r) = p_G(g)p_R(r \mid g)$$

$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4), (\text{d}, 4), (\text{d}, 5), (\text{c}, 1), (\text{c}, 5)\}$$

Intuitive strategy: Estimate each local conditional distribution (p_G and p_R) separately

θ:

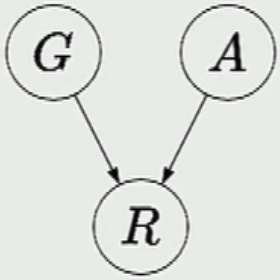
g	$p_G(g)$
d	3/5
c	2/5

g	r	$p_R(r \mid g)$
d	4	2/3
d	5	1/3
c	1	1/2
c	5	1/2

Example: v-structure

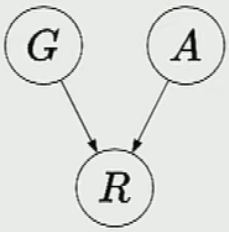
Variables:

- Genre $G \in \{\text{drama}, \text{comedy}\}$
- Won award $A \in \{0, 1\}$
- Rating $R \in \{1, 2, 3, 4, 5\}$



$$\mathbb{P}(G = g, A = a, R = r) = p_G(g)p_A(a)p_R(r \mid g, a)$$

Example: v-structure



$$\mathcal{D}_{\text{train}} = \{(\text{d}, 0, 3), (\text{d}, 1, 5), (\text{d}, 0, 1), (\text{c}, 0, 5), (\text{c}, 1, 4)\}$$

Parameters: $\theta = (p_G, p_A, p_R)$

θ:

g	$p_G(g)$
d	3/5
c	2/5

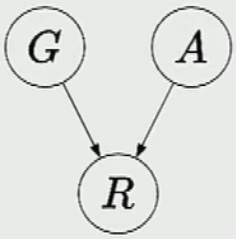
a	$p_A(a)$
0	3/5
1	2/5

g	a	r	$\text{count}_R(g, a, r)$
d	0	1	1
d	0	3	1
d	1	5	1
c	0	5	1
c	1	4	1

Be careful when u normalize P(r | g, a)

When normalize, we focus on r only For every possible unique setting (g, a), I have a different distribution So (d, 0) is a distribution over 1 or 3

Example: v-structure



$$\mathcal{D}_{\text{train}} = \{(\text{d}, 0, 3), (\text{d}, 1, 5), (\text{d}, 0, 1), (\text{c}, 0, 5), (\text{c}, 1, 4)\}$$

Parameters: $\theta = (p_G, p_A, p_R)$

θ:

g	$p_G(g)$
d	3/5
c	2/5

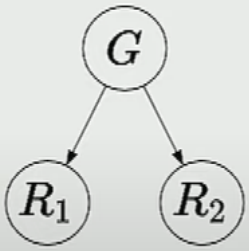
a	$p_A(a)$
0	3/5
1	2/5

g	a	r	$p_R(r \mid g, a)$
d	0	1	1/2
d	0	3	1/2
d	1	5	1
c	0	5	1
c	1	4	1

Example: inverted-v structure

Variables:

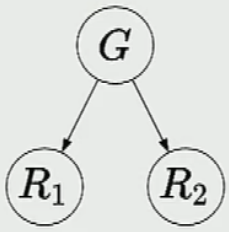
- Genre $G \in \{\text{drama}, \text{comedy}\}$
- Jim's rating $R_1 \in \{1, 2, 3, 4, 5\}$
- Martha's rating $R_2 \in \{1, 2, 3, 4, 5\}$



$$\mathbb{P}(G = g, R_1 = r_1, R_2 = r_2) = p_G(g) p_{R_1}(r_1 \mid g) p_{R_2}(r_2 \mid g)$$

For R1 and R2, we are normalize based on g

Example: inverted-v structure



$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4, 5), (\text{d}, 4, 4), (\text{d}, 5, 3), (\text{c}, 1, 2), (\text{c}, 5, 4)\}$$

Parameters: $\theta = (p_G, p_{R_1}, p_{R_2})$

θ:

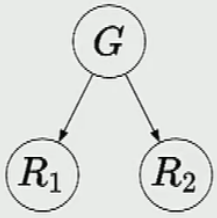
g	$p_G(g)$
d	3/5
c	2/5

g	r_1	$p_{R_1}(r \mid g)$
d	4	2/3
d	5	1/3
c	1	1/2
c	5	1/2

g	r_2	$p_{R_2}(r \mid g)$
d	3	1/3
d	4	1/3
d	5	1/3
c	2	1/2
c	4	1/2

We now want a pR to generally represent all generally rating, cuz we don't want a single distribution for each rating So we can combine R1 and R2

Example: inverted-v structure



$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4, 5), (\text{d}, 4, 4), (\text{d}, 5, 3), (\text{c}, 1, 2), (\text{c}, 5, 4)\}$$

Parameters: $\theta = (p_G, p_R)$

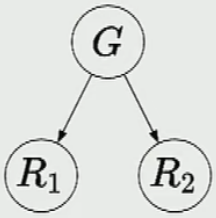
θ:

g	$p_G(g)$
d	3/5
c	2/5

g	r	$\text{count}_R(g, r)$
d	3	1
d	4	3
d	5	2
c	1	1
c	2	1
c	4	1
c	5	1

Stanford

Example: inverted-v structure



$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4, 5), (\text{d}, 4, 4), (\text{d}, 5, 3), (\text{c}, 1, 2), (\text{c}, 5, 4)\}$$

Parameters: $\theta = (p_G, p_R)$

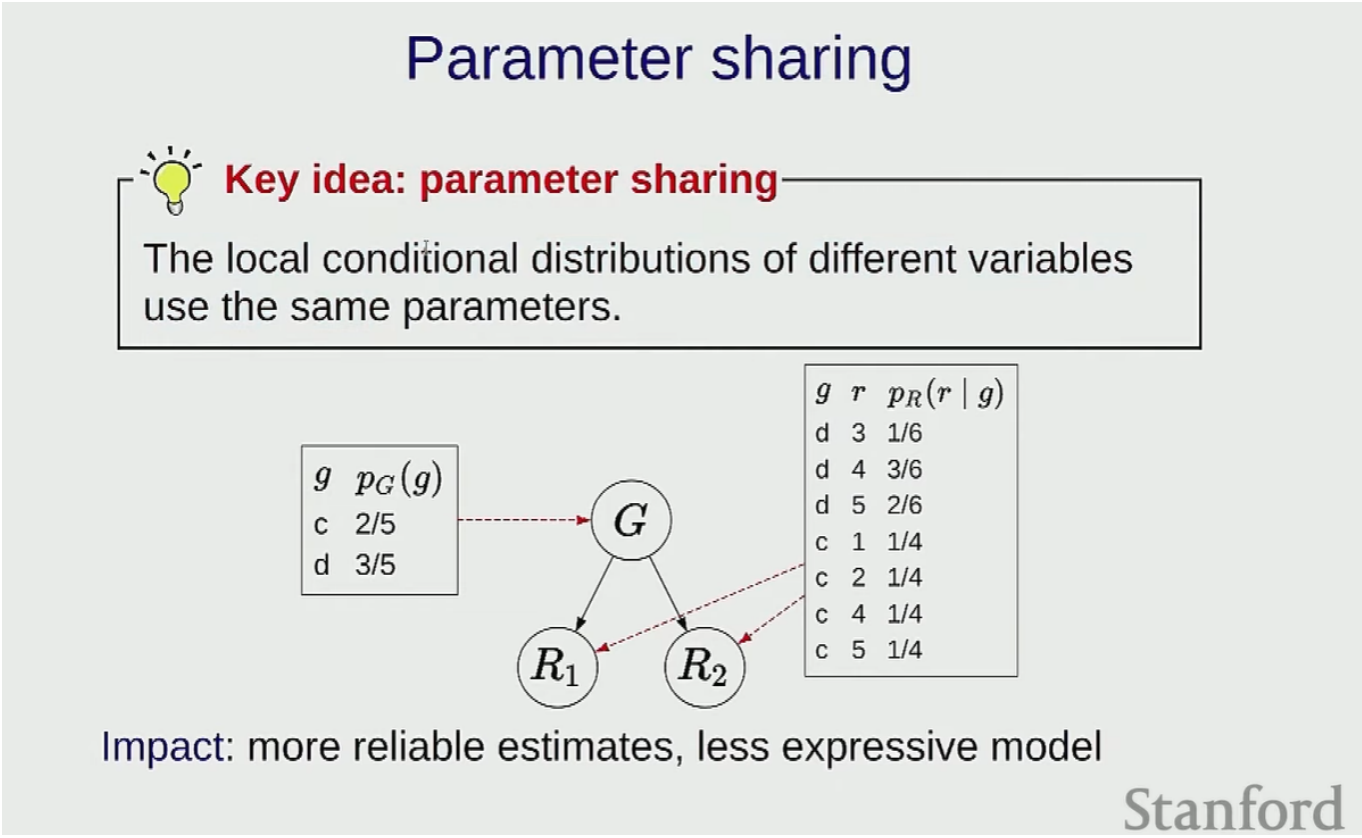
θ:

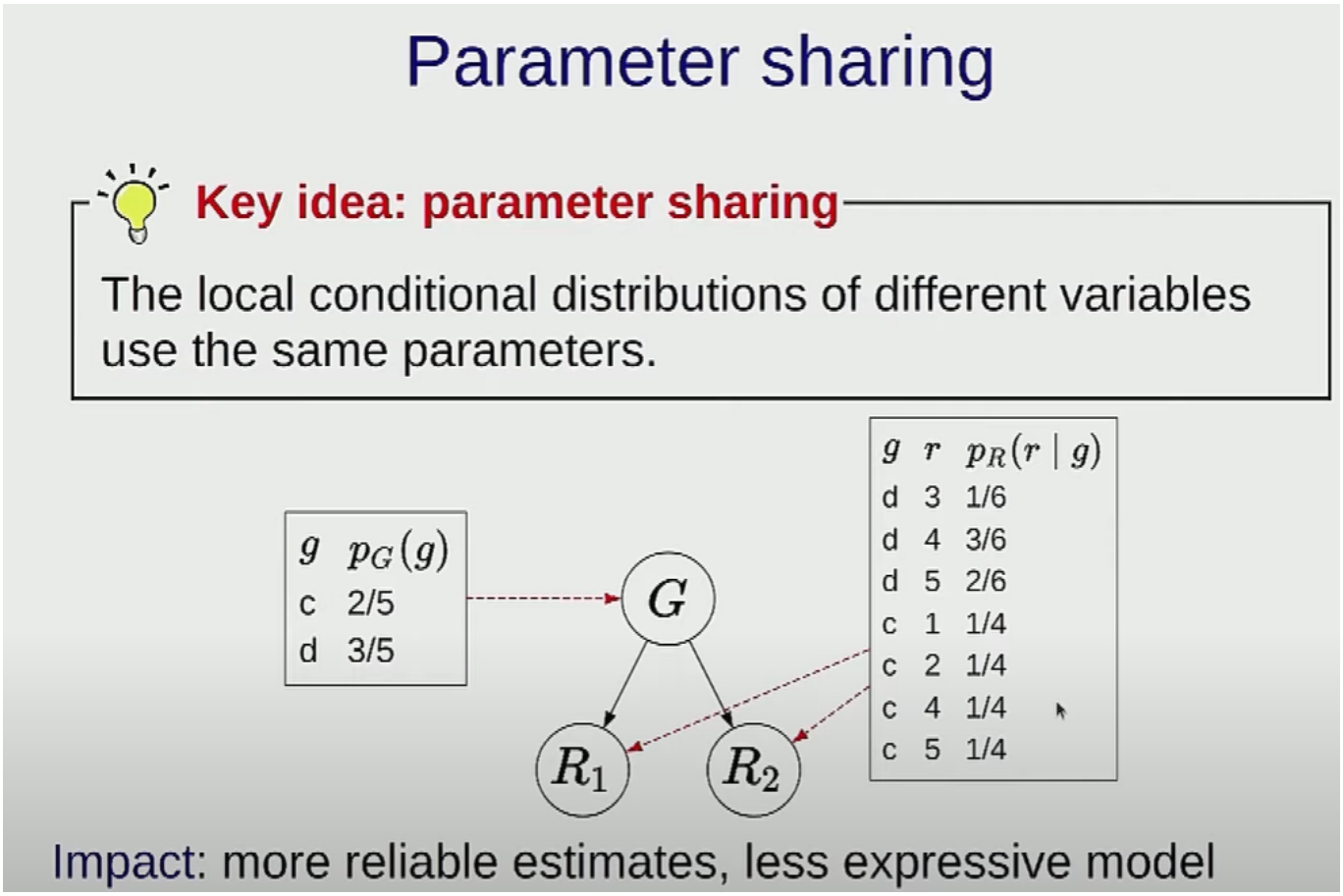
g	$p_G(g)$
d	3/5
c	2/5

g	r	$p_R(r \mid g)$
d	3	1/6
d	4	3/6
d	5	2/6
c	1	1/4
c	2	1/4
c	4	1/4
c	5	1/4

Stanford

We are merging previous table R1 and R2 together Note that we are assuming all data in dataset are independent Also condition on g, R1 and R2 are independent





Each node is gonna be powered by some table, So R1 and R2 now both powered by their joint table

Laplace smoothing

Unsupervised learning with EM