Cluster Allocation After Parameter Estimation

- **Parameters** of the mixture model: $\Lambda=(\{\theta_i\}; \{p(\theta_i)\})$, $i \in [1,k]$
  - Each θ_i represents the **content of cluster i** : **$p(w|\ \theta_i)$**
  - **$p(\theta_i)$** indicates the **size of cluster i**
  - Note that unlike in PLSA, $p(\theta_i)$ doesn't depend on d!
- Which cluster should document d belong to? c_d=?
  - **Likelihood only**: Assign d to the cluster corresponding to the topic θ_i that most likely has been used to generate d

$$c_d = \arg\max_i p(d \mid \theta_i)$$

  - **Likelihood + prior $p(\theta_i)$ (Bayesian):** favor large clusters

$$c_d = \arg\max_i p(d \mid \theta_i)p(\theta_i)$$